S

Saddlepoint Approximations

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán and Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Argentina

Introduction

It is often required to approximate to the distribution of some statistics whose exact distribution cannot be conveniently obtained. When the first few moments are known, a common procedure is to fit a law of the Edgeworth type having the same moments as far as they are given. This method is often satisfactory in practice, but has the drawback that error in the "tail" regions of the distribution are sometimes comparable with the frequencies themselves. Notoriously, the Edgeworth approximation can assume negative values in such regions.

The characteristic function of the statistic may be known, and the difficulty is then the analytical one of inverting a Fourier transform explicitly. It is possible to show that for some statistics a satisfactory approximation to its probability density, when it exists, can be obtained nearly always by the method of steepest descents. This gives an asymptotic expansion in powers of n^{-1}, where n is the sample size, whose dominant term, called the saddlepoint approximation, has a number of desirable features. The error incurred by its use is $O(n^{-1})$ as against the more usual $O(n^{-1/2})$ associated with the normal approximation.

The Saddlepoint Approximation

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be a vector of observations of n random variables with joint density $f(\mathbf{y})$. Suppose that the real random variable $S_n = S_n(\mathbf{y})$ has a density with respect to Lebesgue measure which depends on integer $n > N$ for some positive N. Let $\phi_n(z) = E(e^{izS_n})$ be the characteristic function of S_n where i is the imaginary unit. The cumulant generating function of S_n is $\psi_n(z) = \log \phi_n(z) = K_n(T)$ with $T = iz$. Whenever the appropriate derivatives exist, let $\partial^j \psi_n(\widetilde{z})/\partial z^j$ denote the jth order derivative evaluated at $z = \widetilde{z}$. The jth cumulant κ_{nj} of S_n, where it exists, satisfies the relation

$$i^j \kappa_{nj} = \frac{\partial^j \psi_n(0)}{\partial z^j}. \tag{1}$$

It is assumed that the derivatives $\partial^j \psi_n(z)/\partial z^j$ exist and are $O(n)$ for all z and $j = 1, 2, \ldots, r$ with $r \geq 4$. We use here partial derivatives because the functions involved may depend on something else, a parameter vector for example.

Let $h_n(x)$ be the density of the statistics $X_n = n^{-1/2}\{S_n - E(S_n)\}$. The characteristic function of X_n is

$$\begin{aligned}\phi_n^*(z) = E(e^{izX_n}) &= E\left(\exp\left\{i\frac{z}{\sqrt{n}}\{S_n - E(S_n)\}\right\}\right) \\ &= e^{-i\frac{z}{\sqrt{n}}E(S_n)} E\left\{e^{i\frac{z}{\sqrt{n}}S_n}\right\} \\ &= e^{-i\frac{z}{\sqrt{n}}E(S_n)} \phi_n\left(\frac{z}{\sqrt{n}}\right),\end{aligned} \tag{2}$$

where ϕ_n is the characteristic function of S_n.

Without loss of generality assume that $E(S_n) = 0$, therefore

$$\phi_n^*(z) = E(e^{izX_n}) = \phi_n\left(\frac{z}{\sqrt{n}}\right). \tag{3}$$

The cumulant generating function of X_n is

$$\psi_n^*(z) = \log \phi_n^*(z) = K_n^*(T), \tag{4}$$

with $T = iz$.

Let $\widehat{T} = i\widehat{z}$ be the root of the equation

$$\frac{\partial K_n^*(T)}{\partial T} = X_n. \tag{5}$$

The density function $h_n(x)$ of the statistics X_n is given by the usual Fourier inversion formula

$$\begin{aligned}h_n(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_n^*(z) e^{-izX_n} dz \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left\{\psi_n^*(z) - izX_n\right\} dz.\end{aligned} \tag{6}$$

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

where $\psi_n^*(z)$ was given in (4). It is convenient here to employ the equivalent inversion formula

$$h_n(x) = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \exp\left\{K_n^*(T) - TX_n\right\}dT, \qquad (7)$$

where $-c_1 < a < c_2$, $0 \le c_1 < \infty$, $0 \le c_2 < \infty$, but $c_1 + c_2 > 0$, thus either c_1 or c_2 may be zero, though not both, and $K_n^*(T)$ was defined in (4).

Let us write $T = \widehat{T} + iw$, where $\widehat{T}$ is the root of the Eq. (5). The argument then proceeds formally as follows. On the contour near $\widehat{T}$, the exponent of (7) can be written as

$$\begin{aligned} K_n^*(T) - TX_n &= K_n^*(\widehat{T}) - \widehat{T}X_n + iw\frac{\partial}{\partial T}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{2}(iw)^2\frac{\partial^2}{\partial T^2}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{6}(iw)^3\frac{\partial^3}{\partial T^3}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{24}(iw)^4\frac{\partial^4}{\partial T^4}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} + \cdots \\ &= K_n^*(\widehat{T}) - \widehat{T}X_n - \frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2} \\ &\quad - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3} \\ &\quad + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4} + \cdots, \end{aligned} \qquad (8)$$

where $\frac{\partial}{\partial T}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} = 0$ because $\widehat{T}$ is the root of (5).

Because of (8), the integrand of (7) can be written as

$$\begin{aligned} &\exp\left\{K_n^*(T) - TX_n\right\} \\ &= \exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\} \\ &\quad \times\left\{1 - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3} + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4}\right. \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2 + \cdots\right\}. \end{aligned} \qquad (9)$$

Using $T = \widehat{T} + iw$, we can transform from T to w in (7) resulting that

$$\begin{aligned} h_n(x) &= \frac{1}{2\pi}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad \times\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}\left\{1 - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right. \\ &\quad + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4} \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2 + \cdots\right\}dw. \end{aligned} \qquad (10)$$

The odd powers of w vanish on integration. On the other hand, for $j = 2, 3, \ldots$ and since $\frac{\partial^j}{\partial T^j}K_n(T)$ is $O(n)$

$$\begin{aligned} \frac{\partial^j K_n^*(T)}{\partial T^j} &= \frac{\partial^j}{\partial T^j}K_n\left(\frac{T}{\sqrt{n}}\right) = \frac{\partial^j}{\partial T^{*j}}K_n(T^*)\left(\frac{1}{\sqrt{n}}\right)^j \\ &= O\left(n^{-\frac{j}{2}+1}\right), \end{aligned} \qquad (11)$$

where $\mathbf{T}^* = \frac{\mathbf{T}}{\sqrt{n}}$. Therefore

$$\begin{aligned} h_n(x) &= \frac{1}{\sqrt{2\pi}}\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad \times\left\{1 + \frac{1}{n}Q_4\left(\widehat{T}\right) + \cdots\right\}, \end{aligned} \qquad (12)$$

where

$$\begin{aligned} Q_4\left(\widehat{T}\right) &= \frac{n\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\} \\ &\quad \times\left\{\frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4}\right. \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2\right\}dw. \end{aligned} \qquad (13)$$

Clearly, $Q_4\left(\widehat{T}\right)$ defined in (13) is n times the sum of two terms. The first of these terms is, apart from a multiplicative constant, $\frac{\partial^4 K_n^*(T)}{\partial T^4}$ times fourth order moments of a normal random variable with zero mean and variance $\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-1}$; and the second term is also a constant times $\left(\frac{\partial^3 K_n^*(T)}{\partial T^3}\right)^2$ and sixth order moments of a normal random variable with zero mean and variance $\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-1}$. Thus, because of (11), $Q_4\left(\widehat{T}\right) = O(1)$. Consequently, we write (12) as

$$h_n(x) = \widehat{h}_n(x)\left\{1 + O(n^{-1})\right\}, \qquad (14)$$

where

$$\widehat{h}_n(x) = \frac{1}{\sqrt{2\pi}}\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\}. \qquad (15)$$

The expression (15) receives the name of saddlepoint approximation to $h_n(x)$, been the error of approximation of order n^{-1}.

Daniels (1956) pointed out that when the constant term in the saddlepoint approximation is adjusted to make the integral over the whole sample space equal to unity, the order of magnitude of the error is reduced in a certain sense from n^{-1} to $n^{-3/2}$. He called this process renormalization.

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadística* (Journal of the Argentinean Statistical Society).

Cross References

►Approximations to Distributions
►Dispersion Models
►Edgeworth Expansion
►Exponential Family Models
►Inverse Sampling

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD Thesis, The London School of Economics and Political Science, University of London, England

Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312

Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650

Daniels HE (1956) The approximate distribution of serial correlation coefficients. Biometrika 43:169–185

Durbin J (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333

Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York

Phillips PCB (1978) Edgeworth and saddlepoint approximations in a first order autoregression. Biometrika 65:91–98

Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Sample Size Determination

Michael P. Cohen
Adjunct Professor
NORC at the University of Chicago, Chicago, IL, USA
Adjunct Professor
George Mason University, Fairfax, VA, USA

A common problem arising in statistics is to determine the smallest sample size needed to achieve a specified inference goal. Examples of inference goals include finding a 95% confidence interval for a given statistic of width no larger than a specified amount, or performing a hypothesis test at the 5% significance level with power no smaller than a specified amount. These examples and others are discussed more fully below.

Sample Size to Achieve a Given Variance or Relative Variance

One may want to estimate a parameter θ by an estimator $\hat{\theta}$ based on a sample of size n. Often the variance of $\hat{\theta}$, $\mathrm{var}(\hat{\theta})$, will have the form $\mathrm{var}(\hat{\theta}) = b/n$ for some known constant b. To achieve a variance of $\hat{\theta}$ no larger than a specified amount A, one simply sets $A = b/n$ and solves for n: $n = b/A$. The value of n must be an integer, so one takes n to be the smallest integer no smaller than b/A. Note that n is inversely related to the desired precision A.

It is more typically the case that b will depend on unknown parameters, usually including θ. Because the sample has not been selected yet, one must estimate the parameters from a previous sample or from other outside information. Precise values are not needed as one is usually satisfied with a conservative (that is, high) estimate for the required sample size n.

It is common to be interested in the *relative variance* $\frac{\mathrm{var}(\hat{\theta})}{\theta^2}$, also known as the square of the coefficient of variation or CV^2. In this case, one has

$$\frac{\mathrm{var}(\hat{\theta})}{\theta^2} = \frac{b}{\theta^2 n}$$

so to keep CV^2 less than a desired amount A, one sets $n = \frac{b}{\theta^2 A}$. Again, b and θ may need to be estimated from a previous sample or some outside source.

The variance of an estimated proportion $\hat{p}$ from a ►simple random sample of size n (from an infinite population) is

$$\mathrm{var}(\hat{p}) = \frac{p(1-p)}{n} = \frac{1}{4n} - \frac{(1/2-p)^2}{n} \le \frac{1}{4n}.$$

Therefore, to achieve a variance of $\hat{p}$ of at most A, it suffices that n be at least $\frac{1}{4A}$. For this conservation determination of the sample size, no estimation of unknown parameters is needed.

One can also consider the estimation of an estimated proportion $\hat{p}$ from a simple random sample of size n from a finite population of size N. In this case,

$$\mathrm{var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n} \le \left(1 - \frac{n}{N}\right)\frac{1}{4n}.$$

To achieve a variance of $\hat{p}$ of at most A as a conservative estimate, n must be at least

$$\frac{1}{4A + 1/N}.$$

Sample Size to Achieve a Given Power in a Hypothesis Test

In hypothesis testing, the probability of *type I error* (the probability of rejecting a null hypothesis when it, in fact, holds) is typically fixed at a predetermined level, called alpha (α). The value $\alpha = 5\%$ is very common. A sample size n is sought so that the test achieves a certain *type II error* rate (the probability of not rejecting the null hypothesis when a specific alternative actually holds), called beta (β). The *power* of a test is $1 - \beta$, the probability of rejecting the null hypothesis when a specific alternative holds. So sample size determination can be described as finding the smallest value of n so that for the predetermined α the power achieves some desired level for a fixed alternative. The term *statistical* ▶*power analysis* is frequently used as a synonym for sample size determination.

To be specific, suppose one wants to test that the mean μ of independent, identically normally distributed data is equal to μ_0 versus the alternative that the mean is greater than μ_0. One can write this as $H_0 : \mu = \mu_0$ versus $H_\alpha : \mu > \mu_0$. Suppose also that $\mu' > \mu_0$ is sufficiently far from μ_0 that the difference is deemed to be of practical significance in the subject-matter area of the test. Let Z be a standard normal random variable, Φ be its cumulative distribution function, and z_α be defined by $P(Z \geq z_\alpha) = \alpha$. Then it can be calculated that the type II error at $\mu', \beta(\mu')$, is

$$\begin{aligned}\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\ &= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

where σ^2 is the known variance of the data and n is the sample size. It follows from this that

$$-z_\beta = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}.$$

Solving for n, one gets

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2.$$

This sample size (adjusted upward to an integer value, if necessary) is needed to achieve a significance level of α and power of $1 - \beta(\mu')$ at μ'. The same sample size n applies when the alternative hypothesis is $H_\alpha : \mu < \mu_0$. For the two-sided alternative hypothesis $H_\alpha : \mu \neq \mu_0$, one has by a similar argument (involving an approximation) that

$$n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2.$$

For this testing problem, one is able to get explicit solutions. It is typical, however, to have to resort to complicated tables or, more recently, software, to get a solution.

Sample Size to Achieve a Given Width for a Confidence Interval

A $100(1-\alpha)\%$ ▶confidence interval for the mean μ of a normal population with known variance σ^2 is

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean. When n is reasonably large, say 30 or greater, this interval with σ replaced by $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ holds approximately when σ^2 is unknown.

The width of the interval is $w = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. So, solving for n, the sample size needed to achieve an interval of width w and confidence level $100(1-\alpha)\%$ is $n = 4\sigma^2\left(\frac{z_{\alpha/2}}{w}\right)^2$ (or $n = 4S^2\left(\frac{z_{\alpha/2}}{w}\right)^2$ when σ^2 unknown and $n \geq 30$).

As with hypothesis testing, the sample size problem for confidence intervals more typically requires tables or software to solve.

The Scope of Statistical Procedures for Sample Size Determination

Sample size determination arises in one sample problems, two sample problems, ▶analysis of variance, regression analysis, ▶analysis of covariance, multilevel models, survey sampling, nonparametric testing, ▶logistic regression, survival analysis, and just about every area of modern statistics. In the case of multilevel models (e.g., hierarchical linear models), one must determine the sample size at each level in addition to the overall sample size (Cohen 2005). A similar situation arises in sample size determination for complex sample surveys.

Software for Sample Size Determination

The use of software for sample size determination is highly recommended. Direct calculation is difficult (or impossible) in all but the simplest cases. Tables are cumbersome and often incomplete. Specific software products will not be recommended here, but we mention some to indicate the wide range of products available.

Statisticians who use SAS® should be aware that versions 9.1 and later include releases of PROC POWER and PROC GLMPOWER (PROC means "procedure" in SAS® and GLM stands for "general linear model") that are full featured.

SPSS has a product called SamplePower® that also has many features. Other commercial products include nQuery Advisor and PASS. G*Power is a free product. Sampsize is also free with an emphasis on survey sampling sample size calculations. A Web search will reveal many other products that should suit particular needs.

About the Author

Biography of Cohen is in ►Stratified Sampling.

Cross References

►Confidence Interval
►Power Analysis
►Significance Testing: An Overview
►Statistical Evidence

References and Further Reading

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale
Cohen MP (2005) Sample size considerations for multilevel surveys. Int Stat Rev 73:279–287
Dattalo P (2008) Determining sample size. Oxford University Press, New York

Sample Survey Methods

Peter Lynn
Professor of Survey Methodology
University of Essex, Colchester, UK

A sample survey can be broadly defined as an exercise that involves collecting standardised data from a sample of *study units* (e.g., persons, households, businesses) designed to represent a larger population of units, in order to make quantitative inferences about the population. Within this broad definition there is a large variety of different types of survey. Surveys can differ in terms of the type of data collected, the methods used to collect the data, the design of the sample, and whether data is collected repeatedly, either on the same sample or on different samples. Key features of a sample survey are:

Survey objectives must be clear and agreed at the outset, so that all features of the survey can be designed with these objectives in mind;
The target population – about which knowledge is required – must be defined. For example, it might be all persons usually resident in a particular town, or all farms within a national boundary;
The survey sample must be designed to represent the target population;
Relevant concepts must be addressed by the survey measures, so that the survey data can be used to answer important research questions;
The survey measures – which typically include questions, but could also include anthropometric measures, soil samples, etc – must be designed to provide accurate indicators of the concepts of interest;
Survey implementation should achieve the co-operation of a high proportion of sample members in a cost-efficient and timely manner.

The aim is to obtain relevant data that are representative, reliable and valid.

Representation concerns the extent to which the units in the data set represent the units in the target population and therefore share the pertinent characteristics of the population as a whole. This will depend on the identification of a sampling frame, the selection of a sample from that frame, and the attempts made to obtain data for the units in the sample.

Sampling frame. Ideally, this is a list of all units in the population, from which a sample can be selected. Sometimes the list pre-exists, sometimes it must be constructed especially for the survey, and sometimes a sampling method can be devised that does not involve the creation of an explicit list but is equivalent (Lynn 2002).

Sample design. In 1895 Anders Kiaer, founding Director of Statistics Norway, proposed sampling as a way of learning about a population without having to study every unit in the population. The basic statistical theory of probability sampling developed rapidly in the first half of the twentieth century and underpinned the growth of surveys. The essence is that units must be selected at random with known and non-zero selection probabilities. This enables unbiased estimation of population parameters and estimation of the precision (standard errors) of estimates (Groves et al. 2004, Chap. 4). Design features such as stratified sampling and multi-stage (clustered) sampling are commonly used within a probability sampling framework. Some surveys, particularly in the commercial sector, use non-probability methods such as quota sampling.

Non-response. Once a representative sample has been selected, considerable efforts are usually made to achieve the highest possible *response rate* (Lynn 2008). In many countries, high quality surveys of the general population typically achieve response rates in the range 60–80%, with rates above 80% being considered outstanding. The main reasons for non-response are usually *refusal* (unwillingness of sample member to take part) and *non-contact* (inability of the survey organisation to reach the sample member). Other reasons include an *inability* to take part, for example

due to language or ill health. Different strategies are used by survey organizations to minimize each of these types of non-response. Ultimately, non-response can introduce *bias* to survey estimates if the non-respondents differ from respondents in terms of the survey measures. Adjustment techniques such as *weighting* (Lynn 2004) can reduce the bias caused by non-response.

Obtaining reliable and valid data from respondents depends upon the measurement process. This includes development of *concepts* to be measured, development of *measures* of those concepts (e.g., survey questions), obtaining *responses* to the measures, and post-fieldwork *processing* (such as editing, coding, and combining the answers to a number of questions to produce derived variables). Failure of the obtained responses to correctly reflect the concept of interest is referred to as *measurement error* (Biemer et al. 1991). To minimise measurement error, survey researchers pay attention to cognitive response theory (Tourangeau et al. 2000), which describes four steps in the process of answering a survey question:

Understanding. The sample member must understand the question as intended by the researcher. This requires the question and the required response to be clear, simple, unambiguous and clearly communicated.

Recall. The sample member must be able to recall all the information that is required in order to answer the question. Question designers must be realistic regarding what respondents can remember and should provide tools to aid memory, if appropriate.

Evaluation. The sample member must process the recalled information in order to form an answer to the question.

Reporting. The sample member must be willing and able to communicate the answer. Various special techniques are used by survey researchers to elicit responses to questions on sensitive or embarrassing issues.

Two fundamental survey design issues with considerable implications are the following:

Data collection modes. There are several available methods to collect survey data (Groves et al. 2004, Chap. 5). An important distinction is between interviewer-administered methods (face-to-face personal interviewing, telephone interviewing) and self-completion methods (paper self-completion ►questionnaires, web surveys). With self-completion methods, the researcher usually has less control over factors such as who is providing the data and the order in which questions are answered, as well as having a limited ability to address respondent concerns and to provide help. Self-completion methods also require a higher degree of literacy and cognitive ability than interviews and so may be inappropriate for certain study populations. On the other hand, respondents may be more willing to reveal sensitive or embarrassing answers if there is no interviewer involved. There are often large differences in survey costs between the possible modes. This consideration often leads to surveys being carried out in a mode which might otherwise be thought sub-optimal.

Longitudinal designs. It is often beneficial to collect repeated measures from the same sample over time. This allows the measurement of change and identification of the ordering of events, which can shed light on causality. Surveys which collect data from the same units on multiple occasions are known as longitudinal surveys (Lynn 2009) and involve additional organisation and complexity. Some longitudinal social surveys have been running for several decades and are highly valued data sources.

About the Author

Dr. Peter Lynn is Professor of Survey Methodology, Institute for Social and Economic Research, University of Essex. He is Vice-President of the International Association of Survey Statisticians (2009–2011). He is Editor-in-Chief, *Survey Research Methods* (since 2005), and Director, UK Survey Resources Network (since 2008). He was Joint Editor, *Journal of the Royal Statistical Society Series A* (Statistics in Society) (2002–2005), and Editor of *Survey Methods Newsletter* (1996–2001). Dr Lynn is founding board member (since 2005) of the European Survey Research Association, Elected full member of the International Statistical Institute (since 2002) and Fellow of the Royal Statistical Society (since 1986). He has published widely on topics including survey non-response, weighting, data collection mode effects, respondent incentives, dependent interviewing, sample design, and survey quality. His recent publications include the book *Methodology of Longitudinal Surveys* (Editor, Wiley, 2009) and a chapter, *The Problem of Nonresponse*, in the *International Handbook of Survey Methodology* (Erlbaum, 2008). He was awarded the 2004 Royal Statistical Society Guy Medal in Bronze.

Cross References

►Balanced Sampling
►Business Surveys
►Census
►Cluster Sampling
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Inference Under Informative Probability Sampling
►Internet Survey Methodology: Recent Trends and Developments

►Multistage Sampling
►Non-probability Sampling Survey Methods
►Panel Data
►Questionnaire
►Repeated Measures
►Representative Samples
►Sampling From Finite Populations
►Superpopulation Models in Survey Sampling
►Telephone Sampling: Frames and Selection Techniques
►Total Survey Error

References and Further Reading

Biemer P, Groves RM et al (1991) Measurement errors in surveys. Wiley, New York
Groves RM, Fowler FJ et al (2004) Survey methodology. Wiley, New York
Lynn P (2002) Sampling in human studies. In: Greenfield T (ed) Research methods for postgraduates, 2nd edn. Arnold, London, pp 195–202
Lynn P (2004) Weighting. In: Kempf-Leonard K (ed) Encyclopedia of social measurement. Academic, New York, NY, pp 967–974
Lynn P (2008) The problem of nonresponse. In: deLeeuw E, Hox J, Dillman D (eds) The international handbook of survey methodology. Lawrence Erlbaum Associates, Mahwah, NJ, pp 35–55
Lynn P (ed) (2009) Methodology of longitudinal surveys. Wiley, New York
Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge

Sampling Algorithms

Yves Tillé
Professor
University of Neuchâtel, Neuchâtel, Switzerland

A sampling algorithm is a procedure that allows us to select randomly a subset of units (a sample) from a population without enumerating all the possible samples of the population.

More precisely, let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population and $s \subset U$ a sample or a subset of U. A sampling design $p(s)$ is a probability distribution on the set of all the subsets $s \subset U$, i.e., $p(s) \geq 0$ and

$$\sum_{s \subset U} p(s) = 1.$$

The inclusion probability $\pi_k = pr(k \in s)$ of a unit k is its probability of being selected in the sample s. The sum of the inclusion probabilities is equal to the expectation of the sample size n.

In many sampling problem, the number of possible samples is generally very large. For instance, if $N = 50$ and $n = 10$, the number of possible samples already equals 10,272,278,170. The selection of a sample by enumerating all the possible samples is generally impossible. A sampling algorithm is a method that allows bypassing this enumeration. There exists several class of methods:

- *Sequential algorithms*. In this case, there is only one reading of the population file. Each unit is successively examined and the decision of selection is irremediably taken.
- *One by one algorithms*. At each step, a unit is selected from the population until obtaining the fixed sample size.
- *Eliminatory algorithms*. At each step, a unit is removed from the population until obtaining the fixed sample size.
- *Rejective methods*. For instance, sample with replacement are generated until obtaining a sample without replacement. Rejective methods can be interesting if there exists a more general sampling design that is simpler than the design we want to implement.
- *Splitting methods*. This method described in Deville and Tillé (1998) starts with a vector of inclusion probability. At each step, this vector is randomly replaced by another vector until obtaining a vector containing only zeros and ones i.e., a sample.

The same sampling design can generally be implemented by using different methods. For instance, Tillé (2006) gives sequential, one by one, eliminatory algorithms for several sampling designs like simple random sampling with and without replacement and multinomial sampling.

The main difficulties however appears when the sample is selected with unequal inclusion probabilities without replacement and fixed sample size. The first proposed method was systematic sampling with unequal inclusion probabilities (Madow 1949). For this sequential algorithm, first compute the cumulated inclusion probabilities V_k. Next units such that

$$V_{k-1} \leq u + i - 1 < V_k, \qquad i = 1, 2, \ldots, n,$$

are selected, where u is a uniform continuous random variable in [0,1) and n is the sample size.

An interesting rejective procedure was proposed by Sampford (1967). Samples are selected with replacement. The first unit is selected with probability π_k/n, the $n-1$ other units are selected with probability

$$\frac{\pi_k}{n(1-\pi_k)} \left\{ \sum_{\ell=1}^{N} \frac{\pi_\ell}{n(1-\pi_\ell)} \right\}^{-1}.$$

The sample is accepted if n distinct units are selected, otherwise another sample is selected.

Chen et al. (1994) discussed the sampling design without replacement and fixed sample size that maximizes the ►entropy given by

$$I(p) = -\sum_{s \in U} p(s) \log p(s).$$

They gave a procedure for selecting a sample according this sampling design. Several other efficient algorithms that implement this sampling design are given in Tillé (2006).

Other methods have been proposed by Brewer (1975), Deville and Tillé (1998). A review is given in Brewer and Hanif (1983) and Tillé (2006). Other sampling algorithms allows us to solve more complex problems. For instance, the cube method (Deville and Tillé 2004) allows selecting balanced samples (see ►Balanced Sampling) in the sense that the ►Horvitz-Thompson estimator are equal or approximately equal to the population totals for a set of control variables.

About the Author

Yves Tillé is a professor and Director of the Institute of Statistics of the University of Neuchâtel. He was Associate editor of the *Journal of the Royal Statistical Society B* (2008–2009), *Survey Methodology* (2002–2009) and of *Metron* (2008–). He is author of several books in French and English, including *Sampling Algorithms*, Springer, 2006.

Cross References

►Balanced Sampling
►Entropy
►Horvitz–Thompson Estimator
►Randomization
►Sample Survey Methods
►Sequential Sampling

References and Further Reading

Brewer K (1975) A simple procedure for πpswor. Aust J Stat 17: 166–172
Brewer K, Hanif M (1983) Sampling with unequal probabilities. Springer-Verlag, New York
Chen S, Dempster A, Liu J (1994) Weighted finite population sampling to maximize entropy. Biometrika 81:457–469
Deville J-C, Tillé Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85:89–101
Deville J-C, Tillé Y (2004) Efficient balanced sampling: the cube method. Biometrika 91:893–912
Madow W (1949) On the theory of systematic sampling, II. Ann Math Stats 20:333–354
Sampford M (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54:499–513
Tillé Y (2006) Sampling algorithms. Springer-Verlag, New York

Sampling Distribution

David W. Stockburger
Deputy Director of Academic Assessment
US Air Force Academy
Emeritus Professor of Psychology
Missouri State University, Springfield, MO, USA

What is it?

The sampling distribution is a distribution of a sample statistic. When using a procedure that repeatedly samples from a population and each time computes the same sample statistic, the resulting distribution of sample statistics is a sampling distribution of that statistic. To more clearly define the distribution, the name of the computed statistic is added as part of the title. For example, if the computed statistic was the sample mean, the sampling distribution would be titled "the sampling distribution of the sample mean."

For the sake of simplicity let us consider a simple example when we are dealing with a small *discrete* population consisting of the first ten integers $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let us now repeatedly take random samples without replacement of size $n = 3$ from this population. The random sampling might generate sets that look like $\{8, 3, 7\}, \{2, 1, 5\}, \{6, 3, 5\}, \{10, 7, 5\} \ldots$ If the mean $(\overline{X})$ of each sample is found, the means of the above samples would appear as follows: 6, 2.67, 4.67, 7.33 ... How many different samples can we take, or put it differently, how many different sample means can we obtain? In our artificial example only 720, but in reality when we analyze very large populations, the number of possible different samples (of the same size) can be for all practical purposes treated as countless.

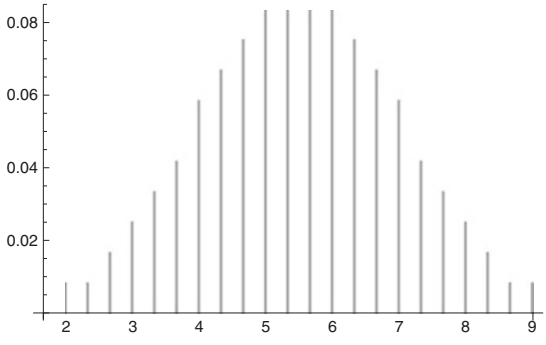
Once we have obtained sample means for all samples, we have to list all their different values and number of their occurrences (frequencies). Finally, we will divide each frequency with the total number of samples to obtain *relative frequencies* (empirical probabilities). In this way we will come up to a list of all possible sample means and their relative frequencies. When the population is discrete, that list is called the *sampling distribution* of that statistic. Generally, the sampling distribution of a statistic is a probability distribution of that statistic derived from all possible samples having the same size from the population.

When we are dealing with a *continuous* population it is impossible to enumerate all possible outcomes, so we have to rely on the results obtained in mathematical statistics (see section "►How Can Sampling Distributions be

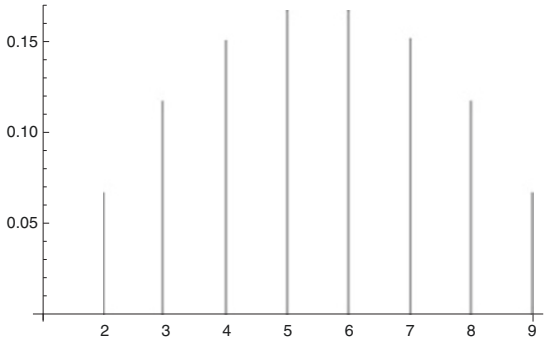
Constructed Mathematically?" of this paper for an example). Still, we can imagine a process that is similar to the one in the case of a discrete population. In that process we will take repeatedly thousands of different samples (of the same size) and calculate their statistic. In that way we will come to the relative frequency distribution of that statistic. The more samples we take, the closer this relative frequency distribution will come to the sampling distribution. Theoretically, as the number of samples approaches infinity our frequency distribution will approach the sampling distribution.

Sampling distribution should not be confused with a *sample* distribution: the latter describes the distribution of values (elements) in a *single* sample.

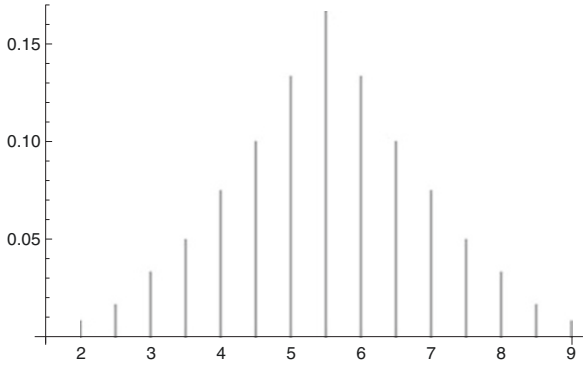
Referring back to our example, we can graphically display the sampling distribution of the mean as follows:



Every statistic has a sampling distribution. For example, suppose that instead of the mean, *medians* (M_d) were computed for each sample. That is, within each sample the scores would be rank ordered and the middle score would be selected as the median. Using the samples above, the medians would be: $7, 2, 5, 7 \ldots$ The distribution of the medians calculated from all possible different samples of the same size is called the sampling distribution of the median and could be graphically shown as follows:



It is possible to make up a new statistic and construct a sampling distribution for that new statistic. For example, by rank ordering the three scores within each sample and finding the mean of the highest and lowest scores a new statistic could be created. Let this statistic be called the mid-mean and be symbolized by $\overline{M}$. For the above samples the values for this statistic would be: $5.5, 3, 4.5, 7.5 \ldots$ and the sampling distribution of the mid-mean could be graphically displayed as follows:



Just as the population distributions can be described with parameters, so can the sampling distribution. The expected value and variance of any distribution can be represented by the symbols μ (mu) and σ^2 (Sigma squared), respectively. In the case of the sampling distribution, the μ symbol is often written with a subscript to indicate which sampling distribution is being described. For example, the expected value of the sampling distribution of the mean is represented by the symbol $\mu_{\overline{X}}$, that of the median by μ_{M_d}, and so on. The value of $\mu_{\overline{X}}$ can be thought of as the theoretical mean of the distribution of means. In a similar manner the value of μ_{M_d}, is the theoretical mean of a distribution of medians.

The square root of the variance of a sampling distribution is given a special name, the *standard error*. In order to distinguish different sampling distributions, each has a name tagged on the end of "standard error" and a subscript on the σ symbol. The theoretical *standard deviation* of the sampling distribution of the mean is called the standard error of the mean and is symbolized by $\sigma_{\overline{X}}$. Similarly, the theoretical standard deviation of the sampling distribution of the median is called the standard error of the median and is symbolized by σ_{M_d}.

In each case the standard error of the sampling distribution of a statistic describes the degree to which the computed statistics may be expected to differ from one another when calculated from a sample of similar size and selected from similar population models. The larger

the standard error of a given statistic, the greater the differences between the computed statistics for the different samples. From the example population, sampling method, and statistics described earlier, we would find $\mu_{\overline{X}} = \mu_{M_d} = \mu_{\overline{M}} = 5.5$ and $\sigma_{\overline{X}} = 1.46$, $\sigma_{M_d} = 1.96$, and $\sigma_{\overline{M}} = 1.39$.

Why is the Sampling Distribution Important – Properties of Statistics

Statistics have different properties as estimators of a population parameters. The sampling distribution of a statistic provides a window into some of the important properties. For example if the expected value of a statistic is equal to the expected value of the corresponding population parameter, the statistic is said to be unbiased. In the example above, all three statistics would be unbiased estimators of the population parameter μ_X.

Consistency is another valuable property to have in the estimation of a population parameter, as the *statistic* with the smallest standard error is preferred as an *estimator* of the corresponding population parameter, everything else being equal. Statisticians have proven that the standard error of the mean is smaller than the standard error of the median. Because of this property, the mean is generally preferred over the median as an estimator of μ_X.

Hypothesis Testing

The sampling distribution is integral to the hypothesis testing procedure. The sampling distribution is used in hypothesis testing to create a model of what the world would look like given the null hypothesis was true and a statistic was collected an infinite number of times. A single sample is taken, the sample statistic is calculated, and then it is compared to the model created by the sampling distribution of that statistic when the null hypothesis is true. If the sample statistic is unlikely given the model, then the model is rejected and a model with real effects is more likely. In the example process described earlier, if the sample $\{3, 1, 4\}$ was taken from the population described above, the sample mean (2.67), median (3), or mid-mean (2.5) can be found and compared to the corresponding sampling distribution of that statistic. The probability of finding a sample statistic of that size or smaller could be found for each e.g. mean ($p < .033$), median ($p < .18$), and mid-mean ($p < .025$) and compared to the selected value of alpha (α). If alpha was set to .05, then the selected sample would be unlikely given the mean and mid-mean, but not the median.

How Can Sampling Distributions be Constructed Mathematically?

Using advanced mathematics statisticians can prove that under given conditions a sampling distribution of some statistic must be a specific distribution. Let us illustrate this with the following theorem (for the proof see for example Hogg and Tanis (1997, p. 256)):

If $X_1, X_2, \ldots, X_n$ are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

then

$$\frac{(n-1)S^2}{\sigma^2} \text{ is } \chi^2(n-1).$$

The given conditions describe the assumptions that must be made in order for the distribution of the given sampling distribution to be true. For example, in the above theorem, assumptions about the sampling process (random sampling) and distribution of X (a normal distribution) are necessary for the proof.

Of considerable importance to statistical thinking is the sampling distribution of the mean, a theoretical distribution of sample means. A mathematical theorem, called the Central Limit Theorem, describes the relationship of the parameters of the sampling distribution of the mean to the parameters of the probability model and sample size. The Central Limit Theorem also specifies the form of the sampling distribution (Gaussian) in the limiting case.

Selection of Distribution Type to Model Scores

The sampling distribution provides the theoretical foundation to select a distribution for many useful measures. For example, the central limit theorem describes why a measure, such as intelligence, that may be considered a summation of a number of independent quantities would necessarily be (approximately) distributed as a normal (Gaussian) curve.

Monte Carlo Simulations

It is not always easy or even possible to derive the exact nature of a given sampling distribution using mathematical derivations. In such cases it is often possible to use Monte Carlo simulations to generate a close approximation to the true sampling distribution of the statistic. For example, a non-random sampling method, a non-standard

distribution, or may be used with the resulting distribution not converging to a known type of probability distribution. When much of the current formulation of statistics was developed, Monte Carlo techniques, while available, were very inconvenient to apply. With current computers and programming languages such as Wolfram Mathematica (Kinney 2009), Monte Carlo simulations are likely to become much more popular in creating sampling distributions.

Summary

The sampling distribution, a theoretical distribution of a sample statistic, is a critical concept in statistical thinking. The sampling distribution allows the statistician to hypothesize about what the world would look like if a statistic was calculated an infinite number of times.

About the Author

Dr. David W. Stockburger is currently the Deputy Director of Academic Assessment at the US Air Force Academy. He is an emeritus professor of psychology at Missouri State University where he taught from 1973 to 2002. His online introductory statistics text http://www.psychstat.missouristate.edu/IntroBook2/sbk.htm has been continuously available since 1996 and an intermediate text http://www.psychstat.missouristate.edu/multibook2/mlt.htm appeared in 1997. His online probability calculator (2001) replaced statistical tables and provided a visual representation of probability distributions, saving students countless hours of learning how to use statistical tables and providing an exact significance level. He has entries in "Encyclopedia of Measurement and Statistics" and "Encyclopedia of Research Design."

Cross References

►Bootstrap Methods
►Central Limit Theorems
►Cornish-Fisher Expansions
►Mean Median and Mode
►Monte Carlo Methods in Statistics
►Nonparametric Statistical Inference
►Significance Testing: An Overview
►Statistical Inference: An Overview

References and Further Reading

Hogg RV, Tanis EA (1997) Probability and statistical inference. 5th edn. Prentice Hall, Upper Saddle River, NJ

Kinney JJ (2009) A probability and statistics companion. Wiley, Hoboken, NJ

Sampling From Finite Populations

Jill M. Montaquila, Graham Kalton
Westat, Rockville, MD, USA

Introduction

The statistical objective in survey research and in a number of other applications is generally to estimate the parameters of a finite population rather than to estimate the parameters of a statistical model. As an example, the finite population for a survey conducted to estimate the unemployment rate might be all adults aged 18 or older living in a country at a given date. If valid estimates of the parameters of a finite population are to be produced, the finite population needs to be defined very precisely and the sampling method needs to be carefully designed and implemented. This entry focuses on the estimation of such finite population parameters using what is known as the *randomization* or *design-based approach*. Another approach that is particularly relevant when survey data are used for analytical purposes, such as for regression analysis, is known as the *superpopulation approach* (*see* ►Superpopulation Models in Survey Sampling).

This entry considers only methods for drawing probability samples from a finite population; *Nonprobability Sampling Methods* are reviewed in another entry. The basic theory and methods of probability sampling from finite populations were largely developed during the first half of the twentieth century, motivated by the desire to use samples rather than censuses (see ►Census) to characterize human, business, and agricultural populations. The paper by Neyman (1934) is widely recognized as a seminal contribution because it spells out the merits of *probability sampling* relative to purposive selection. A number of full-length texts on survey sampling theory and methods were published in the 1950's and 1960's including the first editions of Cochran (1977), Deming (1960), Hansen et al. (1953), Kish (1965), Murthy (1967), Raj (1968), Sukhatme et al. (1984), and Yates (1981). Several of these are still widely used as textbooks and references. Recent texts on survey sampling theory and methods include Fuller (2009), Lohr (2010), Pfeffermann and Rao (2009), Särndal et al. (1992), Thompson (1997), and Valliant et al. (2000).

Let the size of a finite population be denoted by N and let Y_i $(i = 1, 2, \ldots, N)$ denote the individual values of a variable of interest for the study. To carry forward the example given above, in a survey to estimate the unemployment rate, Y_i might be the labor force status of person (element) i. Consider the estimation of the population total

$Y = \sum_i^N Y_i$ based on a probability sample of n elements drawn from the population by sampling without replacement so that elements cannot be selected more than once. Let π_i denote the probability that element i is selected for the sample, with $\pi_i > 0$ for all i, and let π_{ij} denote the probability that elements i and j are jointly included in the sample. The sample estimator of Y can be represented as $\hat{Y} = \sum_i^N w_i Y_i$ where w_i is a random variable reflecting the sample selection, with $w_i = 0$ for elements that were not selected. The condition for $\hat{Y}$ to be an unbiased estimator of Y is that $E(w_i) = 1$. Now $E(w_i) = \pi_i w_i + (1-\pi_i)0$ so that for $\hat{Y}$ to be unbiased $w_i = \pi_i^{-1}$. The reciprocal of the selection probability, $w_i = \pi_i^{-1}$, is referred to as the *base weight*. The unbiased estimator for Y, $\hat{Y} = \sum_i^n w_i Y_i$, is widely known as the ▶Horvitz-Thompson estimator. The variance of $\hat{Y}$ is given by

$$
\begin{aligned}
V(\hat{Y}) &= \sum_i^N V(w_i) Y_i^2 + 2 \sum_i^N \sum_{j>i}^N Cov(w_i, w_j) Y_i Y_j \\
&= \sum_i^N \pi_i^{-1} (1 - \pi_i) Y_i^2 \\
&\quad + 2 \sum_i^N \sum_{j>i}^N \pi_i^{-1} \pi_j^{-1} (\pi_{ij} - \pi_i \pi_j) Y_i Y_j
\end{aligned}
$$

These general results cover a range of the different sample designs described below depending on the values of π_j and π_{ij}. The selection probabilities π_i appear in the estimator and, in addition, the joint selection probabilities π_{ij} appear in the variance. Note that when estimating the parameters of a finite population using the design-based approach for inference, the Y_i values are considered fixed; it is the w_i's that are the random variables.

The selection of a probability sample from a finite population requires the existence of a *sampling frame* for that population. The simplest form of sampling frame is a list of the individual population elements, such as a list of business establishments (when they are the units of analysis). The frame may alternatively be a list of clusters of elements, such as a list of households when the elements are persons. The initial frame may be a list of geographical areas that are sampled at the first stage of selection. These areas are termed *primary sampling units* (PSUs). At the second stage, subareas, or *second stage units*, may be selected within the sampled PSUs, etc. This design, which is known as an *area sample*, is a form of multistage sampling (see below).

The quality of the sampling frame has an important bearing on the quality of the final sample. An ideal sampling frame would contain exactly one listing for each element of the target population and nothing else. Sampling frames used in practice often contain departures from this ideal, in the form of noncoverage, duplicates, clusters, and ineligible units (see Kish 1965, Section 2.7, for a discussion of each of these frame problems). Issues with the sampling frames used in telephone surveys are discussed in the entry ▶Telephone Sampling: Frames and Selection Techniques. Sometimes, two or more sampling frames are used, leading to dual- or multiple-frame designs.

Sampling frames often contain auxiliary information that can be used to improve the efficiency of the survey estimators at the sample design stage, at the estimation stage, or at both stages. Examples are provided below.

Simple Random Sampling

A *simple random sample* is a sample design in which every possible sample of size n from the population of N elements has an equal probability of selection (*see* ▶Simple Random Sample). It may be selected by taking random draws from the set of numbers $\{1, 2, \ldots, N\}$. With simple random sampling, elements have equal probabilities of selection and simple random sampling is therefore *an equal probability selection method* (*epsem*).

Simple random sampling with replacement (SRSWR), also known as *unrestricted sampling*, allows population elements to be selected at any draw regardless of their selection on previous draws. Since elements are selected independently with this design, $\pi_{ij} = \pi_i \pi_j$ for all i, j. Standard statistical theory and analysis generally assumes SRSWR; this is discussed further in the entry ▶Superpopulation Models in Survey Sampling.

In *simple random sampling without replacement* (SRSWOR), also simply known as simple random sampling, once an element has been drawn, it is removed from the set of elements eligible for selection on subsequent draws. Since SRSWOR cannot select any element more than once (so that there are n distinct sampled elements), it is more efficient than SRSWR (i.e., the variances of the estimators are lower under SRSWOR than under SRSWR).

Systematic Sampling

In the simple case where the *sampling interval* $k = N/n$ is an integer, a *systematic sample* starts with a random selection of one of the first k elements on a list frame, and then selects every kth element thereafter. By randomly sorting the sampling frame, systematic sampling provides a convenient way to select a SRSWOR. Kish (1965, Section 4.1B)

describes various techniques for selecting a systematic sample when the sampling interval is not an integer.

If the sampling frame is sorted to place elements that are similar in terms of the survey variables near to each other in the sorted list, then systematic sampling may reduce the variances of the estimates in much the same way as proportionate stratified sampling does. Systematic sampling from such an ordered list is often described as *implicit stratification*. A general drawback to systematic sampling is that the estimation of the variances of survey estimates requires some form of model assumption.

Stratified Sampling

Often, the sampling frame contains information that may be used to improve the efficiency of the sample design (i.e., reduce the variances of estimators for a given sample size). *Stratification* involves using information available on the sampling frame to partition the population into L classes, or *strata*, and selecting a sample from each stratum. (*See* ►Stratified Sampling).

With *proportionate stratification*, the same sampling fraction (i.e., the ratio of sample size to population size) is used in all the strata, producing an epsem sample design. Proportionate stratification reduces the variances of the survey estimators to the extent that elements within the strata are homogeneous with respect to the survey variables.

With *disproportionate stratification*, different sampling fractions are used in the various strata, leading to a design in which selection probabilities vary. The unequal selection probabilities are redressed by the use of the base weights in the analysis. One reason for using a disproportionate stratified design is to improve the precision of survey estimates when the element standard deviations differ across the strata. Disproportionate stratified samples are widely used in business surveys for this reason, sampling the larger businesses with greater probabilities, and even taking all of the largest businesses into the sample (*see* ►Business Surveys). The allocation of a given overall sample size across strata that minimizes the variance of an overall survey estimate is known as *Neyman allocation*. If data collection costs per sampled element differ across strata, it is more efficient to allocate more of the sample to the strata where data collection costs are lower. The sample allocation that maximizes the precision of an overall survey estimate for a given total data collection cost is termed an *optimum allocation*.

A second common reason for using a disproportionate allocation is to produce stratum-level estimates of adequate precision. In this case, smaller strata are often sampled at above average sampling rates in order to generate sufficiently large sample sizes to support the production of separate survey estimates for them.

Cluster and Multistage Sampling

In many surveys, it is operationally efficient to sample clusters of population elements rather than to sample the elements directly. One reason is that the sampling frame may be a list that comprises clusters of elements, such as a list of households for a survey of persons (the elements). Another reason is that the population may cover a large geographical area; when the survey data are to be collected by face-to-face interviewing, it is then cost-effective to concentrate the interviews in a sample of areas in order to reduce interviewers' travel. The selection of more than one element in a sampled cluster affects the precision of the survey estimates because elements within the same cluster tend to be similar with respect to many of the variables studied in surveys. The homogeneity of elements within clusters is measured by the *intraclass correlation* (*see* ►Intraclass Correlation Coefficient). A positive intraclass correlation decreases the precision of the survey estimates from a cluster sample relative to a SRS with the same number of elements.

When the clusters are small, it is often efficient to include all the population elements in selected clusters, for example, to collect survey data for all persons in sampled households. Such a design is termed a *cluster sample* or more precisely a *single-stage cluster sample* (*see* ►Cluster Sampling).

Subsampling, or the random selection of elements within clusters, may be used to limit the effect of clustering on the precision of survey estimates. Subsampling is widely used when the clusters are large as, for example, is the case with areal units such as counties or census enumeration districts, schools, and hospitals. A sample design in which a sample of clusters is selected, followed by the selection of a subsample of elements within each sampled cluster is referred to as a *two-stage sample*. *Multistage sampling* is an extension of two-stage sampling, in which there are one or more stages of subsampling of clusters within the *first-stage units* (or primary sampling units, PSUs) prior to the selection of elements. In multistage sample designs, a key consideration is the determination of the sample size at each stage of selection. This determination is generally based on cost considerations and the contribution of each stage of selection to the variance of the estimator (*See* ►Multistage Sampling).

In general, large clusters vary considerably in the number of elements they contain. Sampling unequal-sized clusters with equal probabilities is inefficient and, with an overall epsem design, it fails to provide control on the

sample size. These drawbacks may be addressed by sampling the clusters with *probability proportional to size* (PPS) *sampling*. By way of illustration, consider a two-stage sample design. At the first stage, clusters are sampled with probabilities proportional to size, where size refers to the number of elements in a cluster. Then, at the second stage, an equal number of population elements is selected within each PSU. The resulting sample is an epsem sample of elements. This approach extends to multi-stage sampling by selecting a PPS sample of clusters at each stage through to the penultimate stage. At the last stage of selection, an equal number of population elements is selected within each cluster sampled at the prior stage of selection. In practice, the exact cluster sizes are rarely known and the procedure is applied with estimated sizes, leading to what is sometimes called *sampling with probability proportional to estimated size* (PPES).

Two-Phase Sampling

It would be highly beneficial in some surveys to use certain auxiliary variables for sample design, but those variables are not available on the sampling frame. Similarly, it may be beneficial to use certain auxiliary variables at the estimation stage, but the requisite data for the population are not available. In these cases, *two-phase sampling* (also known as *double sampling*) may be useful. As an example, consider the case where, if frame data were available for certain auxiliary variables, stratification based on these variables with a disproportionate allocation would greatly improve the efficiency of the sample design. Under the two-phase sampling approach, at the first phase, data are collected on the auxiliary variables for a larger preliminary sample. The first-phase sample is then stratified based on the auxiliary variables, and a second phase subsample is selected to obtain the final sample. To be effective, two-phase sampling requires that the first phase data collection can be carried out with little effort or resource requirements.

Estimation

As noted above, differential selection probabilities must be accounted for by the use of base weights in estimating the parameters of a finite population. In practice, adjustments are usually made to the base weights to compensate for sample deficiencies and to improve the precision of the survey estimates.

One type of sample deficiency is *unit nonresponse*, or complete lack of response from a sampled element. Compensation for unit nonresponse is typically made by inflating the base weights of similar responding elements in order to also represent the base weights of nonresponding eligible elements (see ▶Nonresponse in Surveys, Groves et al. 2001, and Särndal and Lundström 2005).

A second type of deficiency is *noncoverage*, or a failure of the sampling frame to cover some of the elements in the population. Compensation for noncoverage requires population information from an external source. Noncoverage is generally handled through a weighting adjustment using some form of *calibration* adjustment, such as post-stratification (see Särndal 2007). Calibration adjustments also serve to improve the precision of survey estimates that are related to the variables used in calibration.

A third type of deficiency is *item nonresponse*, or the failure to obtain a response to a particular item from a responding element. Item nonresponses are generally accounted for through *imputation*, that is, assigning values for the missing responses (see ▶Imputation and Brick and Kalton 1996).

In practice, samples from finite populations are often based on complex designs incorporating stratification, clustering, unequal selection probabilities, systematic sampling, and sometimes, two-phase sampling. The estimation of the variances of the survey estimates needs to take the complex sample design into account. There are two general methods for estimating variances from complex designs, known as the *Taylor Series* or *linearization* method and the *replication* method (including balanced repeated replications, jackknife repeated replications, and the bootstrap). See Wolter (2007) and Rust and Rao (1996). There are several software programs available for analyzing complex sample survey data using each method.

About the Authors

Jill Montaquila is an Associate Director of the statistical staff and a senior statistician at Westat. Dr. Montaquila is also a Research Associate Professor in the Joint Program in Survey Methodology at the University of Maryland. Her statistical interests cover various aspects of complex sample survey methodology, including random digit dialing survey methodology, address based sampling, and evaluations of nonsampling error. Dr. Montaquila has served as President of the Washington Statistical Society and is a Fellow of the American Statistical Association.

Graham Kalton is Chairman of the Board and Senior Vice President at Westat. He is also a Research Professor in the Joint Program in Survey Methodology at the University of Maryland. He has been at Westat since 1992. Earlier positions include: research scientist at the Survey Research Center of the University of Michigan, where he also held titles of Professor of Biostatistics and Professor of Statistics; Leverhulme Professor of Social Statistics at the University of Southampton; and Reader in Social Statistics

at the London School of Economics. Dr. Kalton has wide ranging interests in survey methodology and has published on several aspects of the subject. He has served as President of the International Association of Survey Statisticians and President of the Washington Statistical Society. He is a Fellow of the American Association for the Advancement of Science, a Fellow of the American Statistical Association, a National Associate of the National Academies, and an elected member of the International Statistical Institute.

Cross References

►Cluster Sampling
►Estimation
►Estimation: An Overview
►Horvitz–Thompson Estimator
►Imputation
►Intraclass Correlation Coefficient
►Multiple Imputation
►Multistage Sampling
►Non-probability Sampling Survey Methods
►Nonresponse in Surveys
►Sample Survey Methods
►Simple Random Sample
►Stratified Sampling
►Superpopulation Models in Survey Sampling
►Telephone Sampling: Frames and Selection Techniques
►Total Survey Error

References and Further Reading

Brick JM, Kalton G (1996) Handling missing data in survey research. Stat Methods Med Res 5:215–238

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Deming WE (1960) Sample design in business research. Wiley, New York

Fuller WA (2009) Sampling statistics. Wiley, New York

Groves RM, Dillman DA, Eltinge JA, Little RJA (eds) (2001) Survey nonresponse. Wiley, New York

Hansen MH, Hurwitz WN, Madow WG (1953) Sample survey methods and theory, vols I and II. Wiley, New York

Kish L (1965) Survey sampling. Wiley, New York

Lohr S (2010) Sampling: design and analysis, 2nd edn. Brooks/Cole, Pacific Grove, CA

Murthy MN (1967) Sampling theory and methods. Statistical Publishing Society, Calcutta, India

Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc 97(4):558–625

Pfeffermann D, Rao CR (eds) (2009) Handbook of statistics. Volume 29A, sample surveys: design, methods and application and volume 29B, sample surveys: inference and analysis. Elsevier, New York

Raj D (1968) Sampling theory. McGraw-Hill, New York

Rust KF, Rao JNK (1996) Variance estimation for complex surveys using replication techniques. Stat Methods Med Res 5:283–310

Särndal CE (2007) The calibration approach in survey theory and practice. Surv Methodol 33:99–119

Särndal CE, Lundström S (2005) Estimation in surveys with nonresponse. Wiley, New York

Särndal CE, Swensson B, Wretman J (1992) Model-assisted survey sampling. Springer-Verlag, New York

Sukhatme PV, Sukhatme BV, Sukhatme S, Asok C (1984) Sampling theory of surveys with applications, 3rd Rev. edn. Iowa State University Press and Indian Society of Agricultural Statistics, Ames, Iowa and New Delhi

Thompson ME (1997) Theory of sample surveys. Chapman and Hall, London

Valliant R, Dorfman AH, Royall RM (2000) Finite population sampling and inference: a prediction approach. Wiley, New York

Wolter K (2007) Introduction to variance estimation, 2nd edn. Springer, New York

Yates F (1981) Sampling methods for censuses and surveys, 4th edn. Charles Griffin, London

Sampling Problems for Stochastic Processes

Masayuki Uchida[1], Nakahiro Yoshida[2]
[1]Professor
Osaka University, Osaka, Japan
[2]Professor
University of Tokyo, Tokyo, Japan

Let $X = (X_t)_{t\in[0,T]}$ be a d-dimensional diffusion process defined by the following stochastic differential equation

$$dX_t = b(X_t, \alpha)dt + \sigma(X_t, \beta)dw_t, \qquad t \in [0, T], \quad X_0 = x_0,$$

where w is an r-dimensional Wiener process, $(\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta$, Θ_α and Θ_β are subsets of $\mathbf{R}^p$ and $\mathbf{R}^q$, respectively. Furthermore, b is an $\mathbf{R}^d$-valued function on $\mathbf{R}^d \times \Theta_\alpha$ and σ is an $\mathbf{R}^d \otimes \mathbf{R}^r$-valued function on $\mathbf{R}^d \times \Theta_\beta$. The drift function b and the diffusion coefficient function σ are known apart from the parameters α and β.

In the asymptotic theory of diffusion processes, the following two types of data are treated: (1) the continuously observed data and (2) the discretely observed data of diffusion processes. Concerning the first order asymptotic theory of diffusion processes based on the continuously observed data, Kutoyants extended Ibragimov and Has'minskii's approach (1981) to semimartingales, and many researchers made contributions to establish the asymptotic theory of semimartingales; see Kutoyants (1984, 1994, 2004) and Küchler and Sørensen (1997), Prakasa Rao (1999a, b) and references therein.

On the other hand, parametric estimation for discretely observed diffusion processes is highly important for

practical applications and now developing progressively. The data are discrete observations at regularly spaced time point on the fixed interval $[0, T]$, that is, $(X_{kh_n})_{0\le k\le n}$ with $nh_n = T$ and h_n is called a discretization step. The discretely observed data are roughly classified into the following three types:

(i) decreasing step size on a fixed interval: the observation time $T = nh_n$ is fixed and the discretization step h_n tends to zero as $n \to \infty$.
(ii) constant step size on an increasing interval: the discretization step is fixed ($h_n = \Delta$) and the observation time $T = nh_n = n\Delta$ tends to infinity as $n \to \infty$.
(iii) decreasing step size on an increasing interval: the discretization step h_n tends to zero and the observation time $T = nh_n$ tends to infinity as $n \to \infty$.

For the setting of type (i), Genon-Catalot and Jacod (1993) proposed estimators of the diffusion coefficient parameter β and they showed that the estimators are consistent, asymptotic mixed normal and asymptotic efficient. For the linearly parametrized case of diffusion coefficient, Yoshida (1997) obtained the asymptotic expansion for the estimator by means of the Malliavin calculus. Gobet (2001) proved the local asymptotic mixed normality for likelihoods by using the Malliavin calculus. On the other hand, for the drift parameter α, we can not generally construct even a consistent estimator under the setting of type (i). However, under the situation where diffusion term is very small, which is called a small diffusion process, we can estimate the drift parameter α. Genon-Catalot (1990) and Laredo (1990) proposed estimators of the drift parameter under the assumption that the diffusion coefficient is known, and they proved that the estimators have consistency, ►asymptotic normality and asymptotic efficiency. Uchida (2008) investigated asymptotic efficient estimators under the general asymptotics. Sørensen and Uchida (2003) obtained estimators of both the drift and the diffusion coefficient parameters simultaneously and investigated the asymptotic properties of their estimators. Gloter and Sørensen (2009) developed the result of Sørensen and Uchida (2003) under the general asymptotics.

As concerns the type (ii), Bibby and Sørensen (1995) proposed martingale estimating functions and obtained the estimators of the drift and the diffusion coefficient parameters from the martingale estimating functions. They proved that both estimators have consistency and asymptotic normality under ergodicity. Masuda (2005) showed the asymptotic normality of the moment estimator for a state space model involving jump noise terms.

Under the setting of type (iii), Prakasa-Rao (1983, 1988) are early work. As seen in Yoshida (1992a), the estimators of α and β jointly converge, and they are asymptotically orthogonal, however their convergence rates are different. Those authors' estimators are of maximum likelihood type in their settings. Kessler (1997) improved the condition on the sampling scheme and gave generalization. Gobet (2002) showed local asymptotic normality for the likelihood. A polynomial type large deviation inequality for an abstract statistical random field, which includes likelihood ratios of stochastic processes, enables to obtain the asymptotic behaviors of the Bayes and maximum likelihood type estimators; see Yoshida (2010) for details. For the asymptotic theory of diffusion processes with jumps, see for example Shimizu and Yoshida (2006).

Regarding the higher order asymptotic theory of diffusion processes, the asymptotic expansions have been studied; see Yoshida (1992b, 1997), Sakamoto and Yoshida (2004) and recent papers.

About the Author

Professor Nakahiro Yoshida was awarded the first Research Achievement Award by the Japan Statistical Society, for studies in the theory of statistical inference for stochastic processes and their applications (2007). He has also received the Analysis Prize, Mathematical Society of Japan (2006) and the Japan Statistical Society Award, Japan Statistical Society (2009). Professor Yoshida is Section Editor and Scientific Secretary, Bernoulli Society for Mathematical Statistics and Probability. Professors Nakahiro Yoshida and Masayuki Uchida are Associate editors of the *Annals of the Institute of Statistical Mathematics*.

Cross References

►Asymptotic Normality
►Brownian Motion and Diffusions
►Local Asymptotic Mixed Normal Family
►Stochastic Differential Equations
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Bibby BM, Sørensen M (1995) Martingale estimating functions for discretely observed diffusion processes. Bernoulli 1:17–39
Genon-Catalot V (1990) Maximum contrast estimation for diffusion processes from discrete observations. Statistics 21:99–116
Genon-Catalot V, Jacod J (1993) On the estimation of the diffusion coefficient for multidimensional diffusion processes. Ann Inst Henri Poincaré Prob Stat 29:119–151

Gloter A, Sørensen M (2009) Estimation for stochastic differential equations with a small diffusion coefficient. Stoch Process Appl 119:679–699
Gobet E (2001) Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach. Bernoulli 7:899–912
Gobet E (2002) LAN property for ergodic diffusions with discrete observations. Ann Inst Henri Poincaré Prob Stat 38:711–737
Ibragimov IA, Has'minskii RZ (1981) Statistical estimation. Springer Verlag, New York
Kessler M (1997) Estimation of an ergodic diffusion from discrete observations. Scand J Stat 24:211–229
Kutoyants YuA (1984) In: Prakasa Rao BLS (ed) Parameter estimation for stochastic processes. Heldermann, Berlin
Kutoyants YuA (1994) Identification of dynamical systems with small noise. Kluwer Dordrecht
Kutoyants YuA (2004) Statistical inference for ergodic diffusion processes. Springer-Verlag, London
Küchler U, Sørensen M (1997) Exponential families of stochastic processes. Springer, New York
Laredo CF (1990) A sufficient condition for asymptotic sufficiency of incomplete observations of a diffusion process. Ann Stat 18:1158–1171
Masuda H (2005) Classical method of moments for partially and discretely observed ergodic models. Stat Inference Stoch Proc 8:25–50
Prakasa Rao BLS (1983) Asymptotic theory for nonlinear least squares estimator for diffusion processes. Math Oper Forsch Stat Ser Stat 14:195–209
Prakasa Rao BLS (1988) Statistical inference from sampled data for stochastic processes. In: Contemporary mathematics, vol 80. American Mathematical Society, Providence, RI, pp 249–284
Prakasa Rao BLS (1999a) Statistical inference for diffusion type processes. Arnold, London
Prakasa Rao BLS (1999b) Semimartingales and their statistical inference. Boca Rotan, FL, Chapman & Hall/CRC
Sakamoto Y, Yoshida N (2004) Asymptotic expansion formulas for functionals of ϵ-Markov processes with a mixing property. Ann Inst Stat Math, 56:545–597
Shimizu Y, Yoshida N (2006) Estimation of parameters for diffusion processes with jumps from discrete observations. Stat Inference Stoch Process 9:227–277
Sørensen M, Uchida M (2003) Small-diffusion asymptotics for discretely sampled stochastic differential equations. Bernoulli 9:1051–1069
Uchida M (2008) Approximate martingale estimating functions for stochastic differential equations with small noises. Stoch Process Appl 118:1706–1721
Yoshida N (1992a) Estimation for diffusion processes from discrete observation, J Multivar Anal 41:220–242
Yoshida N (1992b) Asymptotic expansions of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe. Probab Theory Relat Field 92:275–311
Yoshida N (1997) Malliavin calculus and asymptotic expansion for martingales. Probab Theory Relat Fields 109:301–342
Yoshida N (2010) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Ann Inst Stat Math, doi:10.1007/s10463-009-0263-z

Scales of Measurement

Karl L. Wuensch
Professor
East Carolina University, Greenville, NC, USA

Measurement involves the assignment of scores (numbers or other symbols) to entities (objects or events) in such a way that the scores carry information about some characteristic of the measured entities. With careful consideration of the method by which the scores have been assigned, one can classify the method of measurement as belonging to one or more "scales of measurement." S.S. Stevens (1951) defined four scales of measurement: nominal, ordinal, interval, and ratio. Membership in one or more of these categories depends on the extent to which empirical relationships among the measured entities correspond to numerical relationships among the scores.

If the method of measurement produces scores that allow one to determine whether the measured entities are or are not equivalent on the characteristic of interest, then the scale is referred to as "nominal." For example, I ask the students in my class to take out all of their paper money, write their university identification number on each bill, and deposit all the bills in a bag. I then shake the bag and pull out two bills. From the identification numbers on the bills, I can determine whether or not the same student contributed both bills. The attribute of interest is last ownership of the bill, and the scores allow one to determine whether or not two bills are equivalent on that characteristic – accordingly, the identification number scores represent a nominal scale. "Nominal" derives from the Latin "nomen," name. Nominal scores may be no more than alternative names for entities.

If the scores can be employed to determine whether two entities are equivalent or not on the measured characteristic and, if they are not equivalent, which entity has the greater amount of the measured characteristic, then the scale is "ordinal." The order of the scores is the same as the order of the true amounts of the measured attribute. The identification numbers my students wrote on their bills would not allow one to determine whether "004387" represents more or less of something than does "093752." Imagine that I throw all the money out the window and then invite the students to retrieve the bills. My associate, outside, assigns to the students the ordinal scores shown in Table 1. The measured attribute is time taken to retrieve

Scales of Measurement. Table 1 Relationship between true scores and observed scores

Entity	A	B	C	D	E
True Score	1.0	2.0	4.0	8.0	9.0
Ordinal Score	0.5	0.6	0.7	1.1	1.5
Interval Score	12.0	14	18.0	26.0	28.0
Ratio Score	2.0	4.0	8.0	16.0	18.0

a bill, and the order of the scores is the same as the order of the magnitudes of the measured attribute. If Student A obtains a score of .5 and Student B a score of .6, I am confident that they differ on retrieval time and that Student B took longer than Student A.

Scale of measurement can be inferred from the nature of the relationship between the "observed scores" (the measurements) and the "true scores" (the true amounts of the measured characteristic) (Winkler and Hays 1975, pp. 277–282). If that relationship is positive monotonic, then the scale of measurement is ordinal. Notice that the ordinal scores in Table 1 are related to the true scores in a positive monotonic fashion.

The ordinal scores in Table 1 do not allow one to establish the equivalence of differences or to order differences. Consider the differences between A and B and between D and E. The true scores show that the differences are equivalent, but the ordinal scores might lead one to infer that the difference between D and E is greater than the difference between A and B. Also, the ordinal scores might lead one to infer that the difference between C and D (0.4) is equivalent to the difference between D and E (0.4), but the true scores show that not to be true.

If the relationship between the observed scores and the true scores is not only positive monotonic but also linear, then one will be able to establish the equivalence of differences and will be able to order differences. Such a scale is called "interval." My hypothetical associate used a mechanical device to measure the retrieval times, obtaining the interval scores in Table 1. From these observed scores, one would correctly infer that the difference between A and B is equivalent to the difference between D and E and that the difference between C and D is greater than the difference between D and E.

For the interval scores in Table 1, the function relating the measurements (m) to the true scores (t) is $m = 10 + 2t$. This hypothetical interval scale does not have a "true zero point." That is, it is not true that an entity that has absolutely none of the measured characteristic will obtain a measurement of zero. In this case, it will obtain a measurement of 10. This is problematic if one wishes to establish the equivalences of and orders of ratios of measurements. With the interval data one might infer that the ratio D/C > C/B > B/A, but the true scores show that these ratios are all equivalent. To achieve a ratio scale, the function relating the measurements to the true scores must not only be positive linear but also must have an intercept of zero. For the hypothetical ratio data in Table 1, that function is $m = 0 + 2t$. With the ratio scale the ratios of observed scores are identical to the corresponding ratios of the true scores.

Stevens (1951) argued that scale of measurement is an important consideration when determining the type of statistical analysis to be employed. For example, the mode was considered appropriate for any scale, even a nominal scale. If a fruit basket contains five apples, four oranges, and nine bananas, the modal fruit is a banana. The median was considered appropriate for any scale that was at least ordinal. Imagine that we select five fruits, identified as A, B, C, D, and E. Their true weights are 1.5, 3, 4.5, 9, and 27, and their ordinal scores are 1, 2, 3, 4, and 5. The entity associated with the median is C regardless of whether you use the true scores of the ordinal scores. Interval scores 4, 7, 10, 19, and 55 have a linear relationship with the true scores, $m = 1 + 2t$. The mean true score, 9, is associated with Entity D, and the mean interval score, 19, is also associated with Entity D. With the ordinal scores, however, the mean score, 3, is associated with Entity B.

There has been considerable controversy regarding the role that scale of measurement should play when considering the type of statistical analysis to employ. Most controversial has been the suggestion that parametric statistical analysis is appropriate only with interval or ratio data, but that nonparametric analysis can be employed with ordinal data. This proposition has been attacked by those who opine that the only assumptions required when employing parametric statistics are mathematical, such as homogeneity of variance and normality (Gaito 1980; Velleman and Wilkinson 1993). Defenders of the measurement view have argued that researchers must consider scale of measurement, the relationship between true scores and observed scores, because they are interested in making inferences about the constructs underlying the observed scores (Maxwell and Delaney 1985; Townsend and Ashby 1984). Tests of hypotheses that groups have identical means

on an underlying construct or that the Pearson ρ between two underlying constructs is zero do not require interval level data given the usual assumptions of homogeneity of variance and normality, but with non-interval data the effect size estimates will not apply to the underlying constructs (Davison and Sharma 1988).

When contemplating whether the observed scores to be analyzed represent an interval scale or a non-interval, ordinal scale, one needs makes a decision about the nature of the relationship between the true scores and the observed scores. If one conceives of true scores as part of some concrete reality, the decision regarding scale of measurement may come down to a matter of faith. For example, how could one know with certainty whether or not the relationship between IQ scores and true intelligence is linear? One way to avoid this dilemma is to think of reality as something that we construct rather than something we discover. One can then argue that the results of parametric statistical analysis apply to an abstract reality that is a linear function of our measurements. Conceptually, this is similar to defining a population on a sample rather than the other way around – when we cannot obtain a true random sample from a population, we analyze the data we can obtain and then make inferences about the population for which our data could be considered random.

About the Author

For biography *see* the entry ►Chi-Square Tests.

Cross References

►Rating Scales
►Scales of Measurement and Choice of Statistical Methods
►Variables

References and Further Reading

Davison ML, Sharma AR (1988) Parametric statistics and levels of measurement. Psychol Bull 104:137–144
Gaito J (1980) Measurement scales and statistics: resurgence of an old misconception. Psychol Bull 87:564–567
Maxwell SE, Delaney HD (1985) Measurement and statistics: an examination of construct validity. Psychol Bull 97:85–93
Stevens SS (1951) Mathematics, measurement, and psychophysics. In: Stevens SS (ed) Handbook of experimental psychology. Wiley, New York, pp 1–49
Townsend JT, Ashby FG (1984) Measurement scales and statistics: the misconception misconceived. Psychol Bull 96:394–401
Velleman PF, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47:65–72
Winkler RL, Hays WL (1975) Statistics: probability, inference, and decision, 2nd edn. Holt Rinehart and Winston, New York

Scales of Measurement and Choice of Statistical Methods

Donald W. Zimmerman
Professor Emeritus
Carleton University, Ottawa, ON, Canada

During the last century, it was conventional in many disciplines, especially in psychology, education, and social sciences, to associate statistical methods with a hierarchy of levels of measurement. The well-known classification proposed by Stevens (1946) included nominal, ordinal, interval, and ratio scales, defined by increasingly stronger mathematical restrictions. It came to be generally believed that the use of statistical significance tests in practice required choosing a test to match the scale of measurement responsible for the data at hand. Classes of appropriate statistical methods were aligned with the hierarchy of levels of measurement.

In research studies in psychology and education, the most relevant distinction perhaps was the one made between interval scales and ordinal scales. The Student *t* test (see ►Student's *t* Tests), the ANOVA *F* test, and regression methods were deemed appropriate for interval measurements, and nonparametric tests, such as the ►Wilcoxon–Mann–Whitney test and the Kruskal–Wallis test were appropriate for ordinal measurements.

Despite the widespread acceptance of these ideas by many statisticians and researchers, there has been extensive controversy over the years about their validity (see, for example, Cliff and Keats 2003; Maxwell and Delaney 1985; Michell 1986; Rozeboom 1966; Velleman and Wilkinson 1993; Zimmerman and Zumbo 1993). The mathematical theory eventually included more refined definitions of scales of measurement and additional types of scales (Luce 2001; Narens 1981), but the fourfold classification persisted for a long time in textbooks and research articles.

Scales of Measurement and Distributional Assumptions

The derivation of all significance tests is based on assumptions about probability distributions, such as independence, normality, and equality of the variances of separate groups, and some tests involve more restrictive assumptions than others. In many textbooks and research papers, the requirement of a specific level of measurement was placed on the same footing as these

distributional assumptions made in the mathematical derivation of a test statistic. For example, the Student *t* test and ANOVA *F* test were widely believed to assume three things: normality, homogeneity of variance, and interval measurement, while a nonparametric test such as the Wilcoxon–Mann–Whitney test is presumably free from the two distributional assumptions and requires only ordinal measurement. The assumption of within-sample independence is part of the definition of random sampling, and it is typically taken for granted that the data at hand meets that requirement before a test is chosen.

Many researchers believed that the parametric methods are preferable when all assumptions are satisfied, because nonparametric tests discard some information in the data and have less power to detect differences. Furthermore, the parametric methods were considered to be robust in the sense that a slight violation of assumptions does not lessen their usefulness in practical research. Early simulation studies, such as the one by Boneau (1960), were consistent with these ideas.

Some complications arose for the orderly correspondence of scales and statistics when researchers began to investigate how the Type I and Type II errors of both parametric and nonparametric significance tests depend on properties of standard probability densities. It was found that the nonparametric tests were often more powerful than their parametric counterparts for quite a few continuous densities, such as the exponential, lognormal, mixed-normal, Weibull, extreme value, chi-square, and others familiar in theoretical statistics. The power advantage of the nonparametric tests often turned out to be quite large (see, for example, Blair and Higgins 1980; Lehmann 1975; Randles and Wolfe 1979; Sawilowsky and Blair 1992; Zimmerman and Zumbo 1993). The superiority of nonparametric rank methods for many types of non-normal data has been extensively demonstrated by many simulation studies.

It can be argued that samples from one of these continuous densities by definition conform to interval measurement. That is, equal intervals are assumed in defining the parameters of the probability density. For this reason it is legitimate to employ *t* and *F* tests of location on sequences of random variates generated by different computer programs and obtain useful information. Similarly, the scaling criteria imply that calculation of means and variances is appropriate only for interval measurement, but it has become clear that slight violations of "homogeneity of variance" have severe consequences for both parametric and nonparametric tests.

Rank Transformations and Appropriate Statistics

In the controversies surrounding the notion of levels and measurement, theorists have tended to overlook the implications of a procedure known as the *rank transformation*. It was discovered that the *large-sample* normal approximation form of the Wilcoxon–Mann–Whitney test is equivalent to the Student *t* test performed on ranks replacing the original scores and that the Kruskal–Wallis test is equivalent to the ANOVA *F* test on ranks (Conover and Iman 1981). In the Wilcoxon–Mann–Whitney test, two samples of scores of size n_1 and n_2 are combined and converted to a single series of ranks, that is, integers from 1 to n_1+n_2. Similarly, in one-way ANOVA, scores in *k* groups are combined and converted to $n_1 + n_2 + \ldots + n_k$ ranks. Then, the scores in the original samples are replaced by their corresponding ranks in the combined group.

The above equivalence means that this rank transformation followed by the usual Student *t* test on the ranks replacing the initial scores leads to the same statistical decision as calculating and comparing rank sums, as done by a Wilcoxon–Mann–Whitney test. The Type I and Type II error probabilities turn out to be the same in both cases. That is true irrespective of the distributional form of the original data. If a Student *t* test performed on ranks is not appropriate for given data, then the Wilcoxon–Mann–Whitney test is not appropriate either, and vice versa.

Considered together with the power superiority of nonparametric tests for various non-normal densities, these findings imply that the power of *t* and *F* tests often can be increased by transforming interval data to ordinal data. Arguably, the main benefit of converting to ranks is not a change in scale, but rather augmentation of the robustness of the *t* and *F* tests. At first glance it seems paradoxical that statistical power can be increased, often substantially, by discarding information. However, one should bear in mind that conversion to ranks not only replaces real numbers by integers, but also alters the shape of distributions. Whatever the initial form of the data, ranks have a rectangular distribution, and, as noted before, the shape of non-normal distributions, especially those with heavy tails and extreme outlying values, certainly influences the power, or the extent of the loss of power, of significance tests.

Otherwise expressed, changing the distributional form of the data before performing a significance test appears to be the source of the power advantages, not the details of calculating rank-sums and finding quantiles of the resulting test statistic from a unique formula.

The rank transformation concept, together with the fact that unequal variances of scores in several groups is inherited by unequal variances of the corresponding ranks in the same groups, also provides a rationale for the dependence of both parametric and nonparametric tests on homogeneity of variance (Zimmerman 1996).

Another finding that is difficult to reconcile with notions of scaling is the fact that the beneficial properties of rank tests can be maintained despite alteration of the ranks in a way that modifies the scale properties, sometimes substantially. For example, small random numbers can be added to ranks, or the number of ranks can be reduced in number, with little effect on the power of the *t* and *F* tests under a rank transformation. That is, combining ranks 1, 2, 3, and 4 all into the value 1, ranks 5, 6, 7, and 8 into the value 2, and so on, has little influence on the power of the test when sample sizes are moderately large.

A quick illustration of these properties of scores and ranks is provided by Table 1, which gives the probability of rejecting H_0 by three significance tests at the 0.05 level. These computer simulations consisted of 50,000 pairs of independent samples of size 50 from normal and seven non-normal distributions, generated by a *Mathematica* program. The columns, labeled *t* represent the Student *t* test, those labeled *W* are the Wilcoxon–Mann–Whitney test, and those labeled *m* are the *t* test performed on *modified ranks*.

In this modification, all scores from both groups were combined and ranked as usual. Then, instead of transforming to integers, each original score was replaced by the median of all higher scores in the ranking; that is, the lowest score, ranked 1, was replaced by the median of all the higher scores ranked from 2 to $n_1 + n_2$, the score ranked 2 was replaced by the median of scores ranked from 3 to $n_1 + n_2$, and so on. Finally, the scores in the two initial groups were replaced by their corresponding modified ranks, and the significance test was performed.

This procedure resulted in a kind of hybrid ordinal/interval data not too different from ordinary ranks, whereby the real values of the original scores were retained, the distribution shape was compressed, and ▶outliers were eliminated. Table 1 shows that the Type I error rates of the *t* test on these modified ranks were close to those of ordinary ranks for the various distributions. Moreover, the *t* test on the modified values was nearly as powerful as the Wilcoxon–Mann–Whitney test for two distributions where the ordinary *t* test is known to be superior, and it was considerably more powerful than the *t* test and somewhat more powerful than the Wilcoxon–Mann–Whitney test for distributions for which the nonparametric test is known to be superior.

All these facts taken together imply there is not a one-to-one correspondence between the hierarchy of levels of measurement and methods that are appropriate for making correct statistical decisions. Transforming data so that it conforms to the assumptions of a significance test is not itself unusual, because for many years statisticians employed square-root, reciprocal, and logarithmic

Scales of Measurement and Choice of Statistical Methods. Table 1 Type I error rates and power of Student t test, Wilcoxon–Mann–Whitney test, and t test on modified ranks, 50,000 iterations at 0.05 level, samples from normal and seven non-normal distributions

	$\mu_1 - \mu_2 = 0$			$\mu_1 - \mu_2 = 0.3\sigma$			$\mu_1 - \mu_2 = 0.6\sigma$		
Distribution	t	W	m	t	W	m	t	W	m
Normal	0.051	0.051	0.053	0.314	0.298	0.295	0.847	0.831	0.812
Exponential	0.049	0.049	0.050	0.331	0.615	0.689	0.842	0.978	0.995
Mixed-normal	0.052	0.051	0.050	0.336	0.952	0.967	0.842	1.000	1.000
Lognormal	0.041	0.051	0.051	0.393	0.913	0.962	0.841	0.999	1.000
Extreme value	0.048	0.048	0.049	0.329	0.380	0.426	0.842	0.899	0.934
Uniform	0.049	0.048	0.049	0.311	0.294	0.310	0.845	0.798	0.840
Half-normal	0.049	0.050	0.051	0.318	0.385	0.420	0.837	0.890	0.943
Chi-square	0.049	0.049	0.050	0.326	0.489	0.551	0.845	0.958	0.987

transformations. The rank transformation can be regarded as a member of the same broad class of methods as those procedures. Unlike those methods, it is not continuous and has no inverse. That can be an advantage, because, by substituting small integers for intervals of real numbers, it lessens skewness and eliminates outliers.

As we have seen, the rank transformation in several instances is *equivalent* to a corresponding nonparametric test, in the sense that both either reject or fail to reject H_0 for given data. The earlier normalizing transformations do not possess such equivalences with well-known nonparametric methods. Each is best suited to a specific problem, such as stabilizing variances or changing the shape of a particular distribution, whereas conversion to ranks is an omnibus transformation that always brings data into a rectangular form with no outliers. Also, it is possible to reverse the perspective and regard the Wilcoxon–Mann–Whitney test and the Kruskal–Wallis test as having an affinity with those normalizing transformations, because the conversion to ranks, not the specific formula used in calculations, is apparently what makes the difference.

Conclusion

When all is said and done, *the theory of scales of measurement, although interesting and informative in its own right, is not closely related to practical decision-making in applied statistics*. Present evidence suggests that the mathematical property most relevant to choice of statistics in research is the probability distribution of the random variable that accounts for the observed data.

Caution is needed in making choices, and the rationale for a decision is likely to be more subtle and complex than the prescriptions in textbooks and software packages. In practice, the shape of a population distribution is not usually known with certainty. The degree of violation of assumptions fluctuates from sample to sample along with the estimates of the parameters, no matter what the population may be and what measurement procedures are used. Basing the choice of an appropriate test on inspection of samples, or even on preliminary significance tests performed to assess the validity of assumptions, can lead to incorrect statistical decisions with high probability.

About the Author

Dr. Donald W. Zimmerman is Professor Emeritus of Psychology at Carleton University in Ottawa, Ontario, Canada, and is currently living in Vancouver, British Columbia, Canada. He is the author of over 160 papers in peer-reviewed psychological and statistical journals, and he has served as an editorial consultant and reviewed manuscripts for 13 journals. His teaching and research interests have been in the areas of test theory, statistics, learning theory and conditioning, and the philosophy of science.

Cross References

►Analysis of Variance
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Rank Transformations
►Scales of Measurement
►Significance Testing: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Student's *t*-Tests
►Validity of Scales
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Blair RC, Higgins JJ (1980) The power of *t* and Wilcoxon statistics: a comparison. Eval Rev 4:645–655

Boneau CA (1960) The effects of violation of assumptions underlying the *t*- test. Psychol Bull 57:49–64

Cliff N, Keats JA (2003) Ordinal measurement in the behavioral sciences. Mahwah, Erlbaum, NJ

Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 35:124–129

Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco

Luce RD (2001) Conditions equivalent to unit representations of ordered relational structures. J Math Psychol 45:81–98

Maxwell SE, Delaney HD (1985) Measurement and statistics: an examination of construct validity. Psychol Bull 97:85–93

Michell J (1999) Measurement in psychology – a critical history of a methodological concept. Cambridge University Press, Cambridge

Narens L (1981) On the scales of measurement. J Math Psychol 24:249–275

Randles RH, Wolfe DA (1979) Introduction to the theory of nonparametric statistics. Wiley, New York

Rozeboom WW (1966) Scaling theory and the nature of measurement. Synthese 16:170–233

Sawilowsky SS, Blair RC (1992) A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychol Bull 111:352–360

Stevens SS (1946) On the theory of scales of measurement. Science 103:677–680

Velleman PJ, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47:65–72

Zimmerman DW (1996) A note on homogeneity of variance of scores and ranks. J Exp Educ 4:351–362

Zimmerman DW, Zumbo BD (1993) The relative power of parametric and nonparametric statistical methods. In: Keren G, Lewis C (eds) A handbook for data analysis in the behavioral sciences. Erlbaum, Mahwah, NJ

Seasonal Integration and Cointegration in Economic Time Series

Svend Hylleberg
Professor, Dean of the Faculty of Social Sciences
University of Aarhus, Aarhus C, Denmark

Introduction

A simple filter often applied in empirical econometric work is the seasonal difference filter $(1 - L^s)$, where s is the number of observations per year, typically $s = 2, 4, 12$ or 52. The seasonal differencing assumes that there are unit roots at all the seasonal frequencies. The seasonal difference filter can be written as the product of $(1 - L)$ and the seasonal summation filter $S(L)$, which for quarterly data is $S(L) = (1+L+L^2+L^3)$. The quarterly seasonal summation filter has the real root -1 and the two complex conjugate roots $\pm i$.

The existence of seasonal unit roots in the data generating process implies a varying seasonal pattern where "Summer may become Winter." In most cases, such a situation is not feasible and the findings of seasonal unit roots should be interpreted with care and taken as an indication of a varying seasonal pattern, where the unit root model is a parsimonious approximation and not the true DGP.

The idea that the seasonal components of a set of economic time series are driven by a smaller set of common seasonal features seems a natural extension of the idea that the trend components of a set of economic time series are driven by common trends. In fact, the whole business of seasonal adjustment may be interpreted as an indirect approval of such a view.

If the seasonal components are integrated, the idea immediately leads to the concept of seasonal cointegration, introduced in the paper by Hylleberg et al. (1990). In case the seasonal components are stationary, the idea leads to the concept of seasonal common features, see Engle and Hylleberg (1996), while so-called periodic cointegration considers cointegration season by season, introduced by Birchenhal et al. (1989). For a recent survey see Brenstrup et al. (2004).

Seasonal Integration

In general, consider the autoregressive representation $\phi(L)\, y_t = \varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma^2)$, where $\phi(L)$ is a finite lag polynomial. Suppose $\phi(L)$ has all its roots outside the unit circle except for possible unit roots at the long-run frequency $\omega = 0$ corresponding to $L = 1$, semiannual frequency $\omega = \pi$ corresponding to $L = -1$, and annual frequencies $\omega = \left\{\frac{\pi}{2}, \frac{3\pi}{2}\right\}$ corresponding to $L = \pm i$.

Dickey et al. (1984) suggested a simple test for seasonal unit roots in the spirit of the ▶Dickey – Fuller test for long-run unit roots. They suggested estimating the auxiliary regression $\left(1 - L^4\right) y_t = \pi_0 y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma^2)$. The DHF test statistic is the "t-value" corresponding to π_0, which has a non-standard distributed tabulated in Dickey et al. (1984). This test, however, is a joint test for unit roots at the long-run and all the seasonal frequencies.

In order to construct a test for each individual unit root and overcome the lack of flexibility in the DHF test, Hylleberg et al. (1990) refined this idea. By use of the result that any lag polynomial of order p, $\phi(L)$, with possible unit roots at each of the frequencies $\omega = 0, \pi, [\pi/2, 3\pi/2]$, can be written as $\phi(L) = \sum_{k=1}^{4} \frac{\xi_k \Delta(L)(1-\delta_k(L))}{\delta_k(L)} + \phi^*(L)\Delta(L)$, $\delta_k(L) = 1 - \frac{1}{\varsigma_k} L$, $\varsigma_k = 1, -1, i, -i$, $\Delta(L) = \Pi_{k=1}^{4} \delta_k(L)$, where ξ_k is a constant and $\phi^*(z) = 0$ has all its roots outside the unit circle, it can be shown that the autoregression can be written in the equivalent form

$$\phi^*(L) y_{4t} = \pi_1 y_{1t-1} + \pi_2 y_{2t-1} + \pi_3 y_{3t-2} + \pi_4 y_{3t-1} + \varepsilon_t. \quad (1)$$

where $y_{1t} = \left(1 + L + L^2 + L^3\right) y_t = (1 + L)(1 + L^2) y_t$, $y_{2t} = -\left(1 - L + L^2 - L^3\right) y_t = -(1 - L)(1 + L^2) y_t$, $y_{3t} = -\left(1 - L^2\right) y_t = -(1 - L)(1 + L) y_t$, and $y_{4t} = \left(1 - L^4\right) y_t = (1-L)(1+L)(1+L^2) y_t$. Notice that, in this representation, $\phi^*(L)$ is a stationary and finite polynomial if $\phi(L)$ only has roots outside the unit circle except for possible unit roots at the long-run, semiannual , and annual frequencies.

The HEGY tests of the null hypothesis of a unit root are now conducted by simple "t-value" tests on π_1 for the long-run unit root, π_2 for the semiannual unit root, and "F-value" tests on π_3, π_4 for the annual unit roots. As in the Dickey–Fuller and DHF models, the statistics are not t or F distributed but have non-standard distributions. Critical values for the "t" tests are tabulated in Fuller (1976) while critical values for the "F" test are tabulated in Hylleberg et al. (1990).

Tests for combinations of unit roots at the seasonal frequencies are suggested by Ghysels et al. (1994). See also Ghysels and Osborn (2001), who correctly argue that if the null hypothesis is four unit roots, i.e., the proper transformation is $(1 - L^4)$, the test applied should be an "F-test" of π_i, $i = 1, 2, 3, 4$, all equal to zero.

As in the Dickey–Fuller case the correct lag-augmentation in the auxiliary regression (1) is crucial. The errors need to be rendered white noise in order for the size to be close to the stipulated significance level, but the use of too many lag coefficients reduces the power of the tests.

Obviously, if the data generating process, the DGP, contains a moving average component, the augmentation of the autoregressive part may require long lags, see Hylleberg (1995). As is the case for the Dickey-Fuller test, the HEGY test may be seriously affected by moving average terms with roots close to the unit circle, but also one time jumps in the series, often denoted structural breaks in the seasonal pattern, and noisy data with ▶outliers may cause problems.

A straightforward extension of the HEGY test for quarterly data produces tests for semiannual and monthly data, see Franses (1991) However the extension to weekly or daily data is not possible in practice due to number of regressors in the auxiliary regressions.

The results of a number of studies testing for seasonal unit roots in economic data series suggest the presence of one or more seasonal unit roots, but often not all required for the application of the seasonal difference filter, $(1-L^s)$, or the application of the seasonal summation filter, $S(L)$. Thus, these filters should be modified by applying a filter which removes the unit roots at the frequencies where they were found, and not at the frequencies where no unit roots can be detected. Another and maybe more satisfactory possibility would be to continue the analysis applying the theory of seasonal cointegration.

Seasonal Cointegration

Seasonal cointegration exists at a particular seasonal frequency if at least one linear combination of series, which are seasonally integrated at the particular frequency, is integrated of a lower order. For ease of exposition we will concentrate on quarterly time series integrated of order 1. Quarterly time series may have unit roots at the annual frequency $\pi/2$ with period 4 quarters, at the semiannual frequency π with period 2 quarters, and/or at the long-run frequency 0. The cointegration theory at the semiannual frequency, where the root on the unit circle is real, is a straightforward extension of the cointegration theory at the long run frequency. However, the complex unit roots at the annual frequency leads to the concept of polynomial cointegration, where cointegration exists if one can find at least one linear combination including a lag of the seasonally integrated series which is stationary.

In Hylleberg et al. (1990) seasonal cointegration was analyzed along the path set up in Engle and Granger (1987). Consider the quarterly VAR model $\Pi(L)X_t = \varepsilon_t, t = 1,2,....T$, where $\Pi(L)$ is a $p \times p$ matrix of lag polynomials of finite dimension, X_t is a $p \times 1$ vector of observations on the demeaned variables, while the $p \times 1$ disturbance vector is $\varepsilon_t \sim NID(0,\Omega)$. Under the assumptions that the p variables are integrated at the frequencies $0, \pi/2, 3\pi/2$, and π, and that cointegration exists at these frequencies as well, the VAR model can be rewritten as a seasonal error correction model

$$\begin{aligned}\Phi(L)X_{4t} &= \Pi_1 X_{1,t-1} + \Pi_2 X_{2,t-1} + \Pi_3 X_{3,t-2} + \Pi_4 X_{3,t-1} + \varepsilon_t,\\ \Pi_1 &= \alpha_1\beta_1', \Pi_2 = \alpha_2\beta_2', \Pi_3 = \alpha_4\beta_4' - \alpha_3\beta_3',\\ \Pi_3 &= \alpha_4\beta_3' + \alpha_3\beta_4', \end{aligned} \tag{2}$$

where the transformed $p \times 1$ vectors $X_{j,t}, j = 1,2,3,4$, are defined as in a similar way as $y_{j,t}, j = 1,2,3,4$ above, and where $Z_{1t} = \beta_1' X_{1t}$ and $Z_{2t} = \beta_2' X_{2t}$ contain the cointegrating relations at the long-run and semiannual frequencies, respectively, while $Z_{3t} = \left(\beta_3' + \beta_4' L\right) X_{3t}$ contains the polynomial cointegrating vectors at the annual frequency. In Engle et al. (1993) seasonal and non-seasonal cointegrating relations were analyzed between the Japanese consumption and income, estimating the relations for $Z_{jt}, j = 1,2,3$, in the first step following the Granger-Engle two step procedure.

The well known drawbacks of this method, especially when the number of variables included exceeds two, is partly overcome by Lee (1992) who extended the maximum likelihood based methods of Johansen (1988) for cointegration at the long run frequency, to cointegration at the semiannual frequency π.

To adopt the ML based cointegration analysis at the annual frequency $\pi/2$ with the complex pair of unit roots $\pm i$, is somewhat more complicated, however.

To facilitate the analysis, a slightly different formulation of the seasonal error correction model is given in Johansen and Schaumburg (1999). In our notation the formulation is

$$\begin{aligned}\Phi(L)X_{4t} &= \alpha_1\beta_1' X_{1,t-1} + \alpha_2\beta_2' X_{2,t-1} + \alpha_*\beta_*' X_{*,t}\\ &\quad + \alpha_{**}\beta_{**}' X_{**,t} + \varepsilon_t\\ 2\alpha_* &= \alpha_3 + i\alpha_4, 2\alpha_{**} = \alpha_3 - i\alpha_4, \beta_* = \beta_3 + i\beta_4, \beta_{**}\\ &= \beta_3 - i\beta_4\\ X_{*,t} &= (X_{t-2} - X_{t-4}) + i(X_{t-1} - X_{t-3})\\ &= -X_{3,t-2} - iX_{3,t-1}\\ X_{**,t} &= (X_{t-2} - X_{t-4}) - i(X_{t-1} - X_{t-3})\\ &= -X_{3,t-2} + iX_{3,t-1}. \end{aligned} \tag{3}$$

The formulation in (3), writes the error correction model with two complex cointegrating relations, $Z_{*,t} = \beta_*' X_{*,t}$ and $Z_{**,t} = \beta_{**}' X_{**,t}$, corresponding to the complex pair of roots $\pm i$. Notice that (2)) can be obtained from (3) by inserting the definitions of $\alpha_*, \beta_*, X_{*,t}$, and their complex conjugates $\alpha_{**}, \beta_{**}, X_{**,t}$, and order the terms.

Note that (2) and (3) show the isomorphi between polynomial lags and complex variables. The general results may be found in Johansen and Schaumburg (1999) and Cubbada (2001). The relation between the cointegration vector β_m and polynomial cointegration vector $\beta_m(L)$ is $\beta_m(L) = \beta_m$ for $\omega_m = 0, \pi$, and $\beta_m(L) = [Re(\beta_m) - Im(\beta_m)]\frac{\cos(\omega_m)-L}{\sin(\omega_m)}$ for $\omega_m \in (0, \pi)$.

Based on the extension of the ►canonical correlation analysis to the case of complex variables by Brillinger (1981), Cubbada applies the Johansen ML approach based on canonical correlations to obtain tests for cointegration at all the frequencies of interest, i.e., at the frequencies 0 and π with the real unit roots ±1 and at the frequency $\pi/2$ with the complex roots $\pm i$.

Hence, for each of the frequencies of interest the likelihood function is concentrated by a regression of X_{4t} and $X_{1,t-1}$, $X_{2,t-1}$ or the complex pair $(X_{*,t}, X_{**,t})$ on the other regressors, resulting in the complex residual matrices $U_{*,t}$ and $V_{*,t}$ with complex conjugates $U_{**,t}$ and $V_{**,t}$, respectively. After having purged X_{4t} and $X_{1,t-1}$, $X_{2,t-1}$ or the complex pair $(X_{*,t}, X_{**,t})$ for the effects of the other regressors, the cointegration analysis is based on a canonical correlation analysis of the relations between $U_{*,t}$ and $V_{*,t}$. The product matrices are $S_{UU} = T^{-1}\sum_{t=1}^{T} U_{*,t}U'_{**,t}$, $S_{VV} = T^{-1}\sum_{t=1}^{T} V_{*,t}V'_{**,t}$, and $S_{UV} = T^{-1}\sum_{t=1}^{T} U_{*,t}V'_{**,t}$, and the trace test of r or more cointegrating vectors is found as $TR = -2T\sum_{i=r+1}^{p}\ln(1-\widehat{\lambda}_i)$, where $\widehat{\lambda}_1 > \widehat{\lambda}_2 > .....\widehat{\lambda}_p$ are the ordered eigenvalues of the problem $\left|\lambda S_{VV} - S_{VU}S_{UU}^{-1}S_{UV}\right| = 0$ .The corresponding (possibly complex) eigenvectors properly normalized are v_j, $j = 1, 2...., p$, where the first r vectors form the cointegrating matrix β.

Critical values of the trace tests for the complex roots are supplied by Johansen and Schaumburg (1999) and Cubadda (2001), while the critical values for cointegration at the real root cases are found in Lee (1992) and Osterwald-Lenum (1992).

Furthermore, tests of linear hypotheses on the polynomial cointegration vectors may be executed as χ^2 test, similar to the test applied in the long-run cointegration case.

Although economic time series often exhibit nonstationary behavior, stationary economic variables exist as well, especially when conditioned on some deterministic pattern such as linear trends, seasonal dummies, breaks etc. However, a set of stationary economic times series may also exhibit common behavior, and for instance share a common seasonal pattern. The technique for finding such patterns, known as *Common Seasonal Features* were introduced by Engle and Hylleberg (1996) and further developed by Cubadda (1999).

About the Author

Dr. Svend Hylleberg is Professor of economics at the School of Economics and Management, The Social Science Faculty, Aarhus University. Currently he is Dean of the Social Science Faculty. He has authored or co-authored numerous papers, including some leading papers on the existing standard economic theory of seasonality as well as papers which apply newer statistical tools to the modeling of seasonal phenomena. He is a co-author (with Robert Engle, Clive Granger and B. Sam Yoo) of the seminal paper *Seasonal Integration and Cointegration* (Journal of Econometrics, 44, 215-238, 1990). Professor Hylleberg has written or edited five books including *Seasonality in Regression* (Academic Press, 1986), and *Modelling Seasonality* (Oxford University Press, 1992). He is a member of Econometric Society and Royal Economic Society. He was an Associate editor for *Econometric Review* (1994–2005) and *Scandinavian Journal of Economics* (1995–2006). Currently, he is Associate editor for *Macroeconomic Dynamics* (1997–).

Cross References

►Bayesian Approach of the Unit Root Test
►Dickey-Fuller Tests
►Econometrics
►Seasonality
►Time Series
►Trend Estimation
►Vector Autoregressive Models

References and Further Reading

Birchenhal CR, Bladen-Howell RC, Chui APL, Osborn DR, Smith JP (1989) A seasonal model of consumption. Econ J 99:837–843
Brendstrup B, Hylleberg S, Nielsen MØ, Skipper L, Stentoft L (2004) Seasonality in economic models. Macroeconomic Dyn 8(3):362–394
Brillinger DR (1981) Time series: data analysis and theory. Holden Day, San Francisco
Cubadda G (1999) Common cycles in seasonal non-stationary time series. J Appl Econ 13:273–291
Cubadda G (2001) Complex reduced rank models for seasonally cointegrated time series. Oxf Bull Econ Stat 63:497–511
Dickey DA, Hasza DP, Fuller WA (1984) Testing for unit roots in seasonal time series. J Am Stat Assoc 79:355–367
Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation and testing. Econometrica 55: 251–276
Engle RF, Granger CWJ, Hylleberg S, Lee H (1993) Seasonal cointegration: the Japanese consumption function. J Econom 55: 275–298
Engle RF, Hylleberg S (1996) Common seasonal features: global unemployment. Oxf Bull Econ Stat 58:615–630
Franses PH (1991) Seasonality, nonstationarity and the forecasting of monthly time series. Int J Forecast 7:199–208
Fuller WA (1976) Introduction to statistical time series. Wiley, New York

Ghysels E, Lee HS, Noh J (1994) Testing for unit roots in seasonal time series. Some theoretical extensions and a Monte Carlo investigation. J Econom 62:415–442
Ghysels E, Osborn DR (2001) The econometric analysis of seasonal time series. Cambridge University Press, Cambridge
Hylleberg S, Engle RF, Granger CWJ, Yoo BS (1990) Seasonal integration and cointegration. J Econom 44:215–238
Hylleberg S (1995) Tests for seasonal unit roots. General to Specific or Specific to General. J Econom 69:5–25
Johansen S (1995) Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford
Johansen S, Schaumburg E (1999) Likelihood analysis of seasonal cointegration. J Econom 88:301–339
Lee HS (1992) Maximum likelihood inference on cointegration and seasonal cointegration. J Econom 54:1–47
Osterwald-Lenum M (1992) Recalculated and extended tables of the asymptotic distribution of some important maximum likelihood cointegration test statistics. Oxf Bull Econ Stat 54: 6461–6472

Seasonality

Robert M. Kunst
Professor
University of Vienna, Vienna, Austria

Introduction

Seasonality customarily refers to the annual cycle in time series sampled at intervals that are integer fractions of the annual, such as quarterly or monthly observations. The concept can easily be generalized to analogous features, such as the daily cycle in hourly observations.

The characteristics of seasonality are most easily visualized in the frequency-domain representation of the time series. Denoting the number of observations per year by S, the seasonal cycle is represented by peaks in the spectral density at $2\pi/S$ and at integer multiples of this frequency $2k\pi/S, 1 \le k \le S/2$. Seasonal cycles are distinct from other cycles by their time-constant length, though their shapes often change over time. These shapes often differ strongly from pure sine waves, and two peaks and troughs over the year are not uncommon.

The occurrence of seasonal cycles in time series has generated two related but distinct strands of literature, which can be roughly labeled as *seasonal modeling* and *seasonal adjustment*.

Seasonal modeling is concerned with typically parametric time-series models that describe the seasonal behavior of the observed variable as well as the remaining characteristics. In the spectral density interpretation, a seasonal model captures the spectral mass at the seasonal frequencies as well as the remaining characteristics of the spectral density, for example the low frequencies that represent the long run.

Seasonal adjustment builds on the concept of a decomposition of the data-generating process into a seasonal and a non-seasonal component. This decomposition can be additive ($X = X^s + X^{ns}$) or multiplicative ($X = X^s \cdot X^{ns}$). The aim of adjustment is to retrieve the non-seasonal part X^{ns} from the observed X.

Seasonal Adjustment

Seasonality is not confined to economics data. Examples for seasonal variables range from river-flow data to incidences of flu epidemics. The practice of seasonal adjustment, however, is mainly restricted to economic aggregates.

In economics, seasonal adjustment is so popular that many variables – for example, some variables of national accounts – are only available in their adjusted form, that is as an estimate of X^{ns}. It has often been pointed out that this preference tacitly assumes that X^s is generated by forces outside the economic world, such that the seasonal component of a variable does not contain useful information on the non-seasonal component of the same and of other variables. A famous citation by Svend Hylleberg (Hylleberg 1986) sees seasonal cycles as affected by cultural traditions, technological developments, and the preferences of economic agents, which can be viewed as a critique of this approach.

Currently, seasonal adjustment of economic data is mainly enacted by standardized methods, typically *X-12* in the U.S. and *TRAMO-SEATS* in Europe. The conceptual basis of X-12 is a sequence of two-sided linear filters, outlier adjustments, and the application of linear time-series models to isolate the components (see Findley et al. 1998). TRAMO-SEATS aims at isolating the components using the concepts of unobserved-components representations and of signal extraction. The assessment of the strengths and weaknesses of these procedures is difficult, as the true components are never observed.

Seasonal Modeling

The current literature on seasonal modeling builds on the SARIMA (seasonal autoregressive integrated moving-average) models by Box and Jenkins (1970), who

recommend usage of the *seasonal difference* $X_t - X_{t-S}$, followed by traditional linear modeling of the filtered series. The application of this filter assumes the existence of the factor $1-B^S$ in the generalized ARMA representation of the original series, where B denotes the lag operator. This factor has zeros at S equidistant points around the unit circle, hence the name *seasonal unit roots*. Apart from +1 and possibly −1, these unit roots come in complex pairs, such that the S roots correspond to $[S/2]+1$ frequencies or unit-root events, if $[.]$ denotes the largest integer.

The 1980s saw an increasing interest in replacing the Box-Jenkins visual analysis on differencing by statistical hypothesis tests. An offspring of the unit-root test by Dickey and Fuller is the test for seasonal unit roots by Hylleberg et al. (1990), the HEGY test. A regression is run for seasonal differences of the variable on S specific transforms. F– and t–statistics allow investigating the unit-root events separately. Under the null of seasonal unit roots will the HEGY statistics follow non-standard distributions that can be represented as Brownian motion integrals or as mixtures of normal distributions.

For example, consider quarterly data ($S = 4$). In the HEGY regression, $X_t - X_{t-4}$ is regressed on four lagged 'spectral' transforms, i.e., on $X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}$, on $-X_{t-1} + X_{t-2} - X_{t-3} + X_{t-4}$, on $X_{t-1} - X_{t-3}$ and on $X_{t-2} - X_{t-4}$. The t–statistic on the first regressor tests for the unit root at +1, the t on the second regressor for the root at −1, and an F–statistic on the latter two terms tests for the complex root pair at $\pm i$.

Testing for seasonal unit roots can be interpreted as testing whether seasonal cycles experience persistent changes over time or whether seasonal differencing is really necessary to yield a stationary variable. A process with seasonal unit roots is often called *seasonally integrated*. A variable transformed into white noise by seasonal differencing is a special seasonally integrated process and is called a *seasonal* ▶*random walk*.

The HEGY test was generalized to multivariate models, to cointegration testing, and recently to panel analysis. Other tests for seasonal unit roots have been developed, some of them with unit roots as the alternative (for example, Canova and Hansen 1995). A detailed description of many of these tests and also of other issues in seasonality can be found in Ghysels and Osborn (2001).

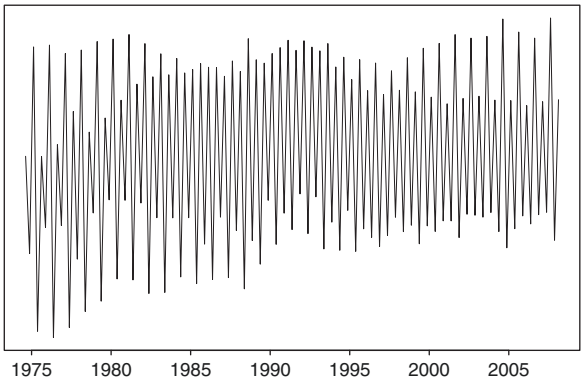
While the seasonal unit-root analysis is confined to extensions of the Box-Jenkins SARIMA class, more sophisticated seasonal models have been suggested, for example models with evolving seasonality, seasonal long memory, and seasonality in higher moments. The most intensely investigated class among them is the *periodic model* (see Franses 1996).

An Example

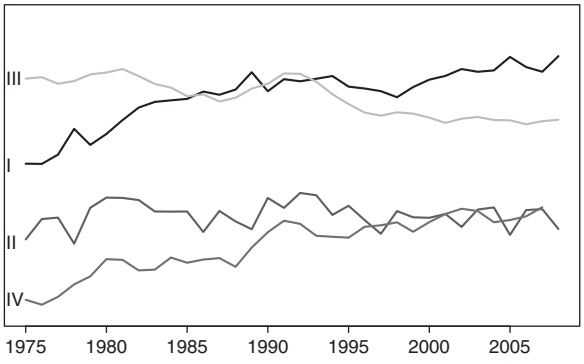
The time series variable is the quarterly number of overnight stays in the Austrian region of Tyrol for the years 1975 to 2008, which is constructed from the Austrian WIFO data base. The time-series plot in Fig. 1 shows the seasonal structure clearly.

It is a common and recommended practice to plot such series by quarters. The changes of ranks of quarters reflect the changes in the seasonal cycle. Figure 2 shows the increasing importance of winter tourism (skiing) over the observation period.

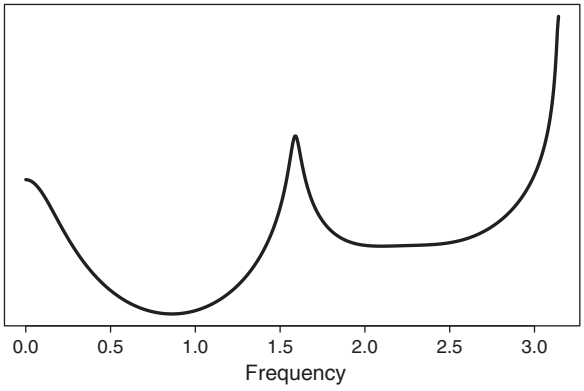
In an estimate of the spectral density (see Fig. 3), the seasonal peaks at π and $\pi/2$ are recognizable, as is another non-seasonal peak at the zero frequency (the



Seasonality. Fig. 1 Overnight stays in Tyrol, quarterly observations 1975–2008



Seasonality. Fig. 2 Overnight stays in Tyrol, plotted by quarters. Curves represent quarters I (solid), II (dashes), III (dots), and IV (dash-dotted)



Seasonality. Fig. 3 Spectral density estimate for the series on Tyrolean overnight stays

trend). Similar information is provided by the correlogram. Statistical tests confirm that this variable appears to have 'seasonal unit roots'. For example, the HEGY regression introduced above, with quarterly dummies, a trend, and a lagged $X_{t-1} - X_{t-5}$ as additional regressors, delivers t–statistics of −2.34 and −3.04, and an F–statistic of 2.56. All of these values are insignificant at 5%. The seasonal differencing operator is required to yield a stationary variable.

About the Author

Robert M. Kunst is a Professor at the University of Vienna, Austria, and a consultant and lecturer at the Institute for Advanced Studies Vienna. He is the coordinating editor of Empirical Economics. He is a Fellow of the Royal Statistical Society. He has authored or co-authored various articles on the topic of seasonality.

Cross References

►Bayesian Approach of the Unit Root Test
►Box–Jenkins Time Series Models
►Exponential and Holt-Winters Smoothing
►Moving Averages
►Seasonal Integration and Cointegration in Economic Time Series
►Time Series

References and Further Reading

Box GEP, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco, CA

Canova F, Hansen BE (1995) Are seasonal patterns constant over time? A test for seasonal stability. J Bus Econ Stat 13:237–252

Findley DF, Monsell BC, Bell WR, Otto MC, Chen B-C (1998) New capabilities and methods of the X-12-ARIMA seasonal adjustment program. J Bus Econ Stat 16:127–177

Franses PH (1996) Periodicity and stochastic trends in economic time series. Oxford University Press, Oxford

Ghysels E, Osborn DR (2001) The econometric analysis of seasonal time series. New York, Cambridge University Press

Hylleberg S (1986) Seasonality in regression. Academic Press, New York

Hylleberg S, Engle RF, Granger CWJ, Yoo BS (1990) Seasonal integration and cointegration. J Econom 44:215–238

Selection of Appropriate Statistical Methods in Developing Countries

Raymond Zepp
Dewey International University, Battambang, Cambodia

Statistical procedures are dictated by the nature of the research design. To the extent that comparisons of group means, searching for trends, or measuring central tendency and dispersion are universal objectives in all societies, it might be argued that the choice of statistical methods should be independent of the country or culture in question.

On the other hand, research in developing countries presents several challenges that are not as prevalent in developed countries, and therefore, the appropriateness of the statistical treatment may vary according to the type of data available.

First, data collected in developing countries can suffer from deficiencies of reliability. Industries, for example, may submit their production figures to the national statistics office in a variety of units of measurement (kilograms, tons, pounds), and these discrepancies are not always noticed by untrained workers in the statistics office.

As a result, statistics should be kept simple and transparent, so that problems of reliability can surface and be spotted easily. Research reports should include ►sensitivity analysis, that is, an analysis of how much variation in outputs could be caused by small variations in inputs.

Second, experimenters may find it more difficult in developing countries to control all variables. For example, social research may find it difficult to control the socioeconomic status of the subjects of a study. In this case, it may be more difficult to identify the real variable that gives rise to group differences. Thus, factoring out extraneous variables, for example by the ►analysis of covariance, may be a primary focus of research designs in developing countries.

Third, probability distributions may stray from the normal bell-shaped curve. Many developing countries have not only widely disparate populations, but may have two or three subpopulations such as tribal cultures or rich-poor splits that can yield bimodal distributions, or even distributions with most of the data occurring at the extremes of the curve rather than in the middle.

For this reason, there may be a tendency to use nonparametric statistical models in the analysis of data. Or, if parametric methods are to be used, careful study of the robustness of the procedure should be taken into account. If slight discrepancies from normality can result in large deviations in results, then the use of the parametric statistics should be called into question.

Fourth, technical and educational facilities in developing countries may limit the capacity to use more sophisticated statistical methods. For one thing, computer capability may be limited in either hardware or software, or else local statisticians may not be fully conversant with statistical software packages. In either case, it is probably more appropriate to adopt statistical methods that are as simple as possible.

A note needs to be made concerning statistical education in developing countries. Because schools and even universities lack the necessary computers, statistics as a subject is often taught by the old-fashioned method of calculations by hand-held calculators or even by pencil-and-paper. In such an educational system, the emphasis is often on the calculation algorithms of, say, means and standard deviations, rather on the interpretation of results. In developed countries where the entire class has unlimited access to computers with statistical software, the calculations can be done very easily, so that the emphasis can be placed on interpreting the results, or on assessing the appropriateness of the statistical method in question. In developing countries, however, students often "lose sight of the forest for the trees," that is, their academic assessment is entirely dependent on their ability to calculate algorithms that they do not focus on design of experiments and interpretation of results.

A second point about education in developing countries is the lack of teachers trained in locally appropriate methods. A university teacher quite likely has been trained in the developed world, and therefore wishes to teach students the most sophisticated and up-to-date methods, even though those methods may not be the most appropriate in the local context.

Related to the above point is the fact that the publication of research results is often biased by the complexity of the statistical methods used. A journal editor may reject a research study simply because the statistics used do not appear sophisticated enough to merit publication. Thus, a researcher may reject a simple but appropriate method in favor of a more complicated one in order to impress the readers.

One may summarize the above points in four recommendations:

1. When in doubt, opt for the simpler statistical procedure.
2. Be prepared to use nonparametric statistics.
3. Sensitivity Analysis should be carried out to compensate for possibilities of unreliable data.
4. Students should be trained in the appropriateness of statistical design and interpretation of results, not just in the calculation of statistical algorithms.

About the Author

Raymond Zepp holds a Bachelor's Degree in Mathematics from Oberlin College, a Master's Degree in Mathematics from the University of Cincinnati, and a Ph.D. in Mathematics Education from the Ohio State University. He is Vice President of the newly-opened Dewey International University in Battambang, Cambodia. As founder of DIU (www.diucambodia.org), he has incorporated his vision of "Learning by Doing" into a strong emphasis on community service learning and research. Dr. Zepp has taught statistics in developing universities, governments, and non-governmental organizations around the world, for example, in Nigeria, Lesotho, Macau, Papua New Guinea, Micronesia, Mozambique, Uganda, Qatar, and Cyprus, and as Fulbright Professor in the Ivory Coast. He has set up research institutes at the University of Cambodia and at Qatar University, and has designed new universities in Nigeria (Maiduguri) and Papua New Guinea (Goroka), and of course Cambodia (Dewey International). He has done statistical consulting for USAID, UNDP, Asia Development Bank, the World Bank, and others. Dr. Zepp has authored or co authored over 40 books (e.g., *Business Research and Statistics*, Hong Kong: Asia Pacific International Press, 1988) and over 100 journal articles, conference papers, etc. He currently resides in Battambang, Cambodia.

Cross References

►African Population Censuses
►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Nonparametric Statistical Inference
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics: Developing Country Perspective
►Sensitivity Analysis

Semiparametric Regression Models

Yingcun Xia
Professor
National University of Singapore, Singapore, Singapore

In statistics, semiparametric regression includes regression models that combine parametric and nonparametric models. They are often used in situations where the fully nonparametric model may not perform well or when the researcher wants to use a parametric model but the functional form with respect to a subset of the regressors or the density of the errors is not known. Suppose Y is a response and $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ are covariates. A basic goal is to estimate $m(x) = E(Y|X = x)$ or the model $Y = m(X) + \varepsilon$ with $E(\varepsilon|X) = 0$ almost surely. Without any information about the structure of the function, it is difficult to estimate $m(x)$ well when $p > 1$, and as a consequence many parametric and semiparametric models have been proposed that impose structural constraints or special functional forms upon $m(x)$. Popular semiparametric models include *partially linear models*, see for example Speckman (1988), in which

$$Y = \beta_1 \mathbf{x}_1 + ... + \beta_{p-1}\mathbf{x}_{p-1} + g_p(\mathbf{x}_p) + \varepsilon,$$

additive models, see for example Hastie and Tibshirani (1990), in which

$$Y = g_1(\mathbf{x}_1) + g_2(\mathbf{x}_2) + ... + g_p(\mathbf{x}_p) + \varepsilon,$$

single-index models, see for example Ichimura (1993), in which

$$Y = g(\beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p) + \varepsilon,$$

varying coefficient models, see for example Chen and Tsay (1993) and Hastie and Tibshirani (1993), in which

$$Y = g_1(\mathbf{x}_1) + g_2(\mathbf{x}_1)\mathbf{x}_2 + ... + g_p(\mathbf{x}_1)\mathbf{x}_p + \varepsilon.$$

and *extended partially linear single-index model*, see Xia et al. (1999), in which

$$Y = \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p + g(\theta_1 \mathbf{x}_1 + ... + \theta_p \mathbf{x}_p) + \varepsilon.$$

In all the above models, $g_1, ..., g_p$ and g are unknown functions and $\beta_1, ..., \beta_p, \theta_1, ..., \theta_p$ are parameters need to be estimated. A general form of the semiparametric model including all the models above is

$$\mu\{E(Y|\mathbf{x}_1, ..., \mathbf{x}_p)\} = G(g, \beta, X),$$

where $g = (g_1, ..., g_q)^T$ are unknown smooth functions, G is known up to a parameter vector β, function μ is known and usually monotonic.

Both splines smoothing and Kernel smoothing can be used to estimate these models. The general model can be estimated by the method proposed by Xia et al. (2002). Theoretically, all these models can avoid the "curse of dimensionality" in the estimation. The estimators of the unknown functions $g_1, ..., g_p$ and g can achieve the optimal consistency rate of univariate function, and the parameters such as $\beta_1, ..., \beta_p$ and θ are root-n consistent.

These models have been found very useful in application; see for example Hastie and Tibshirani (1990), Fan and Gijbels (1996) and Ruppert et al. (2003).

About the Author

Yingcun Xia is Professor of statistics at the National University of Singapore. He was elected member of the International Statistics Institute (2005–). He was Associated Editor for the *Annals of Statistics* (2007–2009). His research interest includes semiparametric modeling, nonlinear time series analysis and statistical modeling of infectious diseases. His work on nonlinear dimension reduction (called MAVE) and on the modeling of transmission of infectious diseases based on gravity mechanism has received wide recognition.

Cross References

►Absolute Penalty Estimation
►Bayesian Semiparametric Regression
►Nonparametric Regression Using Kernel and Spline Methods
►Smoothing Splines

References and Further Reading

Chen R, Tsay R (1993) Functional coefficient autoregressive models: estimation and tests of hypotheses. J Am Stat Assoc 88:298–308
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall/CRC, Boca Rotan, FL
Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econometrics 58:71–120
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression Cambridge University Press, UK
Speckman P (1988) Kernel smoothing in partial linear models. J Roy Stat Soc Ser B 50:413–436
Xia Y, Tong H, Li WK (1999) On extended partially linear single-index models. Biometrika 86:831–842
Xia Y, Tong H, Li WK, Zhu L (2002) An adaptive estimation of dimension reduction space (with discussion). J Roy Stat Soc Ser B 64:363–410

Semi-Variance in Finance

Vijay K. Rohatgi
Professor Emeritus
Bowling Green State University, Bowling Green, OH, USA

For any random variable X with finite variance, and any constant t

$$E\{(X-t)\}^2 = E\{(X-t)^-\}^2 + E\{(X-t)^+\}^2.$$

If $t = \mu = EX$, then $E\{(X-t)\}^2 = \sigma^2$, the variance of X. The quantity $E\{(X-\mu)^-\}^2$ is called the (lower) semi-variance of X whereas $E\{(X-\mu)^+\}^2$ is called the upper semi-variance of X. In financial applications where X represents return on an investment, σ is widely used as a measure of risk of an investment (portfolio). In that context σ is called volatility since it measures volatility of returns. Risk-averse investors like consistency of returns and hence lower volatility. In order to compare two or more investments one compares their returns per unit of risk, that is, μ/σ = 1/coefficient of variation. A modified version of this measure is due to Sharpe (1994) who uses the ratio excess returns (over risk free returns) divided by volatility. Another widely used measure of investors' risk is beta, the coefficient of linear regression of returns over some benchmark returns such as Standard and Poor 500 index. Thus, a value of beta over 1 means that the investment under consideration is more volatile (risky) than the benchmark.

For risk-averse investors neither of these two measures fits their need. They are more interested in the downside risk, the risk of losing money or falling below the target return. For instance, variance assigns equal weight to both deviations, those above the mean and those below the mean. In that sense it is more suitable for symmetric return distributions in which case $\sigma^2 = 2E\{(X-\mu)^-\}^2$. In practice the return distributions are often skewed to the right. No investor is averse to returns in excess of the target. He or she prefers positive skewness because the chance of large deviations from the target rate is much less.

Markowitz (1959) introduced

$$\sigma_D^2(t) = E\{(X-t)^-\}^2$$

as a measure of downside risk. Here t may be called the target rate of return which could be the riskless rate such as the three month T-bill rate or the Libor rate. Recall that $E\{(X-t)\}^2$ is minimized for $t = \mu$. On the other hand $\sigma_D^2(t)$ is an increasing function of t and a Chebyshev type inequality holds:

$$P\left(X < \mu - k\sigma_D(t)\right) \le 1/k^2 \quad \text{for } k \ge 1.$$

As an estimate of $\sigma_D^2(t)$ one generally uses the substitution principle estimator

$$(1/n)\sum_{i=1}^{n}\{(x_i-t)^-\}^2$$

and when $t = \mu$ we use the estimator

$$(1/n)\sum_{1}^{n}\left\{(x_i-\bar{x})^-\right\}^2.$$

Markowitz (1952) was the first to propose a method of construction of portfolios based on mean returns, and their variances and covariances (*see* ►Portfolio theory). In 1959 he proposed semivariance as a measure of downside risk and advocated its use in portfolio selection. Due to computational complexity of semivariance and semicovariance, however, he used the variance measure of risk instead. After the advent of desktop computers and their computational power in 1980s the focus shifted to portfolio selection based on semivariance as a measure of downside risk. See for example Markowitz et al. (1993).

Both $\sigma_D(t)$ and $\sigma_U(t)$ $\left(\sigma_U^2(t) = E\{(X-t)^+\}^2\right)$ have been used in Quality Control (*see* ►Statistical Quality Control) in constructing process capability indices. See for example, Kotz and Cynthia (1998). Other uses are in spatial statistics and in construction of confidence intervals in simulation output analysis Coobineh and Branting (1991). The semi-standard deviation $\sigma_D(\mu)$ can also be used in setting up dynamic stop loss points in security trading.

About the Author

Vijay K. Rohatgi is a Professor Emeritus, Department of Mathematics and Statistics, Bowling Green State Univerity (BGSU), Ohio. He was Chair of the Department (1983–1985) and played a key role in the creation of the doctoral program of the Department. He is internationally well-known as an author/co-author of five successful books on statistics, probability and the related, including *An Introduction to Probability and Statistics* (with A.K.Md. Ehsanes Saleh, Wiley, 2nd edition, 2001) and *Statistical Inference* (Dover Publications, 2003).

Cross References

►Banking, Statistics in
►Coefficient of Variation

►Portfolio Theory
►Variance

References and Further Reading

Coobineh F, Branting D (1991) A split distribution method for constructing confidence intervals for simulation output analysis. Int J Sys Sci 22:367–374
Kotz S, Cynthia L (1998) Process capability indices in theory and practice. Arnold, New York
Markowitz HM (1952) Portfolio selection. J Finance 7:77–91
Markowitz HM (1959) Portfolio selection, efficient diversification of investments. Cowles Foundation Monograph 16. Yale University Press, New Haven
Markowitz H, Todd P, Xu G (1993) Computation of mean-semivariance efficient set by the critical line algorithm. Ann Oper Res 45:307–317
Sharpe WF (1994) The sharpe ratio. J Portfolio Manag 21:49–58

Sensitivity Analysis

Andrea Saltelli[1], Paola Annoni[2]
[1]Head of the Unit of Econometrics and Applied Statistics Joint Research Centre of the European Commission, Institute for the Protection and the Security of the Citizen, Ispra, Italy
[2]Joint Research Centre of the European Commission, Institute for the Protection and the Security of the Citizen, Ispra, Italy

Existing guidelines for impact assessment recommend that mathematical modeling of real or man-made system be accompanied by a 'sensitivity analysis' - SA (EC 2009; EPA 2009; OMB 2006). The same recommendation can be found in textbooks for practitioners (e.g., Kennedy 2007, Saltelli et al. 2008). Mathematical models can be seen as machines capable of mapping from a set of assumptions (data, parameters, scenarios) into an inference (model output).

In this respect modelers should tackle:

- **Uncertainty.** Characterize the empirical probability density function and the confidence bounds for a model output. This can be viewed as the numerical equivalent of the measurement error for physical experiments. The question answered is "How uncertain is this inference?"
- **Sensitivity.** Identify factors or groups of factors mostly responsible for the uncertainty in the prediction. The question answered is "Where is this uncertainty coming from?"

The two terms are often used differently, with sensitivity analysis used for both challenges (e.g., Leamer 1990). We focus on sensitivity analysis proper, i.e., the effect of individual factors or group of factors in driving the output and its uncertainty.

Basic Concepts

The ingredients of a sensitivity analysis are the model's uncertain input factors and model's outputs. Here and in the following we shall interpret as factor all that can be plausibly changed at the level of model formulation or model input in the quest to map the space of the model predictions. Thus a factor could be an input datum acquired with a known uncertainty, as well as a parameter estimated with known uncertainty in a previous stage of modeling, as well a trigger acting on the model's structure (e.g., a mesh size choice), or a trigger selecting the choice of a model versus another, or the selection of a scenario. Modelers usually have considerable latitude of choice as to how to combine factors in a sensitivity analysis, e.g., what to vary, what to keep fixed. Also a modeler's choice is, to some extent, whether to treat factors as dependent upon one another or as independent. The design and the interpretation of this ensemble of the model simulations constitute a sensitivity analysis.

Use of Sensitivity Analysis

Sensitivity analysis is a tool to test the quality of a model or better the quality of an inference based on a model. This is investigated by looking at the robustness of an inference. There is a trade off here between how scrupulous an analyst is in exploring the input assumptions and how wide the resulting inference will be. Edward E. Leamer (1990) calls this an organized sensitivity analysis:

> *I have proposed a form of organized sensitivity analysis that I call 'global sensitivity analysis' in which a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified. Conclusions are judged to be sturdy only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inferences is narrow enough to be useful.*

In fact it is easy to invalidate a model demonstrating that it is fragile with respect to the uncertainty in the assumptions. Likewise one can criticize a sensitivity analysis by showing that its assumptions have not been taken 'wide enough.'

Examples of application of SA are: robustness assessment in the context of impact assessment; model simplification in the context of complex and computer demanding models; quality assurance for detecting coding errors or

misspecifications. Sensitivity analysis can also highlight the region in the space of input factors for which the model output assumes extreme values, as can be relevant in ▶risk analysis. Likewise it can identify model instability regions within the space of the factors for use in a subsequent calibration study.

Local Vs Global Methods

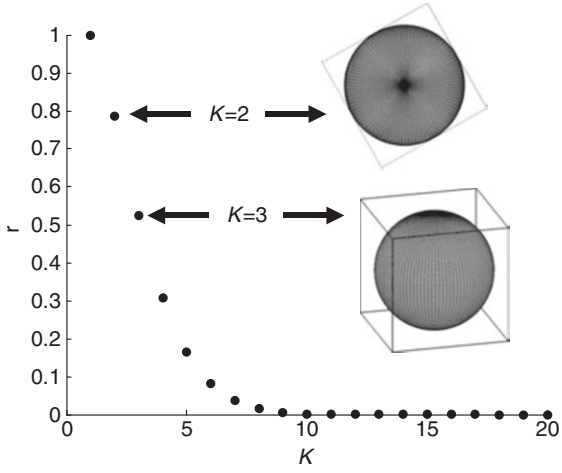
In the model $Y = f(X_1, X_2, \ldots, X_k)$ Y is the output and X_is are the input factors. The model is linear if each factor X_i enters linearly in f. The model is additive if the function f may be decomposed into a sum of k functions $f_i \equiv f_i(X_i)$, each f_i depending only on its own factor X_i.

There are 'local' and 'global' methods for SA. If the model is *linear*, a *local approach* based on first derivatives of the output with respect to the input factors will provide all the information that is needed for SA. If the model is *non linear but additive*, i.e., there are no interactions among factors, then *derivatives of higher and cross order* will be needed. When a-priori information on the nature of the model is not available (*model-free setting*) or the model is acknowledged to be non additive, then *global methods* are needed whereby all the space of the uncertain input factors is explored. Note that often modelers cannot assume linearity and additivity as their models come in the form of computer programs, possibly including several computational steps. In this situation it is better to use 'global' methods (EPA 2009; Saltelli et al. 2008).

A Very Popular Practice: OAT-SA

Most sensitivity analysis met in the literature are realized by varying one factor at a time – OAT approaches. Modelers have many good reasons to adopt OAT, including the use of a common '*baseline*' value from which all factors are moved. Derivative based approaches - when the derivatives stop at the first order - are a particular case of OAT. Typical arguments in favor of OAT are: (1) The baseline vector is a safe starting point where the model properties are well known; (2) Whatever effect is detected on the output, this is solely due to that factor which was moved and to none other; (3) The chances of the model to crash or to give unacceptable results are minimized as these generally increase with the distance from the baseline.

Despite all these points in favor to an OAT sensitivity analysis we would like to discourage as much as possible this practice (Saltelli and Annoni 2010). OAT is inefficient in exploring the input space as the coverage of the design space is extremely poor already with few input factors. The issue of uniformly covering the hyperspace in high dimensions is a well known and widely discussed matter under the name *curse of dimensionality* (Hastie et al. 2001). There



Sensitivity Analysis. Fig. 1 Curse of dimensionality-horizontal axis = number of dimensions; vertical axis = volume of the inscribed unitary sphere

are various ways to visualize this 'curse'. Figure 1 may be effective. It shows that, as the number of dimensions k increases, the volume of the hyper-sphere inscribed in the unitary hyper-cube goes rapidly to zero (it is less than 1% already for $k = 10$).

The OAT approach – moving always one step away from the same baseline – can be shown to generate points inside the hyper-sphere. Of course when one throws a handful of points in a multidimensional space these points will be sparse, and in no way the space will be fully explored. Still, even if one has only a handful of points at disposal, there is no reason why one should concentrate all these points in the hyper-sphere, i.e., closer to the origin on average than randomly generated points in the cube.

An additional shortcoming of OAT is that it cannot detect factor interactions. It may be the case that a factor is detected as no influential while it is actually relevant but only through its interaction with the other factors. In a model free setting, OAT is by no means the winning choice.

Design and Estimators

Unlike OAT, a good experimental design will tend to change more factors simultaneously. This design can be realized using the same techniques used for experimental design (e.g., a saturated two-level design or an unsaturated design with more levels). A practical alternative for numerical experiments is a Monte Carlo method. Beside design, sensitivity analysis needs sensitivity estimators which will translate the function values computed at the design points into sensitivity coefficients for the various factors.

Model's predictions have to be evaluated at different points within the parameter space, whose dimensionality is equal to the number k of input factors. To explore the k-dimensional factor space (the hyperspace) the first step in usually to reduce the problem to traveling across the k-dimensional unit cube by using the inverse cumulative distribution function of input factors. The input space can be explored using ad hoc trajectories (such as in the elementary effects method below), random numbers or quasi-random numbers. Quasi-random numbers are specifically designed to generate samples from the space of input factors as uniformly as possible. For a review on quasi random sequences and their properties see Bratley and Fox (1988).

After sampling the space of input factors, various methods may be applied to compute different sensitivity measures. Selected practices are given next.

Morris' Elementary Effects

The Elementary Effect method (Morris 1991) provides a ranking of input factors according to a sensitivity measure simply based on averages of derivatives over the space of factors. In the Morris setting each input factor is discretized into p levels and the exploration of the input space is carried out along r trajectories of $(k+1)$ points, where each point differs from the previous one in only one component. Each trajectory provides rough sensitivity measures for each factor called elementary effect *EE*. The elementary effect of trajectory j for factor i is:

$$EE_i^{(j)} = \frac{Y(X_1,\ldots,X_{i-1},X_i+\Delta,X_{i+1},\ldots,X_k) - Y(X_1,\ldots,X_k)}{\Delta} \tag{1}$$

where convenient choices for p and Δ are p even and Δ equal to $p/[2(p-1)]$. The point $(X_1,\ldots,X_k)$ is any point in the input space such that the incremental point $(X_1,\ldots,X_{i-1},X_i+\Delta,X_{i+1},\ldots,X_k)$ still belongs to the input space (for each $i=1,\ldots,k$). Elementary effect $EE_i^{(j)}$ provides a sensitivity index which highly depends on the particular trajectory, being in this sense *local*. To compute a more *global* sensitivity measure, many trajectories are chosen and the average value of $EE_i^{(j)}$ over j is computed. Following a recent revision of original Morris' measure, factors may be ranked according to μ^* (Campolongo et al. 2007):

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r}\left|EE_i^{(j)}\right| \tag{2}$$

The elementary effects sensitivity measure is an efficient alternative to OAT. It is used for factor screening, especially with large and complex models. When modellers are constrained by computational costs, a recommended practice is to perform a preliminary analysis by means of Morris' trajectories to detect possible non influential factors. More computationally intensive methods may be then applied to a smaller set of input factors.

Monte Carlo Filtering

An alternative setting for sensitivity analysis is the 'factor mapping' which relates to situations when there is a special concern towards a particular portion of the distribution of the output Y, e.g., one is concerned with Y above or below a given threshold – e.g., an investment loss or a toxicity level not to be exceeded. This is the typical setting of Monte Carlo Filtering MCF (see Saltelli et al. 2004 for a review). The realizations of Y are classified into 'good' – behavioral – and 'bad - non-behavioral depending on the value of Y with respect to the threshold. A MCF analysis is divided into the following steps:

1. Compute different realizations of Y corresponding to different sampled points in the space of input factor by means of a Monte Carlo experiment;
2. Classify each realization as either behavioral (B) or non behavioral $(\overline{B})$;
3. For each X_i define two subsets, one including all the values of X_i which give behavioral Y, denoted $(X_i|B)$, the other including all the remaining values $(X_i|\overline{B})$;
4. Compute the statistical difference between the two empirical distribution functions of $(X_i|B)$ and $(X_i|\overline{B})$. A factor is considered influential if the two distribution functions are statistically different. Classical statistical tests, such as Smirnov two-sample test may be used to the purpose.

Variance-Based Sensitivity Measures

With variance-based sensitivity analysis (VB-SA) input factors can be ranked according to their contribution to the output variance. VB-SA also tackles interaction effects instructing the analyst about cooperative behavior of factors. Interactions can lead to extremal values of model output and are thus relevant to the analysis. In VB-SA sensitivity analysis the two most relevant measures are 'first order' and 'total order' indices.

The best systematization of the theory of variance-based methods is due to Sobol' (Sobol 1990), while total sensitivity indices were introduced by Homma and Saltelli (1996). For reviews see also Saltelli et al. (2005) or Helton et al. (2006). Variance-based SA uses measures as

$$S_i = \frac{V_{X_i}\left(E_{\mathbf{X}_{\sim i}}(Y|X_i)\right)}{V(Y)} \tag{3}$$

and

$$S_{T_i} = \frac{E_{\mathbf{X}_{\sim i}}\left(V_{X_i}(Y|\mathbf{X}_{\sim i})\right)}{V(Y)} = 1 - \frac{V_{\mathbf{X}_{\sim i}}\left(E_{X_i}(Y|\mathbf{X}_{\sim i})\right)}{V(Y)} \quad (4)$$

where $\mathbf{X}_{\sim i} = \{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k\}$.

$E_{\mathbf{X}_{\sim i}}(Y|X_i)$ is the value of Y obtained by averaging over all factors but X_i, and is thus a function of X_i alone. $V_{X_i}\left(E_{\mathbf{X}_{\sim i}}(Y|X_i)\right)$ is the variance of this function over X_i itself. Intuitively a high value of this statistics implies an influent factor.

The quantity S_i corresponds to the fraction of $V(Y)$ that can be attributed to X_i alone. It can be viewed as a measure of how well $E_{\mathbf{X}_{\sim i}}(Y|X_i)$ fits Y: if the fitting is optimal then $S_i \cong 1$ and factor X_i is highly relevant. The quantity S_{T_i} corresponds to the fraction of $V(Y)$ that can be attributed to X_i and all its interactions with other factors. For additive models the two measures S_i and S_{T_i} are equal to one another for each factor X_i. For an interacting factor the difference $S_{T_i} - S_i$ is a measure of the strength of the interactions.

The estimation of S_i and S_{T_i} requires the computation of k-dimensional integrals. They are generally approximated assuming independency among input factors and using Monte-Carlo or quasi-Monte-Carlo sampling from the joint distribution of the space of input factors. Alternative procedures for the computation of S_i and S_{T_i} are available which use direct calculations. They all derive from metamodels, which provide cheap emulators of complex and large computational models (see for example Oakley and O'Hagan 2004; Storlie et al. 2009).

About the Author

Andrea Saltelli, has worked on physical chemistry, environmental sciences and applied statistics. His main disciplinary foci are sensitivity analysis and composite indicators. He is the leading author of three successful volumes published by Wiley on sensitivity analysis: *Sensitivity Analysis: Gauging the Worth of Scientific Models* (2000), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models* (2004), and *Global Sensitivity Analysis: The Primer* (2008), and of several papers on the same subject. Paola Annoni has produced original work in the field of sensitivity analysis and partial ordering. Both work at the Joint Research Centre of the European Commission in Ispra, Italy.

Cross References

►Bayesian Statistics
►Bias Analysis
►Composite Indicators
►Design of Experiments: A Pattern of Progress
►Interaction
►Misuse of Statistics
►Model Selection
►Monte Carlo Methods in Statistics
►Selection of Appropriate Statistical Methods in Developing Countries

References and Further Reading

Bratley P, Fox BL (1988) Algorithm 659 implementing Sobol's quasi-random sequence generator. ACM Trans Math Soft 14(1):88–100

Campolongo F, Cariboni J, Saltelli A (2007) An effective screening design for sensitivity analysis of large models. Environ Model Soft 22:1509–1518

EC 2009 Impact assessment guidelines. SEC, p 24. http://ec.europa.eu/governance/impact/docs/key_docs/iag_2009_en.pdf. Accessed 15 Jan 2009

EPA 2009 Guidance on the development, evaluation, and application of environmental models. Technical Report, Office of the science advisor, Council for Regulatory Environmental Modeling. EPA /100/K-09/003, p 26. http://www.epa.gov/crem/library/cred_guidance_0309.pdf

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York

Helton JC, Johnson JD, Salaberry CJ, Storlie CB (2006) Survey of sampling based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 91:1175–1209

Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of model output. Reliab Eng Syst Saf 52(1):1–17

Kennedy P (2007) A guide to econometrics, 5th edn. Blackwell Publishing, Oxford

Leamer E (1990) Let's take the con out of econometrics, and sensitivity analysis would help. In: Granger C (ed) Modelling economic series. Clarendon Press, Oxford

Morris MD (1991) Fractional sampling plan for preliminary computational experiments. Technometrics 33:161–174

Oakley JE, O'Hagan A (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. J Roy Stat Soc B 66: 751–769

OMB – Office of Management and Budget (2006) Proposed risk assessment bulletin (http://www.whitehouse.gov/omb/inforeg/proposed_risk_assessment_bulletin_010906.pdf)

Saltelli A, Annoni P (2010) How to avoid a perfunctory sensitivity analysis, Environ Model Softw, doi:10.1016/j.envsoft.2010.04.012

Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) Sensitivity analysis in practice. A guide to assessing scientific models. Wiley, Chichester

Saltelli A, Ratto M, Tarantola S, Campolongo F (2005) Sensitivity analysis for chemical models. Chem Rev 105(7):2811–2828

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global sensitivity analysis. The primer. Wiley, Chichester

Sobol' IM (1990) Sensitivity estimates for nonlinear mathematical models. Matem Mod 2:112–118. (in Russian). (trans: Sobol' IM (1993) Sensitivity analysis for non-linear mathematical models. Math Model Comp Exper 1:407–414)

Storlie CB, Swiler LP, Helton JC, Sallaberry CJ (2009) Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. Reliab Eng Syst Saf 94:1735–1763

Sensometrics

Per Bruun Brockhoff
Professor, Head of Statistics Section
Technical University of Denmark, Lyngby, Denmark

Introduction

The use of humans as measurement instruments is playing an increasing role in product development and user-driven innovation in many industries. This ranges from the use of experts and trained human test panels to market studies where the consumer population is tested for preference and behavior patterns. This calls for improved understanding on one side of the human measurement instrument itself and on the other side the modeling and empirical treatment of data. The scientific grounds for obtaining improvements within a given industry span from experimental psychology to mathematical modeling, statistics, chemometrics, and machine learning together with specific product knowledge be it food, TVs, hearing aids, mobile phones, or whatever.

In particular in the food industry, sensory and consumer data is frequently produced and applied as the basis for decision making. And in the field of food research, sensory and consumer data is produced and used similar to the industrial use, and academic environments specifically for sensory and consumer sciences exist worldwide. The development and application of statistics and data analysis within this area is called sensometrics.

Sensory Science and Sensometrics

As the name indicates, sensometrics really grew out of and is still closely linked to sensory science, where the use of trained sensory panels plays a central role. Sensory science is the cross-disciplinary scientific field dealing with human perception of stimuli and the way they act upon sensory input. Sensory food research focuses on better understanding of how the senses react during food intake, but also how our senses can be used in quality control and innovative product development. Historically it can be viewed as a merger of simple industrial product testing with psychophysics as originated by G.T. Fechner and S.S. Stevens in the nineteenth century. Probably the first exposition of the modern sensory science is given by Amerine et al. (1965). Rose Marie Pangborn (1932–1990) was considered one of the pioneers of sensory analysis of food and the main global scientific conference in sensory science is named after her. The first Pangborn Symposium was held in Helsinki, Finland, in 1992 and these conferences are approaching in the order of 1,000 participants – the ninth was planned for in Bangkok, Thailand, in 2011. Jointly with this, international sensometrics conferences have been held also since 1992, where the first took place in Leiden, Holland (as a small workshop), and the tenth took place in Rotterdam, Holland, in 2010. The sensometrics conferences have a participation level of around 120–150. Both conferences are working together with the Elsevier Journal *Food Quality and Preference*, which is also the official membership journal for the Sensometrics Society (www.sensometric.org).

Sensometrics: Statistics, Psychometrics, or Chemometrics?

The "sensometrician" is faced with a vast collection of data types from a large number of experimental settings ranging from a simple one-sample binomial outcome to complex dynamical and/or multivariate data sets; see, e.g., Bredie et al. (2010) for a recent review of quantitative sensory methodology. So what is really (good) sensometrics? The answer will depend on the background of the sensometrician, who for the majority, if not a food scientist, is coming from one of the following fields: generic statistics, psychophysics/experimental psychology, or chemometrics.

The generic statistician arch type would commonly carry out the data analysis as a purely "empirical" exercise in the sense that methods are not based on any models for the fundamental psychological characteristics underlying the sensory phenomena that the measurements express. The advantage of a strong link to the generic scientific fields of mathematical and applied statistics is the ability to employ the most modern statistical techniques when relevant for sensory data and to be on top of sampling uncertainty and formal statistical inferential reasoning. And this is certainly needed for the sensory field as for any other field producing experimental data. The weakness is that the lack of proper psychophysical models may lead to inadequate interpretations of the analysis results. In, e.g., MacKay (2005) the first sentence of the abstract is expressing this concern rather severely: "Sensory and hedonic variability are fundamental psychological characteristics that must be explicitly modeled if one is to develop meaningful statistical models of sensory phenomena." A fundamental challenge of this ambitious approach is that the required psychophysical (probabilistic) models of behavior are on one hand only vaguely verifiable, since they are based on models of a (partly) unobserved system, the human brain and perceptual system, and on the other hand may lead to rather complicated statistical models. MacKay (2005) is published in a special sensory data issue of *The Journal of Chemometrics*; see Brockhoff et al. (2005). Chemometricians are the third and final arch type

of a sensometrician. In chemometrics the focus is more on multivariate data analysis (see ►Multivariate Data Analysis: An Overview) and for some the explorative principle is at the very heart of the field; see, e.g., Munck (2007) and Martens and Martens (2001). The advantage of the chemometrics approach is that usually all multivariate features of the data are studied without forcing certain potentially inadequate model structures on the data. The weakness is exactly also this lack of modeling rendering potentially certain well-understood psychophysical phenomena for the explorative modeling to find out by itself. Also, linked with the explorative approach, the formal statistical inferential reasoning is sometimes considered less important by the chemometrician.

Now, none of these arch types are (at their best) unintelligent and they would, all three of them, understand (some of) the limitations of their pure versions of analysis approach. And they all have ways of dealing with (some of) these concerns for practical data analysis, such that often, at the end of the day, the end results may not differ that much. There is though, in the point of view of this author, a lack of comprehensive comparisons between these different approaches where they all are used at their best.

Example 1: Sensory Profile Data

As an example, consider the so-called descriptive sensory analysis, also called sensory profiling. In sensory profiling the panelists develop a test vocabulary (defining attributes) for the product category and rate the intensity of these attributes for a set of different samples within the category. Thus, a sensory profile of each product is provided for each of the panelists, and most often this is replicated; see Lawless and Heymann (1999). Hence, data is inherently multivariate as many characteristics of the products are measured.

The statistics arch type would focus on the ANOVA structure of the setting and perform univariate and multivariate analyses of variance (ANOVA) and would make sure that the proper version of a mixed model ANOVA is used; see, e.g., Lea et al. (1997) and Næs et al. (2010). For studying the multivariate product structure the Canonical Variates Analysis (CVA) within the Multivariate ANOVA (MANOVA) framework would be the natural choice (see, e.g., Schlich (1998)) since it would be an analysis that incorporates the within-product (co)variability.

The chemometrics arch type would begin with principal components analysis (PCA) on averaged and/or unfolded data. For more elaborate analysis maybe three-way methods (see Brockhoff et al. (1996), Bro et al. (2002)) or other more ANOVA-like extensions would be used (see, e.g., Luciano and Næs (2008)). Analysis accounting for within-product (co)variability could be provided by extensions as presented in Bro et al. (2002) or in Martens et al. (2003).

In MacKay (2005) the approach for this type of data is that of probabilistic multidimensional scaling (PROSCAL). In short, a formal statistical model for product differences is expressed as variability on the (low-dimensional) underlying latent sensory scale. It is usually presented as superior to the use of, e.g., standard PCA, focusing on the point that it naturally includes models for different within-product variability, which in the standard PCA could be confounded with the "signal" – the inter-product distances.

Example 2: Sensory Difference and Similarity Test Data

The so-called difference and/or similarity tests are a commonly used sensory technique resulting in binary and/or categorical frequency data – the so-called triangle test is a classical example. In the triangle test an individual is presented with three samples, two of which are the same, and then asked to select the odd sample. The result is binary: correct or incorrect. Such sensory tests were already in the 1950s treated by the statistical community; see, e.g., Hopkins (1950) and Bradley (1958). These types of tests and results have also been treated extensively from a more psychophysical approach, often here denoted a Thurstonian approach. The focus in the Thurstonian approach is on quantifying/estimating the underlying sensory difference d between the two products that are compared in the difference test. This is done by setting up mathematical/psychophysical models for the cognitive decision processes that are used by assessors in each sensory test protocol see; e.g., Ennis (1993). For the triangle test, the usual model for how the cognitive decision process is taking place is that the most deviating product would be the answer – sometimes called that the assessors are using a so-called tau-strategy. Using basic probability calculus on three realizations from two different normal distributions, differing by exactly the true underlying sensory difference d, one can deduce the probability of getting the answer correct for such a strategy. This function is called the psychometric function and relates the observed number of correct answers to the underlying sensory difference d. Different test protocols will then lead to different psychometric functions. InBock and Jones (1968) probably the first systematic exposition of the psychological scaling theory and methods by Thurstone was given. This included a sound psychological basis as well as a statistical one with the use and theory of maximum likelihood methods. Within the field known as signal detection theory (see, e.g., Green and

Swets (1966) or Macmillan and Creelman (2005)), methods of this kind were further developed, originally with special emphasis on detecting weak visual or auditory signals. Further developments of such methods and their use within food testing and sensory science have developed over the last couple of decades with the numerous contributions of D. Ennis as a corner stone; see, e.g., Ennis (2003). In Brockhoff and Christensen (2010) it was emphasized and exploited that the Thurstonian-based statistical analysis of data from the basic sensory discrimination test protocols can be identified as ►generalized linear models using the inverse psychometric functions as link functions. With this in place, it is possible to extend and combine designed experimentation with discrimination/similarity testing and combine standard statistical modeling/analysis with Thurstonian modeling.

Summary

One recurrent issue in sensometrics is the monitoring and/or accounting for individual differences in sensory panel data, also called dealing with panel performance. A model-based approach within the univariate ANOVA framework was introduced in Brockhoff and Skovgaard (1994), leading to multiplicative models for interaction effect expressing the individual varying scale usage. In Smith et al. (2003) and in Brockhoff and Sommer (2008) random effect versions of such analyses were put forward leading to either a multiplicative (nonlinear) mixed model or a linear random coefficient model. Another recurring issue is the relation of multivariate data sets, e.g., trying to predict sensory response by instrumental/spectroscopic and/or chemical measurements. Similarly there is a wish to be able to predict how the market (consumers) will react to sensory changes in food products – then called Preference Mapping (McEwen 1996). This links the area closely to the chemometrics field and also naturally to the (machine) learning area, which in part is explored in Meullenet et al. (2007). Another commonly used sensory and consumer survey methodology is to use rankings or scoring on an ordinal scale. In Rayner et al. (2005) standard and extended rank-based non-parametrics is presented specifically for sensory and consumer data.

As indicated, there are yet many other examples of sensory and consumer data together with other purposes of analysis challenging the sensometrician whoever he or she is. Recently some open-source dedicated sensometrics software have appeared: the R-based SensoMiner (Lê and Husson 2008), the stand-alone tool PanelCheck (Tomic et al. 2007), and the R-package sensR (Christensen and Brockhoff 2009).

About the Author

Per Bruun Brockhoff is Professor in statistics at the Informatics Department at the Technical University of Denmark (since 2004), and Head of the Statistics Section (since 2008). He was the Chairman of DSTS, the Danish Society for Theoretical Statistics (2003–2007), and Chairman of the International Sensometrics Society (2006–2010). Professor Brockhoff co-authored around 60 peer reviewed scientific papers and 2 books. The books are both on Sensometrics: Statistics for Sensory and Consumer Science (with T. Næs and O. Tomic, John Wiley & Sons, 2010), and Nonparametrics for Sensory Science: A More Informative Approach (with J.C.W. Rayner, D.J. Best and G.D. Rayner, Blackwell Publishing, USA, 2005). He is an Elected member of ISI (2005) and currently member of the editorial boards of the two international journals: Food Quality and Preference and Journal of Chemometrics.

Cross References

►Analysis of Variance
►Chemometrics
►Multidimensional Scaling
►Nonlinear Mixed Effects Models
►Random Coefficient Models
►Random Coefficient Models

References and Further Reading

Amerine MA, Pangborn RM, Roessler EB (1965) Principles of sensory evaluation of food. Academic, New York

Bock DR, Jones LV (1968) The measurement and prediction of judgment and choice. Holden-Day, San Francisco

Bradley RA (1958) Triangle, duo-trio, and difference-from-control tests in taste testing. Biometrics 14:566

Bredie WLP, Dehlholm C, Byrne DV, Martens M (2010) Descriptive sensory analysis of food: a review. Submitted to Food Qual Prefer

Bro R, Sidiropoulos ND, Smilde AK (2002) Maximum likelihood fitting using ordinary least squares algorithms. J Chemometr 16(8–10):387–400

Bro R, Qannari EM, Kiers HA, Næs TA, Frøst MB (2008) Multiway models for sensory profiling data. J Chemometr 22: 36–45

Brockhoff PM, Skovgaard IM (1994) Modelling individual differences between assessors in sensory evaluations. Food Qual Prefer 5:215–224

Brockhoff PB, Sommer NA (2008) Accounting for scaling differences in sensory profile data. Proceedings of Tenth European Symposium on Statistical Methods for the Food Industry. pp 283–290, Louvain-La-Neuve, Belgium

Brockhoff P, Hirst D, Næs T (1996) Analysing individual profiles by three-way factor analysis. In: Næs T, Risvik E (eds) Multivariate analysis of data in sensory science, vol 16, Data handling in science and technology. Elsevier Science, B.V., pp 71–102

Brockhoff PB, Næs T, Qannari M (2005) Editorship. J Chemometr 19(3):121
Brockhoff PB, Christensen RHB (2010) Thurstonian models for sensory discrimination tests as generalized linear models. Food Qual Pref 21:330–338
Christensen RHB, Brockhoff PB (2009) sensR: An R-package for thurstonian modelling of discrete sensory data. R-package version 1.1.0. (www.cran.r-project.org/package=sensR/)
Ennis DM (1993) The power of sensory discrimination methods. J Sens Stud 8:353–370
Ennis DM (2003) Foundations of sensory science. In: Moskowitz HR, Munoz AM, Gacula MC (eds) Viewpoints and Controversies in Sensory Science and Consumer Product Testing. Food and Nutrition, Trumbull, CT
Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
Hopkins JW (1950) A Procedure for quantifying subjective appraisals of odor, flavour and texture of foodstuffs. Biometrics 6(1):1–16
Lawless HT, Heymann H (1999) Sensory evaluation of food. Principles and Practices. Chapman and Hall, New York
Lê S, Husson F (2008) SensoMineR: a package for sensory data analysis. J Sens Stud 23(1):14–25
Lea P, Næs T, Rødbotten M (1997) Analysis of variance of sensory data. Wiley, New York
Luciano G, Næs T (2008) Interpreting sensory data by combining principal component analysis and analysis of variance. Food Qual Pref 20:167–175
MacKay DB (2005) Probabilistic scaling analyses of sensory profile, instrumental and hedonic data. J Chemometr 19(3):180–190
Macmillan NA, Creelman CD (2005) Detection theory, a user's guide, 2nd edn. Mahwah, N.J.: Lawrence Erlbaum Associates
Martens H, Martens M (2001) Multivariate analysis of quality: an introduction. Wiley, Chicester, UK
Martens H, Hoy M, Wise B, Bro R, Brockhoff PB (2003) Prewhitening of data by covariance-weighted pre-processing. J Chemometr 17(3):153–165
McEwen JA (1996) Preference mapping for product optimization. In: Næs T, Risvik E (eds) Multivariate analysis of data in sensory science, vol 16, Data handeling in science and technology. Elsevier Science, B.V., pp 71–102
Meullenet J-F, Xiong R, Findlay CJ (2007) Multivariate and probabilistic analysis of sensory science problems. Blackwell, Ames, USA
Munck L (2007) A new holistic exploratory approach to Systems biology by near infrared spectroscopy evaluated by chemometrics and data inspection. J Chemometr 21:406–426
Næs T, Tomic O, Brockhoff PB (2010) Statistics for sensory and consumer science. Wiley, New York
Rayner JCW, Best DJ, Brockhoff PB, Rayner GD (2005) Nonparametrics for Sensory science: a more informative approach. Blackwell, USA
Schlich P (1998) What are the sensory differences among coffees? Multi-panel analysis of variance and flash analysis. Food Qual Prefer 9:103
Smith A, Cullis B, Brockhoff P, Thompson R (2003) Multiplicative mixed models for the analysis of sensory evaluation data. Food Qual Prefer 14(5–6):387–395
Tomic O, Nilsen AN, Martens M, Næs T (2007) Visualization of sensory profiling data for performance monitoring. LWT – Food Sci Technol 40:262–269

Sequential Probability Ratio Test

Walter W. Piegorsch[1], William J. Padgett[2]
[1]Professor, Chair
University of Arizona, Tucson, AZ, USA
[2]Distinguished Professor Emeritus of Statistics
University of South Carolina, Columbia, SC, USA

Introduction: Sequential Testing and Sequential Probability Ratios

An important topic in statistical theory and practice concerns the analysis of data that are sampled sequentially. The development of powerful mathematical and statistical tools for the analysis of sequential data is a critical area in statistical research. Our emphasis in this short, introductory exposition is on sequential testing, and in particular on the best-known version for such testing, the *sequential probability ratio test*.

Suppose we are given two hypotheses about the underlying distribution of a random variable X: $H_0 : X \sim f_0(x)$ *vs* $H_a : X \sim f_1(x)$, for two probability density functions (pdfs) or probability mass functions (pmfs) $f_i(x), i = 0, 1$. To perform a sequential test of H_0 *vs.* H_a, we sample individual observations one at a time, and assess in a series of separate steps whether or not the accumulated information favors departure from H_0:

Step 0: Begin by setting two constants, A and B, such that $0 < A < 1 < B$.
Step 1: Observe X_1. Compute the probability ratio $\Lambda_1 = f_1(x_1)/f_0(x_1)$. Since very large values of this ratio support H_a, reject H_0 if $\Lambda_1 \geq B$. Alternatively, since very small values of this ratio support H_0, accept H_0 if $\Lambda_1 \leq A$. The sequential approach also allows for an indeterminate outcome, so if $A < \Lambda_1 < B$, continue sampling and go to Step 2.
Step 2: Observe X_2. Compute the probability ratio $\Lambda_2 = f_1(x_1, x_2)/f_0(x_1, x_2)$. As in Step 1, if $\Lambda_2 \geq B$, reject H_0, while if $\Lambda_2 \leq A$, accept H_0. If $A < \Lambda_2 < B$, continue sampling and observe X_3.

⋮

Step n: Observe X_n. Compute the probability ratio $\Lambda_n = f_1(x_1, x_2, \ldots, x_n)/f_0(x_1, x_2, \ldots, x_n)$. As in Step 1, if $\Lambda_n \geq B$, reject H_0, while if $\Lambda_n \leq A$, accept H_0. If $A < \Lambda_n < B$, continue sampling and observe X_{n+1}. (etc.)

This is known as a *Sequential Probability Ratio Test* (*SPRT*), due to Wald (1945a; 1945b).

Notice that in the typical setting where the individual observations are sampled independently from $f_0(x)$ or $f_1(x)$, the probability ratios take the form

$\Lambda_n = \prod_{i=1}^{n}\{f_1(x_i)/f_0(x_i)\}$. Then, the continuance condition $A < \Lambda_n < B$ is equivalent to $\log\{A\} < \log\left\{\prod_{i=1}^{n}[f_1(x_i)/f_0(x_i)]\right\} < \log\{B\}$. For $D_i = \log\{f_1(x_i)\} - \log\{f_0(x_i)\}$ at any $i = 1, 2, \ldots$, this simplifies to

$$\log\{A\} < \sum_{i=1}^{n} D_i < \log\{B\}. \tag{1}$$

An idealized schematic of this procedure can be given, analogous to Fig. 6–13 of Lindgren (1976), for example. For specific choices of f_0 and f_1, one can often simplify (1) even further. Example 1 illustrates the approach.

Example 1 *The Exponential Family*

Suppose we test the simple hypotheses $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_1$. Let the X_is be independent and identically distributed (i.i.d.) with underlying pdf or pmf taken from the *exponential family* of probability functions (Pierce 1998): $f(x) = h(x)c(\theta)e^{\omega(\theta)t(x)}$. Then, the continuance condition simplifies to $\log\{A\} < n\log\{c(\theta_1)/c(\theta_0)\} + [\omega(\theta_1) - \omega(\theta_0)]\sum_{i=1}^{n} t(X_i) < \log\{B\}$, which if $\omega(\theta_1) - \omega(\theta_0) > 0$ becomes

$$a_n < \sum_{i=1}^{n} t(X_i) < b_n, \tag{2}$$

where

$$a_n = \frac{\log\{A\} - n\log\left[\dfrac{c(\theta_1)}{c(\theta_0)}\right]}{\omega(\theta_1) - \omega(\theta_0)} \quad \text{and}$$

$$b_n = \frac{\log\{B\} - n\log\left[\dfrac{c(\theta_1)}{c(\theta_0)}\right]}{\omega(\theta_1) - \omega(\theta_0)}.$$

[If $\omega(\theta_1) - \omega(\theta_0) < 0$, then the inequalities in (2) are reversed.] Notice that the central quantity in (2) is the sufficient statistic $T_n = \sum_{i=1}^{n} t(X_i)$.

For instance, suppose we sample randomly from the single-parameter exponential distribution with mean $\theta, X_i \sim$i.i.d. Exp(θ), and wish to test $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_1$, where $\theta_1 > \theta_0$. The pdf has the form $f(x|\theta) = \theta^{-1}\exp\{-x/\theta\}I_{(0,\infty)}(x)$, which is a member of the exponential family with $c(\theta) = \theta^{-1}$, $\omega(\theta) = -\theta^{-1}$, and $t(x) = x$. Thus $\log\{\Lambda_n\} = n\log\{\theta_0/\theta_1\} + \left[\theta_0^{-1} - \theta_1^{-1}\right]\sum_{i=1}^{n} X_i$. The continuance region's form can be simplified here by noting that since $\theta_1 > \theta_0$, we have $\omega(\theta_1) - \omega(\theta_0) = \theta_0^{-1} - \theta_1^{-1} > 0$, so (2) applies: continue sampling when $a_n < \sum_{i=1}^{n} X_i < b_n$, for

$$a_n = \frac{\log\{A\} - n\log\left[\dfrac{\theta_0}{\theta_1}\right]}{\theta_0^{-1} - \theta_1^{-1}} \quad \text{and}$$

$$b_n = \frac{\log\{B\} - n\log\left[\dfrac{\theta_0}{\theta_1}\right]}{\theta_0^{-1} - \theta_1^{-1}}.$$

Otherwise, reject H_0 when $\sum_{i=1}^{n} X_i \geq b_n$, or accept H_0 when $\sum_{i=1}^{n} X_i \leq a_n$.

Choosing the Sequential Limits *A* and *B*

For most hypothesis tests, concern centers on the testing error rates, i.e., the Type I error rate, $\alpha = P[\text{reject } H_0 | H_0 \text{ true}]$, and the Type II error rate, $\beta = P[\text{accept } H_0 | H_0 \text{ false}]$. For the *SPRT* these quantities will both be functions of A and B, thus one could in principle invert the relationships and select A and B as functions of α and β. Unfortunately, *SPRT* error rates in these forms are difficult to evaluate. It is possible to approximate them, however, as the following theorem shows.

Theorem 1 The *SPRT* as defined above relates its continuance limits and Type I and II error rates via

$$B \leq (1-\beta)/\alpha \qquad \text{and} \qquad A \geq \beta/(1-\alpha). \tag{3}$$

See, e.g., Wald (1947, §3.2) for a proof. The Theorem may be used to define A and B as functions of α and β by choosing A and B to satisfy the equalities in (3): given nominal error rates α^* and β^*, use (3) to set

$$B = (1-\beta^*)/\alpha^* \qquad \text{and} \qquad A = \beta^*/(1-\alpha^*). \tag{4}$$

Of course, these choices of A and B do not ensure that the actual underlying Type I and Type II error rates, α and β, respectively, will attain the nominally-chosen rates α^* and β^*. However, one can produce a series of upper bounds using (3) and (4) to obtain $\alpha + \beta \leq \alpha^* + \beta^*$, $\alpha \leq \alpha^*/(1-\beta^*)$ and $\beta \leq \beta^*/(1-\alpha^*)$. Wald (1947, §3.3) notes that for most typical values of α^* and β^* these bounds are often rather tight and may even be negligible in practice.

Example 2 Suppose we set the nominal error rates to $\alpha^* = 0.01$ and $\beta^* = 0.05$. Then we find $\alpha + \beta \leq 0.06$, while the individual error rates are bounded as $\alpha \leq (0.01)/(0.95) = 0.0105$ and $\beta \leq (0.05)/(0.99) = 0.0505$.

Finite Termination and Average Sample Number (ASN)

Notice that the (final) sample size N of any sequential test procedure is not a fixed quantity, but is in fact a random

variable determined from the data. As such, an obvious concern with any form of sequential test is whether or not the method eventually terminates. Luckily, for i.i.d. sampling the *SPRT* possesses a finite termination characteristic in that $P[N < \infty] = 1$. This holds under either H_0 or H_a, and is based on a more general result given by Wald (1944); also see Lehmann (1959, §3.10). The larger literature on finite termination of sequential tests is quite diverse; some historically interesting expositions are available in, e.g., David and Kruskal (1956), Savage and Savage (1965), or Wijsman (1967).

When $P[N < \infty] = 1$, it is reasonable to ask what the *expected* sample size, $E[N]$, is for a given *SPRT*. This is known as the *average sample number* (*ASN*) or *expected sample number* (*ESN*). A basic result for the *ASN* is available via the following theorem (Wald 1945b):

Theorem 2 (Wald's Equation): Let $D_1, D_2, \ldots$ be a sequence of i.i.d. random variables with $E[|D_i|] < \infty$. Let $N > 0$ be an integer-valued random variable whose realized value, n, depends only on $D_1, \ldots, D_n$, with $E[N] < \infty$. Then $E[D_1 + D_2 + \ldots + D_N] = E[N] \cdot E[D_1]$.

A consequence of Wald's Equation is the immediate application to the *SPRT* and its *ASN*. Clearly $\log\{\Lambda_N\} = \log\{f_1(x_1)/f_0(x_1)\} + \cdots + \log\{f_1(x_N)/f_0(x_N)\} = \sum_{i=1}^{N} D_i$. So, applying Wald's equation yields $E[N] = E[\log\{\Lambda_N\}]/E[D]$, where $D = \log\{f_1(X)/f_0(X)\}$. This result lends itself to a series of approximations. For instance, if H_0 is rejected at some N, $\log\{\Lambda_N\} \approx \log\{B\}$. Or, if H_0 is accepted at some N, $\log\{\Lambda_N\} \approx \log\{A\}$. Thus, under H_0, $E[\log\{\Lambda_N\}|H_0] \approx \alpha \cdot \log\{B\} + (1-\alpha)\log\{A\}$, so $E[N|H_0] \approx [\alpha \cdot \log\{B\} + (1-\alpha)\log\{A\}]/E[D|H_0]$. Similarly, $E[N|H_a] \approx [(1-\beta)\log\{B\} + \beta \cdot \log\{A\}]/E[D|H_a]$. For any given parametric configuration, these relationships may be used to determine approximate values for *ASN*. Wald (1946) gives some further results on ways to manipulate the *ASN*.

An important reason for employing the *SPRT*, at least for the case of testing simple hypotheses, is that it achieves optimal *ASN*s: if the X_is are i.i.d., then for testing $H_0 : \theta = \theta_0$ *vs.* $H_a : \theta = \theta_1$ both $E[N|H_0]$ and $E[N|H_a]$ are minimized among all sequential tests whose error probabilities are at most equal to those of the *SPRT* (Wald and Wolfowitz 1948). For testing composite hypotheses, the theory of *SPRT*s is more complex, although a variety of interesting results are possible (Stuart et al. 1999, §24.23–24; Lai 2001, §2). In his original article, Wald (1945a) himself discussed the problem of sequential testing of composite hypotheses on a binomial parameter; also see Siegmund (1985, §II.3). For testing with normally distributed samples, various forms of sequential t-tests have been proposed; see Jennison and Turnbull (1991) and the references therein for a useful discussion on sequential t-tests (and sequential χ^2- and F-tests) that includes the important problem of *group sequential testing*.

Since Wald's formalization of the *SPRT*, a number of powerful, alternative formulations/constructions have led to wide application of the method. We provide here a short introduction to the basic mathematical underpinnings; however, comprehensive reviews on the larger area of sequential analysis date as far back as Johnson (1961), along with more modern expositions given by Lai (1998, 2001, 2004) and Ghosh (2004). For a perspective emphasizing ▶sequential sampling, see Mukhopadhyay (2002). Also see the book-length treatments by Siegmund (1985), Ghosh and Sen (1991), or Mukhopadhyay and de Silva (2008), along with Wald's (1947) classic text. For cutting-edge developments a dedicated scientific journal exists: *Sequential Analysis*, with more information available online at the website http://www.informaworld.com/smpp/title~db=all~content=t713597296.

Acknowledgments

Thanks are due the Editor and an anonymous referee for their helpful suggestions on an earlier draft of the manuscript. The first author's research was support in part by grant #RD-83241902 from the U.S. Environmental Protection Agency and by grant #R21-ES016791 from the U.S. National Institute of Environmental Health Sciences. The contents herein are solely the responsibility of the authors and do not necessarily reflect the official views of these agencies.

About the Authors

Walter W. Piegorsch is Chair of the Graduate Interdisciplinary Program (GIDP) in Statistics at the University of Arizona, Tucson, AZ. He is also a Professor of Mathematics, a Professor of Public Health, and Director of Statistical Research & Education at the University's BIO5 Institute for Collaborative Bioresearch. Professor Piegorsch has held a number of professional positions, including Chairman of the American Statistical Association Section on Statistics & the Environment (2004), Vice-Chair of the American Statistical Association Council of Sections Governing Board (1997–1999), and election to the Council of the International Biometric Society (2002–2005). He serves as Editor-in-Chief of *Environmetrics*, and also has served as Joint-Editor of the *Journal of the American Statistical Association* (Theory & Methods Section).

Dr. Piegorsch was named a Fellow of the American Statistical Association (1995), a Member (by Election, 1995) of the International Statistical Institute, and has received the Distinguished Achievement Medal of the American Statistical Association Section on Statistics and the Environment (1993), and was a Co-recipient of The Ergonomics Society/Elsevier Ltd. Applied Ergonomics Award (2007).

For biography of William J. Padgett *see* the entry ►Weibull distribution.

Cross References

►Exponential Family Models
►Optimal Stopping Rules
►Sequential Sampling

References and Further Reading

David HT, Kruskal WH (1956) The WAGR sequential *t*-test reaches a decision with probability one. Ann Math Stat 27: 797–805
Ghosh BK (2004) Sequential analysis. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B (eds) Encyclopedia of statistical sciences vol 11, 2nd edn. Wiley, New York, pp 7605–7613
Ghosh BK, Sen PK (eds) (1991) Handbook of sequential analysis, M. Dekker, New York
Jennison C, Turnbull BW (1991) Exact calculations for sequential *t*, χ^2 and *F* tests. Biometrika 78:133–141
Johnson NL (1961) Sequential analysis: a survey. J Roy Stat Soc Ser A (General) 124:372–411
Lai TL (1998) Sequential analysis. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 5. Wiley, New York, pp 4074–4079
Lai TL (2001) Sequential analysis: some classical problems and new challenges (with discussion). Stat Sin 11:303–408
Lai TL (2004) Likelihood ratio identities and their applications to sequential analysis (with discussion). Sequential Anal 23: 467–556
Lehmann EL (1959) Testing statistical hypotheses, 1st edn. Wiley, New York
Lindgren BW (1976) Statistical theory, 3rd edn. Macmillan, New York
Mukhopadhyay N (2002) Sequential sampling. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, vol 4. Wiley, Chichester, pp 1983–1988
Mukhopadhyay N, de Silva BM (2008) Sequential methods and their applications. Chapman & Hall/CRC, Boca Raton, FL
Pierce DA (1998) Exponential family. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 2. Wiley, New York, pp 1448–1455
Savage IR, Savage LJ (1965) Finite stopping time and finite expected stopping time. J Roy Stat Soc Ser B (Meth) 27:284–289
Siegmund D (1985) Sequential analysis: tests and confidence intervals. Springer-Verlag, New York
Stuart A, Ord JK, Arnold S (1999) Kendall's advanced theory of statistics: Volume 2A-Classical inference and the linear model, 6th edn. Arnold, London
Wald A (1944) On cumulative sums of random variables. Ann Math Stat 15:283–296
Wald A (1945a) Sequential tests of statistical hypotheses. Ann Math Stat 16:117–186
Wald A (1945b) Some generalizations of the theory of cumulative sums of random variables. Ann Math Stat 16:287–293
Wald A (1946) Some improvements in setting limits for the expected number of observations required by a sequential probability ratio test. Ann Math Stat 17:466–474
Wald A (1947) Sequential analysis. Wiley, New York
Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339
Wijsman RA (1967) General proof of termination with probability one of invariant sequential probability ratio tests based on multivariate normal observations. Ann Math Stat 38: 8–24

Sequential Ranks

Estate V. Khmaladze
Professor
Victoria University of Wellington, Wellington, New Zealand

To discuss sequential ranks it will be more helpful to present them in comparison with ordinary ranks.

Suppose $X_1, \ldots, X_n$ is a sequence of random variables. Denote by $\mathbb{I}_A$ the indicator function of an event A. For each X_i consider now what one can call its "ordinary" rank:

$$R_{in} = \sum_{j=1}^{n} \mathbb{I}_{\{X_j \le X_i\}}.$$

So, R_{in} counts the number of our random variables that take values not exceeding X_i. For example, if X_i happens to be the smallest, its rank will be 1, and if it happens to be the largest, its rank will be n. If the joint distribution of $X_1, \ldots, X_n$ is absolutely continuous, then with probability 1 all values of our random variables will be different. Therefore, for any integer $k = 1, \ldots, n$ there will be one and only one random variable with rank equal to k. For example, for $n = 5$, if our X_i-s happened to be

$$-1.31,\ 0.24,\ -3.52,\ 4.11\ and\ 2.25,$$

their ranks will be

$$2,\quad 3,\quad 1,\quad 5\quad and\quad 4.$$

Hence, the vector of "ordinary" ranks $\mathbb{R}_n = \{R_{1n}, \ldots, R_{nn}\}$ is a random permutation of the numbers $\{1, \ldots, n\}$. Thus, its distribution possesses a certain degeneracy. In particular, even if $X_1, \ldots, X_n$ are independent and identically distributed, the ordinary ranks are dependent random variables – for example, if $R_{in} = 3$ it precludes any other rank $R_{jn}, j \ne i$, from taking the value 3, so that the conditional probability $P(R_{jn} = 3 | R_{in} = 3) = 0$, while without this condition $P(R_{jn} = 3)$ does not need to be 0 at all.

Moreover, any symmetric statistic from the vector $\mathbb{R}_n$ is not random and, for given n, must be constant: if ψ is a symmetric function of its n arguments, then

$$\psi(R_{1n},\ldots,R_{nn}) = \psi(1,\ldots,n), \quad \text{e.g.,} \quad \sum_{i=1}^{n}\phi(R_{in}) = \sum_{i=1}^{n}\phi(i).$$

The definition of sequential ranks is slightly different, but the difference in their properties is quite remarkable. Namely, the sequential rank of X_i is defined as

$$S_i = \sum_{j=1}^{i}\mathbb{I}_{\{X_j\le X_i\}}.$$

Therefore, it is the rank of X_i among only "previous" observations, including X_i itself, but not "later" observations $X_{i+1},\ldots,X_n$. For the sample values given above, their sequential ranks are

$$1,\quad 2,\quad 1,\quad 4,\quad 4.$$

The relationship between the vectors of ordinary ranks and sequential ranks is one-to-one. Namely, given vector $\mathbb{R}_n = \{R_{1n},\ldots,R_{nn}\}$ of ordinary ranks, the sums

$$S_i = \sum_{j=1}^{i}\mathbb{I}_{\{R_{jn}\le R_{in}\}}$$

return sequential ranks of $X_1,\ldots,X_n$ and the other way around, given a vector of sequential ranks $\mathbb{S}_n$, if

$$S_{i,i+1} = S_i + \mathbb{I}_{\{S_i\ge S_{i+1}\}}, S_{i,i+2} = S_{i,i+1} + \mathbb{I}_{\{S_{i,i+1}\ge S_{i+2}\}},\ldots,$$

then finally

$$S_{i,n} = R_{in}.$$

Because of this one-to-oneness, the vector $\mathbb{S}_n$ also must have some sort of degeneracy. It does, but in a very mild form: S_1 is always 1.

Assume that $X_1,\ldots,X_n$ are independent and identically distributed random variables with continuous distribution function F. Then $U_1 = F(X_1),\ldots,U_n = F(X_n)$ are independent uniformly distributed on $[0,1]$ random variables. The values of R_{in} and S_i will not change, if we replace X_i-s by U_i-s. Therefore, the distribution of both ranks must be independent of F – they both are "distribution free." We list some properties of $\mathbb{S}_n$ in this situation – they can be found, e.g., in Barndolf-Nielsen (1963), Renyi (1962, 1976), Sen (1981).

The distribution of each S_i is $P(S_i = k) = 1/i, k = 1,\ldots,i$, and, therefore, the distribution function of $S_i/(i+1)$ quickly converges to the uniform distribution function:

$$P\left(\frac{S_i}{i+1} = \frac{k}{i+1}\right) = \frac{1}{i}, \text{ and } \left|P\left(\frac{S_i}{i+1}\le x\right) - x\right| \le \frac{1}{i+1}.$$

Recall that, similarly, for ordinary ranks $P(R_{in} = k) = 1/n, k = 1,\ldots,n$, see, e.g., Hajek and Shidak (1975). However, unlike ordinary ranks, sequential ranks $S_1,\ldots,S_n$ are independent random variables. Hence symmetric statistics from sequential ranks are non-degenerate random variables. For example,

$$\sum_{i=1}^{n}\phi(S_i)$$

is a sum of independent random variables. Also unlike ordinary ranks, with arrival of a new observation X_{n+1} sequential ranks $S_1,\ldots,S_n$ stay unchanged and only one new rank S_{n+1} is to be calculated.

Therefore, asymptotic theory of sequential ranks is relatively simple and computationally they are very convenient.

The ordinary ranks are used in testing problems, usually, through the application of two types of statistics–widely used linear rank statistics and goodness of fit statistics, based on the empirical field

$$z_R(t,u) = \sum_{i=1}^{nt}\left[\mathbb{I}_{\{R_{in}\le u(n+1)\}} - \frac{[nu]}{n+1}\right], \quad (t,u)\in[0,1]^2.$$

Linear rank statistics can also be thought of as based on the field $z_R(t,u)$, and, more exactly, are linear functionals from it:

$$\psi(\mathbb{R}_n) = \int \psi(t,u)z_R(dt,du) = \sum_{i=1}^{n}\left[\psi\left(\frac{i}{n},\frac{R_{in}}{n+1}\right) - E\psi\left(\frac{i}{n},\frac{R_{in}}{n+1}\right)\right]$$

(the term "linear" would not be very understandable otherwise). Without loss of generality one can assume that $\int_0^1\psi(t,u)dt = 0$.

One of the central results in the theory of rank tests, see Hajek and Shidak (1975), is the optimality statement about linear rank statistics. If under the null hypothesis the sample is i.i.d.(F) while under the alternative hypothesis the distribution A_i of each X_i is such that

$$\frac{dA_i(x)}{dF(x)} = 1 + \frac{1}{\sqrt{n}}a\left(\frac{i}{n},F(x)\right) + \textit{smaller terms}, \quad \text{as} n\to\infty, \tag{1}$$

where $\int_0^1 a(t,F(x))dt = 0$, then the linear rank statistic, with ψ equal to a from (1),

$$a(\mathbb{R}_n) = \sum_{i=1}^{n}a\left(\frac{i}{n},\frac{R_{in}}{n+1}\right),$$

is asymptotically optimal against this alternative. Indeed, the statistic

$$\sum_{i=1}^{n}a\left(\frac{i}{n},F(X_i)\right)$$

is the statistic of the asymptotically optimal test for our alternative, based on the observations $X_1, \ldots, X_n$ "themselves," and $R_{in}/(n+1)$ is a "natural" approximation for $F(X_i)$.

Returning to sequential ranks, one can again consider the empirical field

$$z_S(t,u) = \sum_{i=1}^{nt} \left[\mathbb{I}_{\{S_i \le u(i+1)\}} - \frac{[iu]}{i+1} \right], \quad (t,u) \in [0.1]^2,$$

and sequential linear rank statistics, based on it:

$$\phi(\mathbb{S}_n) = \int \phi(t,u) z_S(dt,du) = \sum_{i=1}^{n} \left[\phi\left(\frac{i}{n}, \frac{S_i}{i+1}\right) - E\phi\left(\frac{i}{n}, \frac{S_i}{i+1}\right) \right].$$

Although $S_i/(i+1)$ is no less "natural" an approximation for $F(X_i)$, the statistic

$$a(\mathbb{S}_n) = \sum_{i=1}^{n} a\left(\frac{i}{n}, \frac{S_i}{i+1}\right)$$

is not optimal for the alternative (1) any more. The papers (Khmaladze 1986) and (Pardzhanadze 1986) derived the form of this optimal statistic, and hence established the theory of sequential ranks to the same extent as the theory of "ordinary" rank statistics.

More Specifically, it was shown that the empirical fields z_R and z_S are asymptotically linear transformations of each other and, as a consequence, the two linear rank statistics $\psi(\mathbb{R}_n)$ and $\phi(\mathbb{S}_n)$ have the same limit distribution under the null hypothesis and under any alternative (1) as soon as functions ψ and ϕ are linked as below:

$$\psi(t,u) - \frac{1}{t}\int_0^t \psi(\tau,u)d\tau = \phi(t,u) \quad \text{or}$$
$$\phi(t,u) - \int_t^1 \frac{1}{\tau}\phi(\tau,u)d\tau = \psi(t,u).$$

In particular, both linear rank statistics

$$\sum_{i=1}^{n} a\left(\frac{i}{n}, \frac{R_{in}}{n+1}\right) \quad \text{and} \quad \sum_{i=1}^{n} \left[a\left(\frac{i}{n}, \frac{S_i}{i+1}\right) - \frac{n}{i}\int_0^{i/n} a\left(\tau, \frac{S_i}{i+1}\right) d\tau \right] \tag{2}$$

are asymptotically optimal test statistics against alternative (1).

Two examples of particular interest should clarify the situation further.

Example 1 (Wilcoxon rank (or rank-sum) statistic). In the two-sample problem, when we test if both samples came from the same distribution or not, the following Wilcoxon rank statistic

$$\sum_{i=1}^{m} \frac{R_{in}}{n+1}$$

is most widely used (see ►Wilcoxon–Mann–Whitney Test). Its sequential analogue is not mentioned often, but according to (2) there is such an analogue, which is

$$-\sum_{i=m+1}^{n} \frac{m}{i}\frac{S_i}{i+1}.$$

In general, the following two statistics are asymptotically equivalent:

$$\sum_{i=1}^{m} a\left(\frac{R_{in}}{n+1}\right) \quad \text{and} \quad -\sum_{i=m+1}^{n} \frac{m}{i} a\left(\frac{S_i}{i+1}\right).$$

Note again, that if the size m of the first sample is fixed, but we keep adding new observations to the second sample, so that $n-m$ keeps increasing, we would only need to add new summands to the sequential rank statistics, on the right, without changing the previous summands.

Example 2 (Kendall's τ and Spearman's ρ rank correlation coefficients). The latter correlation coefficient has the form

$$\rho_n = \sum_{i=1}^{n} \frac{i}{n}\left(\frac{R_{in}}{n+1} - \frac{1}{2}\right)$$

while the former is

$$\tau_n = \sum_{i=1}^{n} \frac{i}{n}\left(\frac{S_i}{i+1} - \frac{1}{2}\right).$$

These two coefficients are usually perceived as different statistics. However, from (2) it follows that they also are asymptotically equivalent.

Among other papers that helped to form and advance the theory of sequential ranks we refer to Müller-Funk (1983), Renyi (1962, 1976), and Reynolds (1975). Among more recent papers and applications to change-point problem we would point to Bhattacharya and Zhou (1994), Gordon and Pollak (1994), and Malov (1993).

About the Author

For biography *see* the entry ►Testing Exponentiality of Distribution.

Cross References

►Kendall's Tau
►Measures of Dependence
►Record Statistics
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Barndolf-Nielsen O (1963) On limit behaviour of extreme order statistics. Ann Math Stat 34:992–1002

Bhattacharya PK, Zhou H (1994) A rank cusum procedure for detecting small changes in a symmetric distribution, Change-Point problems. IMS Lecture notes, vol 23

Gordon L, Pollak M (1994) An efficient sequential nonparametric scheme for detecting a change in distribution. Ann Stat 22: 763–804

Hajek J, Shidak Z (1975) Theory of rank tests Academic, CSZV, Prague

Khmaladze EV, Parjanadze AM (1986) Functional limit theorems for linear statistics of sequential ranks. Probab theor relat fields 73:1285–1295

Malov SV (1993) Sequential ranks and order statistics. Notes Sci Seminars POMI 204:115–125

Müller-Funk U (1983) Sequential signed rank statistics. Sequential Anal Design Meth Appl 2:123–148

Pardzhanadze AM, Khmaladze EV (1986) On the asymptotic theory of statistics based on sequential ranks. Theor Probab Appl 31:669–682

Renyi A (1962, 1976) On the extreme elements of observations Academiai Kiado. Selected papers of Alfred Renyi

Reynolds M (1975) A sequential rank test for symmetry. Ann Stat 3:382–400

Sen PK (1981) Sequential non-parametrics. Wiley, New York

Sequential Sampling

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

Sequential sampling entails observing data in a sequence. How long should one keep observing data? That will largely depend on the preset levels of errors that one may be willing to live with and the optimization techniques that may be required. In the early 1940s, Abraham Wald developed the theory and practice of the famous *sequential probability ratio test* (SPRT) to decide between a simple null hypothesis and a simple alternative hypothesis (Wald 1947). Wald and Wolfowitz (1948) proved optimality of Wald's SPRT within a large class of tests, including Neyman and Pearson's (1933) UMP test, in the sense that the SPRT needs on an average fewer observations under either hypothesis. These were mentioned in another chapter.

For a comprehensive review, one should refer to the *Handbook of Sequential Analysis*, a landmark volume that was edited by Ghosh and Sen (1991). This nearly 20 years old handbook is still one of the most prized resource in this whole field.

Section ▶"Why Sequential Sampling?" explains with Examples 1 and 2 why one must use sequential sampling strategies to solve certain statistical problems. We especially highlight the Stein (1945, 1949) path-breaking two-stage and the Ray (1957) and Chow and Robbins (1965) purely sequential fixed-width confidence interval procedures in sections ▶"Stein's Two-stage Sampling" and "Purely Sequential Sampling" respectively.

Sections ▶"Two-stage Sampling" and "Purely Sequential Sampling" analogously highlight the Ghosh and Mukhopadhyay (1976) two-stage and the Robbins (1959) purely sequential bounded-risk point estimation procedures respectively. Both sections ▶"Two-stage and Sequential Fixed-width Confidence Interval" and "Two-stage and Sequential Bounded Risk Point Estimation" handle the problems of estimating an unknown mean of a normal distribution whose variance is also assumed unknown.

Section ▶"Which Areas Are Hot Beds for Sequential Sampling?" briefly mentions applications of sequential and multi-stage sampling strategies in concrete problems that are in the cutting edge of statistical research today.

Why Sequential Sampling?

There is a large body of statistical inference problems that cannot be solved by any fixed-sample-size procedure. We will highlight two specific examples. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ where $-\infty < \mu < \infty, 0 < \sigma^2 < \infty$ are both unknown parameters, and $n(\geq 2)$ is fixed.

Example 1 We want to construct a confidence interval I for μ such that (i) the length of I is $2d(> 0)$ where d is preassigned, and (ii) the associated confidence coefficient, $P_{\mu,\sigma^2}\{\mu \in I\} \geq 1 - \alpha$ where $0 < \alpha < 1$ is also preassigned. Dantzig (1940) showed that this problem has no solution regardless of the form of the confidence interval I when n is fixed in advance.

Example 2 Suppose that $\overline{X}_n$, the sample mean, estimates μ and we want to claim its bounded-risk property, namely that $\sup_{\mu,\sigma^2} E[(\overline{X}_n - \mu)^2] \leq \omega$ where $\omega(> 0)$ is a preassigned risk-bound. This problem also has no solution regardless of the form of the estimator of μ.

Theorem 1 *Suppose that $X_1, \cdots, X_n$ are iid with a probability density function $\frac{1}{\sigma} f\left(\sigma^{-1}(x - \theta)\right)$ where $-\infty < \theta < \infty, 0 < \sigma < \infty$ are two unknown parameters. For estimating θ, let the loss function be given by $W(\theta, \delta(\boldsymbol{x})) = H(|\delta(\boldsymbol{x}) - \theta|)$ where $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a realization of $\boldsymbol{X} = (X_1, \cdots, X_n)$. Assume that $H(|u|) \uparrow |u|$, and let $M =$*

$\sup_{-\infty<u<\infty} H(|u|)$, *which may be infinite. Then, for any fixed* $L < M$, *there does not exist an estimator* $\delta(\boldsymbol{X})$ *such that* $\sup_{\theta,\sigma} E_{\theta,\sigma}\{W(\theta, \delta(\boldsymbol{X}))\} \le L$.

This statement is similar to that of Theorem 3.7.1 in Ghosh et al. (1997) and Theorem 2.3.1 in Mukhopadhyay and de Silva (2009). It was originally proved in Lehmann (1951).

Theorem 1 proves immediately the non-existence of a fixed-sample-size methodology to solve the problems mentioned in Examples 1–2 exactly. There are these and numerous other inference problems where we have no fixed-sample-size procedure at all to talk about. In order to address this class of important inference problems, an appropriately designed sequential sampling procedure is a must.

Two-Stage and Sequential Fixed-Width Confidence Interval

In the context of Example 1, we first summarize Stein's (1945, 1949) two-stage procedure and then the purely sequential procedure due to Ray (1957) and Chow and Robbins (1965).

Stein's Two-Stage Sampling

Stein (1945, 1949) gave his path-breaking two-stage sampling design to solve *exactly* the problem mentioned in Example 1. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\ge 2)$. Let $a_{m-1} \equiv a_{m-1,\alpha/2}$ be the upper $50\alpha\%$ point of the Student's t distribution with $m-1$ degrees of freedom. Now, based on $X_1, \cdots, X_m$, we obtain the sample variance, $S_m^2 = (m-1)^{-1}\Sigma_{i=1}^m (X_i - \overline{X}_m)^2$ which estimates unknown σ^2. Let us denote $\langle u \rangle$ = the largest integer $< u$, $u > 0$.

We define the final sample size as

$$N \equiv N(d) = \max\left\{m, \left\langle \frac{a_{m-1}^2 S_m^2}{d^2} \right\rangle + 1\right\}. \tag{1}$$

It is easy to see that N is finite with probability one. This two-stage procedure is implemented as follows:

If $N = m$, it indicates that we already have too many observations at the pilot stage. Hence, we do not need any more observations at the second stage.

But, if $N > m$, it indicates that we have started with too few observations at the pilot stage. Hence, we sample the difference at the second stage by gathering new observations $X_{m+1}, \cdots, X_N$ at the second stage.

Case 1. If $N = m$, the final dataset is $X_1, \cdots, X_m$

Case 2. If $N > m$, the final dataset is $X_1, \cdots, X_m, X_{m+1}, \cdots, X_N$

Combining the two possibilities, one can say that the final dataset is composed of N and $X_1, \cdots, X_N$. This gives rise to the sample mean $\overline{X}_N$ and the associated fixed-width interval $I_N = \left[\overline{X}_N \pm d\right]$.

It is clear that (i) the event $\{N = n\}$ depends only on the random variable S_m^2, and (ii) $\overline{X}_n, S_m^2$ are independent random variables, for all fixed $n(\ge m)$. So, any event defined only through $\overline{X}_n$ must be independent of the event $\{N = n\}$. Using these tools, Stein (1945, 1949) proved the following result that is considered a breakthrough. More details can be found in Mukhopadhyay and de Silva (2009, Sect. 6.2.1).

Theorem 2 $P_{\mu,\sigma^2}\left\{\mu \in \left[\overline{X}_N \pm d\right]\right\} \ge 1 - \alpha$ *for all fixed* $d > 0$, $0 < \alpha < 1$, μ, *and* σ^2.

It is clear that the final sample size N from (1) tried to mimic the optimal fixed sample size C, the smallest integer $\ge z_{\alpha/2}^2 \sigma^2 d^{-2}$, had σ^2 been known. This procedure, however, is known for its significant oversampling on an average.

Purely Sequential Sampling

In order to overcome significant oversampling, Ray (1957) and Chow and Robbins (1965) proposed a purely sequential procedure. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\ge 2)$, and then proceed by taking one additional observation at-a-time until the sampling process terminates according to the following stopping rule: With $\overline{X}_n = n^{-1}\Sigma_{i=1}^n X_i$ and $S_n^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X}_n)^2$, let

$$N \equiv N(d) = \inf\left\{n \ge m : n \ge \frac{z_{\alpha/2}^2 S_n^2}{d^2}\right\}. \tag{2}$$

It is easy to see that N is finite with probability one. Based on the final dataset composed of N and $X_1, \cdots, X_N$, one finds $\overline{X}_N$ and proposes the associated fixed-width interval $I_N = \left[\overline{X}_N \pm d\right]$. Now, one can prove that asymptotically, $P_{\mu,\sigma^2}\{\mu \in \left[\overline{X}_N \pm d\right]\} \to 1 - \alpha$ for all fixed $0 < \alpha < 1$, μ, and σ^2 as $C \to \infty$ when $m \ge 2$.

One can also prove that $E_{\sigma^2}[N - C] = -1.1825$ if $m \ge 4$. This property is referred to as the *asymptotic second-order efficiency* according to Ghosh and Mukhopadhyay (1981). One has to employ mathematical tools from nonlinear renewal theory to prove such a property. The nonlinear renewal theory has been fully developed by Woodroofe (1977) and Lai and Siegmund (1977, 1979).

Two-Stage and Sequential Bounded Risk Point Estimation

In the context of Example 2, we first summarize a two-stage procedure from Ghosh and Mukhopadhyay (1976) followed by a purely sequential procedure along the line of Robbins (1959).

Two-Stage Sampling

Ghosh and Mukhopadhyay (1976) discussed a two-stage sampling design analogous to (1) to solve *exactly* the problem mentioned in Example 2. We again start with pilot observations $X_1, \cdots, X_m$ where $m(\geq 4)$ is the pilot size and obtain S_m^2. Define the final sample size as:

$$N \equiv N(\omega) = \max\left\{m, \left\langle \frac{b_m S_m^2}{\omega} \right\rangle + 1\right\} \qquad (3)$$

where $b_m = \frac{m-1}{m-3}$. It is easy to see that N is finite with probability one.

The two-stage sampling scheme is implemented as before.

Case 1. If $N = m$, the final dataset is $X_1, \cdots, X_m$

Case 2. If $N > m$, the final dataset is $X_1, \cdots, X_m, X_{m+1}, \cdots, X_N$

Combining the two situations, one can see that the final dataset is again composed of N and $X_1, \cdots, X_N$ which give rise to an estimator $\overline{X}_N$ for μ.

Now, we recall that $\overline{X}_n$ is independent of the event $\{N = n\}$ for all fixed $n(\geq m)$. Hence, we can express the risk associated with the estimator $\overline{X}_N$ as follows:

$$E_{\mu,\sigma^2}\left\{(\overline{X}_N - \mu)^2\right\} = \sigma^2 E_{\mu,\sigma^2}\left[N^{-1}\right],$$

which will not exceed the set risk-bound ω for all fixed μ and σ^2. More details can be found in Mukhopadhyay and de Silva (2009, Sect. 6.3).

It is clear that the final sample size N from (3) tried to mimic the optimal fixed sample size n^*, the smallest integer $\geq \sigma^2\omega^{-1}$, had σ^2 been known. This procedure is also well-known for its significant oversampling on an average.

For either problem, there are more efficient two-stage, three-stage, accelerated sequential, and other estimation methodologies available in the literature. One may begin by reviewing this field from Mukhopadhyay and Solanky (1994), Ghosh et al. (1997), Mukhopadhyay and de Silva (2009), among other sources.

Purely Sequential Sampling

In order to overcome significant oversampling, along the line of Robbins (1959), one can propose the following purely sequential procedure. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\geq 2)$, and then proceed by taking one additional observation at-a-time until the sampling process terminates according to the following stopping rule: Let

$$N \equiv N(\omega) = \inf\left\{n \geq m : \; n \geq \frac{S_n^2}{\omega}\right\}. \qquad (4)$$

It is easy to see that N is finite with probability one. Based on the final dataset composed of N and $X_1, \cdots, X_N$, one finds $\overline{X}_N$ and proposes the associated estimator $\overline{X}_N$ for μ. Now, one can prove that asymptotically, $\omega^{-1}E_{\mu,\sigma^2}\left\{(\overline{X}_N - \mu)^2\right\} \to 1$ for all fixed μ, and σ^2 as $n^* \to \infty$ when $m \geq 2$.

One can again prove that $E_{\sigma^2}[N - C]$ is bounded by appealing to nonlinear renewal theory. This property is referred to as the *asymptotic second-order efficiency* according to Ghosh and Mukhopadhyay (1981).

Which Areas Are Hot Beds for Sequential Sampling?

First, we should add that all computer programs necessary to implement the sampling strategies mentioned in sections ▶"Two-stage and Sequential Fixed-width Confidence Interval" and "Two-stage and Sequential Bounded Risk Point Estimation" are available in conjunction with the recent book of Mukhopadhyay and de Silva (2009).

Sequential and multi-stage sampling techniques are implemented practically in all major areas of statistical science today. Some modern areas of numerous applications *include* change-point detection, clinical trials, computer network security, computer simulations, ▶data mining, disease mapping, educational psychology, financial mathematics, group sequential experiments, horticulture, infestation, kernel density estimation, longitudinal responses, multiple comparisons, nonparametric functional estimation, ordering of genes, ▶randomization tests, reliability analysis, scan statistics, selection and ranking, sonar, surveillance, survival analysis, tracking, and water quality.

In a majority of associated statistical problems, sequential and multi-stage sampling techniques are absolutely essential in the sense of our prior discussions in section ▶"Why Sequential Sampling?". In other problems, appropriate sequential and multi-stage sampling techniques are more efficient than their fixed-sample-size counterparts, if any.

For an appreciation of concrete real-life problems involving many aspects of sequential sampling, one may refer to *Applied Sequential Methodologies*, a volume edited by Mukhopadhyay et al. (2004).

About the Author

Dr. Nitis Mukhopadhyay is professor of statistics, Department of Statistics, University of Connecticut, USA. He is Editor-in-Chief of *Sequential Analysis* since 2004. He is Associate Editor for *Calcutta Statistical Association Bulletin* (since 1998), *Communications in Statistics* (since 2002) and *Statistical Methodology* (since 2004). He is Chair of the National Committee on Filming Distinguished Statisticians of the American Statistical Association since 2002. In 2002, he has been named IMS fellow for "outstanding contribution in sequential analysis and multistage sampling; pathbreaking research in selection and ranking; authoritative books; exemplary editorial service; innovative teaching and advising; and exceptional dedication to preserve and celebrate statistical history through films and scientific interviews." He is also an Elected Fellow of The American Statistical Association (2003), and Elected Ordinary Member of The International Statistical Institute (2007), and a life member of: the International Indian Statistical Association, the Calcutta Statistical Association and the Statistical Society of Sri Lanka. Professor Mukhopadhyay was elected a Director of the Calcutta Statistical Association for the period 2005–2008. He has authored/coauthored about 170 papers in international journals and 7 books including, *Sequential Methods and Their Applications* (Chapman & Hall/CRC, Boca Raton, 2009).

Cross References

►Acceptance Sampling
►Loss Function
►Optimal Stopping Rules
►Ranking and Selection Procedures and Related Inference Problems
►Sampling Algorithms
►Sequential Probability Ratio Test

References and Further Reading

Chow YS, Robbins H (1965) On the asymptotic theory of fixed width sequential confidence intervals for the mean. Ann Math Stat 36:457–462
Dantzig GB (1940) On the non-existence of tests of Student's hypothesis having power functions independent of σ. Ann Math Stat 11:186–192
Ghosh BK, Sen PK (eds) (1991) Handbook of sequential analysis. Marcel Dekker, New York
Ghosh M, Mukhopadhyay N (1976) On two fundamental problems of sequential estimation. Sankhya B 38:203–218
Ghosh M, Mukhopadhyay N (1981) Consistency and asymptotic efficiency of two-stage and sequential procedures. Sankhya A 43:220–227
Ghosh M, Mukhopadhyay N, Sen PK (1997) Sequential estimation. Wiley, New York
Lai TL, Siegmund D (1977) A nonlinear renewal theory with applications to sequential analysis I. Ann Stat 5:946–954
Lai TL, Siegmund D (1979) A nonlinear renewal theory with applications to sequential analysis II. Ann Stat 7:60–76
Lehmann EL (1951) Notes on the theory of estimation. University of California, Berkeley
Mukhopadhyay N, Datta S, Chattopadhyay S (2004) Applied sequential methodologies, edited volume. Marcel Dekker, New York
Mukhopadhyay N, de Silva BM (2009) Sequential methods and their applications. CRC, New York
Mukhopadhyay N, Solanky TKS (1994) Multistage selection and ranking procedures: second-order asymptotics. Marcel Dekker, New York
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A 231:289–337
Ray WD (1957) Sequential confidence intervals for the mean of a normal population with unknown variance. J R Stat Soc B 19:133–143
Robbins H (1959) Sequential estimation of the mean of a normal Population. In: Grenander U (ed) Probability and statistics (Harald Cramér Volume). Almquist and Wiksell, Uppsala, pp 235–245
Stein C (1945) A two sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat 16:243–258
Stein C (1949) Some problems in sequential estimation. Econometrica 17:77–78
Wald A (1947) Sequential analysis. Wiley, New York
Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339
Woodroofe M (1977) Second order approximations for sequential point and interval estimation. Ann Stat 5:984–995

Sex Ratio at Birth

Johan Fellman
Professor Emeritus
Folkhälsan Institute of Genetics, Helsinki, Finland

Sex Ratio in National Birth Registers

The sex ratio at birth, also called the secondary sex ratio, and here denoted SR, is usually defined as the number of males per 100 females. Among newborns there is almost always a slight excess of boys. Consequently, the SR is greater than 100, mainly around 106.

John Graunt (1620–1674) was the first person to compile data showing an excess of male births to female births and to note spatial and temporal variation in the SR. John Arbuthnot (1667–1735) demonstrated that the excess of males was statistically significant and asserted that the SR is uniform over time and space (Campbell 2001). Referring to christenings in London in the 82 years up to 1710, Arbuthnot suggested that the regularity in the SR and the dominance of males over females

could not be attributed to chance and must be an indication of divine providence. Nicholas Bernoulli's (1695–1726) counter-argument was that Arbuthnot's model was too restrictive. Instead of a fair coin model, the model should be based on an asymmetric coin. Based on the generalized model, chance could give uniform dominance of males over females. Later, Daniel Bernoulli (1700–1782), Pierre Simon de Laplace (1749–1827) and Siméon-Denis Poisson (1781–1840) also contributed to this discussion (David 1962; Hacking 1975).

Some general features of the SR can be noted. Stillbirth rates are usually higher among males than females, and the SR among stillborn infants is markedly higher than normal values, but the excess of males has decreased during the last decades. Hence, the SR among liveborn infants is slightly lower than among all births, but this difference is today very minute. Further, the SR among multiple maternities is lower than among singletons. In addition to these general findings, the SR shows marked regional and temporal variations.

In a long series of papers, attempts have been made to identify factors influencing the SR, but statistical analyses have shown that comparisons demand large data sets. Variations in the SR that have been reliably identified in family data have in general been slight and without notable influence on national birth registers. Attempts to identify reliable associations between SRs and stillbirth rates have been made, but no consistent results have emerged. Hawley (1959) stated that where prenatal losses are low, as in the high standard of living in Western countries, the SRs at birth are usually around 105 to 106. By contrast, in areas with a lower standard of living, where the frequencies of prenatal losses are relatively high, SRs are around 102. Visaria (1967) stressed that available data on late fetal mortality lend at best only weak support for these findings and concluded that racial differences seem to exist in the SR. He also discussed the perplexing finding that the SR among Koreans is high, around 113.

A common pattern observed in different countries is that during the first half of the twentieth century the SR showed increasing trends, but during the second half the trend decreased. Different studies have found marked peaks in the proportion of males during the First and Second World War. It has been questioned whether temporal or spatial variations of the SR are evident, and whether they constitute a essential health event. A common opinion is that secular increases are caused by improved socioeconomic conditions. The recent downward trends in the SRs have been attributed to new reproductive hazards, specifically exposure to environmental oestrogens. However, the turning point of the SR preceded the period of global industrialization and particularly the introduction of pesticides or hormonal drugs, rendering a causal association unlikely.

Sex Ratio in Family Data

In general, factors that affect the SR within families remain poorly understood. In a long series of papers, using family data, attempts have been made to identify factors influencing the SR. Increasing evidence confirms that exposure to chemicals, including pollutants from incinerators, dioxin, pesticides, alcohol, lead and other such workplace hazards, has produced children with reduced male proportion, Variables reported to be associated with an increase in the SR are large family size, high ancestral longevity, paternal baldness, excessive coffee-drinking, intensive coital frequency and some male reproductive tract disorders.

Some striking examples can be found in the literature of unisexual pedigrees extending over several generations. Slater (1943) stated that aberrant SRs tend, to some extent, to run in families. The finding by Lindsey and Altham (1998) that the probability of couples being only capable of having children of one sex is very low contradicts Slater's statement. The variation in the SR that has been reliably identified in family studies has invariable been slight compared with what we have observed in families with X-linked recessive retinoschisis (cleavage of retinal layers). We noted a marked excess of males within such families, in contrast to normal SRs in families with the X-linked recessive disorders haemophilia and color blindness (Eriksson et al. 1967; Fellman et al. 2002). However, with the exception of the X-linked recessive retinoschisis, no unequivocal examples exist of genes in man that affect the SR, and X-linked retinoschisis is universally very rare. Summing up, influential factors, although they have an effect on family data, have not been identified in large national birth registers.

About the Author

For biography *see* the entry ►Lorenz Curve.

Cross References

►Demography
►Sign Test
►Significance Tests, History and Logic of
►Statistics, History of

References and Further Reading

Campbell RB (2001) John Graunt, John Arbuthnott, and the human sex ratio. Hum Biol 73:605–610
David FN (1962) Games, gods and gambling. Charles Griffin, London

Eriksson AW, Vainio-Mattila B, Krause U, Fellman J, Forsius H (1967) Secondary sex ratio in families with X-chromosomal disorders. Hereditas 57:373–381
Fellman J, Eriksson AW, Forsius H (2002) Sex ratio and proportion of affected sons in sibships with X-chromosomal recessive traits: maximum likelihood estimation in truncated multinomial distributions. Hum Hered 53:173–180
Hacking I (1975) The emergence of probability. Cambridge University Press, Cambridge
Hawley AH (1959) Population composition. In: Hauser PM, Duncan OD (ed) The study of population: an inventory and appraisal. University of Chicago, Chicago, pp 361–382
Lindsey JK, Altham PME (1998) Analysis of the human sex ratio by using overdispersion models. Appl stat 47: 149–157
Slater E (1943) A demographic study of a psychopathic population. Ann Eugenic 12:121–137
Visaria PM (1967) Sex ratio at birth in territories with a relatively complete registration. Eugenic Quart 14:132–142

Sign Test

Peter Sprent
Emeritus Professor
University of Dundee, Dundee, UK

The sign test is a nonparametric test for hypotheses about a population median given a sample of observations from that population, or for testing for equality of medians, or for a prespecified constant median difference, given paired sample (i.e., matched pairs) values from two populations. These tests are analogues of the one-sample and matched pairs t-test for means in a parametric test such as the t-test.

The sign test is one of the simplest and oldest nonparametric tests. The name reflects the fact that each more detailed observation is effectively replaced by one of the signs plus (+) or minus (−). This was basically the test used by Arbuthnot (1710) to refute claims that births are equally likely to be male or female. Records in London showed that for each of 81 consecutive years an excess of male over female births. Calling such a difference a plus, Arbuthnot argued that if births were equally likely to be of either gender, then the probability of such an outcome was, $(0.5)^{81}$, or effectively zero.

Given a sample of n observations from any population which may be discrete or continuous and not necessarily symmetric, the test is used to test a hypothesis $H_0 : M = M_0$ where M is the population median. If H_0 holds the number of values less than M_0 will have a binomial distribution with parameters n and $p = 0.5$. The symmetry of the ►binomial distribution when $p = 0.5$ means the number of sample values greater than M_0 (a plus) may be used as an alternative equivalent statistic in a one or two-tail test.

Although not a commonly arising case, the test is still valid if each observation in a sample is from a different population providing each such population has the same median. For example, the populations may differ in ►variance or in ►skewness.

Among tests for location the sign test thus requires fewer assumptions for validity than any other well established test. The main disadvantage of the test is that it often has lower efficiency and lower power than tests that require stronger assumptions when those assumptions are valid. However, when the stronger assumptions are not valid the sign test may have greater power and efficiency. If the sample is from a normal distribution with known variance the asymptotic relative efficiency (ARE) of the sign test relative to the normal theory test is $2/\pi$. However if the sample is from a double exponential distribution the ARE of the sign test is twice that attained using the t-test.

For continuous data except in special cases like samples from a double exponential distribution the sign test is usually less efficient than some parametric test or nonparametric test that makes more use of information about the data. For example, the t-test is preferable for samples from a normal, or near normal, distribution and the ►Wilcoxon-signed-rank test performs better if an assumption of symmetry can be made.

Even when a sign test is less efficient than some other test it may prove economically beneficial if exact data of the type needed for that other test is expensive to collect but it is easy to determine whether such data, if it were available, would indicate a value less than or greater than an hypothesised median value M_0. For example, if in a manufacturing process rods produced should have a median diameter of 40 mm it may be difficult to measure diameters precisely, but easy to determine whether the diameter of each rod is less than 40 mm by attempting to pass it through a circular aperture of diameter 40 mm. Those that pass though have a diameter less than 40 mm (recorded as a minus); those that fail to pass through have a greater diameter (recorded as a plus). If diameters can be assumed to be normally distributed and a sample size of 30 is required to give the required power with a normal theory test when exact measurements are available, the ARE for a sign test (which gives a fairly good idea of the efficiency for a sample of this size) suggests that if we only have information on whether each item has diameter less than (or greater

than) 40 mm, then a sample of size $30 \times \pi/2 \approx 47$ should have similar power. An assumption here is that efficiency for smaller samples is close to the ARE, a result verified in some empirical studies. Thus if the cost of obtaining each exact measurement were twice that of determining only whether or not a diameter exceeded 40 mm there would be a clear cost saving in measuring simply whether diameters were more or less than 40 mm for a sample of 47 compared to that for taking exact measurements for a sample of 30.

Sample values exactly equal to M_0 are usually ignored when using the test and the sample size used in assessing significance is reduced by 1 for each such value.

In the case of matched pair samples from distributions that may be assumed to differ if at all only in their medians, the test may be applied using the signs of the paired differences to test if the difference is consistent with a zero median and by a slight modification to test the hypothesis that the median difference has some specified value θ_0. The test is available in most standard statistical software packages or may be conducted using tables for the binomial distribution when $p = 0.5$ and the relevant n (sample size). For continuous data one may determine confidence intervals based on this test with the aid of such tables. Details are given in most textbooks covering basic nonparametric methods such as Gibbons and Chakraborti (2004) or Sprent and Smeeton (2007).

An interesting case that leads to a test equivalent to the sign test with heavy tying was proposed by McNemar (1947) and is usually referred to as McNemar's test. This test is relevant where observations are made to test if there are nonneutralizing changes in attitudes of individuals before or after exposure to a treatment or stimulus. For example, a group of 200 motorists may be asked whether or not they think the legal maximum permissible level of blood alcohol for drivers should be lowered. The numbers answering *yes* or *no* are recorded. The group are then shown a video illustrating the seriousness of accidents where drivers have exceeded the legal limit. Their answers to the same question about lowering the level are now recorded and tabulated as shown in this table:

		Before video	Lower limit
		Yes	*No*
After video Lower limit	*Yes*	160	24
	No	11	5

If we denote a change from *No* before the video to *Yes* after the video by a plus there are 24 plus, and a change from *Yes* before to *No* afterwards there are 11 minus. Thus, although the video seems to have influenced some changes of opinion in both directions more (24) who did not support a reduction before seeing the video appear to have been persuaded to support a reduction after seeing the video, whereas 11 have switched opinions in the opposite direction, opposing a ban after seeing the video although they supported one before seeing the video.

A sign test may be applied on the basis of 24 plus and 11 minus being observed in an effective sample of size 35. The diagonal values of 160 and 5 represent "ties" in the sense that they represent drivers who are not influenced by the video and so are ignored.

About the Author

Dr. Peter Sprent is Emeritus Professor of Statistics and a former Head of the Mathematics Department at the University of Dundee, Scotland. Previously he worked as a consultant statistician at a horticultural research station in England. He has 28 years teaching experience in Australia and the United Kingdom. He has written or coauthored 12 books on statistics and related topics, the best known of which is *Applied Nonparametric Statistical Methods* (with Nigel C. Smeeton, Chapman and Hall/CRC; 4th edition, 2007). He has been on the editorial boards, or been an associate editor, of several leading statistical journals and served on the Council and various committees of the Royal Statistical Society. He is a Fellow of the Royal Society of Edinburgh and is an Elected member of the International Statistical Institute.

Cross References

►Asymptotic Relative Efficiency in Testing
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Sex Ratio at Birth
►Wilcoxon-Signed-Rank Test

References and Further Reading

Arbuthnot J (1710) An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. Philos Trans R Soc 27:186–190

Gibbons JD, Chakraborti S (2004) Nonparametric statistical inference, 4th edn. Marcel Dekker, New York

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrica 12:153–157

Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods, 4th edn. Chapman & Hall/CRC Press, Boca Raton

Significance Testing: An Overview

Elena Kulinskaya[1], Stephan Morgenthaler[2], Robert G. Staudte[3]
[1]Professor, Aviva Chair in Statistics
University of East Anglia, Norwich, UK
[2]Professor, Chair of Applied Statistics
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3]Professor and Head of Department of Mathematics and Statistics
La Trobe University, Melbourne, VIC, Australia

Introduction

A *significance test* is a statistical procedure for testing a hypothesis based on experimental or observational data. Let, for example, $\overline{X}_1$ and $\overline{X}_2$ be the average scores obtained in two groups of randomly selected subjects and let μ_1 and μ_2 denote the corresponding population averages. The observed averages can be used to test the null hypothesis $\mu_1 = \mu_2$, which expresses the idea that both populations have equal average scores. A *significant result* occurs if $\overline{X}_1$ and $\overline{X}_2$ are very different from each other, because this contradicts or falsifies the null hypothesis. If the two group averages are similar to each other, the null hypothesis is not contradicted by the data. What exact values of the difference $\overline{X}_1 - \overline{X}_2$ of the group averages are judged as significant depends on various elements. The variation of the scores between the subjects, for example, must be taken into account. This variation creates uncertainty and is the reason why the testing of hypotheses is not a trivial matter. Because of the uncertainty in the outcome of the experiment, it is possible that a seemingly significant result is obtained, even though the null hypothesis is true. Conversely, the null hypothesis being false does not mean that the experiment will necessarily result in a significant result.

The significance of a test is usually measured in terms of a tail-error probability of the null distribution of a test statistic. In the above example, assume the groups are normally distributed with common known variance σ^2. The Z-test statistic is $Z = (\overline{X}_1 - \overline{X}_2)/\mathrm{SE}[\overline{X}_1 - \overline{X}_2]$, where $\mathrm{SE}[\overline{X}_1 - \overline{X}_2] = \sigma^2\{1/n_1 + 1/n_2\}$ is the standard error of the difference. Here n_1, n_2 are the respective sample sizes for the two groups. Under the null hypothesis, Z has the standard normal distribution with cumulative distribution $P(Z \le z) = \Phi(z)$. A large observed value $Z = Z_{obs}$ corresponds to a small tail area probability $P(Z \ge Z_{obs}) = \Phi(-Z_{obs})$.The smaller this probability the more the evidence against the null in the direction of the alternative $\mu_1 > \mu_2$. For a two-sided alternative $\mu_1 \ne \mu_2$, a test statistic is $|Z|$ and the evidence against the null is measured by the smallness of $P(|Z| \ge |Z_{obs}|) = 2\Phi(-|Z_{obs}|)$. These tail-error probabilities are examples of p-values for one- and two-sided tests.

To carry out a significance test then one needs, first, a *statistic* $S(X)$ (real function of the data X) that orders the outcomes X of a study so that larger values of $S(X)$ cast more doubt on the null hypothesis than smaller ones; and second, the probability distribution P_0 of $S(X)$ when the null hypothesis is true. One may be interested in simply assessing the evidence in the value obtained for the statistic S in an experiment, the Fisherian approach, or in making a decision to reject the null hypothesis in favor of an alternative hypothesis, the Neyman–Pearson approach.

Significance Tests for Assessing Evidence

By far the most prevalent concept for assessing evidence in S is the p-value, promoted by the influential scientist R.A. Fisher through his many articles and books, see the collection Fisher (1990).

The p-Value

Having observed data $X = x$, and hence $S(x) = S_{\mathrm{obs}}$, the *p-value* is defined by $p = P_0(S \ge S_{\mathrm{obs}})$. It is the probability of obtaining as much or more evidence against the null hypothesis as just observed with S_{obs}, assuming the null hypothesis is true. The p-value is decreasing with increasing S_{obs}, which means that smaller ▶p-values are indicative of a more significant result. Fisher (1973, pp. 80, 82, and 122), offered some rough guidelines for interpreting the strength of evidence measured by the p-value, based on his experience with agricultural experiments. He suggested that a p-value larger than 0.1 was not small enough to be significant, a p-value as small as 0.05 could seldom be disregarded, and a p-value less then 0.01 was clearly significant. Thus according to Fisher "significance testing" is the conducting of an experiment that will give the data a chance to provide evidence S_{obs} against the null hypothesis. Very small values of the p-value correspond to significant evidence, where "significant" is somewhat arbitrarily defined. It is a matter of history that Fisher's rough guideline "a value as small as 0.05 could seldom be disregarded" became a *de facto* necessity for publication of experimental results in many scientific fields. However, despite its usefulness for filtering out many inconsequential results, the p-value is often confused with fixed significance levels (see section ▶"Significance Tests for Making Decisions").

Finding the Null Distribution

It is not always easy to find the null distribution of a test statistic. It must be chosen carefully. For example, in the Z-test example of section ▶"Introduction", three assumptions were made, normality of the observations, equality of the group variances and knowledge of the common variance σ^2. If the first two assumptions hold, but the latter is relaxed to $\sigma^2 > 0$, then the distribution of the Z-test statistic depends on the unknown *nuisance parameter* σ^2, so one does not have a unique null distribution. An appropriate test statistic is the *two-sample pooled t-statistic*, which is just the Z-test statistic with σ replaced by s_{pooled}, where $s^2_{pooled} = \{(n_1-1)s_1^2 + (n_2-1)s_2^2\}/(n_1+n_1-2)$, and s_1^2, s_2^2 are the the respective sample variances. This t statistic has, under the null $\mu_1 = \mu_2$ a Student-t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, which allows for computation of p-values.

If the assumption of normality of the groups is retained, but their variances are not assumed equal, then one can estimate them separately using the respective sample variances. An approximating t distribution for the resulting standardized mean difference is known as the *Welch t-test* see Welch (1938). If the assumption of normality is relaxed to a continuous distribution then a comparison can be based on the sum S of the ranks of one sample within the ranking of the combined sets of observations. The null hypothesis is that each group has the same continuous distribution F and then S has a unique distribution. This test is known as the ▶*Wilcoxon–Mann–Whitney test*. It is an example of a *distribution-free test*, because F is unspecified.

Another way of computing a p-value when the null hypothesis distribution is not uniquely specified is to sample repeatedly from the empirical distribution of the data and for each sample compute the value of the test statistic; the proportion of values greater than the original S_{obs} is a *bootstrap estimate* of the p-value.

Significance Tests for Making Decisions

Neyman and Pearson (1928), Neyman (1933) formulated the significance testing problem as one of decision making. The data X are assumed to have distributions P_θ indexed by the parameter θ known to lie in one of two mutually exclusive sets Θ_0, Θ_1, and one must choose between them, using only X. The parameter sets Θ_0 and Θ_1 are called the *null* and *alternative hypotheses*, respectively. Each may be *simple*, containing only a single value, or *composite*. If $X \sim P_\theta$ for some $\theta \in \Theta_0$, and one chooses Θ_1 a *Type I error*, (or, error of the first kind), is committed. If $X \sim P_\theta$ for some $\theta \in \Theta_1$, and one chooses Θ_0 a *Type II error*, (or, error of the second kind), is committed. Because the consequences of Type I and Type II errors are often incommensurate, see Neyman (1950), the Neyman–Pearson framework places a bound α on Type I error probabilities, called the *level* of the test, and subject to this constraint seeks a decision rule that in some sense minimizes the Type II error probabilities, $\beta(\theta_1)$ for $\theta_1 \in \Theta_1$.

A *decision rule* equals 1 or 0 depending on whether Θ_1 or Θ_0 is chosen, after observing $X = x$. It is by definition the indicator function $I_C(x)$ of the *critical region* C, which is the set of values of X for which Θ_1 is chosen. This region is critical in the sense that if $X \in C$, one rejects the null hypothesis and risks making a Type I error. The *size* of a critical region is $\sup_{\theta\in\Theta_0} P_\theta(X \in C)$. One seeks a critical region (test) for which the size is no greater than the level α and which has large power of detecting alternatives. The size may be set equal to the desired level α by choice of C when the distributions P_θ are continuous, but in the case of discrete P_θ, the size will often be less than α, unless some form of ▶randomization is employed, see Lehmann (1986).

Power Function of a Test and Optimal Test Statistics

The *power* of a test for detecting an alternative $\theta_1 \in \Theta_1$ is defined by $\Pi(\theta_1) = P_{\theta_1}(X \in C) = 1 - \beta(\theta_1)$. It is the probability of making the right decision (rejecting Θ_0) when $\theta_1 \in \Theta_1$; and as indicated, it is also 1 minus the probability of making a Type II error for this θ_1. The *power function* is defined by $\Pi(\theta_1)$, for each $\theta_1 \in \Theta_1$. Let f_θ be the density of P_θ with respect to a dominating measure for the distributions of X. Neyman and Pearson showed that for a simple hypothesis θ_0 and simple alternative θ_1, there exists a most powerful level-α test which rejects the null when the *likelihood ratio* $\lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x)$ is large. That is, the critical region is of the form $C = \{x : \lambda(x) \geq c\}$, where the *critical value* c defining the boundary of the critical region is chosen so $P_{\theta_0}\{\lambda(X) \geq c\} = \alpha$. For composite hypotheses, the *likelihood test statistic* defined by $\lambda(x) = \sup_\Theta f_\theta(x)/\sup_{\Theta_0} f_\theta(x)$ is the basis for many tests, because its large sample distribution is known. A *uniformly most powerful* level-α test maximizes the power for each value of the alternative amongst all level-α tests. Uniformly most powerful tests for composite alternatives are desirable, but such tests do not usually exist. See Lehmann (1986) for a comprehensive development of the theory of hypothesis testing.

Inversion of a Family of Tests to Obtain Confidence Regions

A *confidence region* of level $1-\alpha$ for a parameter θ is a random set $R(X)$ for which $P_\theta\{\theta \in R(X)\} \geq 1-\alpha$ for all $\theta \in \Theta$. When Θ is a subset of the real line, the region is

usually in the form of a random *confidence interval* $[L, U]$, where $L = L(X), U = U(X)$. The inversion procedure, due to Neyman (1935), supposes that for each $\theta_0 \in \Theta$ there is a level–α test with critical region $C_\alpha(\theta_0)$ for testing the simple null hypothesis $\Theta_0 = \{\theta_0\}$ against its complement $\Theta_0^c = \{\theta \in \Theta : \theta \neq \theta_0\}$. This family of tests can be converted into a level $1-\alpha$ confidence region for θ, given by $R(X) = \{\theta_0 \in \Theta : X \notin C_\alpha(\theta_0)\}$. Thus a parameter θ_0 belongs to the confidence region if and only if it is not rejected by the level α test of $\theta = \theta_0$ against $\theta \neq \theta_0$.

On p-Values and Fixed Significance Levels

The purpose of choosing a fixed level α as a prior upper bound on the probability of Type I errors is to avoid making decisions that are influenced by the observed data x. The p-value, on the other hand, requires knowledge of x for its computation, and subsequent interpretation as evidence against the null hypothesis. Thus when used for the separate purposes for which they were designed, there is no confusion. However, having observed $S(x) = S_{\text{obs}}$, the p-value is equal to the level α for which $S_{\text{obs}} = c_\alpha$; that is, the smallest fixed level for which the test rejects the null. For this reason, it is sometimes called the *observed significance level*. One rejects the null at level α if and only if the p-value $\leq \alpha$. It is widespread practice to use the Neyman–Pearson framework to obtain a powerful test of level $\alpha = 0.05$, and then to report the p-value. Thus there has evolved in practice a combination of concepts that can prove confusing to the uninitiated, see Berger (2003) Hubbard and Bayarri (2003) and Lehmann (1993).

Bayesian Hypothesis Testing

The Bayesian framework for significance testing assumes a *prior* probability measure $\pi(\theta)$ over the parameter space $\Theta = \Theta_0 \cup \Theta_1$. This yields prior probabilities $\pi_0 = \pi(\Theta_0)$, $1 - \pi_0$ on the null and alternative hypotheses Θ_0, Θ_1, respectively, and the *prior odds* $\pi_0/(1-\pi_0)$ in favor of the null. It is further assumed that for each θ, the data X has a conditional distribution $f(x|\theta)$ for X, given θ. The *posterior probability of the null* is then $P(\Theta_0|x) = \int_{\Theta_0} f(x|\theta)d\pi(\theta)/f_X(x)$, where $f_X(x) = \int_\Theta f(x|\theta)d\pi(\theta)$. One can, if a decision is required, reject the null in favor of the alternative when $P(\Theta_0|x)$ is less than some preassigned level, as in NP testing; or, one can simply choose to interpret it as a measure of support for Θ_0.

Bayes Factor

It turns out that the posterior odds for Θ_0 are related to its prior odds by $P(\Theta_0|x)/(1-P(\Theta_0|x)) = B_{01}(x)\,\pi_0/(1-\pi_0)$, The *Bayes factor* $B_{01}(x) = f_{\Theta_0}(x)/f_{\Theta_1}(x)$, where $f_{\Theta_i}(x) = \int_{\Theta_i} f(x|\theta)d\pi(\theta)/\pi(\Theta_i)$, $i = 0, 1$. The Bayes factor measures the change in odds for the null hypothesis Θ_0 after observation of $X = x$. It is also often interpreted as a measure of support for Θ_0, but this interpretation is not without controversy; for further discussion see Kass (1995) and Lavine and Schervish (1999).

Significance Tests for Special Purposes

When one wants to adopt the model $X \sim \{P_\theta : \theta \in \Theta\}$ for inference, be it testing or estimation, a *goodness-of-fit test* rejects the entire model if a suitable test statistic $S(X)$ has small p-value. Thus if the data do not cast doubt on the model, the statistician happily proceeds to adopt it. This procedure is informal in that many other models might equally pass such a test, but are not considered. Tests for submodel selection in regression have the same feature; one "backs into" acceptance of a submodel because an F-test does not reject it. All such significance tests are simply informal guides to ►model selection, with little regard for Type II errors, or the subsequent effects on inference with the chosen model. *Equivalence tests*, on the other hand, place great emphasis on formal testing, and do provide evidence for a null hypothesis of no effect. They do this by interchanging the traditional roles of null and alternative hypotheses. For example, if θ represents the mean difference in effects of two drugs, one might be interested in evidence for $|\theta| \leq \theta_0$, where θ_0 defines a region of "equivalence." This is taken as the alternative hypothesis, to a null $|\theta| \geq \theta_1$, where $\theta_1 > \theta_0$ is large, say. One also simultaneously tests the null $\theta \leq -\theta_1$ against the alternative of equivalence. If one rejects both these null hypotheses in favor of the alternative, evidence for equivalence is found. See Wellek (2003) for a complete development.

Final Remarks and Additional Literature

Statistical significance of a test, meaning a null hypothesis is rejected at a pre-specified level such as 0.05, is not evidence for a result which has practical or scientific significance. This has led many practitioners to move away from the simple reporting of p-values to reporting of confidence intervals for effects; see Krantz (1999) for example. A measure of *evidence for* a positive effect that leads to confidence intervals for effects is developed in Kulinskaya et al. (2008). Fuzzy hypothesis tests and confidence intervals are introduced in Dollinger et al. (1996) and explored in Geyer and Meeden (2006).

About the Author

For biography *see* the entry ►Meta-Analysis.

Cross References

►Accelerated Lifetime Testing
►Anderson-Darling Tests of Goodness-of-Fit
►Bartlett's Test

►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests
►Dickey-Fuller Tests
►Durbin–Watson Test
►Effect Size
►Equivalence Testing
►Equivalence Testing
►Fisher Exact Test
►Frequentist Hypothesis Testing: A Defense
►Full Bayesian Significant Test (FBST)
►Jarque-Bera Test
►Kolmogorov-Smirnov Test
►Mood Test
►Most Powerful Test
►Multiple Comparisons Testing from a Bayesian Perspective
►Neyman-Pearson Lemma
►Nonparametric Rank Tests
►Null-Hypothesis Significance Testing: Misconceptions
►Omnibus Test for Departures from Normality
►Parametric Versus Nonparametric Tests
►Permutation Tests
►Presentation of Statistical Testimony
►Psychological Testing Theory
►Psychology, Statistics in
►P-Values
►Randomization Tests
►Rank Transformations
►Scales of Measurement and Choice of Statistical Methods
►Sequential Probability Ratio Test
►Sign Test
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Simes' Test in Multiple Testing
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistical Significance
►Step-Stress Accelerated Life Tests
►Student's *t*-Tests
►Testing Exponentiality of Distribution
►Testing Variance Components in Mixed Linear Models
►Tests for Discriminating Separate or Non-Nested Models
►Tests for Homogeneity of Variance
►Tests of Fit Based on The Empirical Distribution Function
►Tests of Independence
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Berger JO (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? Stat Sci 18(1):1–32, With discussion

Dollinger MB, Kulinskaya E, Staudte RG (1996) Fuzzy hypothesis tests and confidence intervals. In: Dowe DL, Korb KB, Oliver JJ (eds) Information, statistics and induction in science. World Scientific, Singapore, pp 119–128

Fisher RA (1990) Statistical methods, experimental design and scientific inference. Oxford University Press, Oxford. Reprints of Fisher's main books, first printed in 1925, 1935 and 1956, respectively. The 14th edition of the first book was printed in 1973

Geyer C, Meeden G (2006) Fuzzy confidence intervals and p-values (with discussion). Stat Sci 20:258–387

Hubbard R, Bayarri MJ (2003) Confusion over measures of evidence (*p*'s) versus errors (*a*'s) in classical statistical testing. Am Stat 57(3):171–182, with discussion

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795

Krantz D (1999) The null hypothesis testing controversy in psychology. J Am Stat Assoc 94(448):1372–1381

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis: a guide to calibrating and combining statistical evidence. Wiley, Chichester, www.wiley.com/go/meta_analysis

Lavine M, Schervish M (1999) Bayes factors: what they are and what they are not. Am Stat 53:119–122

Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York

Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? J Am Stat Assoc 88:1242–1249

Neyman J (1935) On the problem of confidence intervals. Ann Math Stat 6:111–116

Neyman J (1950) First course in probability and statistics. Henry Holt, New York

Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20A:175–240 and 263–294

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A 231:289–337

Welch BL (1938) The significance of the difference between two means when the variances are unequal. Biometrika 29:350–361

Wellek S (2003) Testing statistical hypotheses of equivalence. Chapman & Hall/CRC Press, New York

Significance Tests, History and Logic of

Henrik Olsson, Mirta Galesic
Max Planck Institute for Human Development, Berlin, Germany

By most accounts, the first significance test was published in 1710 by the Scottish mathematician, physician, and author John Arbuthnot. He believed that, because males were subject to more external accidents than females, they

enjoyed an advantage of a higher birthrate. Arbuthnot calculated the expectation, or the probability, of the data from 82 years of birth records in London given a chance hypothesis of equal birthrates for both sexes. Because this expectation was very low he concluded "that it is Art, not Chance, that governs" (p. 189), and that this result constituted a proof of existence of an active god. Although he never used the terms *significance* or *significant* – these terms were first used at the end of the nineteenth century by Francis Ysidro Edgeworth (1885) and John Venn (1888) – his argument is strikingly similar to the logic underlying modern null hypothesis testing as implemented in Ronald Fisher's significance testing approach (e.g., 1925, 1935).

The beginning of the twentieth century saw the development of the first modern significance tests: Karl Pearson's (1900) *chi-squared test* and William Sealy Gosset's (or Student's 1908) *t-test* (although the term t-test appeared only later, in 1932 in the fourth edition of Fisher's *Statistical Methods for Research Workers*). Both are examples of tail-area significance tests, in which a hypothesis is rejected if the tail of the null distribution beyond the observed value is less than a prescribed small number. Gosset's article was also the beginning of the field of small sample statistics, where the earlier asymptotics ($n \to \infty$) were replaced by exact probabilities.

The use of significance tests really took root among applied researchers after the publication of Fisher's influential books, *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). Fisher rejected the (older) methods of inverse probability (of hypothesis given data) and proposed a method of *inductive inference*, a formal way of getting from data to hypothesis. His approach can be summarized as follows: The researcher sets up a null hypothesis that a sample statistic comes from a hypothetical infinite population with a known sampling distribution. The null hypothesis is rejected or, as Fisher called it, "disproved," if the sample statistic deviates from the mean of the sampling distribution by more than a specified criterion. This criterion – or *level of significance* – is typically set to 5%, although Fisher later recommended reporting the exact probability. In this approach, no claims about the validity of alternative hypotheses are possible. It is nevertheless tempting to view the complement of the null hypothesis as an alternative hypothesis and argue, as Arbouthnot did, that the rejection of the null hypothesis gives credit to an unspecified alternative hypothesis. Fisher's approach is also associated with an epistemic interpretation of significance: A Fisherian *p-value* is thought to measure the strength of evidence against the null hypothesis and to allow the researcher to learn about the truth or falsehood of a specific hypothesis from a single experiment.

The major rival to Fisher's approach was Jerzy Neyman and Egon Pearson's (1928a, 1928b, 1933) approach to hypothesis testing, originally viewed as an extension and improvement of Fisher's ideas. Neyman and Pearson rejected the idea of *inductive inference* and replaced it with the concept of *inductive behavior*. They sought to establish rules for making decisions between different hypotheses regardless of researcher's beliefs about the truth of those hypotheses. They argued for specifying both a null hypothesis and an alternative hypothesis, which allows for the calculation of two error probabilities, *Type I* error and *Type 2* error, based on considerations regarding decision criteria, sample size and effect size. *Type I* error occurs when the null hypothesis is rejected although it is true. The probability of a *Type I* error is called α. *Type II* error occurs when the alternative hypothesis is rejected although it is true. The probability of a *Type II* error is called β and 1-β is called the *power* of the test or the long run frequency of accepting the alternative hypothesis if it is true. The decision to accept or reject hypotheses in the Neyman–Pearson approach depends on the costs associated with *Type I* and *Type II* errors. The cost considerations lie outside of the formal statistical theory and must be based on context-dependent pragmatic personal judgment. The goal, then, for a researcher is to design an experiment that controls for α and β and use a test that minimizes β given a bound on α. In contrast to the data dependent ►*p-values* in Fisher's approach, α is specified before collecting the data. Despite the different conceptual foundations of Fisher's approach and Neyman-Pearson's approach, classical statistical inference, as commonly presented, is essentially an incoherent hybrid of the two approaches (Hubbard and Bayarri 2003; Gigerenzer 1993), although there exist attempts to reconcile them (Lehmann 1993). There is a considerable literature discussing the pros and cons of classical statistical inference, especially null hypothesis significance testing in the Fisherian tradition (e.g., Berger and Wolpert 1988; Royall 1997; Morrison and Henkel 1970). The major alternative to classical significance and hypothesis testing is Bayesian hypothesis testing (Jeffreys 1961; Kass and Raftery 1995).

Cross References

►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Significance Testing: An Overview
►Statistical Significance

References and Further Reading

Arbuthnot J (1710) An argument for divine providence, taken from the constant regularity observed in the births of both sexes. Philos Tr R Soc 27:186–190

Berger JO, Wolpert RL (1988) The likelihood principle, 2nd edn. Institute of Mathematical Statistics, Hayward, CA
Edgeworth FY (1885) Methods of statistics. Jubilee Volume of the Statistical Society. E. Stanford, London, pp 181–217
Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Gigerenzer G (1993) The Superego, the Ego, and the Id in statistical reasoning. In: Keren G, Lewis C (eds) A handbook for data analysis in the behavioral sciences: methodological issues. Erlbaum, Hillsdale, NJ, pp 311–339
Hubbard R, Bayarri M-J (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. Am Stat 57:171–182
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:377–395
Lehmann EL (1993) The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or Two? J Am Stat Assoc 88:1242–1249
Morrison DE, Henkel RE (eds) (1970) The significance test controversy: a reader. Aldine, Chicago
Neyman J, Pearson ES (1928a) On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika 20A:175–240
Neyman J, Pearson ES (1928b) On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika 20A:263–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Tr R Soc S-A 231:289–337
Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London
Venn J (1888) The logic of chance: an essay on the foundations and province of the theory of probability, 3rd edn. Macmillan, London

Significance Tests: A Critique

Bruno Lecoutre
ERIS, Laboratoire de Mathématiques Raphaël Salem, C.N.R.S. and Université de Rouen, Mont Saint Aignan, France

► It is very bad practice to summarise an important investigation solely by a value of P.

(Cox 1982, p 327)

In spite of some recent changes, significance tests are again conventionally used in most scientific experimental publications. According to this publication practice, each experimental result is dichotomized: significant vs. non-significant. But scientists cannot in this way find appropriate answers to their precise questions, especially in terms of effect size evaluation. It is not surprising that, from the outset (e.g., Boring 1919), significance tests have been subject to intense criticism. Their use has been explicitly denounced by the most eminent and most experienced scientists, both on theoretical and methodological grounds, not to mention the sharp controversies on the very foundations of statistical inference that opposed Fisher to Neyman and Pearson, and continue to oppose frequentists to Bayesians. In the 1960s there was more and more criticism, especially in the behavioral and social sciences, denouncing the shortcomings of significance tests: *the significance test controversy* (Morrison and Henkel 1970).

Significance Test Are Not a Good Scientific Practice

► It is foolish to ask 'Are the effects of A and B different?' They are always different - for some decimal place.

(Tukey 1991, p 100)

In most applications, no one can seriously believe that the different treatments have produced no effect: the point null hypothesis is only a *straw man* and a significant result is an evidence against an hypothesis known to be false before the data are collected, but not an evidence in favor of the alternative hypothesis. It is certainly not a good scientific practice, where one is expected to present arguments that support the hypothesis in which one is really interested. The real problem is to obtain estimates of the sizes of the differences.

The innumerable misuses of significance tests

► The psychological literature is filled with misinterpretations of the nature of the tests of significance.

(Bakan 1967, in Morrison and Henkel 1970, p 239)

Due to their inadequacy in experimental data analysis, the practice of significance tests entails considerable distortions in the designing and monitoring of experiments. It leads to innumerable misuses in the selection and interpretation of results. The consequence is the existence of publication biases denounced by many authors: while non-significant results are – theoretically – only statements of ignorance, only the significant results would really deserve publication.

The evidence of distortions is the use of the symbols *NS*, *, **, and *** in scientific journals, as if the degree of significance was correlated with the meaningfulness of research results. Many researchers and journal editors appear to be "star worshippers": see Guttman (1983), who openly attacked the fact that some scientific journals, and *Science* in particular, consider the significance test as a criterion of scientificness. A consequence of this overeliance

on significant effects is that most users of statistics overestimate the probability of replicating a significant result (Lecoutre et al. 2010).

The Considerable Difficulties Due to the Frequentist Approach

> What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.
>
> (Jeffreys 1998/1939, Sect. 7.2)

Since the *p*-value is the proportion of samples that are "at least as extreme" as the observed data (under the null hypothesis), the rejection of the null hypothesis is based on the probability of the samples that *have not been observed*, what Jeffreys ironically expressed in the above terms. This mysterious and unrealistic use of the sampling distribution for justifying null hypothesis significance tests is for the least highly counterintuitive. This is revealed by questions frequently asked by students and statistical users: "why one considers the probability of samples outcomes that are more extreme than the one observed?"

Actually, due to their frequentist conception, significance tests involve considerable difficulties in practice. In particular, many statistical users misinterpret the *p*-values as inverse (Bayesian) probabilities: $1 - p$ is "the probability that the alternative hypothesis is true." All the attempts to rectify this misinterpretation have been a loosing battle.

Significance Tests Users' Dissatisfaction

> Neither Fisher's null hypothesis testing nor Neyman-Pearson decision theory can answer most scientific problems.
>
> (Gigerenzer 2004, p 599)

Several empirical studies emphasized the widespread existence of common misinterpretations of significance tests among students and scientists (for a review, see Lecoutre et al. 2001). Many methodology instructors who teach statistics, including professors who work in the area of statistics, appear to share their students' misinterpretations. Moreover, even professional applied statisticians are not immune to misinterpretations of significance tests, especially if the test is nonsignificant. It is hard to interpret these finding as an individual's lack of mastery: they reveal that significance test do not address the questions that are of primary interest for the scientific research.

In particular, the dichotomous significant/non significant outcome of significance tests strongly suggests binary research decisions: "reject/accept the null hypothesis." "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested" (Rozeboom, in Morrison and Henkel 1970, p. 221). The "reject/accept" attitude is obviously a poor and unfortunate decision practice.

- A statistically significant test provides no information about the departure from the null hypothesis. When the sample is large a descriptively small departure may be significant.
- A nonsignificant test is not evidence favoring the null hypothesis. In particular, a descriptively large departure from the null hypothesis may be nonsignificant if the experiment is insufficiently sensitive.

In fact, in order to interpret their data in a reasonable way, users must resort to a more or less naive mixture of significance tests outcomes and other information. But this is not an easy task! This leads users to make *adaptative distortions*, designed to make an ill-suited tool fit their true needs. Actually, many users explicitly appear to have a real consciousness of the stranglehold of significance tests: in many cases they use them only because they know no other alternative.

Concluding Remarks

> Inevitably, students (and essentially everyone else) give an inverse or Bayesian twist to frequentist measures such as confidence intervals and P values.
>
> (Berry 1997, p 241)

It is not acceptable that statistical inference methods users will continue using nonappropriate procedures because they know no other alternative. Nowadays, proposals for changes in reporting experimental results are constantly made. In all fields these changes, especially in presenting and interpreting effect sizes, are more and more enforced within editorial policies. Unfortunately, academic debates continue and give a discouraging feeling of *déjà-vu*. Rather than stimulating the interest of experimental scientists, this endless controversy is without doubt detrimental to the impact of new proposals, if not to the image of statistical inference.

The majority official trend is to advocate the use of confidence intervals, in addition to or instead of significance tests. However, reporting confidence intervals appears to have very little impact on the way the authors interpret their data. Most of them continue to focus on the statistical significance of the results. They only wonder whether the

interval includes the null hypothesis value, rather than on the full implications of confidence intervals: the steamroller of significance tests cannot be escaped.

Furthermore, for many reasons due to their frequentist conception, confidence intervals can hardly be seen as the ultimate method. We then naturally have to ask ourselves whether the "Bayesian choice" will not, sooner or later, be unavoidable. It can be argued that an *objective Bayes theory* is by no means a speculative viewpoint but on the contrary is perfectly feasible (Rouanet et al. 2000; Lecoutre et al. 2001; Lecoutre 2008).

Cross References

►Frequentist Hypothesis Testing: A Defense
►Null-Hypothesis Significance Testing: Misconceptions
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Statistical Evidence
►Statistical Inference: An Overview
►Statistical Significance

References and Further Reading

Berry DA (1997) Teaching elementary Bayesian statistics with real applications in science. Am Stat 51:241–246

Boring EG (1919) Mathematical versus scientific significance. Psychol Bull 16:335–338

Cox DR (1982) Statistical significance tests. Br J Clin Pharmacol 16:325–331

Gigerenzer G (2004) Mindless statistics. J Socio-Economics 33:587–606

Guttman L (1983) What is not what in statistics? Statistician 26:81–107

Jeffreys H (1961) Theory of probability, 3rd edn (1st edn: 1939). Clarendon, Oxford

Lecoutre B (2008) Bayesian methods for experimental data analysis. In: Rao CR, Miller J, Rao DC (eds) Handbook of statistics: epidemiology and medical statistics, vol 27. Elsevier, Amsterdam, pp 775–812

Lecoutre B, Lecoutre M-P, Poitevineau J (2001) Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? Int Stat Rev 69:399–418

Lecoutre B, Lecoutre M-P, Poitevineau J (2010) Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. Psychol Methods 15:158–171

Morrison DE, Henkel RE (eds) (1970) The Significance test controversy - a reader. Butterworths, London

Rouanet H, Bernard J-M, Bert M-C, Lecoutre B, Lecoutre M-P, Le Roux B (2000) New ways in statistical methodology: from significance tests to Bayesian inference, 2nd edn. Peter Lang, Bern, CH

Tukey JW (1991) The philosophy of multiple comparisons. Stat Sci 6:100–116

Simes' Test in Multiple Testing

Sanat K. Sarkar
Professor
Temple University, Philadelphia, PA, USA

Over the past decade there has been a revival of interest in the field of multiple testing due to its increased relevance in modern scientific investigations, such as DNA microarray and functional magnetic resonance imaging (fMRI) studies. Simes' (1986) test plays an important role in the developments of a number of multiple testing methods. Given a family of null hypotheses $H_1, \ldots, H_n$ and the corresponding p-values $P_1, \ldots, P_n$, it is a global test of the intersection null hypothesis $H_0 : \bigcap_{i=1}^{n} H_i$ based on these p-values. It rejects H_0 at a significance level α if $P_{(i)} \le i\alpha/n$ for at least one $i = 1, \ldots, n$, where $P_{(1)} \le \cdots \le P_{(n)}$ are the ordered p-values.

Simes' test is more powerful than the Bonferroni test. However, to control the Type I error rate at the desired level, it requires certain assumptions about dependence structure of the p-values under H_0, unlike the Bonferroni test. For instance, if p-values are either independent or positively dependent in the following sense:

$$E_{H_0}\{\phi(P_1, \ldots, P_n) | P_i = u\} \text{ is non-decreasing in } u \quad (1)$$

for each $i = 1, \ldots, n$, and any coordinatewise non-decreasing function $\varphi(P_1, \ldots, P_n)$ of $P_1, \ldots, P_n$, then Simes' test controls the Type I error rate at α; that is, the following inequality holds:

$$\Pr_{H_0}\{\text{Rejecting } H_0\} = \Pr_{H_0}\left\{\bigcup_{i=1}^{n}\left(P_{(i)} \le i\alpha/n_0\right)\right\} \le \alpha.$$

Such positive dependence is exhibited by p-values in some commonly encountered multiple testing situations. For instance, p-values generated from (I) dependent standard normal variates with non-negative correlations, (II) absolute values of dependent standard normal variates with a correlation matrix R such that the off-diagonal entries of $-DR^{-1}D$ are non-negative for some diagonal matrix D with diagonal entries ± 1, (III) multivariate t with the associated normal variates having non-negative correlations (under a minor restriction on the range of values of u), and (IV) absolute values of multivariate t with the associated normal variates having a correlation matrix as in (II), satisfy (1) (Sarkar 1998, 2008a; Sarkar and Chang 1997).

For simultaneous testing of $H_1, \ldots, H_n$, the family-wise error rate (FWER), which is the probability of falsely

rejecting at least one null hypothesis, is often used as a measure of overall Type I error. Methods strongly controlling the FWER, that is, with this probability not exceeding a pre-specified value α under any configuration of true and false null hypotheses, have been proposed. Hochberg (1988) suggested such a method. It rejects H_i if $P_i \le P_{(\hat{i})}$, where

$$\hat{i} = \max\left\{i : P_{(i)} \le \alpha/(n-i+1)\right\}$$

provided the maximum exists, otherwise accepts all null hypotheses. This is a stepup method with the critical values $\alpha_i = \alpha/(n-i+1)$, $i = 1,\ldots,n$. For any stepup method with critical values $\alpha_1 \le \cdots \le \alpha_n$, the FWER is 0 if n_0, the number of true null hypotheses, is 0, otherwise it satisfies the following inequality:

$$FWER \le \Pr\left\{\bigcup_{i=1}^{n_0}\left(\hat{P}_{(i)} \le \alpha_{n-n_0+i}\right)\right\},$$

where $\hat{P}_{(1)} \le \cdots \le \hat{P}_{(n_0)}$ are the ordered versions of the p-values corresponding to the n_0 true null hypotheses (Romano and Shaikh 2006). For the Hochberg method, since

$$\alpha_{n-n_0+i} = \alpha/(n_0-i+1) \le i\alpha/n_0 \text{ for } i = 1,\ldots,n_0,$$

its FWER is bounded above by the Type I error rate of the level α Simes' test for the intersection of n_0 null hypotheses based on their p-values. In other words, the Hochberg method controls its FWER in situations where Simes' global test controls its Type I error rate.

The closed testing method of Marcus et al. (1976) is often used to construct multiple testing method with a strong control of the FWER. It operates as follows. Given a finite family of null hypotheses $\{H_i, i = 1,\ldots,n\}$, form the closure of this family by considering all non-empty intersections $H_J = \bigcap_{i\in J} H_i$ for $J \subseteq \{1,\ldots,n\}$. Suppose a level-$\alpha$ global test is available for each H_J. Then, a closed testing method rejects H_J if and only if every H_K with $K \supseteq J$ is rejected by its level-α test. Hommel (1988) used Simes' global test in the closed testing method to construct an improvement of the Hochberg method. It finds

$$\hat{j} = \max\left\{i : P_{(n-i+k)} \ge k\alpha/i \text{ for all } k = 1,\ldots,i\right\},$$

and rejects H_i if $P_i \le \alpha/\hat{j}$, provided the maximum exists, otherwise rejects all null hypotheses.

Benjamini and Hochberg (1995) introduced the ►false discovery rate (FDR), which is a less conservative notion of error rate than the FWER. With R and V denoting the total number rejections and the total number of false rejections, respectively, of null hypotheses, it is defined as follows:

$$FDR = E(V/\max\{R,1\}).$$

The FDR is said to be strongly controlled at α by a multiple testing method if the above expectation does not exceed α, irrespective of the number of true null hypotheses. As noted in Hommel (1988), while making decisions on the individual null hypotheses using the stepup method based on the critical values in the Simes' test, which are $\alpha_i = i\alpha/n$, $i = 1,\ldots,n$, the FWER is not strongly controlled. However, the false discovery rate (FDR) is strongly controlled, again if the p-values are independent or positively dependent in the sense of (1), but with the P_i now representing the p-value corresponding to a null hypothesis (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Sarkar 2002). A proof of this result can be seen in Sarkar (2008b), who gave the following expression for the FDR of a stepup method with critical values $\alpha_1 \le \cdots \le \alpha_n$:

$$FDR = \sum_{i\in J_0} E\left[\frac{I\left(P_i \le \alpha_{R_{n-1}^{(-i)}+1}\right)}{R_{n-1}^{(-i)}+1}\right],$$

where I is the indicator function, J_0 is the set of indices corresponding to the true null hypotheses, $R_{n-1}^{(-i)}$ is the number of rejections in the stepup method based on the $n-1$ p-values other than P_i and the critical values $\alpha_2 \le \cdots \le \alpha_n$. Examples of p-values satisfying this positive dependence condition are those that are generated from test statistics in (I) and (III).

About the Author

Dr. Sanat K. Sarkar is Professor and Senior Research Fellow, Department of Statistics, Temple University. He is a Fellow of the Institute of Mathematical Statistics, a Fellow of the American Statistical Association, and an Elected member of the International Statistical Institute. He is Associate Editor of *Annals of Statistics*, *The American Statistician* and *Sankhya, B*. He has made significant contributions to the development of modern multiple testing techniques. He has received a number of awards and honors from his university for his research contributions.

Cross References

►False Discovery Rate
►Multiple Comparison
►Multiple Comparisons Testing from a Bayesian Perspective

References and Further Reading

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802
Hommel G (1988) A stagewise rejective multiple test procedure based on a modifed Bonferroni test. Biometrika 75:783–786
Marcus R, Peritz E, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660
Romano JP, Shaikh AM (2006) Stepup procedures for control of generalizations of the familywise error rate. Ann Stat 34:1850–1873
Sarkar SK (1998) Some probability inequalities for ordered *MTP*2 random variables: a proof of the Simes conjecture. Ann Stat 26:494–504
Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30:239–257
Sarkar SK (2008a) On the Simes inequality and its generalization. IMS collections beyond parametrics. In: Interdisciplinary research: Festschrift in honor of professor Pranab K. Sen 1: 231–242
Sarkar SK (2008b) On methods controlling the false discovery rate. Sankhya A 70(Part 2):135–168
Sarkar SK, Chang CK (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. J Am Stat Assoc 92:1601–1608
Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Simple Linear Regression

Sung H. Park
Professor
Seoul National University, Seoul, Korea

Regression analysis is a collection of statistical modeling techniques that usually describes the behavior of a random variable of interest by using one or more quantitative variables. The variable of interest may be the crop yield, the price of oil in the world market, the tensile strength of metal wire, and so on. This variable of interest is called the *dependent variable*, or *response* variable and denoted with Y. Other variables that are thought to provide information on the dependent variable are incorporated into the model as *independent* variables. These variables are also called the *predictor*, or *regressor*, or *explanatory* variables, and are denoted by Xs. If the height of a son is affected by the height of his father, then the height of the father is X and the height of the son becomes Y.

The Xs are assumed to be known constants. In addition to the Xs, all models involve unknown constants, called *parameters*, which control the behavior of the model. In practical situations, the statistical models usually fall into the class of models that are *linear in the parameters*. That is, the parameters enter the model as simple coefficients on the independent variables. Such models are referred to as ►*linear regression* models. If there is only one independent variable X for the dependent variable of interest Y, and the functional relationship between Y and X is a straight line, this model is called the *simple linear regression* model.

In a nonstatistical context, the word *regression* means "to return to an earlier place or state." The term "*regression*" was first used by Francis Galton (1822–1911), who observed that children's heights tended to "revert" to the average height of the population rather than diverting from it. Galton applied "a regression line" to explain that the future generations of offspring who are taller than average are not progressively taller than their respective parents, and parents who are shorter than average do not beget successively smaller children. But the term is now applied to any linear or nonlinear functional relationships in general.

In the simple linear model, the true mean of Y changes at a constant rate as the value of X increases or decreases. Thus, the functional relationship between the true mean of Y, denoted by $E(Y)$, and X is the equation of a straight line

$$E(Y) = \beta_0 + \beta_1 X.$$

Here, β_0 is the intercept, the value of $E(Y)$ when $X = 0$, and β_1 is the slope of the line, the rate of change in $E(Y)$ per unit change in X. Suppose we have n observations on Y, say, $Y_1, Y_2, Y_3, \ldots, Y_n$ at $X_1, X_2, X_3, \ldots, X_n$, respectively. The i^{th} observation on the dependent variable Y_i at X_i is assumed to be a random observation with the random error ε_i to give the statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad (1)$$

The random errors ε_i have zero mean and assumed to have common variance σ^2 and to be pairwise independent. The random error assumptions are frequently stated as

$$\varepsilon_i \sim NID(0, \sigma^2)$$

where *NID* stands for normally and independently distributed. The quantities in parentheses denote the mean and the variance, respectively, of the normal distribution.

Once β_0 and β_1 in Eq. 1 have been estimated from a given set of data on X and Y, the following prediction equation results:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \text{ or } \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \quad (2)$$

The "hats" (as they are called) above β_0 and β_1 signify that those parameters are being estimated, but the hat above Y means that the dependent variable is being predicted. Point estimates of β_0 and β_1 are needed to obtain the fitted line given in Eq. 2. One way is to minimize the sum of the absolute values of the vertical distances with each distance measured from each point to the fitted line (see,

e.g., Birkes and Dodge 1993). These vertical distances are called ►residuals. The standard approach, however, is to minimize the sum of the squares of the vertical distances, and this is accomplished by using the *method of least squares*.

The starting point of the method of ►least squares is to write the estimated model as

$$e = \widehat{Y} - (\widehat{\beta}_0 + \widehat{\beta}_1 X)$$

since the residual e represents the vertical distance Y to the line. Then the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen that minimize the sum of the squares of residuals

$$S = \Sigma e_i^2 = \Sigma \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2.$$

To minimize S, we take the partial derivative of S with respect to each of the two estimates and set the resulting expressions equal to zero. Thus we obtain

$$\widehat{\beta}_0 n + \widehat{\beta}_1 \Sigma X_i = 0$$

$$\widehat{\beta}_0 \Sigma X_i + \widehat{\beta}_1 \Sigma X_i^2 = 0$$

which are called the *normal equations*. If we solve these equations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we obtain

$$\widehat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

The method of least squares, on which most methods of estimation for regression analysis are based was apparently first published by Legendre (1805), but the first treatment along the lines now familiar was given by Gauss (for the details regarding history of least squares *see* ►Gauss–Markov theorem). Gauss showed that the least squares method gives estimators of the unknown parameters with minimum variance among unbiased linear estimators. This basic result is now known as the Gauss–Markov theorem, and the least squares estimators as Gauss–Markov estimators. That is, there is no other choice of values for the two parameters β_0 and β_1 that provide a smaller $\sum e_i^2$. If a residual, e_i, is too large compared with the other residuals, the corresponding Y_i may be an outlier or may be an influential observation that influences the estimates of two parameters β_0 and β_1. Detection of an outlier or an influential observation is an important research area, and many books such as Belsley et al. (1980) and Cook and Weisberg (1982), deal with this topic. (see also ►Cook's distance, ►Regression diagnostics, ►Influential observations).

About the Author

Professor Sung Park is Past President of Korean Statistical Society (1997–1998), Korean Society for Quality Management, Vice President of International Society for Business and Industrial Statistics, and Academician of International Academy for Quality. In 2000, he received the prestigious gold medal from the President of the Korean Government for his contribution to quality management in Korea. Recently, he has served as the Dean of the College of Natural Sciences, Seoul National University. He has published more than 30 books on statistics and quality control including three books in English: *Robust Design and Analysis for Quality Engineering* (Chapman & Hall, 1996), and edited the text *Statistical Process Monitoring and Optimization* (with G. Geoffrey Vining, Marcel Dekker, 2000) and *Six Sigma for Quality and Productivity Promotion* (Asian Productivity Organization, Free eBook, 2003.)

Cross References

►Cook's Distance
►Gauss-Markov Theorem
►Influential Observations
►Least Squares
►Linear Regression Models
►Regression Diagnostics
►Residuals

References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Birkes D, Dodge Y (1993) Alternative methods of regression. Wiley, New York

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London

Legendre AM (1805) Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot, Paris

Simple Random Sample

Roger E. Kirk
Distinguished Professor of Psychology and Statistics
Baylor University, Waco, TX, USA

A ►census, surveying every element in a finite population, is used to discover characteristics of the population. If the population is large, a census can be costly, time consuming, or impracticable. Alternatively, a simple random sample can be used to obtain information and draw inferences about the population. It is customary to sample elements without replacement. That is, once an element has been

selected, it is removed from the population so that it cannot be selected a second time. A simple random sampling procedure is used to obtain a simple random sample. The procedure selects a sample of size n from a finite population of size $N < n$ such that each of the ${}_N C_n = N!/[n!(N-n)!]$ possible samples is equally likely to be selected. If sample elements are returned to the population after being selected – sampling with replacement – each of the ${}_{N+n-1}C_n = (N+n-1)!/\{n![(N+n-1)-n]!\}$ possible samples is equally likely to be selected.

Simple random sampling is a type of probability sampling. All probability sampling procedures have three characteristics in common: (a) the elements that compose the population are explicitly defined, (b) every potential sample of a given size that could be drawn from the population can be enumerated, and (c) the probability of selecting any potential sample can be specified. Non-probability sampling procedures do not satisfy one or more of the three characteristics. An example of a non-probability sampling procedure is convenience sampling-elements are selected because they are readily available. For simple random sampling without replacement, the probability of a particular sample being selected is $1/({}_N C_n)$. For sampling with replacement, the probability of a particular sample being selected is $1/({}_{N+n-1}C_n)$. When sampling with replacement the inclusions of the ith and jth $(i \neq j)$ members of the population are statistically independent. However, these events are not statistically independent when sampling without replacement. For this case, the probability of the inclusions of ith and jth population members is $n(n-1)/[N(N-1)]$ (McLeod 1988).

Simple random samples have two interrelated advantages over non-probability samples. First, randomness avoids bias, that is, a systematic or long-run misrepresentation of the population. Second, randomness enables researchers to apply the laws of probability in determining the likely error of sample statistics. A particular random sample rarely yields an estimate of the population characteristic that equals the population characteristic. However, the expected value of the sample estimate will over an indefinitely large number of samples equal the population characteristic. Furthermore, for any simple random sample, it is possible to estimate the magnitude of the error associated with the estimate. For large populations the error depends only on the sample size, a fact that is counterintuitive (Anderson 2001).

The first step in obtaining a simple random sample is to develop a sampling frame: a list of all of the elements in the population of interest. The sampling frame operationally defines the population from which the sample is drawn and to which the sample results can be generalized. Once the sampling frame has been developed, a simple random sample can be obtained in a variety of ways. For example, a researcher can record on a slip of paper the identifying code for each element in the sampling frame. The slips of paper are placed in a container and thoroughly shuffled. The first n unique slips drawn without bias from the container compose the sample. The most common method of obtaining a simple random sample uses random numbers. Tables of random numbers are available in many statistics textbooks. The tables contain a sequence of random digits whose terms are chosen so that each digit is equally likely to be $0, 1, \ldots, 9$ and the choices at any two different places in the sequence are independent. For ease in reading the digits in a random number table, the digits are often grouped with two digits in a group, four digits in a group, and so on. To use a table to select a simple random sample of size, say, $n = 50$ from a population of size $N = 988$, assign the numbers $000, 002, \ldots, 987$ to the elements in the sampling frame. Select a starting point in the table by dropping a pointed object on the table. Choose three-digit numbers beginning at the starting point until 50 distinct numbers between 000 and 987 are obtained. The sample consists of the elements corresponding to the 50 numbers selected. This procedure illustrates sampling without replacement because once a number has been selected, the number is ignored if it is encountered again. Computer packages such as SAS, SPSS, and MINITAB and many hand calculators have routines that produce numbers that in every observable way appear to be random. For an in-depth discussion of sampling procedures, see Schaeffer et al. (2006).

About the Author

Professor Kirk is a Fellow of the American Psychological Association, Association for Psychological Science, American Educational Research Association, and the American Association of Applied and Preventive Psychology. He is the 2005 recipient of the American Psychological Association's Jacob Cohen Award for Distinguished Contributions to Teaching and Mentoring. He is a Founding Associate editor of the *Journal of Educational Statistics*, 42nd president of the Southwestern Psychological Association (1995–1996), and 46th president of Division 5 (Evaluation, Measurement, and Statistics) of the American Psychological Association (1992–1993). Professor Kirk was President of the Society for Applied Multivariate Research (1984–1985).

Cross References

►Handling with Missing Observations in Simple Random Sampling and Ranked Set Sampling
►Non-probability Sampling Survey Methods

►Proportions, Inferences, and Comparisons
►Ranked Set Sampling
►Representative Samples
►Sample Size Determination
►Sampling From Finite Populations
►Uniform Random Number Generators

References and Further Reading

Anderson NH (2001) Empirical directions in design and analysis. Erlbaum, Mahwah

McLeod I (1988) Simple random sampling. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 8. Wiley, New York, pp 478–479

Schaeffer RL, Ott RL, Mendenhall W (2006) Elementary survey sampling 6th edn. Thompson Learning, Belmont

Simpson's Paradox

Zhi Geng
Professor, Director of the Institute of Mathematical Statistics of Peking University
Peking University, Beijing, China

An association measurement between two variables X and Y may be dramatically changed from positive to negative by omitting a third variable Z, which is called Simpson's paradox or the Yule-Simpson paradox (Yule, 1903; Simpson, 1951). A numerical example is shown in Table 1. The risk difference (RD) is defined as the difference between the recovery proportion in the treated group and that in the placebo group, $RD = (80/200) - (100/200) = -0.10$. If the population is split into two populations of male and female, a dramatic change can be seen from Table 2. The risk differences for male and female are both changed to 0.10. Thus we obtain a self-contradictory conclusion that the new drug is effective for both male and female but it is ineffective for the whole population. Should patients in the population take the new drug or not? Should the correct answer depend on whether the doctor know the gender of patients?

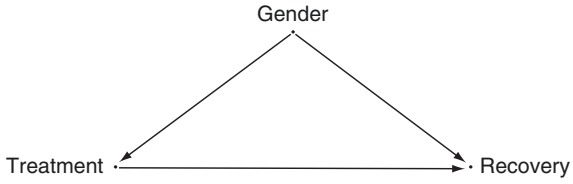
From Table 2, we can see that most males took placebo, but most females took the new drug. As depicted in Fig. 1, there may be a spurious association between treatment and response because gender associates with both treatment and response. Such a factor that is associated with both treatment and response is called a confounding factor or a confounder. If a confounder is known and observed, the bias due to the confounder can be removed by stratification or standardization. If there are unknown or unobserved confounders, in order to remove the confounding bias, we can randomize the treatment assignment such that the association path between the confounders and the treatment is broken. In Fig. 2, we depict a randomized experiment for this example, where 200 males (M) and

Simpson's Paradox. Table 1 Recovery proportions in treatment and placebo groups

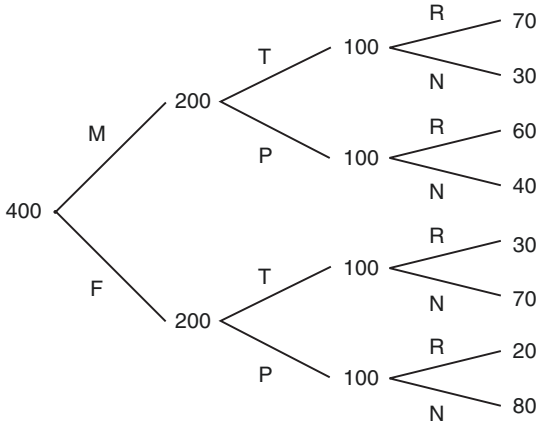
Treatment	Recovery	Non-recovery	Total
New drug	80	120	200
Placebo	100	100	200
	$RD = \frac{80}{200} - \frac{100}{200} = -0.10$		

Simpson's Paradox. Table 2 Populations stratified by gender

	Male		Female	
Treatment	Recovery	Non-recovery	Recovery	Non-recovery
New drug	35	15	45	105
Placebo	90	60	10	40
	$RD_M = 0.10$		$RD_F = 0.10$	



Simpson's Paradox. Fig. 1 A confounding factor: gender



Simpson's Paradox. Fig. 2 Randomized experiment

Simpson's Paradox. Table 3 Subscription renewal rates in 1979

Month	Source of current subscription					Overall
	Gift	Previous renewal	Direct mail	Subscription service	Catalog agent	
January						
Total	3,594	18,364	2,986	20,862	149	45,955
Renewals	2,918	14,488	1,783	4,343	13	23,545
Rate	0.812	0.789	0.597	0.208	0.087	0.512
February						
Total	884	5,140	2,224	864	45	9,157
Renewals	704	3,907	1,134	122	2	5,869
Rate	0.796	0.760	0.510	0.141	0.044	0.641

Simpson's Paradox. Table 4 Total income and total tax (in 10^3 dollars) and tax rate

Adjusted gross income	1974			1978		
	Income	Tax	Tax rate	Income	Tax	Tax rate
Under $5,000	41,651,643	2,244,467	0.054	19,879,622	689,318	0.035
$5,000 to $9,999	146,400,740	13,646,348	0.093	122,853,315	8,819,461	0.072
$10,000 to $14,999	192,688,922	21,449,597	0.111	171,858,024	17,155,758	0.100
$15,000 to $99,999	470,010,790	75,038,230	0.160	865,037,814	137,860,951	0.159
$100,000 or more	29,427,152	11,311,672	0.384	62,806,159	24,051,698	0.383
Total	880,179,247	123,690,314	0.141	1,242,434,934	188,577,186	0.152

200 females (F) are randomly assigned into the new drug group (T) and the placebo group (M). The recovery proportion is 35/50 in the new drug group of males, and thus 70 of 100 treated males recover (R) and the other 30 do not recover (N). From Fig. 2, the total number of recovered people is 70+30=100 and the recovery proportion is 100/200 in the new drug group; the total number is 60+20=80 and the recovery proportion is 80/200 in the placebo group. Thus, we conclude on that the new drug increases recovery proportion by 10%, which is consistent with that shown in Table 2.

Two real-life examples of Simpson's paradox were showed by Wagner (1982). The first example, as shown in Table 3, illustrates that the overall renewal rate of *American History Illustrated* magazine increased from 51.2 percent in January 1979 to 64.1 percent in February 1979, but the renewal rates actually declined in every subscription category. The second example, as shown in Table 4, illustrates that the overall income tax rate increased from 14.1 percent in 1974 to 15.2 percent in 1978, but the tax rate decreased in each income category. Reintjes et al. (2000) gave the following example from hospital epidemiology: 3519 gynecology patients from eight hospitals in a nonexperimental study were used to study the association between antibiotic prophylaxis (AB-proph.) and urinary tract infections (UTI). The eight hospitals were stratified into two groups with a low incidence percentage (< 2.5%) and a high percentage (> 2.5%) of UTI. By Table 5, the relative risk (RR) was (42/1279)/(104/2240) = 0.7 for the overall eight hospitals, which means that AB-proph. had a protective effect on UTI. But the RRs were 2.6 and 2.0 for the low and the high incidence groups, respectively, which means that

Simpson's Paradox. Table 5 Data on UTI and AB-proph. stratified by incidence of UTI per hospital

	Hospitals with low UTI		Hospitals with high UTI		All hospitals	
AB-proph.	UTI	no-UTI	UTI	no-UTI	UTI	no-UTI
Yes	20	1093	22	144	42	1237
No	5	715	99	1421	104	2136
	$RR_L = 2.6$		$RR_H = 2.0$		$RR = 0.7$	

AB-proph. had a risk effect on UTI for both groups. The real effect of AB-proph. on UTI has been shown to be protective in randomized clinical trials, which is consistent with the crude analysis rather than the stratified analysis. This result explains that there were more unidentified confounders which canceled their effects each other out in the crude analysis.

There are many topics related to Simpson's paradox. Collapsibility of association measurements deals with conditions under which association measurements are unchanged by omitting other variables (Cox and Wermuth, 2003; Ma et al. 2006). From the viewpoint of causality, Simpson's paradox occurs because there are confounders such that association measurement is biased from causal effects (Pearl, 2000; Geng et al. 2002). A variation of Simpson's paradox is a surrogate paradox, which means that a treatment has a positive effect on an intermediate variable called a surrogate, which in turn has a positive effect on the true endpoint, but the treatment has a negative effect on the true endpoint (Chen et al. 2007; Ju and Geng, 2010). Moore (1995) describes a real trial of antiarrhythmic drugs in which an irregular heartbeat is a risk factor of early mortality but correction of the heartbeat increased mortality.

About the Author

Dr. Zhi Geng is Professor, School of Mathematical Sciences, Peking University. He obtained his PhD (1989) from Kyushu University, Japan. He has been a member of International Statistical Institute since 1996. Professor Geng is Associate Editor of *Computational Statistics and Data Analysis*.

Cross References

►Causation and Causal Inference
►Collapsibility
►Confounding and Confounder Control
►Statistical Fallacies

References and Further Reading

Chen H, Geng Z, Jia J (2007) Criteria for surrogate endpoints. J R Stat Soc B 69:919–932
Cox DR, Wermuth N (2003) A general condition for avoiding effect reversal after marginalization. J R Stat Soc B 65:937–941
Geng Z, Guo J, Fung WK (2002) Criteria for confounders in epidemiological studies. J R Stat Soc B 64:3–15
Ju C, Geng Z (2010) Criteria for surrogate endpoints based on causal distributions. J R Stat Soc B 72:129–142
Ma ZM, Xie XC, Geng Z (2006) Collapsibility of distribution dependence. J R Stat Soc B 68:127–133
Moore T (1995) Deadly medicine: why tens of thousands of patients died in America's worst drug disaster. Simon & Shuster, New York
Pearl J (2000) Causality: models, reasoning, and inference. University Press, Cambridge
Reintjes R, de Boer A, van Pelt W, Mintjes-de Groot J (2000) Simpson's paradox: an example from hospital epidemiology. Epidemiology 11:81–83
Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc B 13:238–241
Wagner CH (1982) Simpson's paradox in real life. Am Stat 36:46–48
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors

Lennart Hoogerheide[1], Herman K. van Dijk[2]
[1]Econometric and Tinbergen Institutes, Erasmus University Rotterdam, Rotterdam, The Netherlands
[2]Professor, Director of the Tinbergen Institute Econometric and Tinbergen Institutes, Erasmus University Rotterdam, Rotterdam, The Netherlands

The financial market turmoil has been shocking the world since early 2008. As is aptly stated by the president of the European Central Bank, Trichet (2008), the widespread

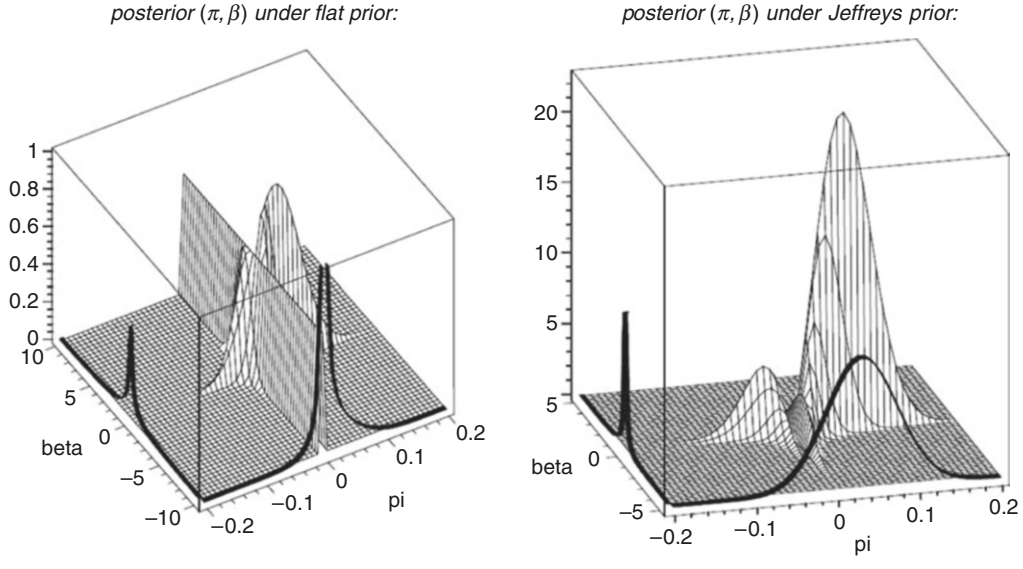
undervaluation of risk is one of the most important issues in this context and appropriate operational risk management is a crucial issue to be investigated. A seemingly unrelated issue is to measure and predict the *treatment effect of education on income*. This issue is crucial for any country that increasingly relies on the "knowledge economy." In recent research by the authors it is stressed that ***these seemingly unrelated issues pose similar questions and have common components from a modeling and statistical viewpoint***.

There exist connections between dynamic time series models used in the first issue and treatment effect models. This common problem structure is explained in research by the authors as follows: the restricted reduced form of the instrumental variable (IV) model and the Vector Error Correction Model (VECM) under cointegration are both instances of the general reduced rank regression model with different variables and parameters playing the same roles, as summarized in the Table 1.

In these models with near reduced rank one may encounter non-elliptical posteriors. In the Bayesian analysis of treatment effects, for example in the instrumental variable (IV) model, we often encounter posterior distributions that display these shapes. The reason for this is

Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors. Table 1 Common model structures

Model	Restricted reduced form (RRF) of instrumental variable (IV) model	Vector Error Correction Model (VECM) under cointegration
Endogenous variables	Dependent variable and (possibly) endogenous regressors	Vector of variables' changes (= current − previous values)
Predetermined variables corresponding to parameter matrix with *reduced rank*	Instrumental variables (having no *direct* effect on the dependent variable, only an *indirect* effect via the (possibly) endogenous regressors)	Vector of previous values
Predetermined variables corresponding to *unrestricted* parameter matrix	Control variables (having a *direct* effect on both the dependent variable and the (possibly) endogenous regressors)	Vector of other explanatory variables and past variables' changes



Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors. Fig. 1 Posterior density of π (expected difference in years of education between children born in April-December and January-March) and β (treatment effect of education on income) for 29,015 data (used by Angrist and Krueger (1991)) from men born in the state of New York

local non-identification: if some of the model parameters (reflecting the strength of the instruments) tend to 0, i.e., the case of weak instruments, other model parameters (corresponding to the relevant treatment effect) become unidentified.

Angrist and Krueger (1991) consider the estimation of the treatment effect β of education on income, which is non-trivial due to unobserved (intellectual) capabilities that not only influence education but also directly affect income, and due to measurement errors in the reported education level. Angrist and Krueger (1991) use American data and suggest using quarter of birth to form ►instrumental variables. These instruments exploit that students born in different quarters have different average education. This results since most school districts require students to have turned age six by a certain date, a so-called "birthday cutoff" which is typically near the end of the year, in the year they enter school, whereas compulsory schooling laws compel students to remain at school until their 16th, 17th or 18th birthday. This asymmetry between school-entry requirements and compulsory schooling laws compels students born in certain months to attend school longer than students born in other months: students born earlier in the year enter school at an older age and reach the legal dropout age after less education. Hence, for students who leave school as soon as the schooling laws allow for it, those born in the first quarter have on average attended school for three quarters less than those born in the fourth quarter. Suppose we use as a single instrument a 0/1 indicator variable with value 0 indicating birth in the first quarter; the strength of this instrument is given by its effect on education, parameter π. The left panel of Fig. 1 shows the posterior density of π and β (under a flat prior) for 29,015 data from men born in the state of New York in 1930–1939. This shows a clear "ridge" around $\pi = 0$, indicating that for π tending to 0 a wide range of values of β becomes possible. An alternative prior, the Jeffreys prior, regularizes the posterior shapes in the sense that it eliminates the asymptote around $\pi = 0$ for the marginal posterior of π, yet the joint posterior shapes in the right panel of Fig. 1 are still far from elliptical. This example illustrates that the weakness of the instruments may imply that even for large data sets posterior distributions may be highly non-elliptical.

Thus for the Bayesian analysis of (non-linear) extensions of the IV model, we need flexible simulation methods. The use of neural network based simulation is then particularly useful. A Bayesian optimal information processing procedure using advanced simulation techniques based on artificial neural networks (ANN) is recently developed and it can be used as a powerful tool for forecasting and policy advice. These simulation methods have already been successfully applied to evaluate risk measures (Value-at-Risk, Expected Shortfall) for a single asset. The procedures proposed by the authors are just one step forward on the path of understanding these issues and these involve a novel manner of processing the information flow on these issues. It is – of course – the intention of this research that its results improve forecasting of risk and uncertainty that influence the effectiveness of interventions and treatments.

About the Author

Herman K. van Dijk is director of the Tinbergen Institute and professor of Econometrics with a Personal Chair at Erasmus University Rotterdam. He is a former Director of the Econometric Institute and Honorary Fellow of the Tinbergen Institute. He has been a visiting fellow and a visiting professor at Cambridge University, the Catholic University of Louvain, Harvard University, Duke University, Cornell University, and the University of New South Wales. He received the Savage Prize for his Ph.D. dissertation and is listed in the journal Econometric Theory in the Econometricians Hall of Fame amongst the top ten European econometricians. He serves on the Editorial Board of major journals in econometrics. His publications consist of several books and more than 160 international scientific journal papers and reports.

Cross References

►Instrumental Variables
►Neural Networks
►Quantitative Risk Management

References and Further Reading

Angrist JD, Krueger AB (1991) Does compulsory school attendance affect schooling and earnings? Quart J Econom 106:979–1014

Ardia D, Hoogerheide LF, Van Dijk HK (2009) To bridge, to warp or to wrap? A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods. TI discusion paper 09-017/4

Ardia D, Hoogerheide LF, Van Dijk HK (2009b) AdMit: adaptive mixtures of student-*t* distributions. The R Journal 1(1):25–30

Ardia D, Hoogerheide LF, Van Dijk HK (2009c) Adaptive mixture of Student-*t* distributions as a flexible candidate distribution for efficient simulation: the R package AdMit. J Stat Softw 29(3): 1–32

Hoogerheide LF (2006) Essays on neural network sampling methods and instrumental variables. Ph.D. thesis, Book nr. 379 of the Tinbergen Institute Research Series, Erasmus University Rotterdam

Hoogerheide LF, Van Dijk HK (2006) A reconsideration of the Angrist-Krueger analysis on returns to education. Report 2006-15 of the Econometric Institute, p 35

Hoogerheide LF, Van Dijk HK (2007) Note on neural network sampling for Bayesian inference of mixture processes. Bulletin of the

International Statistical Institute, Proceedings of the Biennial Sessions in Lisbon 2007, p 8
Hoogerheide LF, Van Dijk HK (2010) Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling. Int J Forecasting, 26:231–247
Hoogerheide LF, Kaashoek JF, Van Dijk HK (2003) Neural network approximations to posterior densities: an analytical approach. In: Proceedings of the section on Bayesian statistical science, American Statistical Association, 2003, p 5
Hoogerheide LF, Kaashoek JF, Van Dijk HK (2007a) On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. J Econom 139(1): 154–180
Hoogerheide LF, Kleibergen F, Van Dijk HK (2007b) Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. J Econom 138(1):63–103
Hoogerheide LF, Kleijn R, Ravazzolo F, Van Dijk HK, Verbeek M (2010) Forecast accuracy and economic gains from Bayesian model averaging using time varying weights. J Forecast, 29:251–269
Hoogerheide LF, Van Dijk HK, Van Oest RD (2009) Simulation based Bayesian econometric inference: principles and some recent computational advances. Chapter 7 in Handbook of Computational Econometrics, Wiley, pp 215–280
Trichet JC (2008) Macroeconomic policy is essential to stability. Financial Times, November 13, 2008, p 13

Singular Spectrum Analysis for Time Series

Anatoly Zhigljavsky
Professor, Chair in Statistics
Cardiff University, Cardiff, UK

Singular spectrum analysis (SSA) is a technique of time series analysis and forecasting. It combines elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. SSA aims at decomposing the original series into a sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and a "structureless" noise. It is based on the singular-value decomposition of a specific matrix constructed upon time series. Neither a parametric model nor stationarity-type conditions have to be assumed for the time series; this makes SSA a model-free technique.

The commencement of SSA is usually associated with publication of the papers (Broomhead and King 1986a, b) by Broomhead and King. Nowadays SSA is becoming more and more popular, especially in applications. There are several hundred papers published on methodological aspects and applications of SSA, see Golyandina et al. (2001), Vautard et al. (1992), Vautard and Ghil (1989), Allen and Smith (1996), and Zhigljavsky (2010) and references therein. SSA has proved to be very successful, and has already become a standard tool in the analysis of climatic, meteorological and geophysical time series; see, for example, Vautard et al. (1992), Vautard and Ghil (1989), and Allen and Smith (1996). More recent areas of application of SSA include engineering, medicine, econometrics and many other fields. Most recent developments in the theory and methodology of SSA can be found in Zhigljavsky (2010). We start with 'Basic SSA', which is the most common version of SSA.

Basic SSA

Let $x_1, \ldots, x_N$ be a time series of length N. Given a window length L $(1<L<N)$, we construct the L-lagged vectors $X_i = (x_i, \ldots, x_{i+L-1})^T$, $i = 1, 2, \ldots, K = N–L+1$, and compose these vectors into the matrix $\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = [X_1 : \ldots : X_K]$. This matrix has size $L \times K$ and is often called "trajectory matrix." It is a Hankel matrix, which means that all the elements along the diagonal $i+j$=const are equal.

The columns X_j of $\mathbf{X}$, considered as vectors, lie in the L-dimensional space $\mathbb{R}^L$. The singular-value decomposition of the matrix $\mathbf{X}\mathbf{X}^T$ yields a collection of L eigenvalues and eigenvectors. A particular combination of a certain number $l < L$ of these eigenvectors determines an l-dimensional subspace in $\mathbb{R}^L$. The L-dimensional data $\{X_1, \ldots, X_K\}$ is then projected onto this l-dimensional subspace and the subsequent averaging over the diagonals gives us some Hankel matrix $\tilde{\mathbf{X}}$ which is considered as an approximation to $\mathbf{X}$. The series reconstructed from $\tilde{\mathbf{X}}$ satisfies some linear recurrent formula which may be used for forecasting.

In addition to forecasting, the Basic SSA can be used for smoothing, filtration, noise reduction, extraction of trends of different resolution, extraction of periodicities in the form of modulated harmonics, gap-filling (Kondrashov and Ghil 2006; Golyandina and Osipov 2007) and other tasks, see Golyandina et al. (2001). Also, the Basic SSA can be modified and extended in many different ways some of which are discussed below.

Extensions of the Basic SSA

SSA for analyzing stationary series (Vautard and Ghil 1989). Under the assumption that the series $x_1, \ldots, x_N$ is stationary, the matrix $\mathbf{X}\mathbf{X}^T$ of the Basic SSA is replaced with the so-called lag-covariance matrix $\mathbf{C}$ whose elements are $c_{ij} = \frac{1}{N-k}\sum_{t=1}^{N-k} x_t x_{t+k}$ with $i,j = 1, \ldots, L$ and $k = |i-j|$. In the terminology of Golyandina et al. (2001), this is "Toeplitz SSA."

Monte-Carlo SSA (Allen and Smith 1996). In the Basic SSA we implicitly associate the "structureless" component of the resulting SSA decomposition with "white noise" (this noise may not necessarily be random). In some applications, however, it is more natural to assume that the noise is "colored". In this case, special tests based on Monte Carlo simulations may be used to test the hypothesis of the presence of a signal.

Improvement or replacement of the singular-value decomposition (SVD) procedure. There are two main reasons why it may be worthwhile to replace the SVD operation in the Basic SSA with some another operation. The first reason is simplicity: in problems where the dimensions of the trajectory matrix is large, SVD may simply be too costly to perform; substitutions of SVD are available, see Golub and van Loan (1996) and Moskvina and Schmidt (2003). The second reason is the analysis of the accuracy of SSA procedures based on the perturbation theory (Zhigljavsky 2010). For example, in the problems of separating signal from noise, some parts of the noise are often found in SVD components corresponding to the signal. As a result, a small adjustment of the eigenvalues and eigenvectors is advisable to diminish this effect. The simplest version of the Basic SSA with a constant adjustment in all eigenvalues was suggested in Van Huffel (1993) and is sometimes called the minimum-variance SSA.

Low-rank matrix approximations, Cadzow iterations, connections with signal processing. As an approximation to the trajectory matrix $\mathbf{X}$, the Basic SSA yields the Hankel matrix $\tilde{\mathbf{X}}$. This matrix is obtained as a result of the diagonal averaging of a matrix of rank l. Hence $\tilde{\mathbf{X}}$ is typically a matrix of full rank. However, in many signal processing applications, when a parametric form of an approximation is of prime importance, one may wish to find a Hankel matrix of size $L \times K$ and rank l which gives the best approximation to $\mathbf{X}$; this is a problem of the structured low-rank approximation (Markovsky et al. 2006). The simplest procedure of finding a solution to this problem (not necessarily the globally optimal one though) is the so-called Cadzow iterations (Cadzow 1988) which are the repeated alternating projections of the matrices (starting at $\mathbf{X}$) to the set of matrices of rank l (by performing the singular-value decompositions) and to the set of Hankel matrices (by making the diagonal averaging). That is, Cadzow iterations are simply the repeats of the Basic SSA. It is not guaranteed however that Cadzow iterations lead to more accurate forecasting formulas than the Basic SSA (Zhigljavsky 2010).

SSA for change-point detection and subspace tracking (Moskvina and Zhigljavsky 2003). Assume that the observations $x_1, x_2, \ldots$ of the series arrive sequentially in time and we apply the Basic SSA to the observations at hand. Then we can monitor the distances from the sequence of the trajectory matrices to the l-dimensional subspaces we construct and also the distances between these l-dimensional subspaces. Significant changes in any of these distances may indicate on a change in the mechanism generating the time series. Note that this change in the mechanism does not have to affect the whole structure of the series but rather only a few of its components.

SSA for multivariate time series. Multivariate (or multichannel) SSA (shortly, MSSA) is a direct extension of the standard SSA for simultaneous analysis of several time series. Assume that we have two series, $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$. The (joint) trajectory matrix of the two-variate series (X, Y) can be defined as either $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ or $\mathbf{Z} = (X, Y)^T$, where $\mathbf{X}$ and $\mathbf{Y}$ are the trajectory matrices of the individual series X and Y. Matrix $\mathbf{Z}$ is block-Hankel rather than simply Hankel. Other stages of MSSA are identical to the stages of the univariate SSA except that we build a block-Hankel (rather than ordinary Hankel) approximation $\tilde{\mathbf{Z}}$ to the trajectory matrix $\mathbf{Z}$.

MSSA may be very useful for analyzing several series with common structure. MSSA may also be used for establishing a causality between two series. Indeed, the absence of causality of Y on X implies that the knowledge of Y does not improve the quality of forecasts of X. Hence an improvement in the quality of forecasts for X which we obtain using MSSA against univariate SSA forecasts for X gives us a family of SSA-causality tests, see Hassani et al. (2010).

Cross References

►Forecasting: An Overview
►Monte Carlo Methods in Statistics
►Statistical Signal Processing
►Time Series

References and Further Reading

Allen MR, Smith LA (1996) Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. J Clim 9:3373–3404

Broomhead DS, King GP (1986a) Extracting qualitative dynamics from experimental data. Physica D 20:217–236

Broomhead DS, King GP (1986b) On the qualitative analysis of experimental dynamical systems. In: Sarkar S (ed) Nonlinear phenomena and chaos. Adam Hilger, Bristol, pp 113–144

Cadzow JA (1988) Signal enhancement a composite property mapping algorithm. IEEE Trans Acoust Speech Signal Process 36: 49–62

Golub G, van Loan C (1996) Matrix computations. Johns Hopkins University Press, London

Golyandina N, Osipov E (2007) The Caterpillar-SSA method for analysis of time series with missing values. J Stat Plan Infer 137:2642–2653

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman & Hall/CRC Press, New York/London

Hassani H, Zhigljavsky A, Patterson K, Soofi A (2010) A comprehensive causality test based on the singular spectrum analysis, causality in science. In: McKay Illary P, Russo F, Williamson J (eds) Causality in Sciences, Oxford University Press

Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Proc Geoph 13: 151–159

Markovsky I, Willems JC, Van Huffel S, De Moor B (2006) Exact and approximate modeling of linear systems: a behavioral approach. SIAM, Philadelphia

Moskvina V, Schmidt KM (2003) Approximate projectors in singular spectrum analysis. SIAM J Matrix Anal Appl 24:932–942

Moskvina VG, Zhigljavsky A (2003) An algorithm based on singular spectrum analysis for change-point detection. Commun Stat Simul Comput 32:319–352

Van Huffel S (1993) Enhanced resolution based on minimum variance estimation and exponential data modeling. Signal process 33:333–355

Vautard R, Ghil M (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Physica D 35:395–424

Vautard R, Yiou P, Ghil M (1992) Singular spectrum analysis: a toolkit for short, noisy and chaotic series. Physica D 58:95–126

Zhigljavsky A (ed) (2010) Statistics and its interface, special issue on the singular spectrun analysis for time series. Springer, New York

SIPOC and COPIS: Business Flow – Business Optimization Connection in a Six Sigma Context

Rick L. Edgeman
Professor, Chair & Six Sigma Black Belt
University of Idaho, Moscow, ID, USA

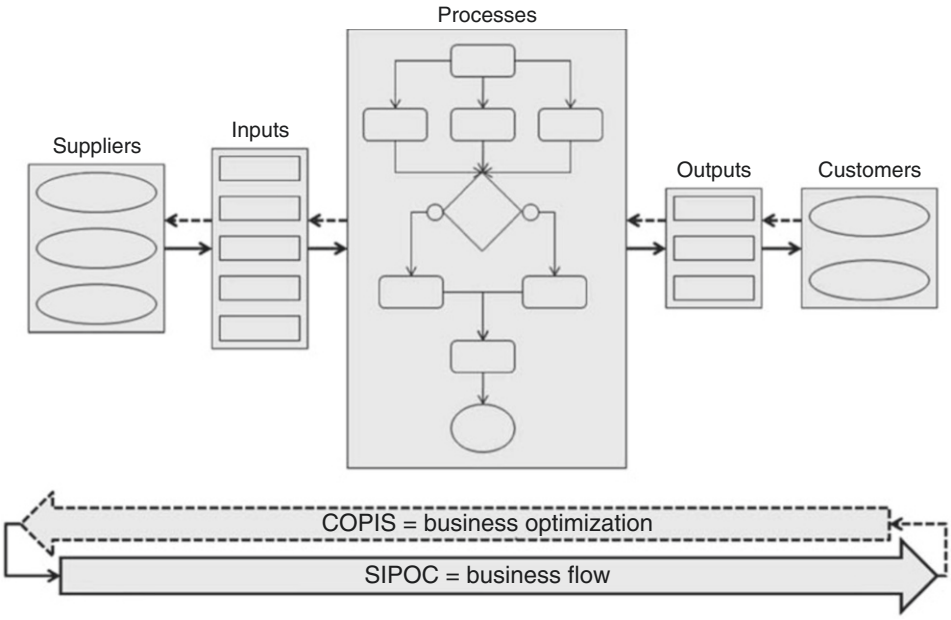
►Six Sigma can be defined as *a highly structured strategy for acquiring, assessing, and applying customer, competitor, and enterprise intelligence in order to produce superior product, system or enterprise innovation and designs* (Klefsjö et al. 2006). Focal to this definition is the customer and indeed the customer functions as the pivot point for this contribution as customer needs and wants drive change in most organizations.

Six Sigma originated at Motorola approximately 3 decades ago as a means of generating near-perfect products via focus on associated manufacturing processes and while initially applied almost exclusively in manufacturing environments, its inherent sensibilities and organization facilitated migration to service operations. Similarly, while Six Sigma was at the outset used to generate significant innovation in and improvement of existing products, those same sensibilities led to its adaptation to new product and process design environments. In statistical terms a process operating at a "true" six sigma level produces an average of only 3.4 defects per million opportunities (DPMO) for defects where this figure is associated with a process with a 12 standard deviation spread between lower and upper specification limits, but wherein the 3.4 DPMO figure is based on allowance for a 1.5 standard deviation non-centrality factor or shift away from "perfect centering" so that, in essence, one specification limit is 4.5 standard deviations away from the targeted or ideal performance level whereas the other specification limit is 7.5 standard deviations away from that performance level.

Within the context of a structured problem-solving context Six Sigma integrates various strategies and tools from Statistics, Quality, Business, and Engineering with the adoption of new ones likely as its use expands to more business sectors and areas of application. Its focus divides into two significant and related branches that share a number of tools, techniques and objectives, but often apply these tools and techniques differently and its use has added multiple billions in any currency to the financial bottom lines of numerous organizations across many sectors of the economy, including financial, healthcare, military, and general manufacturing. Six Sigma's branches are ones that focus on significant innovation/redesign in or of existing products, processes, and systems and a second that is directed at design of new products, processes or systems. Included among the leading companies emphasizing Six Sigma are GE, 3M, Raytheon, Sun Microsystems, DuPont, Bank of America, American Express, Motorola, Rolls Royce, and Boeing.

Central to business flow is the familiar SIPOC model (Suppliers → Inputs → Processes → Outputs → Customers) indicating that, commonly, suppliers provide inputs that are transformed by internal processes into outputs that are in turn provided to customers. While this flow is common and logical, its optimization is far less so, but can be approached application of Stephen Covey's familiar "habit" of "beginning with the end in mind" (Covey 1989), a manifestation of which in the present case is COPIS (Customers → Outputs → Processes → Inputs → Suppliers).

Organizations that practice COPIS – often as part of a quality management or six sigma culture – do so by first carefully elaborating who their customers are as well as the needs and wants of those customers (called the "Voice of the Customer" or "VOC"). Customer-driven organizations will ensure that these needs and wants are reflected in and fulfilled by the outputs of processes that must be optimally



SIPOC and COPIS: Business Flow – Business Optimization Connection in a Six Sigma Context. Fig. 1

configured in order to deliver these outputs by transforming the most appropriate inputs that have been provided by the most apt suppliers. It can be seen from this that, consistent with Covey, "see the end from the beginning," that is, to be customer-driven. In a continuous improvement culture this occurs not once, but cyclically. These ideas are portrayed in Fig. 1.

Statistical and other quantitatively oriented methods that can be brought to bear throughout the COPIS-SIPOC flow include the use of sample survey methods to elicit the VOC and numerous additional analytical techniques from across the statistical spectrum can be used to assess the VOC. Optimal process configuration is not merely a matter of work flow and equipment, but also of ensuring that however those are assembled, that the outputs themselves are optimized. While many tools can be employed, generally outputs can be regarded as response variables, Y, where

$$Y = f(X_1, X_2, \ldots, X_P) + \varepsilon,$$

where $X_1, X_2, \ldots, X_P$ are controllable variables, the optimal combination of settings of which can be determined using response surface methods, steepest ascent methods, and evolutionary operations or EVOP (Myers et al. 2009). In a similar way, such methods can be used to assist in selection of inputs and subsequently the suppliers from whom these should be obtained.

In all, what we see is that as best practice, business is conceived of as COPIS to yield optimal results as determined by the VOC, but subsequently deployed as SIPOC. While SIPOC is common to most business environments, employment of COPIS is practiced far less often and then typically only in customer-driven environments. Practice of COPIS offers rich opportunities for application of statistical methods as well as subsequent rewards.

About the Author

For biography *see* the entry ►Design for Six Sigma.

Cross References

►Business Statistics
►Design for Six Sigma
►Industrial Statistics
►Six Sigma

References and Further Reading

Covey SR (1989) The seven habits of highly effective people. Free, New York

Klefsjö B, Bergquist B, Edgeman, R (2006) Six sigma and total quality management: different day, same soup? Six Sigma and Competitive Advantage 2(2):162–178

Myers RH, Montgomery DC, Anderson-Cook CM (2009) Response surface methodology: process and product optimization using designed experiments, 3rd edn. Wiley, New York

Six Sigma

David M. Levine
Professor Emeritus of Statistics and Computer Information Systems
Baruch College, City University of New York, New York, NY, USA

Six Sigma is a quality improvement system originally developed by Motorola in the mid-1980s. After seeing the huge financial successes at Motorola, GE, and other early adopters of Six Sigma, many companies worldwide have now instituted Six Sigma to improve efficiency, cut costs, eliminate defects, and reduce product variation (see Arndt 2002; Cyger 2006; Hahn et al. 2000; Snee 2000). Six Sigma offers a more prescriptive and systematic approach to process improvement than TQM. It is also distinguished from other quality improvement systems by its clear focus on achieving bottom-line results in a relatively short 3- to 6-month period of time.

The name Six Sigma comes from the fact that it is a managerial approach designed to create processes that result in no more than 3.4 defects per million. The Six Sigma approach assumes that processes are designed so that the upper and lower specification limits are six standard deviations away from the mean. Then, if the processes are monitored correctly with ▶control charts, the worst possible scenario is for the mean to shift to within 4.5 standard deviations from the nearest specification limit. The area under the normal curve less than 4.5 standard deviations below the mean is approximately 3.4 out of a million.

The DMAIC Model

To guide managers in their task of improving short- and long-term results, Six Sigma uses a five-step process known as the *DMAIC model* – named for the five steps in the process:

- *Define*. The problem is defined, along with the costs, the benefits, and the impact on the customer.
- *Measure*. Operational definitions for each critical-to-quality (CTQ) variable are developed. In addition, the measurement procedure is verified so that it is consistent over repeated measurements.
- *Analyze*. The root causes of why defects occur are determined, and variables in the process causing the defects are identified. Data are collected to determine benchmark values for each process variable. This analysis often uses control charts.
- *Improve*. The importance of each process variable on the CTQ variable is studied using designed experiments. The objective is to determine the best level for each variable.
- *Control*. The objective is to maintain the benefits for the long term by avoiding potential problems that can occur when a process is changed.

The Define phase of a Six Sigma project consists of the development of a project charter, performing a SIPOC analysis, and identifying the customers for the output of the process. The development of a project charter involves forming a table of business objectives and indicators for all potential Six Sigma projects. Importance ratings are assigned by top management, projects are prioritized, and the most important project is selected. A SIPOC analysis is used to identify the *S*uppliers to the process, list the *I*nput provided to the suppliers, flowchart the *P*rocess, list the process *O*utputs, and identify the *C*ustomers of the process. This is followed by a Voice of the Customer analysis that involves market segmentation in which different types of users of the process are identified and the circumstances of their use of the process are identified. Statistical methods used in the Define phase include tables and charts, descriptive statistics, and control charts.

In the Measure phase of a Six Sigma project, members of a team first develop operational definitions of each CTQ variable. This is done so that everyone will have a firm understanding of the CTQ. Then studies are undertaken to ensure that there is a valid measurement system for the CTQ that is consistent across measurements. Finally, baseline data are collected to determine the capability and stability of the current process. Statistical methods used in the Measure phase include tables and charts, descriptive statistics, the normal distribution, the Analysis of Variance, and control charts.

The Analyze phase of a Six Sigma project focuses on the factors that affect the central tendency, variation, and shape of each CTQ variable. Factors are identified, related to each CTQ, have operational definitions developed, and have measurement systems established. Statistical methods used in the Analyze phase include tables and charts, descriptive statistics, the ▶Analysis of Variance, regression analysis, and control charts.

In the Improve phase of a Six Sigma project, team members carry out designed experiments to actively intervene in a process. The objective of the experimental design is to determine the settings of the factors that will optimize the central tendency, variation, and shape of each CTQ variable. Statistical methods used in the Improve phase include tables and charts, descriptive statistics, regression

analysis, hypothesis testing, the Analysis of Variance, and designed experiments.

The Control phase of a Six Sigma project focuses on the maintenance of improvements that have been made in the Improve phase. A risk abatement plan is developed to identify elements that can cause damage to a process. Statistical methods used in the Control phase include tables and charts, descriptive statistics, and control charts.

About the Author

Dr. David M. Levine is Professor Emeritus of Statistics and Computer Information Systems at Baruch College, City University of New York (CUNY). David has won a number of awards for his teaching including Teacher of the Year and the Baruch College Dean's Award for Continued Excellence in Teaching. He has also earned the Robert Pearson Award from the Northeast Region of the Decision Science Institute for his development of an innovative Statistical Process Control Course for Business Students and an Honorable Mention for the Decision Science Institute Innovation Teaching Award. He is a co author of the well known introductory books on statistics: *Basic Business Statistics* (with M.L. Berenson and T.C. Krehbiel, 11th edition, Prentice Hall, 2008) and *Statistics for Managers Using Microsoft Excel* (with D.F. Stephan, T.C. Krehbiel and M.L. Berenson, 5th edition, Prentice Hall, 2007). David's most recent specialty is Six Sigma and he is the author of well-received *Statistics for Six Sigma for Green Belts and Champions* and coauthor of *Quality Management*, and *Business Statistics for Quality and Productivity*.

Cross References

►Business Statistics
►Design for Six Sigma
►Industrial Statistics
►SIPOC and COPIS: Business Flow–Business Optimization Connection in a Six Sigma Context

References and Further Reading

Arndt M (2002) Quality isn't just for widgets. BusinessWeek July 22:72–73
Automotive Industry Action Group (AIAG) (1995) Statistical Process Control Reference Manual (Chrysler, Ford, and General Motors Quality and Supplier Assessment Staff
Bothe DR (1997) Measuring process capability. McGraw-Hill, New York
Cyger M (November/December 2006) The last word – riding the bandwagon. iSixSigma Magazine
Davis RB, Krehbiel TC (2002) Shewhart and zone control charts under linear trend. Commun Stat Simulat 31(1):91–96
Deming WE (1986) Out of the crisis. MIT Center for Advanced Engineering Study, Cambridge, MA
Deming WE (1993) The new economics for business, industry, and government. MIT Center for Advanced Engineering Study, Cambridge, MA
Gabor A (1990) The man who discovered quality. Time Books, New York
Gitlow H, Levine D (2005) Six sigma for green belts and champions. Financial Times/Prentice Hall, Upper Saddle River, NJ
Gitlow H, Levine D, Popovich E (2006) Design for six sigma for green belts and champions. Financial Times/Prentice Hall, Upper Saddle River, NJ
Hahn GJ, Doganaksoy N, Hoerl R (2000) The evolution of six sigma. Qual Eng 12:317–326
Lemak DL, Mero NP, Reed R (2002) When quality works: a premature post-mortem on TQM. J Bus Manage 8:391–407
Levine DM (2006) Statistics for six sigma for green belts with minitab and JMP. Financial Times/Prentice Hall, Upper Saddle River, NJ
Microsoft Excel 2007 (Redmond, WA: Microsoft Corp., 2007)
Scherkenbach WW (1987) The deming route to quality and productivity: road maps and roadblocks. CEEP, Washington, DC
Shewhart WA, Economic Control of the Quality of Manufactured Product (New York: Van Nostrand-Reinhard, 1931, reprinted by the American Society for Quality Control, Milwaukee, 1980)
Snee RD (2000) Impact of six sigma on quality. Qual Eng 12:ix–xiv
Vardeman SB, Jobe JM (2009) Statistical methods for quality assurance: basics, measurement, control, capability and improvement. Springer, New York
Walton M (1986) The Deming management method. Perigee Books, New York

Skewness

Paul von Hippel
Assistant Professor
University of Texas, Austin, TX, USA

Skewness is a measure of distributional asymmetry. Conceptually, skewness describes which side of a distribution has a longer tail. If the long tail is on the right, then the skewness is rightward or positive; if the long tail is on the left, then the skewness is leftward or negative. Right skewness is common when a variable is bounded on the left but unbounded on the right. For example, durations (response time, time to failure) typically have right skewness since they cannot take values less than zero; many financial variables (income, wealth, prices) typically have right skewness since they rarely take values less than zero; and adult body weight has right skewness since most people are closer to the lower limit than to the upper limit of viable body weight. Left skewness is less common in practice, but it can occur when a variable tends to be closer to its maximum than its minimum value. For example, scores on an easy

exam are likely to have left skewness, with most scores close to 100% and lower scores tailing off to the left. Well-known right-skewed distributions include the Poisson, chi-square, exponential, lognormal, and gamma distributions. I am not aware of any widely used distributions that always have left skewness, but there are several distributions that can have either right or left skew depending on their parameters. Such ambidextrous distributions include the binomial and the beta.

Mathematically, skewness is usually measured by the third standardized moment $E((X - \mu)/\sigma)^3)$, where X is a random variable with mean μ and standard deviation σ. The third standardized moment can take any positive or negative value, although in practical settings it rarely exceeds 2 or 3 in absolute value. Because it involves cubed values, the third standardized moment is sensitive to ►outliers (Kim and White 2004), and it can even be undefined for heavy-tailed distributions such as the Cauchy density or the Pareto density with a shape parameter of 3. When the third standardized moment is finite, it is zero for symmetric distributions, although a value of zero does not necessarily mean that the distribution is symmetric (Ord 1968; Johnson and Kotz 1970, p. 253). To estimate the third standardized moment from a sample of n observations, a biased but simple estimator is the third sample moment $1/n \sum((x - \bar{x})/s)^3$, where $\bar{x}$ is the sample mean and s is the sample standard deviation. An unbiased estimator is the third k statistic, which is obtained by taking the third sample moment and replacing $1/n$ with the quantity $n/((n-1)(n-2))$ (Rose and Smith 2002).

Although the third standardized moment is far and away the most popular definition of skewness, alternative definitions have been proposed (MacGillivray 1986). The leading alternatives are bounded by –1 and +1, and are zero for symmetric distributions, although again a value of zero does not guarantee symmetry. One alternative is Bowley's (1920) quartile formula for skew: $((q_3 - m) - (m - q_1))/(q_3 - q_1)$, or more simply $(q_1 + q_3 - 2m)/(q_3 - q_1)$, where m is the median and q_1 and q_3 are the first (or left) and third (or right) quartiles. Bowley's skew focuses on the part of the distribution that fits in between the quartiles: if the right quartile is further from the median than is the left quartile, then Bowley's skew is positive; if the left quartile is further from the median than the right quartile, then Bowley's skew is negative. Because it doesn't cube any values and doesn't use any values more extreme than the quartiles, Bowley's skew is more robust to outliers than is the conventional third-moment formula (Kim and White 2004). But the quantities in Bowley's formula are arbitrary: instead of the left and right quartiles – i.e., the 25th and 75th percentiles – Bowley could just as plausibly have used the 20th and 80th percentiles, the 10th and 90th percentiles, or more generally the 100pth and 100$(1 - p)$th percentiles $F^{-1}(p)$ and $F^{-1}(1 - p)$. Substituting these last two expressions into Bowley's formula, Hinkley (1975) proposed the generalized skewness formula $(F^{-1}(1-p) + F^{-1}(p) - 2m)/(F^{-1}(1-p) - F^{-1}(p))$, which is a function of high and low percentiles defined by p. Since it is not clear what value of p is most appropriate, Groeneveld and Meeden (1984) averaged Hinkley's formula across all ps from 0 to 0.5. Groeneveld and Meeden's average was $(\mu - m)/E|X - m|$, which is similar to an old skewness formula that is attributed to Pearson: $(\mu - m)/\sigma$ (Yule 1911).

The Pearson and Groeneveld–Meeden formulas are consistent with a widely taught rule of thumb claiming that the skewness determines the relative positions of the median and mean. According to this rule, in a distribution with positive skew the mean lies to the right of the median, and in a distribution with negative skew the mean lies to the left of the median. If we define skewness using the Pearson or Groeneveld–Meeden formulas, this rule is self-evident: since the numerator of both formulas is simply the difference between the mean and the median, both will give positive skew when the mean is greater than the median, and negative skew when the situation is reversed. But if we define skewness more conventionally, using the third standardized moment, the rule of thumb can fail. Violations of the rule are rare for continuous variables, but common for discrete variables (von Hippel 2005). A simple discrete violation is the ►binomial distribution with $n = 10$ and $\pi = 0.09$ (cf. Lesser 2005). In this distribution, the mean 0.9 is left of the median 1, but the skewness as defined by the third standardized moment is positive, at 0.906, and the distribution, with its long right tail, looks like a textbook example of positive skew. Examples like this one argue against using the Pearson, Groeneveld-Meeden, or Bowley formulas, all of which yield a negative value for this clearly right-skewed distribution. Most versions of Hinkley's skew also contradict intuition here: Hinkley's skew is negative for $0.5 > p > 0.225$, zero for $0.225 \geq p > 0.054$, and doesn't become positive until $p \leq 0.054$.

Since many statistical inferences assume that variables are symmetrically or even normally distributed, those inferences can be inaccurate if applied to a variable that is skewed. Inferences grow more accurate as the sample size grows, with the required sample size depending on the amount of skew and the desired level of accuracy. A useful rule states that, if you are using the normal or t distribution to calculate a nominal 95% confidence interval for the mean of a skewed variable, the interval will have at least 94% coverage if the sample size is at least 25 times the absolute value of the (third-moment) skew (Cochran

1977; Boos and Hughes-Oliver 2000). For example, a sample of 50 observations should be plenty even if the skew is as large as 2 (or −2).

In order to use statistical techniques that assume symmetry, researchers sometimes transform a variable to reduce its skew (von Hippel 2003). The most common transformations for reducing positive skew are the logarithm and the square root, and a much broader family of skew-reducing transformations has been defined (Box and Cox 1964). But reducing skew has costs as well as benefits. A transformed variable can be hard to interpret, and conclusions about the transformed variable may not apply to the original variable before transformation (Levin et al. 1996). In addition, transformation can change the shape of relationships among variables; for example, if X is right-skewed and has a linear relationship with Y, then the square root of X, although less skewed, will have a curved relationship with Y (von Hippel 2010). In short, skew reduction is rarely by itself a sufficient reason to transform a variable. Skew should be treated as an important characteristic of the variable, not just a nuisance to be eliminated.

Cross References

►Box–Cox Transformation
►Heavy-Tailed Distributions
►Mean Median and Mode
►Mean, Median, Mode: An Introduction
►Normality Tests
►Omnibus Test for Departures from Normality

References and Further Reading

Boos DD, Hughes-Oliver JM (2000) How large does n have to be for Z and t intervals? Am Stat 54(2):121–128
Bowley AL (1920) Elements of statistics. Scribner, New York
Box GEP, Cox D (1964) An analysis of transformations. J R Stat Soc B 26(2):211–252
Cochran WG (1977) Sampling techniques. Wiley, New York
Groeneveld RA (1986) Skewness for the Weibull family. Stat Neerl 40:135–140
Groeneveld RA, Meeden G (1984) Measuring skewness and kurtosis. Stat 33:391–399
Hinkley DV (1975) On power transformations to symmetry. Biometrika 62:101–111
Johnson NL, Kotz S (1970) Continuous univariate distributions 1. Houghton Mifflin, Boston
Kim TH, White H (2004) On more robust estimation of skewness and kurtosis. Finance Res Lett 1(1):56–73
Lesser LM (2005) Letter to the editor [comment on von Hippel (2005)]. J Stat Educ 13(2) http://www.amstat.org/publications/jse/v13n3/lesser_letter.html
Levin A, Liukkonen J, Levine DW (1996) Equivalent inference using transformations. Commun Stat Theor Meth 25(5):1059–1072
MacGillivray HL (1986) Skewness and asymmetry: measures and orderings. Ann Stat 14(3):994–1011
Ord JK (1968) The discrete student's t distribution. Ann Math Stat 39:1513–1516
Rose C, Smith M (2002) Mathematical statistics with mathematica. Springer, New York
Sato M (1997) Some remarks on the mean, median, mode and skewness. Aust J Stat 39(2):219–224
von Hippel PT (2003) Normalization. In: Lewis-Beck M, Bryman A, Liao TF (eds) Encyclopedia of social science research methods. Sage, Thousand Oaks
von Hippel PT (2005) Mean, median, and skew: correcting a textbook rule. J Stat Edu 13 (2) www.amstat.org/publications/jse/v13n2/vonhippel.html
von Hippel PT (2010) How to impute skewed variables under a normal model. Unpublished manuscript, under review
Yule GU (1911) Introduction to the theory of statistics. Griffith, London

Skew-Normal Distribution

Adelchi Azzalini
Professor of Statistics
University of Padua, Padua, Italy

In its simplest reading, the term "skew-normal" refers to a family of continuous probability distributions on the real line having density function of form

$$\varphi(z;\alpha) = 2\,\varphi(z)\,\Phi(\alpha z), \qquad (-\infty < z < \infty), \qquad (1)$$

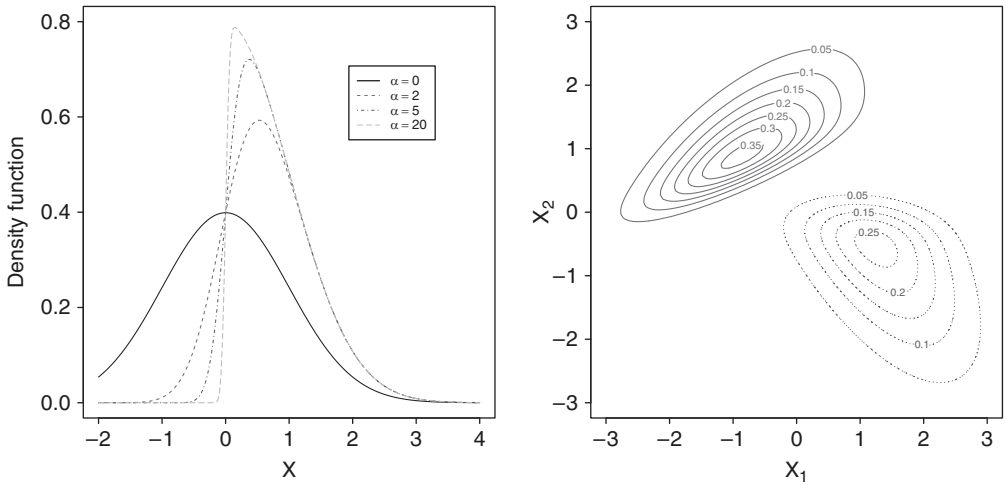
where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the $\mathrm{N}(0,1)$ density and cumulative distribution function, respectively, and α is a real parameter which regulates the shape of the density. The fact that (1) integrates to 1 holds by a more general result, given by Azzalini (1985), where φ and Φ are replaced by analogous functions for any choice of two distributions symmetric around 0.

It is immediate that the choice $\alpha = 0$ lends the $\mathrm{N}(0,1)$ distribution, and that, if Z is a random variable with density (1), denoted $Z \sim SN(\alpha)$, then $-Z \sim SN(-\alpha)$. Figure 1a displays $\varphi(z;\alpha)$ for a few choices of α; only positive values of this parameter are considered, because of the property just stated.

An interesting property is that $Z^2 \sim \chi^2_1$, if $Z \sim SN(\alpha)$, irrespectively of α. The ►moment generating function of Z is

$$M(t) = 2\,\exp(t^2/2)\,\Phi(\delta t), \quad \delta = \alpha/\sqrt{1+\alpha^2}, \qquad (2)$$

and from $M(t)$ it is simple to obtain the mean, the variance, the index of skewness and the index of kurtosis,



Skew-Normal Distribution. Fig. 1 Some examples of skew-normal density function, for the scalar case (*left*) and for the bivariate case in the form of contour level plots (*right*)

which are

$$\mu_\alpha = \sqrt{\frac{2}{\pi}}\,\delta, \quad \sigma_\alpha^2 = 1 - \mu_\alpha^2,$$
$$\gamma_1 = \frac{4-\pi}{2}\frac{\mu_\alpha^3}{\sigma_\alpha^3}, \quad \gamma_2 = 2(\pi-3)\,\frac{\mu_\alpha^4}{\sigma_\alpha^4} \tag{3}$$

respectively. Multiplication of $M(t)$ by $\exp(t^2/2)$ shows another interesting property: if $U \sim \mathrm{N}(0,1)$ independent of Z, then $(Z+U)/\sqrt{2} \sim SN\left(\alpha/\sqrt{2+\alpha^2}\right)$. Additional facts about this distribution are given by Azzalini; Azzalini (1985; 1986), Henze (1986) and Chiogna (1998).

For practical statistical work, we need to consider the three-parameter distribution of $Y = \xi + \omega Z$, where ξ and ω are a location and a scale parameter, respectively ($\omega > 0$). Extension of the above results to the distribution of Y is immediate.

For the d-dimensional version of (1) we introduce directly a location parameter $\xi \in \mathbb{R}^d$ and a scale $d \times d$ matrix Ω which is symmetric and positive definite, and we denote by ω a $d \times d$ diagonal matrix formed by the square roots of the diagonal elements of Ω. The density function of the multivariate skew-normal distribution at x is

$$2\,\varphi_d(x-\xi;\Omega)\,\Phi\left(\alpha^\top \omega^{-1}(x-\xi)\right), \quad (x \in \mathbb{R}^d), \tag{4}$$

where $\varphi_d(x;\Omega)$ denotes the $\mathrm{N}_d(0,\Omega)$ density function, and the shape parameter α is a vector in $\mathbb{R}^d$. Figure 1b displays function 4 for two choices of the parameter set (ξ, Ω, α). Initial results on this distribution have been obtained by Azzalini and Dalla Valle (1996) and by Azzalini and Capitanio (1999).

The multivariate skew-normal distribution enjoys a number of formal properties. If Y is a d-dimensional random variable with density (4), its moment generating function is

$$M(t) = \exp\left(\xi^\top t + \frac{1}{2}t^\top \Omega t\right)\Phi(\delta^\top \omega t),$$
$$\delta = \frac{1}{(1+\alpha^\top \bar{\Omega}\alpha)^{1/2}}\bar{\Omega}\alpha \tag{5}$$

where $\bar{\Omega} = \omega^{-1}\Omega\omega^{-1}$ is the correlation matrix associated to Ω. From $M(t)$ one obtains that

$$\mathbb{E}\{Y\} = \xi + \sqrt{\frac{2}{\pi}}\,\omega\delta, \quad var\{Y\} = \Omega - \frac{2}{\pi}\omega\delta\delta^\top\omega,$$

while the marginal indices of skewness and kurtosis are computed by applying expressions γ_1 and γ_2 in (3) to each component of δ. Another result derived from (5) is that an affine transformation $a + A\,Y$, where $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times d}$, is still of type (4), with suitably modified dimension and parameters. This fact implies closure of this family of distributions with respect to marginalization. Closure of the class under conditioning holds if one extends the class by inserting an additional parameter in the argument of Φ in (4), and adapting the normalizing constant correspondingly; for details on this extended class, see Arnold and Beaver (2000) and Capitanio et al. (2003).

The *chi-square distribution* property stated for the scalar case extends substantially in the multivariate case. If Y has density (4) with $\xi = 0$, then a quadratic form $Y^\top AY$, where A is a symmetric $d \times d$ matrix, has the same distribution of $X^\top AX$ where $X \sim \mathrm{N}_d(0,\Omega)$; for instance

$Y^\top \Omega^{-1} Y \sim \chi^2_d$. This distributional result can be obtained from first principles, but it is mosty simply derived as a special case of the distributional invariance property of the family of *skew-symmetric distributions*, of which the skew-normal distribution is a special instance. According to this property, the distribution of $T(Y)$ is the same of $T(X)$ for any function T, possibly multi-valued, such that $T(x) = T(-x)$ for all $x \in \mathbb{R}^d$.

An attractive feature of this distribution is that it admits various *stochastics representations*, which are relevant for random number generation and also for supporting the adoption of this distribution in statistical modelling work. Here we restrict ourselves to one of these representations, which is related to a *selective sampling* mechanism: if

$$\begin{pmatrix} X_0 \\ X \end{pmatrix} \sim N_{1+d}(0, \Omega^*), \quad \Omega^* = \begin{pmatrix} 1 & \delta^\top \omega \\ \omega\delta & \Omega \end{pmatrix} > 0,$$

where X_0 and X have dimension 1 and d, respectively, then

$$Y = \xi + \begin{cases} X & \text{if} \quad X_0 > 0, \\ -X & \text{otherwise} \end{cases}$$

has density function (4) where $\alpha = (1 - \delta^\top \bar{\Omega}^{-1} \delta)^{-1/2} \bar{\Omega}^{-1} \delta$.

Additional information on the skew-normal distribution and related areas is presented in the review paper of Azzalini (2005), followed by a set of comments of Marc Genton, and rejoinder of the author. Themes considered include: additional properties and types of stochastic representation, aspects of statistical inference, historical development, extensions to skew-ellitical and skew-symmetric type of distributions, and connections with various application areas.

About the Author

Professor Azzalini is an elected fellow of the International Statistical Institute, and a member of various scientific societies. As a member of the "Bernoulli Society", he served as a member of the Council of the Society (1991–94) and as a chairman of the European Regional Committee (2006–2008). He also served as an associate editor for some scholarly journals (*Applied Statisics*, *Scandinavian J. Statistics*, *Metron*). Currently he is on the Advisory Board of *Metron*.

Editor's note: Professor Azzalini was the first to thoroughly set the foundations and provided systematic treatment of skew-normal distribution (in 1985 and 1986) and introduced the multivariate skew-normal distribution, with Dalla Valle, in 1996.

Cross References

►Chi-Square Distribution
►Normal Distribution, Univariate
►Skewness
►Skew-Symmetric Families of Distributions

References and Further Reading

Arnold BC, Beaver RJ (2000) Hidden truncation models. Sankhyā A 62(1):22–35
Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178
Azzalini A (1986) Further results on a class of distributions which includes the normal ones. Statistica 46(2):199–208
Azzalini A (2005) The skew-normal distribution and related multivariate families (with discussion) Scand J Stat 32:159–188 (C/R 189–200)
Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat SocB 61(3):579–602 Full version of the paper at http://arXiv.org (No. 0911.2093)
Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726
Capitanio A, Azzalini A, Stanghellini E (2003) Graphical models for skew-normal variates. Scand J Statist 30:129–144
Chiogna M (1998) Some results on the scalar skew-normal distribution. J Ital Stat Soc 7:1–13
Henze N (1986) A probabilistic representation of the 'skew-normal' distribution. Scand J Stat 13:271–275

Skew-Symmetric Families of Distributions

Adelchi Azzalini
Professor of Statistics
University of Padua, Padua, Italy

The term 'skew-symmetric distributions' refers to the construction of a continuous probability distribution obtained by applying a certain form of perturbation to a symmetric density function.

To be more specific, a concept of *symmetric distribution* must be adopted first, since in the multivariate setting various forms of symmetry have been introduced. The variant used in this context is the one of *central symmetry*, a natural extension of the traditional one-dimensional form to the d-dimensional case: if f_0 is a density function on $\mathbb{R}^d$ and ξ is a point of $\mathbb{R}^d$, central symmetry around ξ requires that $f_0(t - \xi) = f_0(-t - \xi)$ for all $t \in \mathbb{R}^d$, ignoring sets of 0 probability. To avoid notational complications, we shall concentrate on the case with $\xi = 0$; it is immediate to rephrase what follows in the case of general ξ, which simply amounts to a shift of the location of the distribution.

If f_0 is a probability density function on $\mathbb{R}^d$ centrally symmetric around 0, there are two largely equivalent expressions to build skew-symmetric densities. For the first one, introduce a one-dimensional continuous distribution function G such that $G(-x) = 1 - G(x)$ for all $x \in \mathbb{R}$, and $w(\cdot)$ a real-valued function on $\mathbb{R}^d$ such that $w(-t) = -w(t)$ for all $t \in \mathbb{R}^d$. Then it can be shown that

$$f(t) = 2 f_0(t)\, G\{w(t)\} \tag{1}$$

is a density function on $\mathbb{R}^d$. Notice that in general $G\{w(t)\}$ is not a probability distribution. In the second type of formulation, consider a function $\pi(t)$ such that $0 \le \pi(t) \le 1$ and $\pi(t) + \pi(-t) = 1$ for all $t \in \mathbb{R}^d$, which leads to the density function

$$f(t) = 2 f_0(t)\, \pi(t)\,. \tag{2}$$

Formulations (1) and (2) have been obtained independently by Azzalini and Capitanio (2003) and by Wang et al. (2004), who adopted the term 'skew-symmetric distribution'. Each of the two forms has its advantages. Any expression of type $G\{w(t)\}$ in (1) automatically satisfies the requirements for $\pi(t)$ in (2), but it is not unique: there are several forms $G\{w(t)\}$ corresponding to the same $\pi(t)$. On the other hand, any $\pi(t)$ can be written in the form $G\{w(t)\}$. Hence the two sets of distributions coincide.

The proof that (1) and (2) are proper density functions is exceptionally simple. The argument below refers to (1) in the univariate case; the multivariate case is essentially the same with only a minor technical complication. If Y is a random variable with density function f_0 and X is an independent variable with distribution function G, then $w(Y)$ is symmetrically distributed around 0 and

$$\frac{1}{2} = \mathbb{P}\{X - w(Y) \le 0\} = \mathbb{E}_Y\{\mathbb{P}\{X \le w(y) | Y = y\}\}$$
$$= \int_{\mathbb{R}^d} G\{w(y)\} f_0(y) \mathrm{d}y\,.$$

This proof also shows the intimate connection of this formulation with a selective sampling mechanism where a value Y sampled from f_0 is retained with probability $G\{w(t)\}$, and it is otherwise rejected. A refinement of this scheme says that

$$Z = \begin{cases} Y & \text{if } X \le w(Y), \\ -Y & \text{otherwise} \end{cases} \tag{3}$$

has density (1). Since (3) avoids rejection of samples, it is well suited for random numbers generation.

In spite of their name, skew-symmetric distributions are not *per se* linked to any idea of ▶skewness. The name is due to the historical connection with the ▶*skew-normal distribution*, which has been the first construction of this type. The skew-normal density function is

$$2\, \varphi_d(y; \Omega)\, \Phi(\eta^\top y), \qquad (y \in \mathbb{R}^d), \tag{4}$$

where $\varphi_d(y; \Omega)$ is the $\mathrm{N}_d(0, \Omega)$ density function, Φ is the $\mathrm{N}(0,1)$ distribution function and η is a vector parameter. This density is of type (1) with $f_0(y) = \varphi_d(y; \Omega)$ and $G\{w(y)\} = \Phi(\eta^\top y)$. In this case the perturbation of the original density φ_d does indeed lead to an asymmetric density, as it typically occurs when $w(y)$ is a linear function.

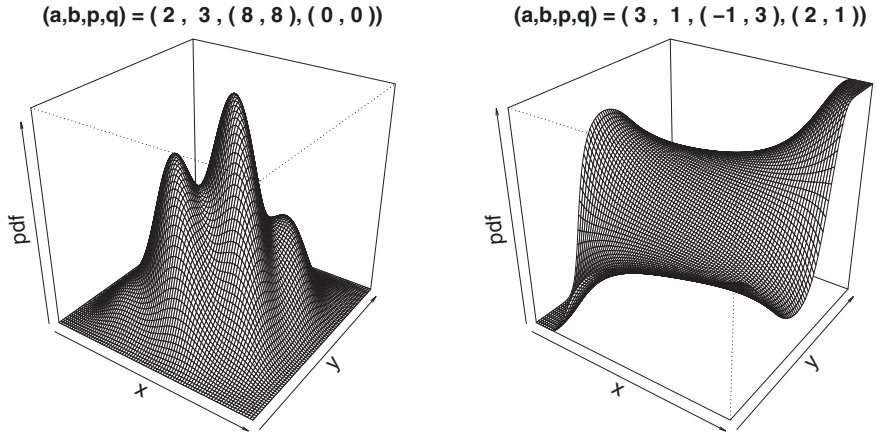
To illustrate visually the flexibility which can be achieved by the perturbation mechanism, consider f_0 to be the product of two symmetric Beta densities of parameters (a, a) and (b, b), say, both shifted and scaled to the interval $(-1, 1)$, G equal to the standard logistic distribution function and

$$w(y) = \frac{\sin(p_1 y_1 + p_2 y_2)}{1 + \cos(q_1 y_1 + q_2 y_2)}, \qquad y = (y_1, y_2)^\top \in (-1, 1)^2$$

where $p = (p_1, p_2)$ and $q = (q_1, q_2)$ are additional parameters. Figure 1 displays a few of the shapes produced with various choices of the parameters a, b, p, q. These skew-symmetric densities do not exhibit any obvious sign of skewness.

An important implication of representation (3) is the following property of distributional invariance: if Y has density f_0 and Z has density (1), then $T(Z)$ and $T(Y)$ have the same distribution for any function $T(\cdot)$ from $\mathbb{R}^d$ to $\mathbb{R}^q$ which is even, in the sense that $T(z) = T(-z)$ for all $z \in \mathbb{R}^d$. For instance, if Z has skew-normal distribution (4), then a quadratic form $T(Z) = Z^\top A Z$ has the same distribution of $T(Y) = Y^\top A Y$ when $Y \sim N_d(0, \Omega)$, for any symmetric matrix A; a further specialization says that $Z^\top \Omega^{-1} Z \sim \chi^2_d$. Other results on skew-elliptical distributions have been given by Arellano-Valle et al. (2006) and Umbach (2008).

An important subset of the skew-symmetric distributions occurs if f_0 in (1) or (2) is an *elliptically contoured density*, or briefly an *elliptical density*, in which case we obtain a skew-elliptical distribution. In fact, this subset was the first considered, in chronological order, starting from the skew-normal distribution, and the formulation evolved via a sequence of successive generalizations. This development is visible in the following sequence of papers, to be complemented with those already quoted: Azzalini and Capitanio (1999), Branco and Dey (2001), Genton and



Skew-Symmetric Families of Distributions. Fig. 1 Densities obtained by perturbation of the product of two symmetric Beta densities for some choices of the parameters a,b,p,q

Loperfido (2005) and the collection of papers in the book edited by Genton (2004).

About the Author

For biography *see* the entry ►Skew-Normal Distribution.

Cross References

►Beta Distribution
►Logistic Distribution
►Skewness
►Skew-Normal Distribution

References and Further Reading

Arellano-Valle RB, Branco MD, Genton MG (2006) A unified view on skewed distributions arising from selections. Canad J Stat 34:581–601

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc B 61(3):579–602. Full version of the paper at http://arXiv.org (No. 0911.2093)

Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution. J R Stat Soc B 65(2):367–389. Full version of the paper at http://arXiv.org (No. 0911.2342)

Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. J Multivariate Anal 79(1):99–113

Genton MG (ed) (2004) Skew-elliptical distributions and their applications: a journey beyond normality. Chapman & Hall/CRC Press, Boca Raton, FL

Genton MG, Loperfido N (2005) Generalized skew-elliptical distributions and their quadratic forms. Ann Inst Stat Math 57: 389–401

Umbach D (2008) Some moment relationships for multivariate skew-symmetric distributions. Stat Probab Lett 78(12): 1619–1623

Wang J, Boyer J, Genton MG (2004) A skew-symmetric representation of multivariate distributions. Stat Sinica 14:1259–1270

Small Area Estimation

Danny Pfeffermann
Professor Emeritus
Hebrew University of Jerusalem, Jerusalem, Israel
Professor
University of Southampton, Southampton, UK

Introduction

Over the past three decades there is a growing demand in many countries for reliable estimates of small domain parameters such as means, counts, proportions or quantiles. Common examples include the estimation of unemployment rates, proportions of people under poverty, disease incidence and use of illicit drugs. These estimates are used for fund allocations, new social or health programs, and more generally, for short and long term planning. Recently, small area estimates are employed for testing, the administrative records used for modern censuses (see ►Census). Although commonly known as "small area estimation" (SAE), the domain of studies may actually consist of socio-demographic subgroups as defined, for example, by gender, age and race, or the intersection of such domains with geographical location.

The problem of SAE is that the sample sizes in at least some of the domains of study are very small, and often there are no samples available for many or even most of these domains. As a result, the direct estimates obtained from the survey are unreliable (large, unacceptable variances), and no direct survey estimates can be computed for areas with no samples. SAE methodology addresses therefore the following two major problems:

1. How to obtain reliable estimates for each of these areas.
2. How to assess the error of the estimators (MSE, confidence intervals, etc.).

Notice in this regard that even if direct survey estimates can be used for areas with samples, no design-based methodology exists for estimating the quantities of interest in areas with no samples. The term "Design-based inference" refers to inference based on the randomization distribution over all the samples possibly selected from the finite population under study, with the population values considered as fixed numbers. Note also that the sample sizes in the various areas are random, unless when some of the domains of study are defined as strata and samples of fixed sizes are taken in these domains.

In what follows I describe briefly some of the basic methods used for SAE, assuming, for simplicity, that the sample is selected by simple random sampling. More advanced methods and related theory, with many examples and references can be found in the book of Rao (2003) and the review papers by Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002), and Rao (2005). See also Chaps. 31 and 32 in the new Handbook of Statistics, 29B (eds. Pfeffermann and Rao 2009).

Design-Based Methods

Let Y define the characteristic of interest and denote by y_{ij} the outcome value for unit j belonging to area i, $i = 1,\ldots,M$; $j = 1,\ldots,N_i$, where N_i is the area size. Let $s = s_1 \cup \cdots \cup s_m$ denote the sample, where s_i of size n_i is the sample observed for area i. Suppose that it is required to estimate the true area mean $\overline{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$. If no auxiliary information is available, the *direct* design unbiased estimator and its design variance over the *randomization distribution* (the distribution induced by the random selection of the sample with the population values held fixed), are given by

$$\hat{\overline{Y}}_i = \sum_{j=1}^{n_i} y_{ij}/n_i; \quad Var_D\left[\hat{\overline{Y}}_i | n_i\right] = \left(S_i^2/n_i\right)\left[1-(n_i/N_i)\right] = S_i^{*2}, \tag{1}$$

where $S_i^2 = \sum_{k=1}^{N_i}(y_{ik}-\overline{Y}_i)^2/(N_i-1)$. Clearly, for small n_i the variance will be large, unless the variability of the *y-values* is sufficiently small. Suppose, however, that values x_{ij} of p concomitant variables $\mathbf{x}_1,\ldots,\mathbf{x}_p$ are measured for each of the sample units and that the area means $\overline{X}_i = \sum_{k=1}^{N_i} x_{ik}/N_i$ are likewise known. Such information may be obtained from a recent census or some other administrative records. In this case, a more efficient design-based estimator is the regression estimator,

$$\hat{\overline{Y}}_{i,reg} = \overline{y}_i + (\overline{X}_i - \overline{x}_i)'\beta_i; \quad Var_D(\overline{y}_{reg,i}|n_i) = S_i^{*2}\left(1-R_i^2\right), \tag{2}$$

where $\overline{y}_i$ and $\overline{x}_i$ are the sample means of Y and X in area i, and β_i and R_i are correspondingly the vector of regression coefficients and the multiple correlation coefficient between Y and $\mathbf{x}_1,\ldots,\mathbf{x}_p$ in area i. Thus, by use of the concomitant variables, the variance is reduced by the factor $\left(1-R_i^2\right)$, illustrating the importance of using auxiliary information with good prediction power for SAE.

In practice, the coefficients β_i are unknown. Replacing β_i by its ordinary least square estimator from the sample s_i may not be effective in the case of a small sample size. If, however, the regression relationships are "similar" across the areas and assuming $x_{ij,1} = 1$ for all (i,j), a more stable estimator is the *synthetic regression* estimator,

$$\hat{\overline{Y}}_{i,syn} = \sum_{j=1}^{N_i} \hat{y}_{ik}/N_i = \overline{X}_i'\hat{B}, \tag{3}$$

where $\hat{y}_{ik} = x_{ik}'\hat{B}$ and $\hat{B} = \left[\sum_{i,j\in S} x_{ij}x_{ij}'\right]^{-1}\sum_{i,j\in S} x_{ij}y_{ij}$ is the ordinary least squares estimator computed from all the sample data. The prominent advantage of synthetic estimation is the substantial variance reduction since the estimator uses all the sample data, but it can lead to severe biases if the regression relationships differ between the areas.

An approximately design-unbiased estimator is obtained by replacing the synthetic estimator by the GREC estimator,

$$\hat{\overline{Y}}_{i,greg} = \sum_{k=1}^{N_i} \hat{y}_{ik}/N_i + \sum_{j\in S_i}(y_{ij}-\hat{y}_{ij})/n_i. \tag{4}$$

However, this estimator may again be very unstable in small samples. The choice between the synthetic estimator and the GREG is therefore a trade off between bias and variance. A compromise is achieved by using a composite estimator of the form,

$$\hat{\overline{Y}}_{i,com} = \alpha_i\hat{\overline{Y}}_{i,greg} + (1-\alpha_i)\hat{\overline{Y}}_{i,syn}, \tag{5}$$

but there is no principled theory of how to determine the coefficients α_i.

Design-based estimators are basically model free but the requirement for approximate design-unbiasedness generally yields estimators with large variance due to the small sample sizes. The construction of confidence intervals requires large sample normality assumptions, which do not generally hold in SAE problems. No design-based theory exists for estimation in areas with no samples.

Model-Dependent Estimators

In view of the problems underlying the use of design-based methods, it is common practice in many applications to use instead statistical models that define how to "borrow strength" from other areas and/or over time in case of repeated surveys. Let θ_i define the parameter of interest in area i, $i = 1, \ldots, M$, and let y_i, x_i denote the data observed for this area. When the only available information is at the area level, y_i is typically the direct estimator of θ_i and $\mathbf{x}_i$ is a vector of area level covariates. When unit level information is available, y_i is a vector of individual outcomes and $\mathbf{x}_i$ is the corresponding matrix of individual covariate information.

A typical small area model consists of two parts: The first part models the distribution of $y_i|\theta_i; \Psi_{(1)}$. The second part models the distribution of $\theta_i|x_i; \Psi_{(2)}$ linking θ_i to the parameters in other areas and to the covariates. The (vector) parameters $\Psi_{(1)}$ and $\Psi_{(2)}$ are typically unknown and are estimated from all the available data $D(s) = \{y_i, x_i; 1, \ldots, m\}$. In what follows I define and discuss briefly three models in common use.

"Unit Level Random Effects Model"

The model, employed originally by Battese et al. (1988), assumes,

$$y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij}, \tag{6}$$

where u_i and ε_{ij} are mutually independent error terms with zero means and variances σ_u^2 and σ_ε^2 respectively. The random term u_i represents the joint effect of area characteristics not accounted for by the concomitant variables. Under the model, the true small area means are $\overline{Y}_i = \overline{X}'_i\beta + u_i + \overline{\varepsilon}_i$, but since $\overline{\varepsilon}_i = \sum_{k=1}^{N_i}\varepsilon_{ik}/N_i \cong 0$ for large N_i, the target parameters are often defined as $\theta_i = \overline{X}'_i\beta + u_i$. For known variances $\left(\sigma_u^2, \sigma_\varepsilon^2\right)$, the Best Linear Unbiased Predictor (BLUP) of θ_i is,

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\overline{X}_i - \bar{x}_i)'\hat{\beta}_{GLS}] + (1-\gamma_i)\overline{X}'_i\hat{\beta}_{GLS}, \tag{7}$$

where $\hat{\beta}_{GLS}$ is the generalized least square (GLS) estimator of β computed from all the observed data and $\gamma_i = \sigma_u^2/\left(\sigma_u^2 + \sigma_\varepsilon^2/n_i\right)$. For areas l with no samples, $\hat{\theta}_l = \overline{X}'_l\hat{\beta}_{GLS}$. Notice that unlike under the randomization distribution, the synthetic estimator $\overline{X}'_i\hat{\beta}_{GLS}$ is unbiased for θ_i under the model in the sense that $E(\overline{X}'_i\hat{\beta}_{GLS} - \theta_i) = 0$.

The BLUP $\hat{\theta}_i$ is also the Bayesian predictor (posterior mean) under normality of the error terms and a diffuse prior for β. In practice, however, the variances σ_u^2 and σ_ε^2 are seldom known. A Bayesian solution to this problem is to set prior distributions for the unknown variances and then compute the corresponding posterior mean and variance of $\theta_i|\{y_k, x_k; k \in s\}$ by aid of Markov Chain Monte Carlo (MCMC) simulations (see ►Markov Chain Monte Carlo). The common procedure under the frequentist approach is to replace the unknown variances in the BLUP formula by standard variance components estimates like Maximum Likelihood Estimators (MLE), Restricted MLE (REML) or Analysis of Variance (ANOVA) estimators. The resulting predictors are known as the Empirical BLUP (EBLUP). See the references listed in the introduction for estimation of the MSE of the EBLUP under different methods of variance estimation.

"Area Level Random Effects Model"

This model is in broad use when the concomitant information is only at the area level. It was used originally by Fay and Herriot (1979) for predicting the mean income in geographical areas of less than 500 inhabitants. Denote by $\tilde{\theta}_i$ the direct sample estimator of θ_i. The model assumes that,

$$\tilde{\theta}_i = \theta_i + e_i; \quad \theta_i = x'_i\beta + u_i, \tag{8}$$

such that e_i represents the sampling error, assumed to have zero mean and known design variance $Var_D(e_i) = \sigma_{Di}^2$, $(= S_i^{*2}$ if $\tilde{\theta}_i = \bar{y}_i$, see Eq. 1). The model integrates therefore a model dependent random effect u_i and a sampling error e_i with the two errors being independent. The BLUP under this model is,

$$\hat{\theta}_i = \gamma_i\tilde{\theta}_i + (1-\gamma_i)x'_i\hat{\beta}_{GLS} = x'_i\hat{\beta}_{GLS} + \gamma_i\left(\tilde{\theta}_i - x'_i\hat{\beta}_{GLS}\right), \tag{9}$$

which again is a composite estimator with coefficient $\gamma_i = \sigma_u^2/\left(\sigma_{Di}^2 + \sigma_u^2\right)$. As with the unit level model, the variance σ_u^2 is usually unknown and is either assigned a prior distribution under the Bayesian approach, or is replaced by a sample estimate in (9), yielding the corresponding EBLUP predictor.

Unit Level Random Effects Model for Binary Data

The previous two models are for continuous measurements. Suppose now that y_{ij} is a binary variable taking the values 0 or 1. For example, $y_{ij} = 1$ if individual j in area i is unemployed (or suffers from a certain disease), and $y_{ij} = 0$ otherwise, such that $p_i = N_i^{-1}\sum_{k=1}^{N_i} y_{ik}$ is the true unemployment rate (true disease incidence). The following model is often used for predicting the proportions p_i:

$$\begin{aligned} y_{ij}|p_{ij} &\overset{indep.}{\sim} Bernoulli(p_{ij}) \\ \text{logit}(p_{ij}) &= \log[p_{ij}/(1-p_{ij})] = x'_{ij}\beta + u_i; \\ u_i &\overset{indep.}{\sim} N\left(0, \sigma_u^2\right), \end{aligned} \tag{10}$$

where as in (6), $\mathbf{x}_{ij}$ is a vector of concomitant values, β is a vector of fixed regression coefficients and u_i is a random effect representing the unexplained variability of the individual probabilities between the areas.

For this model there is no explicit expression for the predictor $\hat{p}_i$. Writing $p_i = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} y_{il}\right]$, predicting p_i by its best predictor is equivalent to the prediction of the sum $\sum_{l\notin s_i} y_{il}$ of the missing observations. See Jiang et al. (2002) for the computation of the empirical best predictor and estimation of its MSE.

About the Author

Danny Pfeffermann is Professor of statistics at the Hebrew University of Jerusalem, Israel, and at the Southampton Statistical Sciences Research Institute, University of Southampton, UK. He has presented and published many articles on small area estimation in leading statistical journals and is teaching this topic regularly. He is consulting the Office for National Statistics in the UK and the Bureau of Labor Statistics in the USA on related problems. Professor Pfeffermann is past president of the Israel Statistical Association (2005–2007), an Elected Fellow of the American Statistical Association (1990), and an Elected member of the International Statistical Institute. He was Associate Editor of *Biometrika* and the *Journal of Statistical Planning and Inference* and is currently Associate Editor for *Survey Methodology*. Professor Pfeffermann has recently completed jointly with Professor C.R. Rao editing the new two-volume Handbook of Statistics on *Survey Samples*, published by North Holland (2009). He is the recipient of the Waksberg award for 2011.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Census
►Estimation
►Estimation: An Overview
►Inference Under Informative Probability Sampling
►Markov Chain Monte Carlo
►Non-probability Sampling Survey Methods
►Sample Survey Methods
►Social Statistics
►Superpopulation Models in Survey Sampling

References and Further Reading

Battese GE, Harter RM, Fuller WA (1988) An error component model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 83:28–36

Fay RE, Herriot R (1979) Estimates of income for small places: an application of James Stein procedures to census data. J Am Stat Assoc 74:269–277

Ghosh M, Rao JNK (1994) Small area estimation: an appraisal (with discussion). Stat Sci 9:65–93

Jiang J, Lahiri P, Wan SM (2002) A unified jackknife theory for empirical best prediction with M-estimation. Ann Stat 30: 1782–1810

Pfeffermann D (2002) Small area estimation – new developments and directions. Int Stat Rev 70:125–143

Pfeffermann D, Rao CR (eds) (2009) Handbook of statistics 29B. Sample surveys: inference and analysis. Elsevier, North Holland

Rao JNK (1999) Some recent advances in model-based small area estimation. Survey Methodol 25:175–186

Rao JNK (2003) Small area estimation. Wiley, New York

Rao JNK (2005) Inferential issues in small area estimation: some new developments. Stat Transit 7:513–526

Smoothing Splines

Grace Wahba
IJ Schoenberg-Hilldale Professor of Statistics, Professor of Biostatistics and Medical Informatics
University of Wisconsin, Madison, WI, USA

Univariate Smoothing Splines

Univariate smoothing splines were introduced by I.J. Schoenberg in the 40s, an early paper is (Schoenberg 1964). Given data $y_i = f(x(i)) + \epsilon_i, i = 1,\cdots,n$, where the ϵ_i are i.i.d samples from a zero mean Gaussian distribution and $0 < x(1) < \cdots < x(n) < 1$, the (univariate) polynomial smoothing spline is the solution to: find f in W_2^m to minimize

$$\frac{1}{n}\sum_{i=0}^{1}(y_i - f(x(i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx,$$

where W_2^m is the Sobolev space of functions with square integral mth derivative. The solution is well known to be a piecewise polynomial of degree $2m-1$ between each pair $\{x(j+1), x(j)\}, j = 1,\cdots,n-1$ and of degree $m-1$ in $[0, x(1)]$ and $[x(n), 1]$, and the pieces are joined so that the function has $2m-1$ continuous derivatives. Figure 1 illustrates the cubic smoothing spline ($m = 2$) and how it depends on the smoothing parameter λ. The dashed line in each of the three panels is the underlying function $f(x)$ used to generate the data. The observations y_i were generated as $y_i = f(x_i) + \epsilon_i$ where the ϵ_i were samples from a zero mean Gaussian distribution with common variance. The wiggly solid line in the top panel was obtained with a λ that is too small. The solid line in the middle panel has λ too large.

If λ had been even larger, the solid line would have tended to flatten out towards the least squares straight line best fitting the data. Note that linear functions are in the *null space* of the penalty functional $\int (f'')^2$, that is, their second derivatives are 0. In the third panel, λ has been chosen by the GCV (Generalized Cross Validation) method (Craven and Wahba 1979; Golub et al. 1979). Generalizations of the univariate smoothing spline include penalties that replace $(f^{(m)})^2$ with $(Lf)^2$, where Lf is a linear differential operator of order m, see Kimeldorf and Wahba (1971) and Ramsay and Silverman (1997). Code for smoothing splines is available in the R library http://cran.r-project.org, for example `pspline` and elsewhere. Other generalizations include replacing the residual sum of squares by the negative log likelihood for Bernoulli, Poisson or other members of the exponential family, by robust or quantile functionals, or by the so-called hinge function to get a Support Vector Machine (Cristianini and Shawe-Taylor 2000). In each case the solution will be a piecewise polynomial of the same form as before as a consequence of the so called representer theorems in Kimeldorf and Wahba (1971). Other tuning criteria are appropriate for the other functionals, for example the GACV (Xiang and Wahba 1996) for Bernoulli data.

Thin Plate Splines

Thin Plate Splines (TPS) appeared in French in 1975 (Duchon 1975) and were combined with the GCV for tuning in Wahba and Wendelberger (1980). The TPS of order 2 in two dimensions is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i))^2 + \lambda J_{2,2}(f)$$

where $J_{2,2}$ is given by

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2 dx_1 dx_2.$$

In this case f is known to have a representation

$$f(x) = d_0 + d_1x_1 + d_2x_2 + \sum_{i=1}^{n} c_i E(x, x(i))$$

where

$$E(x, x(i)) = \|x - x(i)\|^2 log\|x - x(i)\|,$$

where $\|\cdot\|$ is the Euclidean norm.

There is no penalty on linear functions of the components (x_1, x_2) of the attribute vector (the "null space" of $J_{2,2}$). It is known that the c_i for the solution satisfy $\sum_{i=1}^{n} c_i = 0$, $\sum_{i=1}^{n} c_i x_1(i) = 0$ and $\sum_{i=1}^{n} c_i x_2(i) = 0$, and furthermore,

$$J_{2,2}(f) = \sum_{i,j=1,\cdots,n} c_i c_j E(x(i), x(j)).$$

The TPS is available for general d and for any m with $2m - d > 0$. The general TPS penalty functional in d dimensions and m derivatives is

$$J_{d,m} = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{-\infty}^{\infty}\cdots \int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)^2 \prod_j dx_j.$$

See Wahba (1990). Note that there is no penalty on polynomials of degree less than m, so that the TPS with d greater than 3 or 4 is rarely attempted because of the very high dimensional null space of $J_{d,m}$. As λ tends to infinity, the solution tends to its best fit in the unpenalized space, and as λ tends to 0, the solution attempts to interpolate the data. Public codes in R containing TPS codes include `assist, fields, gss, mgcv`. Again, the residual sum of squares may be replaced by other functionals as in the univariate spline and the form of the solution will be the same.

Splines on the Sphere

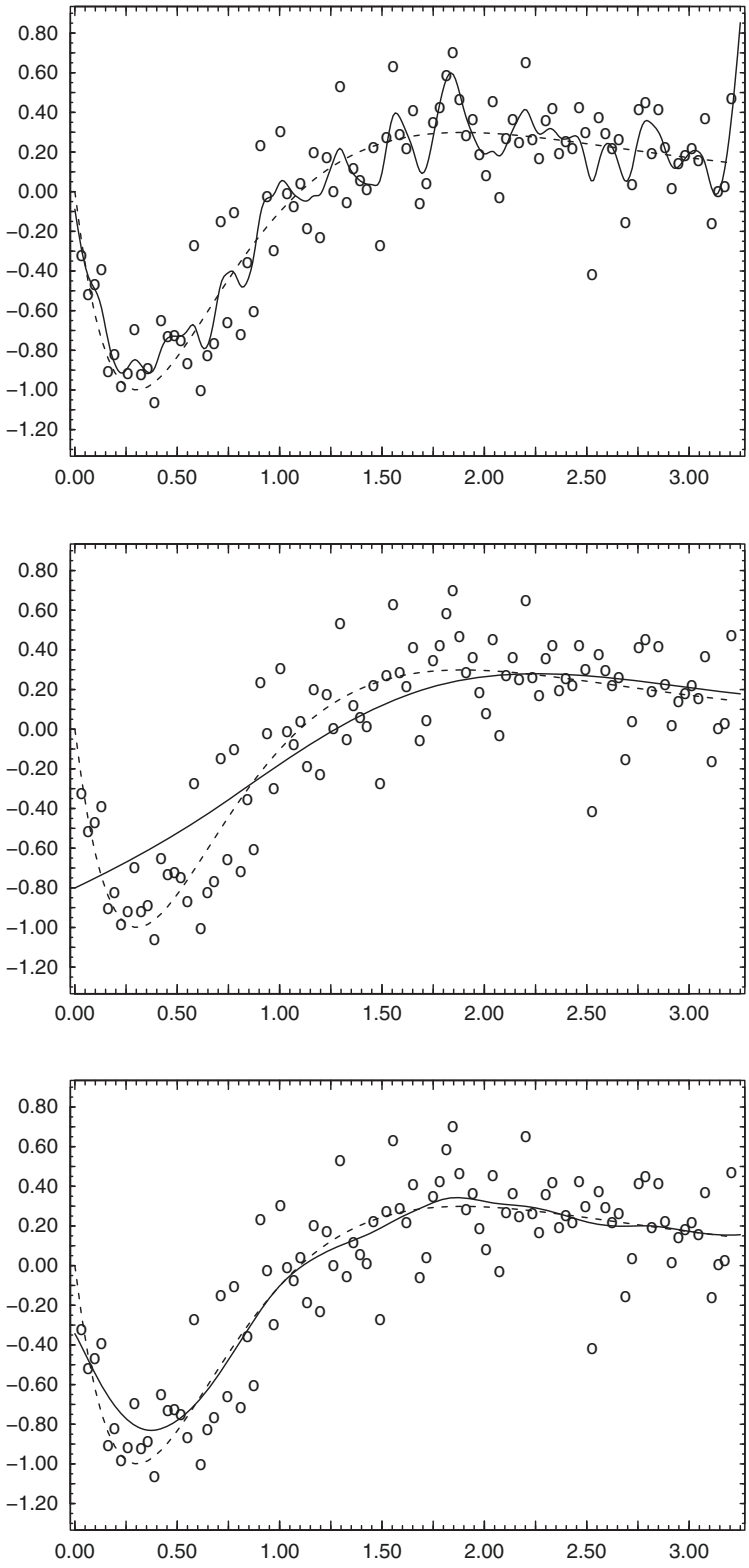
Splines on the sphere were proposed in Wahba; Wahba (1981; 1982). The penalty functional $J(f)$ for splines on the sphere is $J(f) = \int (\Delta)^{m/2} f$ where Δ is the (surface) Laplacian on the the (unit) sphere given by

$$\Delta f = \frac{1}{\cos^2\phi} f_{\theta\theta} + \frac{1}{\cos\phi}(\cos\phi f_\phi)_\phi$$

where θ is the longitude, $(0 \le \theta \le 2\pi)$ and ϕ is the latitude $\left(-\frac{\pi}{2} \le \phi \le \frac{\pi}{2}\right)$. Here we are using subscripts θ and ϕ to indicate derivatives with respect to θ and ϕ. Closed form expressions for the minimizer f are not in general available, but closed form expressions for a close approximation are, see Wahba; Wahba (1981; 1982).

Splines on Riemannian Manifolds

The splines we have mentioned above have penalty functionals associated with the Laplacian (note the form is different for the compact domain cases of splines on the unit interval and splines on the sphere, as opposed to the thin plate spline on the infinite plane). Splines on arbitrary compact Riemannian manifolds can be defined, implicitly or explicitly involving the eigenfunctions and eigenvalues of the m-iterated Laplacian, see Kim (1999), Pesenson (2004), Belkin and Niyogi (2004, Sect. 5.2).



Smoothing Splines. Fig. 1 Cubic smoothing spline with three different tuning parameters

Smoothing Spline ANOVA Models

Let $x = (x_1, \cdots, x_d)$, where $x_\alpha \in \mathcal{T}^{(\alpha)}, \alpha = 1, \cdots, d$ and $y_i = f(x(i)) + \epsilon_i, i = 1, \cdots, n$. where the ϵ_i are as before. The $\mathcal{T}^{(\alpha)}$ can be quite arbitrary domains. It is desired to estimate $f(x)$ for x in some region of interest contained in $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$. f is expanded as $f(x) = C + \sum_\alpha f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, t_\beta) + \cdots$, where the terms satisfy side conditions analogous to those in ordinary ANOVA which guarantee identifiability, and the decomposition is usually truncated at some point. The model is fit by minimizing the residual sum of squares plus

$$J_\lambda(f) = \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots.$$

The $J_\alpha, J_{\alpha\beta}, \cdots$ are composites of penalty functionals on the individual components and closed form expressions are available when they are available for the components. As before, the residual sum of squares may be replaced by the negative log likelihood and other functionals depending on y_i and $f(x(i))$. Details may be found in Wahba et al. (1995) and Gu (2002), and the R codes `assist` and `gss` are available to fit these models.

About the Author

Grace Wahba is IJ Schoenberg–Hilldale Professor of Statistics at the University of Wisconsin, where she has been a faculty member since 1967 after receiving her Ph.D. from Stanford in 1966. She is a Fellow of the International Statistical Institute, Institute of Mathematical Statistics, American Statistical Association, Society for Industrial and Applied Mathematics, and American Association for the Advancement of Science. She is also a Member of the National Academy of Sciences (2000). Dr. Wahba has an international reputation as an innovator in research on the theory and applications of "Spline Models for Observational Data." She was named the "Statistician of the Year" by the Chicago Chapter of ASA in 2004. Among many awards, she has been awarded the First Emanuel and Carol Parzen Prize for Statistical Innovation (1994), Committee of Presidents of Statistical Societies Elizabeth Scott Award (1996), Hilldale Award in the Physical Sciences, University of Wisconsin-Madison (2003). She has authored/coauthored about 130 papers and the book: *Spline Models for Observational Data* (SIAM, 1990). In 2007 she received an Honorary Doctorate from the University of Chicago. In 2009, Professor Wahba received the Gottfried E. Noether Senior Researcher Award for "outstanding contributions to the theory and applications of nonparametric statistics," and became the inaugural recipient of the Distinguished Alumni Award at Cornell University.

"Grace Wahba, the I. J. Schoenberg professor of statistics, University of Wisconsin-Madison, represents the very best of the modern synthesis of applied statistical, mathematical and computational science. Her most influential work has concerned problems in the estimation of curves and surfaces from large, high-dimensional data sets, such as occur frequently in geophysics." (Convocation Session, the University of Chicago Chronicle, Vol. 26, No 18, 2007).

Cross References

►Nonparametric Estimation
►Nonparametric Regression Using Kernel and Spline Methods
►Semiparametric Regression Models
►Smoothing Techniques

References and Further Reading

Belkin M, Niyogi P (2004) Semi-supervised learning on Riemannian manifolds. Mach Learn 56:209–239

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge

Duchon J (1975) Fonctions splines et vecteurs aleatoires. Technical Report 213, Seminaire d'analyse numerique, universite scientifique et medicale, Grenoble

Golub G, Heath M, Wahba G (1979) Generalized cross validation as a method for choosing a good ridge parameter. Technometrics 21:215–224

Gu C (2002) Smoothing spline ANOVA models. Springer, New York

Kim P (1999) Splines on Riemannian manifolds and a proof of a conjecture by Wahba. Report, Department of Mathematics and Statistics, University of Guelph

Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. J Math Anal Appl 33:82–95

Pesenson I (2004) Variational splines on Riemannian manifolds with aplications to integral geometry. Adv Appl Math 33:548–572

Ramsay J, Silverman B (1997) Functional data analysis. Springer, New York

Schoenberg I (1964) Spline functions and the problem of graduation. In: Proceedings of the national academy sciences, vol 52. USA, pp 947–950

Wahba G (1981) Spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 2:5–16

Wahba G (1982) Erratum: Spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 3:385–386

Wahba G (1990) Spline models for observational data. In: CBMS-NSF regional conference series in applied mathematics, vol 59. SIAM, Philadelphia

Wahba G, Wendelberger J (1980) Some new mathematical methods for variational objective analysis using splines and cross-validation. Mon Weather Rev 108:1122–1145

Wahba G, Wang Y, Gu C, Klein R, Klein B (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. Ann Stat 23:1865–1895, Neyman Lecture

Xiang D, Wahba G (1996) A generalized approximate cross validation for smoothing splines with non-Gaussian data. Stat Sinica 6:675–692

Smoothing Techniques

Adrian W. Bowman
Professor
The University of Glasgow, Glasgow, UK

The idea of smoothing techniques is to identify trends, patterns, relationships and shapes in data without adopting strong assumptions about the specific nature of these. The one assumption that *is* made is that any trends and patterns are smooth. The term *nonparametric* is often used in the context of smoothing techniques to distinguish the methods from *parametric* modelling where specific distributional shapes (such as normal) or trends (such as linear) are adopted, leaving only some parameters to be estimated.

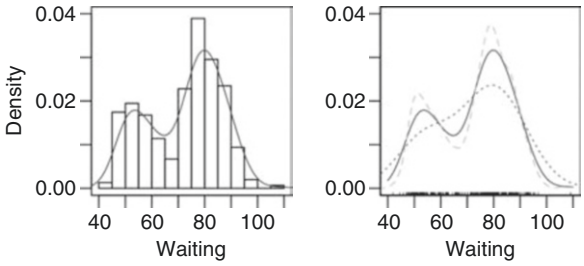
There are many situations where smoothing can be applied and many ways in which it can be implemented. This short article will give some simple examples in just two areas, namely density estimation and regression, and show how the latter techniques can be used in the context of wider regression modelling.

Density Estimation

The histogram is a time-honored way of presenting the shape of the variation in a set of data in graphical form. In fact, when the histogram is scaled to have area 1 it can be viewed as an estimate of the underlying density function $f(y)$. However, from that perspective it can be criticized because of its sharp edges. Instead of building the estimate from rectangular blocks, a kernel density estimate uses smooth functions, called kernels, in the estimate

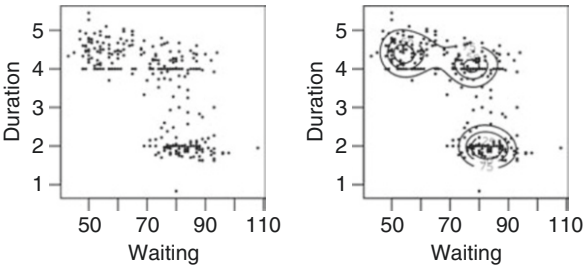
$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h)$$

constructed from a sample of data $\{y_1, \ldots, y_n\}$. The kernel $w(.; h)$ might conveniently be chosen as a normal density function with mean 0 and standard deviation h. It remains to make a choice of the bandwidth, or smoothing parameter, h which is the equivalent of the bin width in a histogram. One effective means of doing this is to estimate the optimal value produced by a theoretical analysis. However, a very simple choice, which can also be very effective, is to use the optimal value associated with a normal distribution. That is the solution used in the examples below.



The left panel of the figure above shows a histogram of data on the waiting times between eruptions of the Old Faithful geyser in Yellowstone National Park. A kernel density estimate has been superimposed for comparison. The right panel shows the same density estimate along with estimates produced with larger (short dashed line) and smaller (long dashed line) degrees of smoothing.

These simple principles extend without difficulty to other types of data, simply by adopting a suitable form of kernel function. For example, the left hand panel below shows a plot of waiting time and the subsequent eruption time. The right panel shows the same plot with the contours of a density estimate superimposed. The kernel function here is simply a two-dimensional normal density function, with two smoothing parameters, one for each dimension. Although the scatterplot clearly shows a cluster of eruptions with shorter durations, the density estimate draws attention to the presence of two clusters in the eruptions with longer durations. In general, smoothing techniques such as density estimation can be helpful in identifying structure which is sometimes obscured by the variation in the data.



Silverman (1986) gave one of the first discussions of density estimation, with Scott (1992) focussing on the multivariate case. Wand and Jones (1995) is a source of very

useful theoretical analysis while Simonoff (1996) is particularly helpful in its broad coverage and extensive references.

Nonparametric Regression

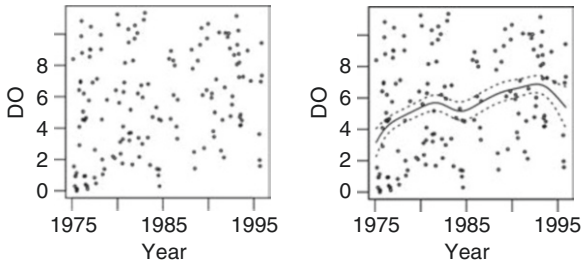
In the case of regression with a single covariate, smoothing techniques assume the model

$$y_i = m(x_i) + \varepsilon_i$$

for observed data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where the ε_i denote errors terms. The smooth function m can be estimated in a wide variety of ways. A kernel approach fits a standard model, such as a linear regression, but does so locally by solving the weighted least squares problem

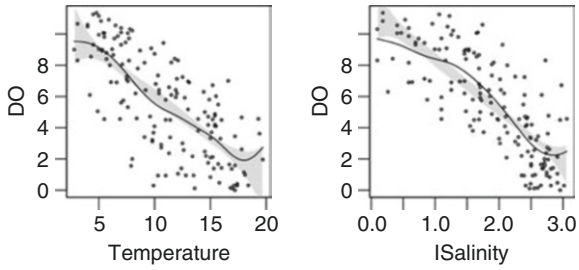
$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h).$$

The solution $\hat{\alpha}$ provides the estimate. However, there are many other approaches, many of these based on splines. For example, ▶smoothing splines arise as the solution of the problem $\min_m \sum_{i=1}^{n} \{y_i - m(x_i)\} + \lambda \int_a^b m''(x)dx$. Regression splines fit a model which is constructed as a linear combination of a set of basis functions while penalized splines place a smoothness penalty on these coefficients. This is a research topic with a large literature. Fan and Gijbels (1996) and Bowman and Azzalini (1997) describe the theory and applications of the kernel approach while Green and Silverman (1994) and Ruppert et al. (2003) focus on spline representations. In broad terms, these different methods have different approaches but a common aim. The method chosen for a particular problem can be a matter of convenience.



The panels above illustrate local linear smoothing on water quality data, expressed in dissolved oxygen (DO) at a particular sampling station on the River Clyde near Glasgow. The left hand panel shows DO against time in years, with little evidence of trend. The right hand plot adds a nonparametric regression curve which suggests that some trend may in fact be present, obscured by the large degree of variation in the data. The vector of fitted values from local linear, and indeed most other, forms of regression smoothing can be represented in vector–matrix form as $\hat{m} = Sy$, where S is an $n \times n$ smoothing matrix. This linear structure gives relatively easy access to standard errors and to the quantification of the level of smoothing through approximate degrees of freedom, by analogy with standard linear models. The right hand panel above has added two standard errors on either side of the nonparametric regression line, to indicate the precision of estimation. Bias is an inevitable consequence of smoothing so this cannot be strictly interpreted as a confidence band.

The two panels below show DO against temperature and Salinity on a log scale. Here the patterns are close to linear and the suitability of this model can be assessed by displaying a reference band around the linear model, based on two standard errors of the difference between a linear and a nonparametric model. Linearity looks reasonable for temperature but less so for Salinity.



The plots above were created by specifying the level of smoothing through the approximate number of degrees of freedom (6). The level of smoothing can also be chosen in a data-adaptive manner, through principles such as cross-validation or AIC.

These methods of nonparametric smoothing can be adapted to a wide variety of situations, such as more than one covariate or other types of response data.

Additive Models

Smoothing techniques can be built into wider models, particularly where several covariates are involved. An attractive framework is provided by additive models, described by Hastie and Tibshirani (1990) with an updated treatment by Wood (2006). Here, the regression model is defined as

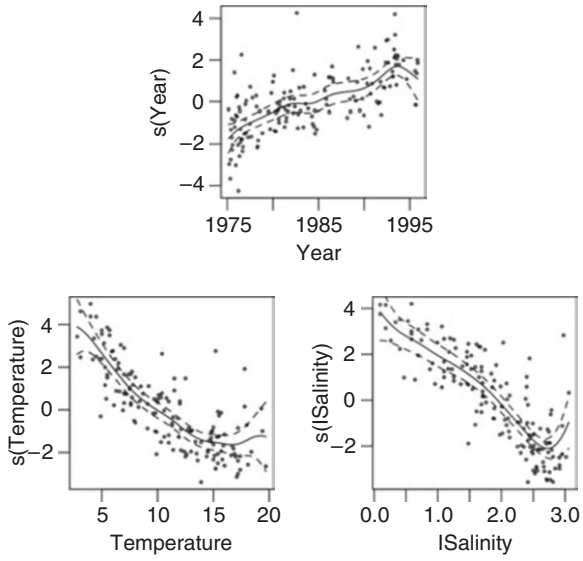
$$y_i = \alpha + m_1(x_{1i}) + \ldots + m_p(x_{pi}) + \varepsilon_i$$

for covariates $x_1, \ldots, x_p$. Each covariate x_j is allowed to influence the response variable through its own regression function m_j, which may be nonparametric but could in fact be linear or some other standard form. The backfitting

algorithm provides a means of fitting this type of model through the iterations defined by

$$\hat{m}_j^{(r+1)} = S_j\left(y - \hat{\alpha}\mathbf{1} - \sum_{k<j}\hat{m}_k^{(r+1)} - \sum_{k>j}\hat{m}_k^{(r)}\right).$$

At each stage, the regression function m_j is estimated by smoothing the partial residuals by S_j, the smoothing matrix associated with covariate j. For identifiability, the constraint that each component sums to 1 over the observed covariate values should also be added.



The panels above illustrate an additive model for the Clyde data. Instead of examining the effects of the covariates separately, they are combined into a single model which estimates the effects of one covariate while adjusting for the effects of the others. This much more powerful description now shows a much clearer time trend. The effects of temperature and salinity remain broadly linear but some unusual behavior is evident at high temperature and high salinity.

Bowman (2008) gives a more extended discussion of this example, using a different sampling station on the Clyde while McMullan et al. (2007) develop a more complex model for the whole river.

Acknowledgments

This work received partial support of grant PBCT-ADI13 of the Chilean Science and Technology Bicentennial Foundation.

About the Author

Adrian Bowman is a Professor of Statistics at the University of Glasgow. He is an elected Fellow of the International Statistical Institute and of the Royal Society of Edinburgh. He served as Joint Editor of *Applied Statistics* (J. Roy. Stat. Soc. Series C) for four years and has at various times served as associate editor for *Biometrika*, *J. Roy. Stat. Soc. Series B* and *Biometrics*. He is currently associate editor for *Biostatistics* and the *Journal of Statistical Software*. Prof. Bowman has acted as chair of the UK Committee of Professors of Statistics. He has also held a wide variety of responsibilities within the Royal Statistical Society, where he is currently an Honorary Officer.

Cross References

►Exponential and Holt-Winters Smoothing
►Median Filters and Extensions
►Moving Averages
►Nonparametric Density Estimation
►Nonparametric Estimation
►Nonparametric Models for ANOVA and ANCOVA Designs
►Nonparametric Regression Using Kernel and Spline Methods
►Smoothing Splines

References and Further Reading

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis. Oxford University Press, Oxford
Bowman AW (2008) Smoothing techniques for visualisation. In: Chen C-H, Härdle W, Unwin A (eds) Handbook of data visualization. Springer, Heidelberg
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London
Green P, Silverman B (1994) Nonparametric regression and generalized linear models. Chapman & Hall, London
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
McMullan A, Bowman AW, Scott EM (2007) Water quality in the river clyde: a case study of additive and interaction models. Environmetrics 18:527–539
Ruppert D, Wand MP, Carroll R (2003) Semi- parametric regression. Cambridge University Press, London
Scott D (1992) Multivariate density estimation: theory, practice, and visualization. Wiley, New York
Silverman B (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
Simonoff JS (1996) Smoothing methods in statistics. Springer, New York
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London
Wood S (2006) Generalized additive models: an introduction with R. Chapman & Hall/CRC Press, London

Social Network Analysis

Tom A. B. Snijders
Professor of Statistics
University of Oxford, UK
Professor of Methodology and Statistics, Faculty of Behavioral and Social Sciences
University of Groningen, Groningen, Netherlands

Social Networks

Social Network Analysis is concerned with the study of relations between social actors. Examples are friendship between persons, collaboration between employees in a firm, or trade between countries. The relation is regarded as a collection of dyadic ties, i.e., ties between pairs of actors. In most cases, data collection is either *sociocentric*, where a given group of actors is specified (in the examples this could be, e.g., a school class, a department of the firm, or all countries in the world), and all ties of the specific kind between actors in this group are considered; or *egocentric*, where a sample of actors is taken, and all ties of the sampled actors are considered. Other types of data collection exist, of which snowball sampling is the main example. The most interesting contributions of network analysis are made by considering indirect ties – in the sense that the way in which actors i and j are tied is better understood by considering the other ties of these two actors. Information about these is obtained much better from sociocentric than from egocentric approaches. Therefore, this article considers only statistical models for sociocentric network data.

The first step for the collection of sociocentric network data is to define the relation and the group of actors. This group will usually be treated as an isolated group, and any ties outside this group are disregarded. This is called the *network boundary problem*. An overview of methods for collecting network data is given by Marsden (2005).

Notation

The group of actors is denoted by $\mathcal{N} = \{1, \ldots, n\}$. Relations under study often are directed, which means that the tie $i \to j$ is distinct from the tie $j \to i$. The relation can then be represented by a nonreflexive directed graph (digraph) on $\mathcal{N}$ or, alternatively, by an $n \times n$ adjacency matrix with a structurally zero diagonal. The actors $i \in \mathcal{N}$ are the nodes of the graph. The adjacency matrix $\mathbf{y} = (y_{ij})$ indicates by $y_{ij} = 1$ or $y_{ij} = 0$, respectively, that there is a tie, or there is no tie, from actor i to actor j. The nonreflexivity means that self-ties are not considered, so that $y_{ii} = 0$ for all i. The variables y_{ij} are referred to as *tie variables*. If the network is nondirected, the representation is by a simple graph, or a symmetric adjacency matrix. Models for social networks in this article will be random graphs or digraphs and denoted by $\mathbf{Y}$.

Exponential Random Graph Models

Exponential families of probability distributions for graphs or digraphs are usually called *Exponential Random Graph Models* or ERGMs. The first model of this kind was the so-called p_1 model proposed by Holland and Leinhardt (1981). In this model the symmetrically positioned pairs (Y_{ij}, Y_{ji}) are assumed to be independent. This very restrictive assumption was lifted in the definition by Frank and Strauss (1986) of *Markov graphs*. This model can represent tendencies toward transitivity. It postulates that edge indicators Y_{ij} and Y_{hk}, when i, j, k, h are four distinct nodes, are independent conditional on the rest of the graph, i.e., conditional on the collection of tie indicators Y_{rs} for $(r,s) \neq (i,j)$, $(r,s) \neq (h,k)$. For non-directed networks with distributions not depending on the node labels, they proved that this property holds if and only if the probability distribution for $\mathbf{Y}$ can be expressed as

$$P_\theta\{\mathbf{Y} = \mathbf{y}\} = \exp\left(\sum_h \theta_h z_h(\mathbf{y}) - \psi(\theta)\right), \qquad (1)$$

where the $z_h(\mathbf{y})$ are functions of $\mathbf{y}$ each of which can be either the number of k-stars embedded in the graph $\mathbf{y}$ (for some $k, 1 \le k \le n-1$) or the number of triangles embedded in $\mathbf{y}$. These are the statistics S_k and T defined by

$$S_1(\mathbf{y}) = \sum_{1 \le i < j \le n} y_{ij} \quad \text{number of edges}$$
$$S_k(\mathbf{y}) = \sum_{1 \le i \le n} \binom{y_{i+}}{k} \quad \text{number of } k\text{-stars } (k \ge 2) \qquad (2)$$
$$T(\mathbf{y}) = \sum_{1 \le i < j < h \le n} y_{ij}\, y_{ih}\, y_{jh} \quad \text{number of triangles.}$$

The Markov model was generalized by Frank (1991) and Wasserman and Pattison (1996) to the Exponential Random Graph Model, in which the statistics $z_h(\mathbf{y})$ in (1) can be any functions of $\mathbf{y}$ and of covariates. Markov chain Monte Carlo (MCMC) methods (see ▶Markov Chain Monte Carlo) for parameter estimation for this model were proposed by Snijders (2002). Some interesting properties of this model are discussed by Robins et al. (2005). It appeared in applications, however, that in most cases the Markov model is not plausible as a model for transitivity. An model specification with more appropriate choices of the functions $z_h(\mathbf{y})$ was proposed in Snijders et al. (2006), and this has turned out to be a very useful model for representing empirically observed networks.

This model can represent dependencies between tie variables Y_{ij} in a reasonable manner. It can be used when the representation of these dependencies (transitivity, hierarchy, brokerage etc.) is an aim in itself; but also when the dependencies are a nuisance and the aim of the statistical analysis is the dependence of tie variables on covariates.

Latent Structure Models

Another way to represent dependencies between tie variables is to postulate a latent space of which the nodes are elements, and which probabilistically determines the ties. This is an application of the ideas of Latent Structure Analysis (Lazarsfeld and Henry 1986), and closely related to Latent Class Analysis. The tie variables Y_{ij} – or sometimes the dyads (Y_{ij}, Y_{ji}) – then are assumed to be conditionally independent given the latent structure.

Various latent space models have been proposed.

- A discrete (categorical) space, where the nodes have 'colors' and the distribution of the dyad (Y_{ij}, Y_{ji}) depends on the colors of i and j: see Nowicki and Snijders (2001).
- A general or Euclidean metric space, where the probability of a tie $Y_{ij} = 1$ depends on the distance between nodes i and j: see Hoff et al. (2002).
- An ultrametric space, where the probability of a tie $Y_{ij} = 1$ depends on the ultrametric distance between nodes i and j: see Schweinberger and Snijders (2003).
- A partially ordered space, where the probability of a tie $Y_{ij} = 1$ depends on how i and j are ordered: see Mogapi (2009).

Compared to Exponential Random Graph Models, these models have less flexibility to represent dependence structures between tie variables, so that they will usually achieve a less satisfactory goodness of fit. However, the representation of the nodes in the latent space can often provide an illuminating representation in itself and may be regarded as a helpful type of data reduction.

Longitudinal Models

Models for longitudinally observed networks were proposed by Snijders (2001). The most usual observational design is a panel design, where the observations of the network are $\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_M)$ for observation moments $t_1, \ldots, t_M$ ($M \geq 2$). A flexible class of models for panel data on networks can be obtained by assuming that the data are momentary observations of a continuous-time Markov process (see ►Markov Processes), in which each tie variable $X_{ij}(t)$ develops in stochastic dependence on the entire network $X(t)$. An actor-based model is often plausible, where tie changes are based on hypothetical choices of the actors. Such a model can be defined by the following steps, formulated in such a way that they can easily be represented by a computer simulation model. To obtain a parsimonious model, it is assumed that only one tie variable can change at any given moment. The model is characterized by so-called *rate functions* $\lambda_i(\mathbf{y})$ and *objective functions* $f_i(\mathbf{y})$, defined on the set of all digraphs.

1. The current state of the network is denoted $\mathbf{y}$.
2. The time until the next change is an exponentially distributed waiting time, with an expected duration of $1/\lambda_+(\mathbf{y})$ where $\lambda_+(\mathbf{y}) = \sum_i \lambda_i(\mathbf{y})$.
3. When this change occurs, the probability that an outgoing tie variable Y_{ij} of actor i can be changed, is $\lambda_i(\mathbf{y})/\lambda(\mathbf{y})$.
4. If actor i can change on outgoing tie variable, the set of new possible states of the network is

$$\mathcal{C}(\mathbf{y}) = \{\mathbf{y}' \mid y'_{hk} \neq y_{hk} \text{ only for } h = i, \text{ and for at most one } k\} .$$

The probability that the new state is $\mathbf{y}'$ is

$$\frac{\exp\left(f_i(\mathbf{y}')\right)}{\sum_{\mathbf{y}'' \in \mathcal{C}(\mathbf{y})} \exp\left(f_i(\mathbf{y}'')\right)} .$$

The model specification is done in the first place by the appropriate definition of the objective function. This is usually specified as a linear combination,

$$f_i(\beta, \mathbf{y}) = \sum_k \beta_k s_{ki}(\mathbf{y}) . \tag{3}$$

The functions $s_{ki}(\mathbf{y})$ represent ways in which the creation and maintenance of ties depend on currently existing ties, e.g.,

$$\begin{aligned} s_{ik}(,\mathbf{y}) &= \sum_j y_{ij} && \text{(outdegree)} \\ &\sum_j y_{ij}\, y_{ji} && \text{(reciprocated ties)} \\ &\sum_{j,k} y_{ij}\, y_{jk}\, y_{ik} && \text{(transitive triplets),} \end{aligned}$$

and they can also depend on combinations of network structure and covariates.

For this model, estimation procedures and algorithms according to a method of moments were proposed by Snijders (2001), Bayesian procedures by Koskinen and Snijders (2007), and an algorithm for maximum likelihood estimation by Snijders et al. (2010).

This model was generalized to a model for the simultaneous dynamics of networks and actor characteristics

("networks and behavior") by Snijders et al. (2007). Statistical procedures for this model are available in the R package RSiena.

About the Author

For biography *see* the entry ►Multilevel Analysis.

Cross References

►Graphical Markov Models
►Markov Chain Monte Carlo
►Markov Processes
►Methods of Moments Estimation
►Network Models in Probability and Statistics
►Network Sampling
►Panel Data
►Probabilistic Network Models

References and Further Reading

Basic information about social network analysis is in Wasserman and Faust (1994) and in Carrington et al (2005) A review of a variety of other statistical procedures and models for network analysis is given by Airoldi et al (2007)

Airoldi E, Blei DM, Fienberg SE, Goldenberg A, Xing EP, Zheng AX (2007) Statistical network analysis: models, issues and new directions (ICML 2006). Lecture notes in computer science, vol 4503. Springer, Berlin

Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge

Frank O (1991) Statistical analysis of change in networks. Stat Neerl 45:283–293

Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81: 832–842

Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. J Am Stat Assoc 97:1090–1098

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65

Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic network data. J Stat Plan Infer 13:3930–3938

Lazarsfeld PF, Henry NW (1968) Latent structure analysis. Houghton Mifflin, Boston

Marsden PV (2005) Recent developments in network measurement. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 8–30

Mogapi O (2009) A latent partial order model for social networks. D.Phil. thesis, Department of Statistics, University of Oxford

Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96:1077–1087

Robins GL, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. Am J Sociol 110:894–936

Schweinberger M, Snijders TAB (2003) Settings in social networks: a measurement model. In: Stolzenberg RM (ed) Sociological methodology, vol 23. Blackwell, Boston, pp 307–341

Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology. Basil Blackwell, Boston and London, pp 361–395

Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3:2

Snijders TAB, Pattison PE, Robins GL, Handcock MS (2006) New specifications for exponential random graph models. Sociol Methodol 36:99–153

Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Mahwah, pp 41–71

Snijders TAB, Koskinen JH, Schweinberger M (2010) Maximum likelihood estimation for social network dynamics. Ann Appl Stat, to be published

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Wasserman S, Pattison PE (1996) Logit models and logistic regression for social networks: I an introduction to Markov graphs and p^*. sychometrika 61:401–425

Social Statistics

Vassiliy Simchera
Director of Rosstat's Statistical Research Institute, Moscow, Russia

Social statistics is one of the largest domains of modern statistical science and practice, the subject of which is the exposure and study of regularity for formation and alteration of social phenomena with statistical techniques.

It has grown and developed at the borders of other sciences (►demography, economics, political science, philosophy, ethics, and psychology) as the discipline that integrates statistical resources and bases of humanitarian information studying human beings and society. It gained intensive development in the 19thand 20th centuries as a science studying social dynamics, which was initiated in the United States by Russian-American sociologist *Pitirim Sorokin*, although the first record of it one can find in ancient origins at the beginning of AD.

Social statistics operates with the branched system of indicators characterizing standards of life and human activities and further groups of people, public societies, nations, and civilizations, their evolution and structure, ways and standards of life, households, culture, education, moral, and human values, freedoms, rights, etc.

In contrast to many other statistical disciplines, its main emphasis is on the study of quantitatively immeasurable indicators as most common in social science.

It also scrutinizes and forecasts unobservable and non-registering "shadow," illegal, and informal social phenomena, by means of analysis techniques of social projects and doctrines, votes, and elections in particular.

In its work, along with the methods of sample surveys and ►public opinion polls, social statistics extensively applies *special methods*, among which are various methods of multivariate factor analysis, cluster analysis (see ►Cluster Analysis: An Introduction), and latent analysis. The particular classes are the methods of social modeling and managerial social analysis, on the basis of which a new section of modern statistics, called sociometrics evolved.

At present time, social statistics is positioned as an instrument of the application of its methods and information about social sciences, the main aim and product of which is qualitative measurement of social and widely spiritual aspects of material production and their integration as superior values and achievements of modern society into the socio-economic context.

There are an extensive collection of models, not only for common but also for applied social changes, in particular, the dynamics of climate change, epidemics, catastrophes, health care and diseases, crime, cloning, psychological and psychotropic conspiracies and wars, application of up-to-date and specialized computer and mathematical methods in demographics, medicine and sanitary statistics, as well as in biology, anthropology and other related sciences.

Social statistics also develops as *social groups statistics*, in particular poverty statistics, behavioral statistics, i.e., behavior of people in the exotic environment, statistics of crime, statistics of fair competition, and statistics on globalization and mass protests.

Another area is a statistics of interethnic conflicts and wars, terrorism, crisis and anthropogenic catastrophes, which threaten the existence of world civilizations.

Social statistics is formed on the basis of sampling surveys and public opinion polls; it actually relies upon opinions about facts rather than on the facts themselves, it characterizes mainly feedback, original responses to events in the surrounding world, rather than the events themselves. Without reliable criteria of estimation for data quality. Social statistics and its indicators, where applicable, require preliminary verification of their results and publications as they are least of all true and acceptable.

Main Social Statistics Centers:

- *Harvard Institute for Quantitative Social Science*
- *Inter-University Consortium for Political and Social Research*
- *Social Statistics Division, School of Social Sciences, University of Southampton, UK*
- *Social Statistics Research Group, University of Auckland, New Zealand*
- *UN Statistics Division - Demographic and Social Statistics*
- *Organization for Economic Co-operation and Development* (*OECD*)

Cross References

►Economic Statistics
►Public Opinion Polls
►Small Area Estimation
►Sociology, Statistics in

References and Further Reading

EuroStat (2006) European social statistics-social protection, Luxembourg
Irvine J, Miles I, Evans J (eds) (1979) Demystifying social statistics (1979). Pluto Press, London

Sociology, Statistics in

Gudmund R. Iversen
Professor Emeritus
Swarthmore College, Swarthmore,
PA, USA

Introduction

Statistics and sociology have a strong relationship that goes back several centuries. As new social theories and methods have been developed, statistics has responded by developing appropriate statistical methods. Also, sociologists have been quick adopting new statistical methods not necessarily developed with them in mind. The same is also the case with other social sciences such as political science, economics and psychology.

A few social sciences have relied more on statistics than others. Perhaps, the heaviest user of statistics has been economics, and the uses of statistics there have led to their own branch of statistics known as *econometrics*. With the abundance of economic data, econometrics has led to new uses of regression analysis. In turn, econometrics has been adopted by other social sciences, such as sociology and psychology.

Psychology is another social science where statistics has led to its own branch of statistics known as *psychometrics*. Psychology has an abundance of scores on

tests administered to college students and people seeking employment as well as psychiatry trying do diagnose people with suspected mental disorders. The most well known statistical methods in psychometrics is known as *factor analysis* of various kinds.

The abundance of survey analysis with the uses of questionnaires (see ►Questionnaire) in sociology would not have been possible without modern statistical *sampling* methods. Needs of sociology have led statisticians to develop sampling methods such as stratified sampling, ►cluster sampling and other sampling procedures. In turn, this has spilled over into the uses of sampling when the goal is obtain a complete ►census of some population. One of the leading organizations in the development of modern sampling methods for the collection of social science data has been the United States Bureau of the Census.

Sampling Theory

Sociologists, as well as others, have long collected data on individuals to study how people feel about issues of the day. In addition, political scientists have used sample surveys to try to predict outcomes of elections to be held sometime in the future. One of the most famous examples of such a prediction being wrong took place during the presidential election in the United States in 1948. On the night of the elections many surveys showed that Thomas E. Dewey had won and the incumbent Harry S. Truman had lost the election. Instead, Truman woke up the next day and found he had been elected president for the next four years. Another famous example took place during the US presidential election of 1936 when a well-known publication predicted on the basis of their poll that Governor Alf Landon would win the election. Instead, Franklin D. Roosevelt won almost two thirds of the popular vote that year and went on to win the next two elections as well.

What went wrong in both of these two cases was that statisticians had not stressed hard enough is that in order to generalize from a sample to a larger population, the sample must have been selected according to proper random statistical methods. In 1936 the sample was drawn from lists of people who owned cars. But this was in the middle of the economic depression years, and only reasonably wealthy people owned cars while most people without cars voted for Roosevelt. In 1948 George Gallup and others made use of the so-called quota sampling method. Each interviewer was told to go out and select respondents in such a way that the sample would reflect the population on characteristics such as gender and age. But that way interviewers would miss people who worked during off hours like a night shift at a factory and slept during the daytime when interviewers were seeking people with the right characteristic to satisfy the quotas they were given. An occasional survey still uses quota sampling for the selection of respondents, in spite of the well-known shortcomings of quota sampling. These days it is much more common to chose respondents by making a random selection of telephone numbers and dial those numbers.

Demography

For centuries, states have wanted to count the number of inhabitants for tax and military purposes. For this purpose, the German word *Statistik* was introduced more than two hundred and fifty years ago to denote matters of state, and the word probably comes from the Latin word *Statisticum*. In principle, a census does not require the use of statistical methods, but it is very difficult to take an accurate census without the use of sampling to count people who otherwise would be hard to include in the final count.

Simultaneous Structural Equations

The analysis of complex sociological models has led to generalizations of simple regressions models to models involving several regression equations where the parameters in all the equations are estimated at the same time. This formulation of a model has led both statisticians and sociologists to fruitful collaborations on how to estimate the parameters and how to interpret the estimates. The estimation procedure has moved from ordinary least squares estimation to what is known as two-stage and even three-stage estimation, depending upon the model. This is a case where theoretical work by economists have made major contributions to statistical theory and major uses in sociology.

Such models also go under the name of causal analysis or path analysis. Path analysis seems to have originated in biology around 1920, and it caught on in sociology in the 1960ies. A leading person in this field was the sociologist Hubert Blalock, perhaps best known for his famous textbook *Social Statistics* in addition to his writings on causal models. Causal modeling using path analysis has lost some of its attraction after people realized that establishing causality using statistical models did not necessarily lead to truly causal connections between variables.

Contingency Table Analysis

Much of the data in sociology consist of nominal (qualitative) variables such as gender (female, male), religious affiliation (protestant, catholic, Muslim, Jewish, etc.) and others. Because there are no meaningful numerical values attached to these categories, such data cannot be analyzed

by using means, standard deviations, single or multiple regression, etc. Instead, perhaps the best-known and oldest statistical method for the analysis of the relationship between two such variables is the chi-square analysis. It is based on the difference between the observed frequencies and expected frequencies computed as what the frequencies would have been if there were no relationship between the two variables.

A more recent development is the multivariate chi-square analysis for more than two categorical variables. This permits the study of interaction effects of the independent variables onto the dependent variable. Also, ►logistic regression has become popular for the case where the dependent variable has only two values. Finally, the use of ►dummy variables for quantitative variables have become possible using software so designed. Any quantitative variable with k different categories can be represented by $k - 1$ dummy variable, each having values of 0 and 1. With the data in this form it is possible to use ordinary linear regression for the study of the relationship between the dependent and the independent variables.

Conclusion

The empirical part of sociology could not exist without the use of statistics. Statistics has become an integral part of empirical sociological research. Any randomly chosen issue of a major sociological journal will have several articles making using of data analysis and statistics.

At one time it looked as if mathematics could play a similar role for sociology, but that effort has not paid off the way it was hoped. This takes us back to the importance of statistics for sociology. However, a major obstacle is that most sociologists lack the necessary background in statistics, partly due to the fact that they do not know enough mathematics to fully understand the statistical methods they are using. Similarly, most statisticians lack the knowledge of sociology needed to understand what statistical methods sociologists need. A few people have been able to bridge this gap, but most sociology students, even sociology graduate students, see the study of statistics as a hard task, perhaps mostly because statistics for sociologist has not been taught very well.

About the Author

For biography *see* the entry ►Analysis of Variance.

Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Confounding and Confounder Control
►Demography
►Event History Analysis
►Factor Analysis and Latent Variable Modelling
►Non-probability Sampling Survey Methods
►Psychology, Statistics in
►Role of Statistics
►Social Statistics
►Structural Equation Models

References and Further Reading

Blalock H (1971) Causal models in the social sciences. Aldine, New York
Hald A (1998) A history of mathematical statistics from 1750 to 1930. Wiley, New York
Stiegler SM (1986) The history of statistics: the measurement of uncertainty before 1900. Belkap Press of Harvard University Press, Cambridge, MA and London

Spatial Point Pattern

Peter J. Diggle
Distinguished Univerisity Professor
Lancaster University, Lancaster, UK
Adjunct Professor
Johns Hopkins University School of Public Health, Baltimore, MD, USA
Adjunct Senior Researcher
Columbia University, New York, NY, USA

Introduction

A spatial point pattern is a set of data consisting of the locations, $x_i : i = 1, \ldots, n$, of all events of a particular kind within a designated spatial region A. Typically, the pattern is assumed to be the outcome of a stochastic point process (see ►Point Processes) whose properties are of scientific interest.

An example would be the locations x_i of all trees in a designated region within a naturally regenerated forest. The observed pattern could be the result of a complex mix of natural processes. For example: regeneration from seedlings around the base of a mature tree could produce clusters of young trees; variation in soil fertility could produce patches of relatively low and high intensity of regeneration; competition for limited nutrient or light could lead to a spatially regular pattern is which only the dominant member of a cluster of seedlings survives.

Complete Spatial Randomness

The simplest statistical model for a spatial point process is the homogeneous Poisson process (see ►Poisson Processes). One of several possible definitions of this process is that:

1. The number of points in any planar region A follows a Poisson distribution with mean $\lambda|A|$, where $|\cdot|$ denotes area and the parameter $\lambda > 0$ is the *intensity*, or mean number of points per unit area.
2. The numbers of events in any two disjoint areas are independent.

Properties (1) and (2) imply that, conditionally on the number of points in A, their locations form an independent random sample from the uniform distribution on A.

Models

The Poisson process provides a standard of complete spatial randomness, but is inadequate as a model for most naturally occurring phenomena. As would be the case in our hypothetical forestry example, we need models to describe a response to an inhomoegeneous environment, or a tendency for points either to cluster together or to inhibit the occurrence of mutually close sets of points.

To model a response to an inhomogeous environment, a first possibility is to replace the constant intensity λ by a function $\lambda(x)$. In practice, this is only useful if we can model $\lambda(x)$ as a function of spatially referenced explanatory variables, for example height above sea-level. In the absence of such information, we can treat $\lambda(x)$ as a realisation of an unobserved stochastic process, so defining the class of Cox processes (Cox 1955).

The first, and still widely used, model for clustering of points is the Neyman–Scott process (Neyman and Scott 1958), in which *parents* form a homogeneous Poisson process and each parent generates a family of *offspring* that are spatially dispersed around their parent. Bartlett (1964) showed that in some cases the resulting process is indistinguishable from a Cox process; specifically, a process in which family sizes are independent Poisson variates and the positions of offspring relative to their parents are an independent random sample from a bivariate distribution with density $f(\cdot)$ is also a Cox proicess with stochastic intensity proportional to $\sum_{i=1}^{\infty} f(x - X_i)$, where the X_i are the points of a homogeneous Poisson process.

The most widely used model for an inhibitory process is a Markov point process (Ripley and Kelly 1977). A Markov point process can be defined by its likelihood ratio with respect to a Poisson process with intensity $\lambda = 1$. A useful sub-class of such processes is the *pairwise interaction process*, in which the likelihood ratio for a realization $\mathcal{X} = \{x_i : i = 1, \ldots, n\}$ is

$$\ell(\mathcal{X}) = \beta^n \prod_{j \neq i} h(||x_i - x_j||),$$

where $||\cdot||$ denotes distance, $h(\cdot)$ is an *interaction function* and $\beta > 0$ determines the intensity of the process. A sufficient condition for validity of the model is that $h(\cdot)$ is inhibitory, meaning that $0 \leq h(u) \leq 1$ for all u. The case $h(u) = 1$ yields a homogeneous Poisson process.

Inference

Until relatively recently, likelihood-based inference was considered intractable for most spatial point process models. Instead, sensible ad hoc methods based on functional summary statistics were used. These included so-called nearest neighbor methods and moment-based methods (Ripley 1977). Recent developments in Monte Carlo methods of inference have made likelihood-based inference a feasible, albeit computationally intensive, alternative (Møller and Waagepetersen 2004).

General accounts of statistical models and methods for spatial point pattern data include Diggle (2003) and Ilian et al. (2008).

About the Author

Peter Diggle is Distinguished University Professor of Statistics and Associate Dean for Research in the School of Health and Medicine, Lancaster University, Adjunct Professor in the Department of Biostatistics, Johns Hopkins University School of Public Health and Adjunct Senior Researcher in the International Research Insitute for Climate and Society, Columbia University. Between 1974 and 1983 he was a Lecturer, then Reader, in Statistics at the University of Newcastle upon Tyne. Between 1984 and 1988 he was Senior, then Principal, then Chief Research Scientist and Chief of the Division of Mathematics and Statistics at CSIRO, Australia. Peter's research interests are in the development of statistical methods for spatial and longitudinal data analysis, motivated by applications in the biomedical, health and environmental sciences. He has published 8 books and around 180 articles on these topics in the open literature. He was awarded the Royal Statistical Society's Guy Medal in Silver in 1997, is a former editor of the Society's Journal, Series B and is a Fellow of the American Statistical Association. Peter was founding co-editor, with his close friend and Johns Hopkins colleague Scott Zeger, of the journal *Biostatistics* between 1999 and 2009. He is a Trustee for Biometrika, and a member of the UK Medical Research Council's Population and Systems Medicine Research Board. Away from work, Peter

plays mixed-doubles badminton with his family (partner Amanda, children Jono and Hannah). He also enjoys music, playing guitar and recorder, and listening to jazz.

Cross References

►Analysis of Areal and Spatial Interaction Data
►Point Processes
►Poisson Distribution and Its Application in Statistics
►Poisson Processes
►Spatial Statistics

References and Further Reading

Bartlett MS (1964) The spectral analysis of two-dimensional point processes. Biometrika 51:299–311
Cox DR (1955) Some statistical methods related with series of events (with discussion). J R Stat Soc B 17:129–57
Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, London
Ilian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistial analysis and modelling of spatial point patterns. Wiley, Chichester
Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. Chapman & Hall, London
Neyman J, Scott EL (1958) Statistical approach to problems of cosmology. J R Stat Soc Ser B 20:1–43
Ripley BD (1977) Modelling spatial patterns (with discussion). J R Stat Soc B 39:172–212
Ripley BD, Kelly FP (1977) Markov point processes. J Lond Math Soc 15:188–92

Spatial Statistics

Jürgen Pilz
Professor, Head of the Institute of Statistics
University of Klagenfurt, Klagenfurt, Austria

Introduction

Spatial statistics is concerned with modeling and analysis of spatial data. By spatial data we mean data where, in addition to the (primary) phenomenon of interest the relative spatial locations of observations are recorded, too, because these may be important for the interpretation of data. This is of primary importance in earth-related sciences such as geography, geology, hydrology, ecology and environmental sciences, but also in other scientific disciplines concerned with spatial variations and patterns such as astrophysics, economics, agriculture, forestry, epidemiology and, at a microscopic scale, medical and health research.

In contrast to non-spatial data analysis, which is concerned with statistical modelling and analysis of data which just happen to phenomena in space and time, spatial statistics focuses on methods and techniques which consider explicitly the importance of the locations, or the spatial arrangement of the objects being analysed. The basic difference from classical statistics is that in spatial statistics we are concerned with non-independence of observations.

In spatial problems, observations come from a spatial random process $\mathcal{Z} = \{Z(s) : s \in S\}$, indexed by a spatial/spatiotemporal set $S \subset \mathbb{R}^d$, with $Z(s)$ taking values in some state space. The positions of observation sites $s \in S$ are either fixed in advance or random. Typically, $S \subset \mathbb{R}^2$, the study of spatial dynamics adds a temporal dimension, i.e., $S \subset \mathbb{R}^2 \times (0, \infty)$. However, S could also be one-dimensional (e.g., field trials along transect lines) or a subset of $\mathbb{R}^3$ (oil and mineral prospection, 3D imaging). In some fields such as Bayesian data analysis and simulation one even requires spaces S of dimension $d \geq 3$, this pertains, in particular, to the design and analysis of computer experiments with a moderate to large number of input variables. Comprehensive treatments of the whole field of spatial statistics are given in Ripley (1988), Cressie (1993) and Gaetan and Guyon (2010). Statistical Methods for spatio-temporal systems are given in Finkenstädt et al. (2007).

Basically, there are four classes of problems which spatial statistics is concerned with: point pattern analysis, geostatistical data analysis, areal/lattice data analysis and spatial interaction analysis. These subproblems are treated separately in a number of papers in this volume: Mase (2010), Kazianka and Pilz (2010), Vere-Jones (2010), Diggle (2010) and Spöck and Pilz (2010). Therefore, in this paper we limit ourselves to a brief overview over the areas comprising spatial statistics.

For a good overview on software for different problem areas of spatial data analysis we recommend the book by Bivand et al. (2008), for the important issue of simulation of spatial models we refer to Lantuéjoul (2002) and Gaetan and Guyon (2010).

Geostatistics

Here, S is a *continuous* subspace of $\mathbb{R}^d$ and the random field is observed at n fixed sites $\{s_1, \ldots, s_n\} \subset S$. Typical examples include rainfall data, data on soil, characteristics (porosity, humidity etc.), oil and mineral exploration data, airquality and groundwater data a.s.o. For $d \geq 2$ the random process $\mathcal{Z} = \{Z(s) : s \in S\}$ is usually termed a *random field*. The mathematical structure and the most important properties of random fields are described in Moklyachuk (2010).

The concept of stationarity is key in the analysis of spatial and/or temporal variation: roughly spoken, stationarity means that the statistical properties. (e.g., mean and

variance) of the variable of interest do not change over the considered area. However, testing for stationarity is not possible. For spatial prediction the performance of a stationary and a nonstationary model could be compared through assessment of the accuracy of predictions.

The random field is characterised by its finite dimensional distributions $P(Z(s_1) \leq z_1, \ldots, Z(s_n) \leq z_n)$ for all $n \in \mathbb{N}$ and $s_j \in S; j = 1, \ldots, n$. If all these distributions are Gaussian then Z is called a *Gaussian random field* (GRF). A GRF is completely determined by its expectation (trend function) $m(s) = E(Z(s))$ and covariance function $C(s_1, s_2) = Cov(Z(s_1), Z(s_2))$. Contrary to traditional statistics, in a geostatistical setting we usually observe only one realization of Z at a finite number of locations $s_1, \ldots, s_n$. Therefore, the distribution underlying the random field cannot be inferred without imposing further assumptions. The most simple assumption is that of (strict) stationarity, which means that the finite dimensional distributions do not change when all positions are translated by the same (lag) vector h, i.e., $(Z(s_1), \ldots, Z(s_n))$ and $(Z(s_1 + h), \ldots, Z(s_n + h))$ are identically distributed for all $n \in \mathbb{N}$ and locations $s_j \in S; j = 1, \ldots, n$. For a GRF this implies that $m(s) = \text{const}$ for all $s \in S$, and $C(s_1, s_2) = C(s_1 - s_2)$ for all $s_1, s_2 \in S$. For arbitrary RF's, the invariance of the first two moments is denoted as the property of *weak stationarity*. In geostatistics it is common to use the so-called semi-variogram $\gamma(s_1, s_2) = 0.5 * Var(Z(s + h) - Z(s))$ instead of the covariance function and to assume *intrinsic* stationarity: $m(s) = \text{const}$ and $\gamma(s, s + h) = \gamma(h)$ for all $s, h \in S$. If $Z(\cdot)$ is weakly stationary then $\gamma(h) = C(0) - C(h)$. Weak stationarity implies intrinsic stationarity, the converse is not true.

For $d = 1$, however, intrinsic stationarity is equivalent to weak stationarity of the first order differences of the underlying random process, a well-known fact from time series analysis. For an intrinsically stationary RF the semi-variogram has the important property of *conditional negative definiteness*, i.e.,

$$Var(a_1 Z(s_1) + \ldots + a_n Z(s_n)) = -\sum_{i=1}^{n} \sum_{j \neq i} a_i a_j \gamma(s_i - s_j) \geq 0$$

for all $n \in \mathbb{N}$ and real numbers $a_1, \ldots, a_n$ such that $\sum a_i = 0$. This is the reason why one usually employs parametric models (e.g., spherical, exponential, Gaussian or Matérn models) for fitting variogram functions to the data. Moreover, fitting is often done under the additional assumption of isotropy: $\gamma(h) = \gamma(|h|), |h| = \text{length of } h \in S$. For "classical" estimation methods for variogram parameters see Mase (2010), for Bayesian approaches we refer to Banerjee et al. (2004) and Kazianka and Pilz (2010). For non-stationary variogram modeling we refer to the review provided by Sampson et al. (2001) and Schabenberger and Gotway (2005).

Now, let us step to predicting Z at an unobserved location $s_0 \in S$, based on the observations $\mathbf{Z} := (Z(s_1), \ldots, Z(s_n))^T$, such that the mean squared error of prediction (MSEP) $E[Z(s_0) - \hat{Z}(s_0)]^2$ is minimized. For a GRF, the optimal predictor is known to be the mean of the conditional distribution of $Z(s_0)$ given the data:

$$\hat{Z}(s_0) = E(Z(s_0)|\mathbf{Z}) = E(Z(s_0)) + c_0^T K^{-1}(\mathbf{Z} - E(\mathbf{Z})) \quad (1)$$

where the vector c_0 has elements $C(s_0 - s_i); i = 1, \ldots, n$; and K is the covariance matrix of the observations. For non-Gaussian RF's, the predictor (1) is the best linear unbiased predictor (BLUP). Inserting the optimal estimators for $EZ(s_0)$ and $E(\mathbf{Z})$ into 1 we get various forms of Kriging predictors: assuming $EZ(s) = m$ to be constant we get $\widehat{EZ(s_0)} = \hat{m} = (\mathbf{1}^T K^{-1} \mathbf{Z})/(\mathbf{1} K^{-1} \mathbf{Z})$ and $E(\mathbf{Z}) = \hat{m}\mathbf{1}$, where $\mathbf{1}$ denotes the n-vector of one's, and this is known as the *ordinary Kriging* predictor. For non-constant m, assuming a linear regression setup for $m(s)$, one arrives at the *universal Kriging* predictor. Clearly, for non-Gaussian data, the best predictor w.r.t. MSEP is no longer linear in the observations. Comprehensive accounts of "classical" linear and nonlinear geostatistics are given in Chilés and Delfiner (1999) and Webster and Oliver (2007).

In a Bayesian setting, assuming a prior distribution for the covariance parameters, one has to determine the predictive density of $Z(s_0)|\mathbf{Z}$ via the posterior distribution of the covariance parameters given $\mathbf{Z}$, from which an optimal predictor and the associated uncertainty can be derived. For non-Gaussian data, the framework of generalized linear models or the copula framework can be used to arrive at optimal predictors (see Banerjee et al. (2004), Diggle and Ribeiro (2007) and Kazianka and Pilz (2010)). This extension of the classical geostatistical methodology has become known under the heading of *model-based geostatistics*. Concerning software for geostatistical analysis, we recommend the freely available R-packages "gstat," "geoR," "geoRglm" and the functions contained in the R-library "intamap." For spatio-temporal analysis and prediction of environmental processes we refer to Le and Zidek (2006) where also software is being described. For geostatistical space-time models particular care is needed for combining spatial and temporal variables (separability versus non-separability), a thorough treatment of this issue is given in Gneiting et al. (2007). A very exciting new development has been opened by Rue et al. (2009) who consider approximate Bayesian inference in latent Gaussian models, using an integrated nested Laplace approximation (INLA). This

approach offers computational advantages, the approximations are accurate and orders of magnitude faster than MCMC algorithms, and its generality also allows the computation of various predictive measures for doing model comparisons.

Point Process Analysis and Random Sets

By a (spatial) point process (PP) or point pattern we mean a random, locally finite collection $\mathcal{Z} = \{s_1, s_2, \ldots\}$ of points $s_i \in S \subset \mathbb{R}^d$ such that $s_i \neq s_j$ for $i \neq j$. Here, locally finite means that the number of points is finite in each bounded subset of S. The process is said to be *marked* if at each site s_i we additionally record a (random) value, for example the length of the material cracks, height or diameter of plants, intensity of earthquakes a.s.o. For statistical analysis, the process is observed in a window $W \subset S$ leading to a realization $z = \{s_1, \ldots, s_n\}$ with a random number $n = n(z)$ of points $s_i \in S$. Thus, contrary to geostatistical data analysis, in point pattern analysis the set of observation sites $\{s_1, \ldots, s_n\}$ is random, along with the number of sites n.

►Point processes are important in a variety of applications, in ecology and forestry (spatial, spatiotemporal distribution of plant/animal species), epidemiology (location of sick individuals, spatiotemporal spread of diseases), seismology (earthquake epicenters), materials science (locations of cracks and porosities), biology and medicine (centers of cells/tumours in histological sections), crime scene analysis (locations and intensities of burglaries) etc.

The probabilistic theory of PP's is quite technical and requires a good knowledge of measure theory, for a good introductory account we refer to the review articles by Møller and Waagepetersen (2007), Vere-Jones (2010) and Diggle (2010).

The PP $\mathcal{Z}$ is characterized through the finite-dimensional distributions $(N(B_1), \ldots, N(B_k))$ for all $k \in \mathbb{N}$ and bounded subsets $B_1, \ldots, B_k$ in $\mathbb{R}^d$, where the random variable $N(B_i)$ counts the number of points in B_i. The point pattern is called *stationary*, iff its finite-dimensional distributions are invariant under translations, and *isotropic* iff all these distributions are invariant under rotations.

One of the major problems is to find out whether a given point pattern can be considered as completely random, or if there is a tendency to clustering or to some "regularity." As the reference model for "no interaction between points" or "complete spatial randomness (CSR)" the *Poisson* Process (see ►Poisson Processes) is chosen (cf. Diggle 2010).

In general the mean structure of the count variables is modelled by a non-negative intensity function $\lambda(\cdot)$ such that $\mu(B) := \int_B \lambda(s)ds$ for all B in $\mathbb{R}^d$. Here the interpretation is that $\lambda(s)ds$ is the probability that there is precisely one point in the ball with center at s and area/volume ds. Likewise, the second order moment measure $\mu_2(A \times B) := E\{N(A)N(B)\}$ is modelled by a second order product density λ_2 such that $\mu_2(A \times B) = \int_A \int_B I_{A\times B}(u,v)\lambda_2(u,v)dudv$. For a Poisson PP one then has: $\mu_2(a \times B) = \mu(A)\mu(B), \lambda_2(u,v) = \lambda(u)\lambda(v)$.

The tendency of attraction or repulsion between points can be characterized by the so-called *pair correlation* function $g(u,v) := \lambda_2(u,v)/[\lambda(u)\lambda(v)]$. If points appear independently from each other then we have $\lambda_2(u,v) = \lambda(u)\lambda(v)$ and thus $g(u,v) = 1$. Thus, there is attraction between points of $\mathcal{Z}$ at locations u and v iff $g(u,v) > 1$ and repulsion iff $g(u,v) < 1$.

The characterization of point patterns becomes relatively easy in case of stationarity and additional isotropy. Then $\lambda(u) = \lambda = const$, $\lambda_2(u,v) = \lambda_2(|u-v|)$, $g(u,v) = g(|u-v|)$ and it suffices to work with the so-called K-function $K(r) = (1/\lambda)E$ {number of extra points within distance r of a randomly chosen point}. This takes the form

$$K(r) = (v_d/\lambda^2) \int_0^r u^{d-1}\lambda_2(u)du$$

where v_d stands for the surface area of the unit sphere in $\mathbb{R}^d$. For the Poisson PP in $\mathbb{R}^2$, for example, we have $K(r) = \pi r^2$. We remark, however, that second order moments and the related K function describe the dependence in point patterns only partly, i.e., the visual appearance of two point patterns may be different even if their first and second order moments are the same. Therefore, other features are considered as well, in particular the *empty space function* F_s and the *nearest neighbour function* G_s. The former is defined as $F_s(r) = P(N(b(s,r)) > 0)$, where $b(s,r)$ is the ball with radius $r > 0$ and centered at a fixed location $s \in \mathbb{R}^d$ (not necessarily $s \in \mathcal{Z}$). For a stationary PP the function F_s does not depend on s. The function G_s is the distribution function of the distance of a given point $s \in \mathcal{Z}$ to its nearest neighbour in $\mathcal{Z}$, i.e., $G_s(r) = P(N(b(s,r) > 1 | s \in \mathcal{Z})$. For the sake of comparison, the functions F and G are compared to those of a homogeneous Poisson (constant intensity) PP, for which $F(r) = 1 - exp(-\lambda|b(0,r)|) = G(r), r > 0$. Popular models of processes with dependence between points include the *Cox* PPs (less regular than Poisson PPs) and the *Gibbs* PPs (more regular than Poisson PPs). The Cox-process is defined by a two-stage model $Z|\zeta$ with random intensity $\mu(B) = \int \zeta(s)ds$ where ζ is a latent (non-observable) non-negative random field. For example, $\mathcal{Z}$ describes the (random) locations of the plants and ζ models the random environmental conditions at these

locations. Therefore, a Cox process is often termed a "doubly stochastic" Poisson PP (Poisson PP with random intensity). Assuming $\log \zeta(\cdot)$ to be a Gaussian RF leads to the widely used *log-Gaussian Cox process*: $\log \zeta(s) = g(s)^T\beta + \varepsilon(s)$, $g(s)$ includes the covariates, β is a parameter vector modeling (random) effects and $\varepsilon(s)$ is a centered Gaussian RF. Choosing $\zeta(s) = \lambda \sum_i k(s - s_i)$, where $\{s_1, s_2, \ldots\}$ form a stationary Poisson PP and $k(\cdot)$ is a density on S centered at $s_i \in \mathbb{R}^d$, we arrive at a so-called *Neyman–Scott process*. This way clustering tendencies can be modelled interpreting the points s_i as cluster centers (positions of parents) around which clusters with random numbers of descendants (children) are formed. Various special cases arise with particular choices of the density function $k(\cdot)$, choosing e.g., a Gaussian density results in a *Thomas* PP. The class of Cox models allows for many generalizations of Thomas and Neyman–Scott processes: different spatial configuration of the parents PP, interdependence (competition) and nonidentical distribution for children (variable fertility of parents) etc., all leading to aggregated PPs which are less regular than the Poisson PP.

One way to "regularize" a spatial point pattern is to disallow close points. This is appropriate for modeling situations such as tree distributions in forests and cell distributions in cellular tissues. These models are special cases of Gibbs models which are conditionally specified through the probabilities that there is a point at location s given the pattern on $\mathbb{R}^d \setminus \{s\}$: $\lambda(s|z)ds := P(N(b(s, ds)) = 1|\mathcal{Z} \cap (\mathbb{R}^d \setminus \{s\}) = z)$. The conditional intensity $\lambda(s|z)$ is usually modelled through some energy functional $U(s, z)$: $\lambda(s|z) = \exp(-U(s, z))$. For example, *Strauss* PP's correspond to the choice $U(s, z) = \exp\big(-a - b \sum_i I(\|s - s_i\| \le r)\big)$ including only the energy of the singletons and pair potentials. For $b > 0$ we have repulsion and, conversely, $b < 0$ implies attraction. We remark that the Strauss PPs are examples of *Markov* PPs since the conditional density $\lambda(s, z)$ depends only on neighboring points of s belonging to the pattern z.

For testing the CSR hypothesis, the parameters and functions introduced before $(\lambda, \lambda_2, K, F, G)$ have to be estimated on the basis of an observation window $W \subset \mathbb{R}^d$ (usually a (hyper-) rectangle). For testing this hypothesis, estimates of the following two summary statistics are in common use: $L(r) = \{K(r)/b_d\}^{1/d}$ and $J(r) = (1 - G(r))/(1-F(r))$, b_d denotes the volume of the unit sphere in $\mathbb{R}^d$. For a stationary PP, $J > 1, J = 1$ and $J > 1$ indicate respectively that the PP is more, equally or less regular than a Poisson PP. For estimation of G the well-known ►Kaplan–Meier-estimator can be used, for a comprehensive discussion of estimators and its properties we refer to Illian et al. (2008). Baddeley et al. (2006) present a number of interesting case studies in spatial point process modeling, in areas as diverse as human and animal epidemiology, materials sciences, social sciences, biology and seismology. For practical estimation and testing we recommend the freely available R-package "spatstat."

Random Sets

These are generalizations of point patterns in such a way that $\mathcal{Z}$ defines an arbitrary random closed subset (RACS) of $\mathbb{R}^d$. Again, stationarity means that the distributions of $\mathcal{Z}$ are invariant w.r.t. translations. In this case, random closed sets can be characterized by some simple numbers and functions, resp., e.g., by (a) the covariance function $C(h) = P(\{s \in \mathcal{Z}\} \cap \{s + h \in \mathcal{Z}\})$ and (b) the contact distribution $H_B(r) = 1 - P(\mathcal{Z} \cap rB = \emptyset)/(1 - P(s \in \mathcal{Z}))$ for some (test) set $B \subset \mathbb{R}^d$, e.g., a ball or polygon.

The most simple models for RACS are Boolean models, $\mathcal{Z} = \bigcup_{i=1}^{\infty} \{Z_i + s_i\}$, where $\{s_1, s_2, \ldots\}$ is a Poisson PP with constant intensity and $Z_1, Z_2, \ldots$ a sequence of i.i.d. RACS which are independent of the PP. For instance, Z_i can be assumed to be spheres with random radii, or segments of random length and direction. In applications, the random sets are not of that simple type. However, more realistic models can be built on the basis of Boolean models using the opening and closure operations of mathematical morphology, see e.g., Serra (1988) and Lantuéjoul (2002); for interesting applications in the materials sciences we refer to Ohser and Mücklich (2000).

Lattice Data Analysis

In areal/lattice data analysis we observe the random field $\mathcal{Z} = \{Z(s) : s \in S\}$ at the points of a fixed, discrete and non-random set $S \subset \mathbb{R}^d$, which is then often called a *lattice*. Then it is sufficient to describe the joint probability function or density on S. Typical examples of such type of data are population characteristics and infections disease numbers at district or country level, remote sensing imagery and image texture data from materials sciences. The lattice may be regularly or irregularly spaced. In areal data analysis, the measurements are aggregated over spatial zones (administrative units, land parcel sections) and the points s_i are geographical regions (areas) represented as a network with a given adjacency graph. In image analysis, the lattice S is a regularly spaced set of pixels or voxels. Goals of the analysis for these types of data include the quantification of spatial correlations, prediction, classification and synthesis of textures and image smoothing and reconstruction.

For areal data analysis usually autoregressive models are employed, the spatial correlation structure is induced by the particular model chosen, e.g., SAR or CAR models. For a detailed account of this type of analysis we refer to

Lloyd (2007) and Anselin and Rey (2010), for an overview and further references see Spöck and Pilz (2010). A particular area of lattice data analysis is image analysis where $d = 2$ (or 3), $S = \{1,\ldots,N\}^d$ and $N = 2^k$ for some integer $k > 1$. For modelling, Markov random fields are widely used. We call $\mathcal{Z} = \{Z(s) : s \in S\}$ a *Markov random field* if the conditional density of $Z(s)$ given $Z(y), y \neq s$, only depends on realizations of $Z(y)$ for which y belongs to some neighbourhood $\mathcal{N}(s)$ of s. As a simple example, consider a Gaussian Markov random field (GMRF). The neighborhood of s is usually defined via a symmetric neighborhood relation $s \sim y$ which is non-reflexive, i.e., $s \not\sim s$. Then the joint density on S can be written as $p(\boldsymbol{z}) \propto \exp(-0.5(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}))$ and the conditional density of $Z(s)$ given $Z(y), y \neq s$, is easily seen to be normal with expectation

$$E(Z(s)|Z(y) = z_y, y \in S \setminus \{s\}) = \mu_s - \frac{1}{a_{ss}} \sum_{y \neq s} a_{sy}(z_y - \mu_y)$$

and variance $1/a_{ss}$, where $\mu_y = E(Z(y))$ and a_{sy} denotes the element of the inverse of $\Sigma = (Cov(Z(s), Z(y)))_{s,y \in S}$. Therefore, a Gaussian RF is Markovian iff $a_{sy} \neq 0 \rightarrow y \in \mathcal{N}(s)$, i.e., iff Σ^{-1} is sparse. For a detailed account of GMRF we refer to Rue and Held (2005). According to the Hammersley–Clifford theorem (see Besag (1974)), MRF can be characterized as *Gibbs* RFs with local interaction potentials. The state space of a Gibbs random field can be rather general: $\mathbb{N}$ for count variables, e.g., in epidemiology, $\mathbb{R}^+$ for a positive-valued RF, e.g., a Gamma RF, a finite set of labels for categorical RFs, as e.g., in texture analysis, $\{0,1\}$ for binary RFs labeling presence or absence or alternative configurations as in *Ising models*, $\mathbb{R}^d$ for GRF, or mixtures of qualitative and quantitative states. Gibbs RFs are associated with families of conditional distributions p_Φ defined w.r.t. interaction potentials $\Phi = \{\phi_A, A \in \mathcal{S}\}$ where $\mathcal{S}$ is a family of finite subsets of S. In Bayesian image restoration, with $k > 2$ qualitative states (e.g., colours, textures or features) and finite set $S = \{0,1,\ldots,255\}^2$ one often uses models of the form $p_\Phi(z) \propto \exp(-U(z))$ where U stands for the energy associated with Φ. In the simplest case one has only one interaction parameter β and $U(z) = \beta \cdot n(z)$, where $n(z)$ is the number of points of neighbouring sites with the same state. Here β plays the role of a regularization parameter: decreasing β leads to more regularity. The central goal in (Bayesian) image and signal processing is then to reconstruct an object z based on a noisy observation y from the posterior $p_\Phi(\cdot|y)$ of $\mathcal{Z}$ given y, e.g., on the basis of the MAP = maximum (mode) of the a posteriori distribution.

A good summary of the theory and applications of image data analysis based on the theory of random fields is given in Li (1995) and Winkler (2003); for description, classification and simulation of 3D-image data we refer to Ohser and Schladitz (2009).

About the Author

For biography *see* the entry ►Statistical Design of Experiments.

Cross References

►Agriculture, Statistics in
►Analysis of Areal and Spatial Interaction Data
►Environmental Monitoring, Statistics Role in
►Geostatistics and Kriging Predictors
►Model-Based Geostatistics
►Point Processes
►Poisson Processes
►Random Field
►Spatial Point Pattern

References and Further Reading

Anselin L, Rey SJ (eds) (2010) Perspectives on spatial data analysis. Springer, Berlin

Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) (2006) Case studies in spatial point process modeling. Lecture notes in statistics, Springer, New York

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC Press, Boca Raton

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 36:192–236

Bivand RS, Pebesma EJ, Gomez-Rubio V (2008) Applied spatial data analysis with R. Springer, Berlin

Chilés J-P, Delfiner P (1999) Geostatistics. Modeling spatial uncertainty. Wiley, New York

Chilés J-P, Lantuéjoul Ch (2005) Prediction by conditional simulation: models and algorithms. In: Bilodeau M, Meyer F, Schmitt M (eds) Space structure and randomness. Lecture notes in statistics, vol 183. Springer, Berlin, pp 39–68

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, London

Diggle P (2010) Spatial pattern, this volume

Diggle PJ, Ribeiro PJ (2007) Model-based Geostatistics. Springer, New York

Finkenstädt B, Held L, Isham V (eds) (2007) Statistical methods for spatio-temporal systems. Chapman & Hall/CRC, Boca Raton

Gaetan C, Guyon H (2010) Spatial statistics and modeling. Springer, New York

Gneiting T, Genton MG, Guttorp P (2007) Geostatistical space-time models, stationarity, separability and full symmetry. In: Finkenstäedt B et al (eds) Statistical methods for spatio-temporal systems. Chapman & Hall/CRC, Boca Raton

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Wiley, New York

Kazianka H, Pilz J (2010) Model-based geostatistics. this volume

Lantuéjoul Ch (2002) Geostatistical simulation. Models and algorithms. Springer, Berlin

Le ND, Zidek JV (2006) Statistical analysis of environmental space-time processes. Springer, New York
Li SZ (1995) Markov random field modeling in computer vision. Springer, Tokyo
Lloyd ChD (2007) Local models for spatial analyis. CRC Press, Boca Raton
Mase S (2010) Geostatistics and kriging predictors. this volume
Moklyachuk MP (2010) Random field. this volume
Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. Chapman & Hall/CRC, Boca Raton
Møller J, Waagepetersen RP (2007) Modern statistics for spatial point processes. Scand J Stat 34:643–684
Ohser J, Mücklich F (2000) Statistical analysis of microstructures in materials science. Statistics in practice. Wiley, Chichester
Ohser J, Schladitz K (2009) 3D Images of materials structures. Wiley, Weinheim
Ripley BD (1981) Spatial statistics. Wiley, New York
Ripley BD (1988) Statistical inference for spatial processes. Cambridge University Press, Cambridge
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC Press, Boca Raton
Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested laplace approximations. J R Stat Soc B 71:1–35
Sampson PD, Damien D, Guttorp P (2001) Advances in modelling and inference for environmental processes with nonstationary spatial covariance. In: Monestiez P, Allard D, Froidevaux R (eds) GeoENV III: Geostatistics for environmental applications. Kluwer, Dordrecht, pp 17–32
Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC Press, Boca Raton
Serra J (ed) (1988) Image analysis and mathematical morphology. Theoretical advances. Academic, London
Spöck G, Pilz J (2010) Analysis of areal and spatial interaction data. this volume
Stein M (1999) Interpolation of spatial data. Springer, New York
Vere-Jones D (2010) Point processes, this volume
Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester
Winkler G (2003) Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction, 2nd edn. Springer, New York

Spectral Analysis

Peter Naeve
Professor Emeritus
University of Bielefeld, Bielefeld, Germany

Introduction

The term *spectral analysis* surely for most of us is connected with the experiment where a beam of sunlight is sent through a prism and split into many components of different colors, the spectrum. What looks nice is the starting point of a deeper understanding of nature, too.

The idea of splitting into components was copied by statisticians when working on time series. At first they proceeded like Kepler, who found his rules by fitting a model to data gathered by Tycho de Brahe. Deterministic modeling is a standard procedure in time series analysis. Given an economic time series x_t, one tries to fit $x_t = G_t + Z_t + S_t + R_t$ where G stands for trend, Z is a cyclic component, S a seasonal component, and R stands for the rest, the so-called noise. Regression is the important tool to study these models. The book by Davis still is a good starter. Unfortunately, this approach is not always as successful as with Kepler, "too many suns," Hotelling once complained.

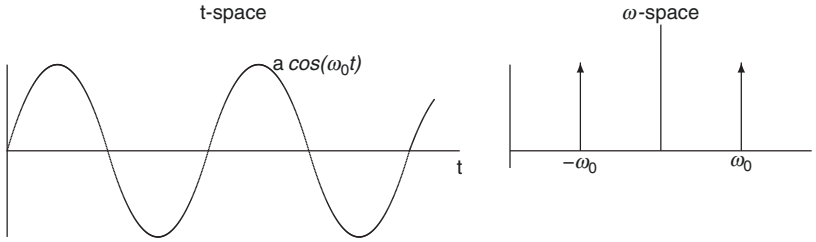
Quite another approach is to interpret a time series $\{x_t\}_{t\epsilon T}$ as a realization of a stochastic process $\{X(t)\}_{t\epsilon T}$. From now on we assume T to be a countable set. Then we might go in the direction of ARIMA-models – see, for instance, the book by Box and Jenkins – or choose spectral analysis as we will do here. So we are looking for a prism to work with.

A stochastic process is based on a system $F_n(u_1, \ldots, u_n; t_1, \ldots, t_n)$ of distribution functions. For these functions certain rules are valid, i.e., symmetric conditions $F_2(u_1, u_2; t_1, t_2) = F_2(u_2, u_1 : t_2, t_1)$, or consistency conditions such as $F_1(u_1; t_1) = F_2(u_1, \infty; t_1, t_2)$. Let E stand for the expectation operator. Then the mean function of the process is defined as $M(t) = E[X(t)]$ and the (auto-)covariance function as $C(t_1, t_2) = E[X(t_1)X(t_2)]$. A process is stationary if $M(t) = m$ and $C(t, s) = C(t-s) = C(\tau)$ for all $t, s\epsilon T$.

For such stationary processes the autocovariance function can be represented as $C(\tau) = \int e^{i\tau\omega} dF(\omega)$. The function $F(\omega)$ is called *spectral distribution*. When we have $dF(\omega) = f(\omega)d\omega$ the function $f(\omega)$ is called *spectral density*. The integration borders are $-\infty, \infty$ for continuous index set T and π, π for countable T. As can be seen by $C(0) = \int dF(\omega)$, the spectral distribution splits the variance into components. $dF(\omega)$ is the contribution to the variance of the frequencies in the interval between ω and $\omega + d\omega$. Such a stationary process can be written as $X(t) = \int e^{it\omega} dZ(\omega)$. For $\omega_1 \neq \omega_j$ $dZ(\omega_i)$, $dZ(\omega_j)$ are orthogonal random variables with $E[dZ(\omega)\overline{dZ(\omega)} = dF(\omega)$. So the process $\{X(t)\}_{t\epsilon T}$ is split into orthogonal components $e^{it\omega} dZ(\omega)$.

What can be gained by spectral analysis may be seen by two simple examples.

Example 1 Firstly, take the process $\{X(t)\} = \{\xi \cos \omega_0 t + \eta \sin \omega_0 t\}$ where ξ and η are random variables with $E[\xi] = E[\eta] = 0$, $E[\xi^2] = E[\eta^2] = c$, and $E[\xi\eta] = 0$. The object is to get information about ω_0. The covariance function of this process is $C(\tau) = c \cos \omega_0 \tau$. In Fig. 1 the function C and the corresponding spectral density, $c\pi\{\delta(\omega - \omega_0) +$



Spectral Analysis. Fig. 1 Covariance function (left) Spectral density (right)

$\delta(\omega + \omega_0)\}$, demonstrate how the latter provides a much clearer picture of the structure of the process.

Example 2 Next let us take a stationary process $\{X(t)\}_{t\epsilon T}$ with autocovariance function $C_X(\tau)$ and spectral density $f_X(\omega)$. $Y(t)_{t\epsilon T}$ is a linear time invariant transformation of $\{X(t)\}_{t\epsilon T}$. If $w(t)$ is the impulse function of the transformation, we have $Y(t) = \int_{-\infty}^{\infty} w(\tau)X(t-\tau)d\tau$. Doing some mathematics, we get for the autocovariance function $C_Y(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(\tau_1)w(\tau_2)C_X(\tau-\tau_1-\tau_2)d\tau_1 d\tau_2$. Turning to the spectral densities of the processes, we get $f_Y(\omega) = |\phi(\omega)|^2 f_X(\omega)$, with $\phi(\omega) = \int_{-\infty}^{\infty} w(\tau)e^{i\tau\omega}d\tau$, a nice, simple multiplication of a spectral density with the square of a Fourier transform.

From now on we assume that we deal with discrete stationary processes. For these the covariance function $C(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega}f(\omega)d\omega$ and the spectral density $f(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} e^{-i\tau\omega}C(\tau)$ are a pair of Fourier transforms that are the base for further steps.

Estimation of the Spectral Density

In applications we usually don't have the full ensemble but only one member – a piece of a member – of the sample space. To go on, we have to assume that the process $\{X(t)\}_{t\epsilon T}$ is ergodic. That is, $\lim_{T_0\to 0}\frac{1}{T_0}\sum_{t=1}^{T_0} X(t) = E(X(t)$ (mean ergodic) and $\lim_{T_0\to 0}\frac{1}{T_0}\sum_{t=1}^{T_0} X(t+\tau)X(t) = E(X(t+\tau)X(t)$ (covariance ergodic). In both cases, the convergence is in quadratic mean. A simple sufficient condition for mean ergodic is $|C(\tau)| < \epsilon$, i.e., events far away are not correlated – might be true in many applications. For covariance ergodic the same must be true for the process $Z(t) = X(t+\tau)X(t)$.

To get an estimate for the spectral density there are two approaches. Either one starts with an estimate of the covariance function and take its Fourier transform as an estimate for the spectral density. Or one starts from the representation $X(t) = \int e^{it\omega}dZ(\omega)$ and $E[dZ(\omega)\overline{dZ(\omega)} = dF(\omega)$. The so-called periodogram $P_n(\omega) = \frac{1}{2\pi n}|\sum_{t=1}^{n} x(t)e^{it\omega}|^2$ combines these features. This approach is backed by the fast Fourier transform (FFT). Cooley and Tukey found this famous algorithm.

In each case, applying spectral analysis to time series of finite length leads to a lot of problems. So we only have estimates $C(\tau)$ for $|\tau| \le \tau_0$. Theory calls for an estimator for all τ. A function $L(\tau)$ with $L(0) = 1$, $L(\tau) = L(-\tau)$ for $|\tau| \le \tau_0$, and $L(\tau) = 0$ elsewhere may be a solution. $\tilde{C}(\tau) = L(\tau)C(\tau)$ is defined for all τ. Further problems emerge immediately. How does one choose τ_0? Is this estimator unbiased, consistent? What is a good $L(\tau)$? And so on. Theoretically, these questions are hard to solve. Simulation is an aid in studying these problems. The book by Jenkins and Watts may be a good introduction to this approach.

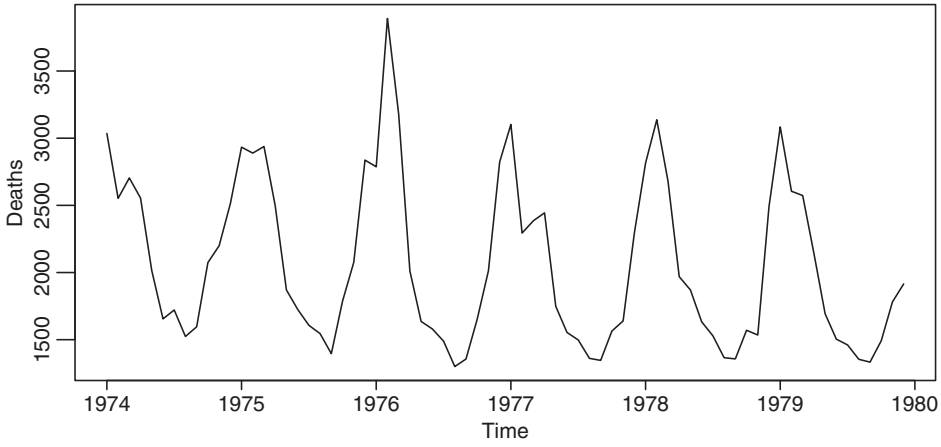
Multivariate Spectral Analysis

The simplest cases of multiple spectral analysis are two stochastic processes, $\{X(t)\}_{t\epsilon T}$ and $\{Y(t)\}_{t\epsilon T}$. The base of our analysis is the cross-variance function $C_{XY}(t_1, t_2) = E[X(t_1)Y(t_2)] = C_{XY}(t_1 - t_2)$. For this function we have the representation $C_{xy}(\tau) = \int e^{i\tau\omega}dF_{XY}(\omega)$. From $C_{xy}(\tau) = \int e^{i\tau\omega}dF_{XY}(\omega)$ we get the complex cross-spectral density $f_{XY}(\omega) = k(\omega) + iq(\omega)$ $k(\omega)$ is called co-spectrum and $q(\omega)$ quadrature spectrum. A number of functions are based on these two spectra, e.g., the amplitude $A(\omega) = \sqrt{\{k(\omega)\}^2 + \{q(\omega)\}^2}$, the phase $\phi(\omega) = \arctan(q(\omega)/k(\omega))$, and the coherence $C(\omega) = \frac{A(\omega)}{f_X(\omega)f_Y(\omega)}$. Plots of these functions are nice tools to study the relation between $\{X(t)\}_{t\epsilon T}$ and $\{Y(t)\}_{t\epsilon T}$.

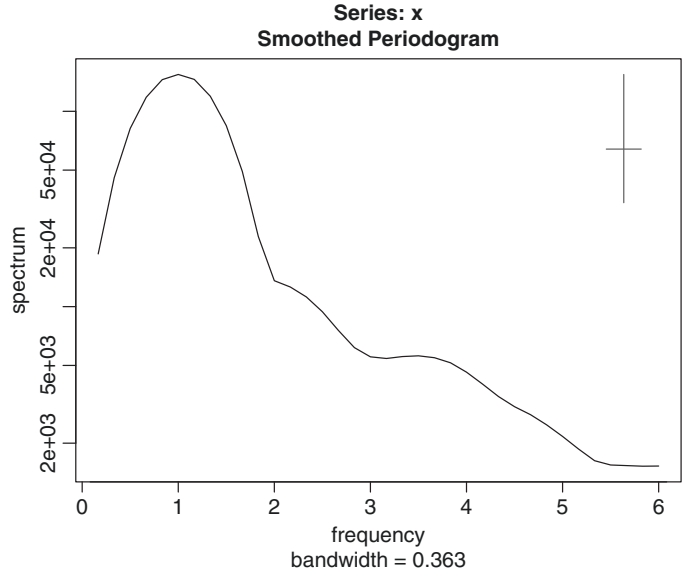
An Application

Finally we will deal with an application of spectral methods. This example is a very short version taken from the book by Venables and Ripley p. 355 f. The details are shown in Figs. 2 and 3. Figure 2 depicts the time series of monthly deaths from lung diseases in the UK 1974–1979. Figure 3 shows one estimate of the spectrum. All calculation were done with R. The function spectrum is based on FFT and smoothing by running means.

The interpretation of spectral functions and graphs calculated in applications is not an easy task. The book by Granger – the late Nobel Prize winner – might be a good starting place.



Spectral Analysis. Fig. 2 Time series



Spectral Analysis. Fig. 3 Spectrum

About the Author

Peter Naeve is Professor Emeritus at the University of Bielefeld since 2002. From 1979 on he held the Chair in Statistics and Computer Science at the Department of Economics. He is member of the International Statistical Institute, American Statistical Association, Royal Statistical Society, and International Association for Statistical Computing (IASC). His main field of interest is Computational Statistics. From the very beginning he was involved in the essential activities in this field, i.e., establishing an organized community (IASC), implementing a series of meetings (COMPSTAT Symposium on Computational Statistics), and providing a journal for this field. Among other positions he served as Co-Editor for the journal *Computational Statistics and Data Analysis* (1991–2000).

Cross References

►Box–Jenkins Time Series Models
►Stochastic Processes
►Stochastic Processes: Classification
►Time Series
►Time Series Regression

References and Further Reading

It is hard to sample a list of references from hundreds of books and an almost uncountable set of articles and discussion papers.

Being a good starting place was the (personally biased) criterion. When I entered the field, the books by Granger, Hatanka and Blackman, and Tukey were my first guides.
Blackman RB, Tukey JW (1959) The measurement of power spectra. Dover, New York
Box GEP, Jenkins GM (1970) Time series analysis forecasting and control. Holden-Day, San Francisco
Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comp 19:297–301
Davis HT (1963) The analysis of economic time series. Principia, San Antonio
Granger CWJ, Hatanaka M (1964) Spectral analysis of economic time series. Princeton University Press, Princeton
Jenkins GM, Watts DG (1968) Spectral analysis and Its applications. Holden-Day, San Francisco
Venables WN, Ripley BD (1994) Modern applied statistics with S-plus. Springer, New York
To assist those who like to Google, here are the names of some other pioneers: Bartlett, Parzen, Hannan, Priestley, Brillinger, Rosenblatt, Bingham

Sport, Statistics in

Stephen R. Clarke[1], John M. Norman[2]
[1]Professor
Swinburne University, Melbourne, VIC, Australia
[2]Emeritus Professor
Sheffield University, Sheffield, UK

Fans love statistics about sport – sets of numbers that describe and summarise what is happening on the field. With developments in computer technology, global positioning systems and the internet, the range and availability of sports statistics is growing at a rapid rate. In tennis majors, for example, an on-court statistician enters the result of every rally, whether the final shot was a forehand or backhand drive or volley, a winner or forced or unforced error, and whether either or both players were at the net. Cumulative results are immediately available to spectators, the media, and the general population through the internet. Only a few years ago, the number of kicks marks and handballs each player obtained in an Australian Rules football match was provided in printed tables two days after the match. Now over 80 statistics are collected in real time and immediately available to coaches and the general public. The science of statistics can be used to add value, to make sense, to discern patterns, to separate random variation from underlying trends in these sports data.

We are discussing here not just the collection and accumulation of statistics, but statistical modeling. Collection of raw statistics is one thing (how long is it since a batsman made over 400 in an international match? how old was Stanley Matthews when he played his last soccer game?) and statistical modeling (how can statistics be used) by analysts is another. If we are interested in the chance a male player might break 60 in a golf tournament next year, past statistics might tell us the percentage of all tournament rounds in which this has occurred. But if we want to estimate the chance Tiger Woods will break 60 in the US masters next year, this is of little use. We need to do some modeling. For example we might use past statistics to obtain Tiger's scores on each hole in previous masters, and by sampling from these use simulation to get a useful estimate.

Cricket has the distinction of being the first sport used for the illustration of statistics. In *Primer in Statistics*, (Elderton and Elderton 1909) used individual scores of batsmen to illustrate frequency distributions and elementary statistics. Some previous work on correlation and consistency resulted in (Wood 1945) and (Elderton 1945) reading separate papers at the same meeting of the Royal Statistical Society. These papers investigated the distribution of individual and pairs of batsmen scores, and have some claim as the first full quantitative papers applying statistics to sport.

The literature now contains hundreds of papers detailing applications of statistical modeling in virtually every sport. Researchers in the area are not confined to Statisticians. Other disciplines include Mathematics, Operational research, Engineering, Economics and Sports Science. Learned societies such as the American Statistical Association, the Australian Mathematical Society and the Institute of Mathematics and its Applications have sections of their membership or conferences devoted to this area. The range of journals which publish articles on sport often makes it difficult to search for previous work in a particular topic.

Much early work in the area is covered in the two texts (Machol et al. 1976) and (Ladany and Machol 1977). More recently (Bennett 1998) gives an excellent overview, with chapters on particular sports: American football, baseball, basketball, cricket, soccer, golf, ice hockey, tennis, track and field; and theme chapters on design of tournaments, statistical data graphics, predicting outcomes and hierarchical models. Later collections of papers include (Butenko et al. 2004) and (Albert and Koning 2008). These provide good examples of the issues currently being investigated by researchers. We discuss here some of these issues.

As mentioned above, fitting known distributions to sporting data was amongst the earliest work performed in this area. If the performance data follow a known distribution, that tells you something about the underlying behavior of the sportsman. If a batsman's cricket scores follow an exponential (or geometric) distribution, then he has a constant hazard, or probability of dismissal, throughout

his innings. If the number of successful shots a basketball player makes in a given number of tries can be modeled by the ▶Binomial distribution, then he has a constant probability of success, and is not affected by previous success or failure. If goals scored each match by a soccer team are Poisson distributed, this implies their form is not variable throughout the season, and they are not affected by early success or failure in a match. Departures from known distributions can be used to investigate the existence of the "hot hand" in basketball or baseball, or "momentum" in tennis or soccer.

Predicting the outcomes of sporting contests is of great interest to modelers and fans alike. Statistical modelers are usually interested in not only predicting the winner, but in estimating the chance of each participant winning and likely scores or margins. These predictions have become increasingly important with the introduction of sports betting. The estimated chances developed from the statistical model can be compared with the bookmaker's odds, and inefficiencies of betting markets investigated (or exploited). If the probabilities of head to head encounters can be estimated, then the chances of various outcomes of whole tournaments or competitions can be estimated via simulation.

A usual by-product of prediction is the rating of individuals or teams. For example a simple model might predict the winning margin between two teams as the difference in their ratings plus a home advantage. ▶Least squares, maximum likelihood or other methods are then used to obtain the ratings and home advantage that give the best fit to previous results. Chess has a rating system based on exponential smoothing that is applicable to past and present players from beginners to world champions. In golf, much effort has gone into developing ratings of players (handicaps) that are fair to players of all standards from all courses.

Home advantage, the degree to which a team performs better at home than away, is present in most sports. (Stefani and Clarke 1992) show that in balanced competitions the home side wins anywhere from 54% (baseball) to 70% (international soccer) of the matches. In scoring terms 1 goal in 3 in international soccer can be attributed to home advantage, while in baseball the home advantage contributes 1 run in 34. While home advantage can be quantified it is more difficult to isolate its causes. Many papers have looked at the effects of travel, crowd, ground familiarity and referee bias without much consensus. Other research has shown that models assuming a different home advantage for different teams or groups of teams provide a better fit to the data than ones with a common home advantage.

There are many different scoring systems in sport, (for example in racquet sports), and researchers are interested in their operating characteristics. To what extent do the scoring systems affect the probabilities of each player winning, and the distribution of the number of rallies in the match? What is the chance of winning from any scoreline? Generally the longer the match the more chance for the better player. For example, a player who wins 52% of the points at tennis, will win 55% of the games, 64% of the sets and 75% of 5 set matches. But the few breaks of serve in men's tennis makes the scoring system relatively inefficient. The better player may win a higher percentage of his serves than his opponent, but the set score still reaches 6 all. Researchers have suggested alternative scoring systems, such as 4-3 tennis, where the server still has to win 4 points to win the game, but the receiver only has to win 3 points. They have also looked at the importance of points – the change in a player's chance of winning the game (or match) resulting by winning or losing the point. (In tennis the most important point in a game is the service break point). The assertion that better players win the important points can then be tested.

What often makes sport interesting is the choice of alternative strategies. Should a baseball player try and steal a base or not? Should a footballer try for a field goal or a touchdown? Should a tennis player use a fast or slow serve? Should an orienteer choose a short steep route or a longer flatter one? When should the coach pull the goalie in ice-hockey? Operational Researchers find this a fertile field for study (Wright 2009), with techniques such as Dynamic Programming and simulation used to determine optimal strategies. (Norman 1995) gives one example of the use of Dynamic Programming in each of 12 sports.

Sport is an important area for the application of statistical modeling. Sport is big business, and occupies an important role in today's society. By the use of a range of modeling and analysis techniques Statisticians can assist players, coaches, administrators and fans to better understand and improve their performance and enjoyment.

About the Authors

Dr. Stephen Clarke is a Professor of Statistics in the faculty of Life and Social Sciences at Swinburne University, Melbourne, Australia. He has authored and co-authored more than 130 papers. He received the (U.K.) Operational Research Society president's medal in 1989 for his paper on one-day cricket.

John M. Norman is an emeritus professor at Sheffield University Management School, UK. He has written two books and fifty papers, several in collaboration with Stephen Clarke.

Cross References

►Binomial Distribution
►Poisson Distribution and Its Application in Statistics
►Record Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Albert J, Koning RH (eds) (2008) Statistical thinking in sports. Chapman & Hall, Boca Raton
Bennett J (ed) (1998) Statistics in sport. Arnold, London
Butenko S, Gil-Lafuente J et al (eds) (2004) Economics, management and optimization in sports. Springer-Verlag, Berlin
Elderton WE (1945) Cricket scores and some skew correlation distributions. J Roy Stat Soc (Ser A) 108:1–11
Elderton WP, Elderton EM (1909) Primer of statistics. Black, London
Ladany SP, Machol RE (1977) Optimal strategies in sports. North Holland, Amsterdam
Machol RE, Ladany SP et al (1976) Management science in sports. North Holland, New York
Norman JM (1995) Dynamic programming in sport: a survey of applications. IMA J Math Appl Bus Ind 6(December):171–176
Stefani RT, Clarke SR (1992) Predictions and home advantage for Australian rules football. J Appl Stat 19(2):251–261
Wood GH (1945) Cricket scores and geometrical progression. J Roy Stat Soc (Ser A) 108:12–22
Wright MB (2009) 50 years of OR in sport. J Oper Res Soc 60(S1):S161–S168

Spreadsheets in Statistics

Rade Stankic, Jasna Soldic-Aleksic
Professors, Faculty of Economics
Belgrade University, Belgrade, Serbia

Spreadsheet is a computer program that manipulates tables consisting of rows and columns of cells. It transforms a computer screen into a ledger sheet or grid of coded rows and columns simulating a paper worksheet. The program environment consists of one or more huge electronic worksheets (each worksheet can contain up to one million rows by a few thousands columns) organized in the form of an electronic workbook.

The general features of such programs are powerful computing and graphical capabilities, flexibility, excellent report generating feature, easy-to-use capability, and compatibility with many other data analytical software tools. These features are responsible for the substantial popularity and wide practical usage of the program. Thus, spreadsheet software is being used in academic, government, and business organizations for tasks that require summarizing, reporting, data analysis, and business modeling.

The spreadsheet concept became widely known in the late 1970s and early 1980s due to the Dan Bricklin's implementation of VisiCalc which is considered to be the first electronic spreadsheet. It was the first spreadsheet program that combined all essential features of modern spreadsheet applications, such as: WYSIWYG (*What You See Is What You Get*), interactive user interface, automatic recalculation, existence of status and formula lines, copy of cell range with relative and absolute references, and formula building by selecting referenced cells. Lotus 1–2–3 was the leading spreadsheet program in the period when DOS (Disk Operating System) prevailed as an operating system. Later on, Microsoft Excel took the lead and became the dominant spreadsheet program in the commercial electronic spreadsheet market.

The basic building blocks of a spreadsheet program are cells that represent the intersections of the rows and columns in a table. Each individual cell in the spreadsheet has a unique column and row identifier that takes specific forms in different spreadsheet programs. Thus, the top left-hand cell in the worksheet may be designated with symbols A1, 11, or 1A. The content of the cell may be a value (numerical or textual data) or a formula. When the formula is entered in a particular cell, it defines how the content of that cell is calculated and updated depending on the content of another cell (or combination of cells) that is/are referenced to in the formula. References can be relative (e.g., `A1`, or `C1:C3`), absolute (e.g., `$B$1`, or `$C$1:$C$3`), mixed row-wise or column-wise absolute/relative (e.g., `$B1` is column-wise absolute and `B$1` is row-wise absolute), three-dimensional (e.g., Sheet1!A1), or external (e.g., [Book1]Sheet1!A1). This well-defined structure of cell addresses enables a smooth data flow regardless whether data are stored in just one or several worksheets or workbooks. In most implementations, a cell (or range of cells) can be "named" enabling the user to refer to that cell (or cell range) by its name rather than by grid reference. Names must be unique within a spreadsheet, but when using multiple sheets in a spreadsheet file, an identically named cell range on each sheet can be used if it is distinguished by adding the sheet name. Name usage is primarily justified by the need for creating and running macros that repeat a command across many sheets.

What makes the spreadsheet program a powerful data analytical tool is the wide range of integrated data processing functions. Functions are organized into logically distinct groups, such as: *Arithmetic functions, Statistical functions*, *Logical functions*, *Financial functions*, *Date and Time functions*, *Text functions*, *Information*, *Mathematical function*, etc. In general, each function is determined by its name (written in uppercase by convention) and

appropriate argument(s) which is/are placed in parenthesis. The arguments are a set of values, separated by semicolons, to which the function applies. Thus, a function called *FUNCTION* would be written as follows: *FUNCTION* (argument1; argument2; etc.).

Spreadsheet software integrates a large number of built-in statistical functionalities, but some caveats about its statistical computations have been observed. A few authors have criticized the use of spreadsheets for statistical data processing and have presented some program shortcomings, such as: no log file or audit trail, inconsistent behavior of computational dialogs, poor handling of missing values, low-level of accuracy of built-in spreadsheet statistical calculations, and no sophisticated data coding techniques for specific statistical calculations. In response to such criticism directed against the statistical "incorrectness" and limitations of spreadsheet programs, many efforts have been made (both in the academic and commercial community) to compensate for them. Thus, many statistics add-ins have appeared, granting robust statistical power to the spreadsheet program environment. These add-ins are usually seamlessly integrated into a spreadsheet program and cover the range of most commonly used statistical procedures, such as: descriptive statistics, ►normality tests, group comparisons, correlation, regression analysis, forecast, etc. Some leading statistical software vendors have provided statistical modules and functionalities for spreadsheet users. For example, the statistical software package PASW Statistics 17.0 offered the following additional techniques and features for Excel spreadsheet program (*SPSS Advantage for Excel 2007*): Recency, Frequency, and Monetary value (RFM) analysis for direct marketing research (where most profitable customers are identified), classification tree analysis for group identification, unusual data detection, procedure for data preparation and transformation, and the option to save spreadsheet data as a statistical software data file.

One of the crucial spreadsheet package features is its capability to carry out "What-if" data analysis. "What-if" analysis is the process of observing and learning how the changes in some cells (as an input) affect the outcome of formulas (as an output) in the other cells in the worksheet. For example, Microsoft Excel provides the following "what-if" analytical tools: scenario manager, data tables, and Goal Seek. Scenario manager and data tables operate in a very simple way: they take sets of input values and determine possible results. While a data table works only with one or two variables, accepting many different values for those variables, a scenario manager can handle multiple variables, but has a limitation of accommodating only up to 32 values. These tools are appropriate for running the *sensitivity analysis*, which determines how a spreadsheet's output varies in response to changes to the input values. Contrary to the functioning of scenario manager and data tables, Goal Seek allows the user to compute a value for a spreadsheet input that makes the value of a given formula match a specified goal.

In the era of the Internet, networked computing, and web applications, online spreadsheet programs also came about. An online spreadsheet is a spreadsheet document edited through a web-based application that allows multiple users to have access, to edit and to share it online (multiple users can work with a spreadsheet, view changes in real time, and discuss changes). Equipped with a rich Internet application user interface, the best web-based online spreadsheets have many of the features seen in desktop spreadsheet applications and some of them have strong multiuser collaboration features. Also, there are spreadsheet programs that offer real time updates from remote sources. This feature allows updating of a cell's content when its value is derived from an external source - such as a cell in another "remote" spreadsheet. For shared, web-based spreadsheets, this results in the "immediate" updating of the content of cells that have been altered by another user and, also, in the updating of all dependent cells.

Cross References

►Statistical Software: An Overview

References and Further Reading

Albright SC, Winston WL, Zappe CJ (2009) Data analysis and decision making with Microsoft Excel, 3rd edn. South-Western Cengage Learning, USA

Monk EF, Davidson SW, Brady JA (2010) Problem-solving cases in Microsoft Access and Excel. Course technology. Cengage Learning, USA

Nash JC (2006) Spreadsheets in statistical practice – another look. Am Stat 60(3):287–289

Turban E, Ledner D, McLean E, Wetherbe J (2007a) Information technology for management: transforming organizations in the digital economy, 6th edn. Wiley, New York

Walkenbach J (2007) Excel *2007 Bible*. Wiley, New York

Winston WL (2004) Microsoft Excel – data analysis and business modeling. Microsoft, Washington

Spurious Correlation

Simon J. Sheather
Professor and Head of Department of Statistics
Texas A&M University, College Station, TX, USA

A well-known weakness of regression modeling based on observational data is that the observed association between

two variables may be because both are related to a third variable that has been omitted from the regression model. This phenomenon is commonly referred to as "spurious correlation." The term spurious correlation dates back to at least Pearson (1897).

Neyman (1952, pp. 143–154) provides an example based on fictitious data which dramatically illustrates spurious correlation. According to Kronmal (1993, p. 379), a fictitious friend of Neyman was interested in empirically examining the theory that storks bring babies and collected data on the number of women, babies born and storks in each of 50 counties. This fictitious data set was reported in Kronmal (1993, p. 383) and it can be found on the web page associated with Sheather (2009), namely, http://www.stat.tamu.edu/~sheather/book.

Figure 1 shows scatter plots of all three variables from the stork data set along with the least squares fits. Ignoring the data on the number of women and fitting the following straight-line regression model produces the output shown below.

$$\text{Babies} = \beta_0 + \beta_1 \text{Storks} + e \tag{1}$$

The regression output for model (1) shows that there is very strong evidence of a positive linear association between the number of storks and the number of babies born (p-value < 0.0001). However, to date we have ignored the data available on the other potential predictor variable, namely, the number of women.

Regression output for model (1)				
	Coefficients			
	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	4.3293	2.3225	1.864	0.068
Storks	3.6585	0.3475	10.528	1.71e-14 ***
Residual standard error: 5.451 on 52 degrees of freedom				
Multiple *R*-Squared: 0.6807, Adjusted *R*-squared: 0.6745				

Next we consider the other potential predictor variable, namely, the number of women. Thus, we consider the following regression model:

$$\text{Babies} = \beta_0 + \beta_1 \text{Storks} + \beta_2 \text{Women} + e \tag{2}$$

Given below is the output from *R* for a regression model (2). Notice that the estimated regression coefficient for the number of storks is zero to many decimal places. Thus, correlation between the number of babies and the number of storks calculated from (1) is said to be spurious as it is due to both variables being associated with the number of women. In other words, a predictor (the number of women) exists which is related to both the other predictor (the number of storks) and the outcome variable (the number of babies), and which accounts for all of the observed association between the latter two variables. The number of women predictor variable is commonly called either an omitted variable or a confounding covariate.

Regression output for model (2)				
	Coefficients			
	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	1.000*e*+01	2.021*e*+00	4.948	8.56*e*−06***
Women	5.000*e*+00	8.272*e*−01	6.045	1.74*e*−07***
Storks	−6.203*e*−16	6.619*e*−01	−9.37*e*−16	1
Residual standard error: 4.201 on 51 degrees of freedom				
Multiple *R*-Squared: 0.814, Adjusted *R*-squared: 0.8067				

We next briefly present some mathematics wish quantifies the effect of spurious correlation due to omitted variables. We shall consider the situation in which an important predictor is omitted from a regression model. We shall denote the omitted predictor variable by v and the predictor variable included in the one-predictor regression model by x. In the fictitious stork data x corresponds to the number of storks and v corresponds to the number of women.

To make things as straightforward as possible we shall consider the situation in which Y is related to two predictors x and v as follows:

$$Y = \beta_0 + \beta_1 x + \beta_2 v + e_{Y \cdot x,v} \tag{3}$$

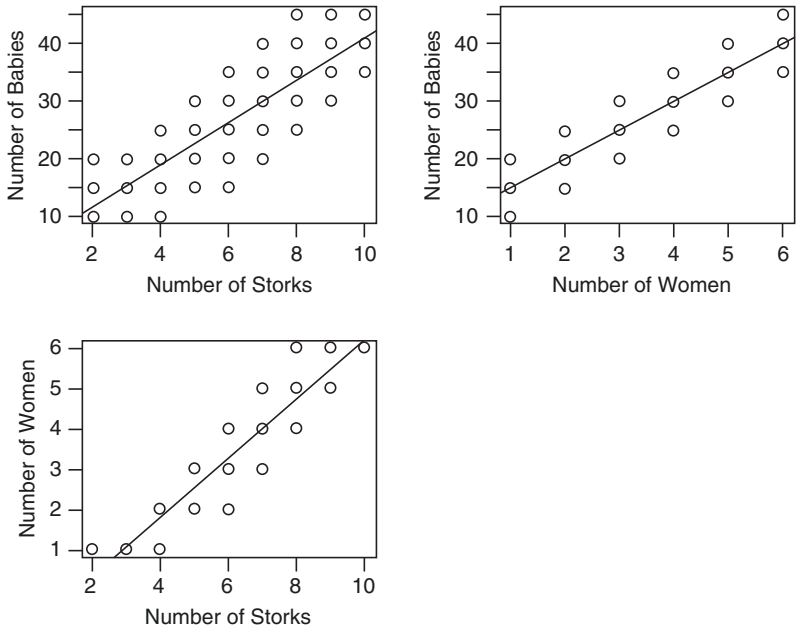
Similarly, suppose that v is related to x as follows:

$$v = \alpha_0 + \alpha_1 x + e_{v \cdot x} \tag{4}$$

Substituting (4) into (3) we will be able to discover what happens if omit v from the regression model. The result is as follows:

$$Y = (\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x + (e_{Y \cdot x,v} + \beta_2 e_{v \cdot x}) \tag{5}$$

Notice that the regression coefficient of x in (5) is the sum of two terms, namely, $\beta_1 + \beta_2 \alpha_1$.



Spurious Correlation. Fig. 1 A plot of the variables from the fictitious data on storks

We next consider two distinct cases:

1. $\alpha_1 = 0$ and/or $\beta_2 = 0$: Then the omitted variable has no effect on the regression model, which includes just x as a predictor.
2. $\alpha_1 \neq 0$ and $\beta_2 \neq 0$: Then the omitted variable has an effect on the regression model, which includes just x as a predictor. For example, Y and x can be strongly linearly associated (i.e., highly correlated) even when $\beta_1 = 0$. (This is exactly the situation in the fictitious stork data.) Alternatively, Y and x can be strongly negatively associated even when $\beta_1 > 0$.

Spurious correlation due to omitted variables is most problematic in observational studies. We next look at a real example, which exemplifies the issues. The example is based on a series of papers (Cochrane et al. 1978; Hinds 1974; Jayachandran and Jarvis 1986) that model the relationship between the prevalence of doctors and the infant mortality rate. The controversy was the subject of a 1978 Lancet editorial entitled "The anomaly that wouldn't go away." In the words of one of the authors of the original paper, Selwyn St. Leger (2001):

> When Archie Cochrane, Fred Moore and I conceived of trying to relate mortality in developed countries to measures of health service provision little did we imagine that it would set a hare running 20 years into the future. . . The hare was not that a statistical association between health service provision and mortality was absent. Rather it was the marked positive correlation between the prevalence of doctors and infant mortality. Whatever way we looked at our data we could not make that association disappear. Moreover, we could identify no plausible mechanism that would give rise to this association.

Kronmal (1993, p. 624) reports that Sankrithi et al. (1991) found a significant negative association ($p < 0.001$) between infant mortality rate and the prevalence of doctors after adjusting for population size. Thus, this spurious correlation was due to an omitted variable. In summary, the possibility of spurious correlation due to omitted variables should be considered when the temptation arises to over interpret the results of any regression analysis based on observational data. Stigler (2005) advises that we "discipline this predisposition (to accept the results of observational studies) by a heavy dose of skepticism."

About the Author

Professor Sheather is Head of the Department of Statistics, Texas A&M University. Prior to that he was the Head of the Statistics and Operations Group and Associate Dean of Research, Australian Graduate School of Management, at the University of New South Wales in Sydney, Australia. He is an Elected Fellow of the American Statistical Association (2001). Professor Sheather is currently listed on

ISIHighlyCited.com among the top one-half of one percent of all mathematical scientists, in terms of citations of his published work.

Cross References

►Causation and Causal Inference
►Confounding and Confounder Control
►Correlation Coefficient
►Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)
►Role of Statistics in Advancing Quantitative Education

References and Further Reading

Cochrane AL, St. Leger AS, Moore F (1978) Health service "input" and mortality "output" in developed countries. J Epidemiol Community Health 32:200–205
Hinds MW (1974) Fewer doctors and infant survival. New Engl J Med 291:741
Jayachandran J, Jarvis GK (1986) Socioeconomic development, medical care and nutrition as determinants of infant mortality in less developed countries. Social Biol 33:301–315
Kronmal RA (1993) Spurious correlation and the fallacy of the ratio standard revisited. J R Stat Soc A 156:379–392
Neyman J (1952) Lectures and conferences on mathematical statistics and probability, 2nd edn. US Department of Agriculture, Washington DC, pp 143–154
Pearson K (1897) Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 60:489–498
Sankrithi U, Emanuel I, Van Belle G (1991) Comparison of linear and exponential multivariate models for explaining national infant and child mortality. Int J Epidemiol 2:565–570
Sheather SJ (2009) A modern approach to regression with R. Springer, New York
St. Leger S (2001) The anomaly that finally went away? J Epidemiol Community Health 55:79
Stigler S (2005) Correlation and causation: a comment. Persp Biol Med 48(1 Suppl.):588–594

St. Petersburg Paradox

James M. Joyce
Chair and Professor of Philosophy and of Statistics
University of Michigan, Ann Arbor, MI, USA

The St. Petersburg "Paradox" concerns a betting situation in which a gambler's fortune will be increased by $\$2^n$ if the first tail appears on the *nth* toss a fair coin. Nicholas Bernoulli introduced this problem in 1713 as a challenge to the then prevailing view that the fair price of a wager (the price at which one should be equally happy to buy or sell it) is equal to its expected monetary payoff. While Bernoulli's wager has an infinite expected payoff, any reasonable person will sell it for $20. By 1727 Gabriel Cramer had recognized that the prevailing view goes wrong because it assumes that people value money linearly. As he wrote, "mathematicians evaluate money in proportion to its quantity while, in practice, people with common sense evaluate money in proportion to the (practical value) they can obtain from it" (Bernoulli 1954, p. 33). Since an extra increment of money buys less happiness for a prince than a pauper, Cramer observed, the St. Petersburg wager can have a finite "practical value" provided that the worth of an extra dollar falls off rapidly enough as a person's fortune grows. In modern terms, Cramer had understood that money has declining marginal utility and that the St. Petersburg wager can have a finite expected utility if the marginal decrease in utility is sufficiently steep. He noted, for example, that a utility function of the form $u(\$x) = x^{1/2}$ produces an expected utility of $\sum_n \left(\frac{1}{2}\right)^n 2^{n/2} \approx 2.41421$ for Bernoulli's wager, which is equivalent to a fair price of $5.83.

Cramer never published, and it was left to Daniel Bernoulli to report Cramer's contributions and to write the definitive treatment (1954) of his cousin Nicholas's problem in the *St. Petersburg Academy Proceedings* of 1738, from which the Paradox derives its name. Daniel, who hit upon the declining utility of money independently of Cramer, went further by advocating the general principle that rational agents should value wagers according to their expected utility. He also argued that a person's marginal utility for an extra sum of money should be both inversely proportional to the person's fortune and directly proportional to the size of the sum. This means that the utility of $x is a function of the form $u(\$x) = k \cdot \ln(x)$. When evaluated using such a utility function, the St. Petersburg wager has a finite expected utility of $k \cdot \ln(4)$.

Bernoulli was also explicit that, as a general matter, the value of any gamble is its expected utility, and not its expected payoff. Specifically, he maintained that if the utility function $u(x)$ measures the "practical value" of having fortune $x, then the value of any wager X is $E(u(X)) = \int_0^1 P(X = x) \cdot u(x)dx$ and its fair price is that sum $\$f$ such that $u(f) = E(u(X))$. Though this was perhaps Bernoulli's deepest insight, its implications were not fully appreciated until the early 1950s when the work of Savage (1954) and von Neumann and Morgenstern (1953) moved the hypothesis of expected utility maximization to the very center of both microeconomics and ►Bayesian statistics.

Until that time, Bernoulli was better known among economists and statisticians for postulating that money has declining marginal utility and for solving the St. Petersburg

Paradox. The thesis that money has declining marginal utility has been immensely influential since it serves as the basis for the standard theory of risk aversion, which explains a wide variety of economic phenomena. In economic parlance, a *risk averse* agent prefers a straight payment of a gamble's expected payoff to the gamble itself. Economists seek to explain risk aversion by postulating *concave* utility functions for money, with greater concavity signaling more aversion. If $u(x)$ is concave for $a \leq x \leq b$, and if a wager X's payouts are confined to $[a, b]$, then it is automatic that $E(u(X)) \geq u(E(X))$. Moreover, if v is a concave transformation of u, the absolute risk aversion associated with v exceeds that associated with u, where absolute risk aversion is measured by the Arrow (1965)–Pratt (1964) coefficient $v''(x)/v'(x)$. Agents with Bernoulli's logarithmic utility are everywhere risk averse, and their absolute level of risk aversion decreases with increases in x since $u''(x)/u'(x) = 1/x$.

Interestingly, the Cramer/Bernoulli solution to the St. Petersburg Paradox failed the test of time. As Karl Menger (1934) first recognized (Basset 1987), if money has *unbounded* utility then one can always construct a "Super St. Petersburg Paradox." For example, using $u(\$x) = \ln(x)$, a wager that pays $e^2, e^4, e^8, \ldots$ if a tail appears first on the 1st, 2nd, 3rd,... toss will have infinite expected utility. One can avoid this either by insisting that realistic utility functions are bounded or by restricting the allowable gambles so that events of high utility are always assigned such low probabilities that gambles with infinite expected utilities never arise. On either view, the St. Petersburg Paradox ceases to be a problem since there is no chance that anyone will ever face it. Most standard treatments, e.g., (Ingersoll 1978), endorse bounded utility functions on the grounds that arbitrarily large payoffs are impossible in a finite economy. Others, who want to leave open the theoretical possibility of unbounded utility, require all realizable wagers to be limits of wagers with uniformly bounded support, where limits are taken in the weak topology. For a well-developed approach of this sort see (Kreps 1988, pp. 63–68).

About the Author

James M. Joyce is Professor of Philosophy and Statistics at the University of Michigan, Ann Arbor. He is the author of *The Foundations of Causal Decision Theory* (Cambridge Studies in Probability, Induction and Decision Theory, Cambridge University Press, 1999), as well as a number of articles on decision theory and Bayesian approaches to epistemology and the philosophy of science.

Cross References

►Statistics and Gambling

References and Further Reading

Arrow KJ (1965) Aspects of the theory of risk-bearing. Markham, Chicago

Basset GW (1987) The St. Petersburg paradox and bounded utility. Hist Polit Econ 19:517–523

Bernoulli D (1738) Specimen theoriae de mensura sortis. Commentarii academiae scientiarum imperialis petropolitanae. In: Proceedings of the royal academy of science, St. Petersburg. English translation (1954) by Louise Sommer with notes by Karl Menger. Exposition of a new theory on the measurement of risk. Econometrica 22:23–36

Ingersoll J (1978) Theory of financial decision making. Rowman and Littlefield, Oxford

Kreps D (1988) Notes on the theory of choice. Westview, Boulder

Menger K (1934) Das unsicherheitsmoment in der wertlehre. Zeitschrift für Nationalöcokonomie 51:459–485

Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136

Savage LJ (1954) The foundations of statistics. Wiley, New York

von Neumann J, Morgenstern O (1953) Theory of games and economic behavior, 3rd edn. Princeton University Press, Princeton

Standard Deviation

Sekander Hayat Khan M.
Professor of Statistics
Institute of Statistical Research and Training
University of Dhaka, Dhaka, Bangladesh

Introduction

Standard deviation is a measure of variability or dispersion. The term *Standard deviation* was first used in writing by Karl Pearson in 1894. This was a replacement for earlier alternative names for the same idea: for example, "mean error" (Gauss), "mean square error," and "error of mean square" (Airy) have all been used to denote standard deviation. Standard deviation is the most useful and most frequently used measure of dispersion. It is expressed in the same units as the data. Standard deviation is a number between 0 and ∞. A large standard deviation indicates that observations/data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

Definition

If X is a random variable with mean value $\mu = E(x)$, the standard deviation of X is defined by

$$\sigma = \sqrt{E(X-\mu)^2}. \quad (1)$$

That is, the standard deviation σ is the square root of the average value of $(X-\mu)^2$. The standard deviation of a continuous real-valued random variable X with probability density function $f(x)$ is

$$\sigma = \sqrt{\int (x-\mu)^2 f(x)dx}, \tag{2}$$

where $\mu = \int xf(x)dx$, and the integrals are the definite integrals taken over the range of X. If the variable X is discrete with probability function $f(x)$, the integral signs are replaced by summation signs.

In the case where X takes random values from a finite data set $x_1, x_2, \cdots, x_N$, the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}, \tag{3}$$

where μ is the mean of X.

Estimation

For estimating the standard deviation from sample observations, μ in Eq. 3 is to be replaced by the sample mean $\bar{x}$ given by $\bar{x} = \sum_{i=1}^{n} x_i/n$, and then it is denoted by s_n. This s_n is the maximum likelihood estimate of σ when the population is normally distributed.

For estimating the standard deviation from a small sample, the sample standard deviation, denoted by s, can be computed by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{4}$$

where $\{x_1, x_2, \cdots, x_n\}$ is the sample, and $\bar{x}$ is the sample mean. This correction (use of $n-1$ instead of n), known as Bessel's correction, makes s^2 an unbiased estimator for the variance σ^2.

It can be shown that $\hat{\sigma}$ = IQR/1.35, where IQR is the interquartile range of the sample, is a consistent estimate of σ. The asymptotic relative efficiency of this estimator with respect to sample standard deviation is 0.37. It is, therefore, better to use sample standard deviation for normal data, while $\hat{\sigma}$ can be more efficient when the distribution of data is with thicker tail[3]. Standard deviation is independent of change of origin but not of scale.

Interpretation and Application

Standard deviation is the most useful and frequently used measure of dispersion. Standard deviation is used both as a separate entity and as a part of other analyses, such as computing confidence intervals and in hypotheses testing. Standard deviation is zero if all the elements of a population or data set are identical. It becomes larger if the data tend to spread over a larger range of values.

In science, researchers use standard deviation of experimental data for testing statistical significance. σ and $\hat{\sigma}$ are used in making certain tests of statistical significance. Standard deviation of a group of repeated measurements gives the precision of those measurements. In finance, it is used as a measure of risk on an investment. Standard deviation can be used to examine if a professional is consistent in his work. Similarly, standard deviation of scores (runs) made by a cricket player in a season tells about the consistency in his performance.

Standard deviation of an estimate, called the *Standard error*, is used to have an idea of the precision of that estimate.

►Chebyshev's inequality, (which enables to find probability without knowing probability function of a random variable), throws light on the connection between standard deviation and dispersion. For all distributions for which standard distribution is defined, it states that at least $\left(1-\frac{1}{k^2}\right)100\%$ of the values are within k standard deviation from the mean.

About the Author

Professor Khan is Editor of the *Journal of Statistical Research* (JSR), official publication of the Institute of Statistical Research and Training, University of Dhaka, Bangladesh.

Cross References

►Chebyshev's Inequality
►Coefficient of Variation
►Portfolio Theory
►Variance

References and Further Reading

Pearson Karl (1894) On the dissection of asymmetrical curves. Philos Tr R Soc S-A 185:719–810

Miller J. Earliest known uses of some of the words of mathematics. http://jeff560.tripod.com/mathword.html

Das Gupta A, Haff L (2006) Asymptotic expansions for correlations between measures of spread. J Stat Plan Infer 136: 2197–2213

Yule GU, Kendall MG (1958) An introduction to the theory of statistics, 14th edn. 3rd Impression. Charles Griffin & Company, London

Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences

David S. Jones
Professor of Biomaterial Science, Chair of Biomaterial Science
School of Pharmacy, Queens University of Belfast, Belfast, UK

Introduction

Essential to the efficacy of performance of drug delivery systems is the ability of the drug to diffuse from the said delivery systems and dissolve within the biological medium. Following this, the drug may diffuse through the biological media and subsequently diffuse across the attendant biological membranes, thereby gaining entry into the systemic circulation. In certain systems, the rate at which the drug dissolves within the biological fluid is the slowest and hence the rate-limiting step whereas in other scenarios the diffusion of the drug across the biological membrane may present the greatest challenge. In light of the importance of drug release, it is essential to ensure that the statistical analysis of the data from such experiments is successfully performed to enable rational conclusions to be drawn.

The conductance and design of drug release experiments is relatively straightforward and is defined within the scientific literature and within Pharmacopoeial monographs, e.g., the British Pharmacopoeia, the United States Pharmacopoeia. However, there is a relative paucity of information concerning methods that may be used to statistically quantify the outcomes of these experiments. Experimentally the analysis of drug release is typically performed by immersion of the dosage form within a defined volume of fluid designed to mimic a particular biological matrix, e.g., simulated gastric fluid, simulated intestinal fluid. The volume of fluid is chosen to ensure that the subsequent dissolution is typically not affected by the concentration of dissolved drug within the fluid. Thereafter, at defined time intervals, a sample of the surrounding fluid is removed and the mass of drug quantified using an appropriate analytical method, e.g., ultraviolet spectroscopy, fluorescence spectroscopy. After this analysis, there are two major challenges to the pharmaceutical scientist to ensure that the interpretation of the data is satisfactorily performed, namely:

(1) Selection of the appropriate mathematical model to define release.

(2) Use of statistical methods to examine formulation effects or release fluid effects on drug release.

The intention of this paper is to define appropriate statistical methods to address the above issues and thereby to define a protocol for the analysis of data that has been derived from drug release experiments.

Drug Release from Pharmaceutical Systems

Since the first publication of papers on the modelling of drug release for drug delivery systems (see Baker 1987, Chien 1992) there have been several papers that have applied mathematical concepts to understand the mechanism of drug release from such systems. For the purpose of this article, these methods may be summarised into three categories defined according to the mechanism of drug release, as follows:

(a) *Controlled (Fickian) release from monolithic devices*
In this method the release of a homogeneously dispersed drug from the delivery system is controlled by conventional diffusion (as initially described by Adolf Fick). Mathematically, Fickian diffusion of a drug from a slab geometry may be defined as follows:

$$\frac{M_t}{M_\infty} = 1 - \sum_{n=0}^{\infty} \frac{8\exp\left[-D(2n+1)^2\pi^2 t/l^2\right]}{(2n+1)^2\pi^2}. \quad (1)$$

At early time approximations $\left(0 \le \frac{M_t}{M_\infty} \le 0.6\right)$, the following approximation may be made:

$$\frac{M_t}{M_\infty} = 4\left(\frac{Dt}{\pi l^2}\right)^{0.5}, \quad (2)$$

where: D is the diffusion coefficient of the drug
t is time
l is the thickness of the slab geometry
M is the mass of drug released.
Accordingly it may be observed that the fraction of drug release is proportional to the square root of time.

(b) *Reservoir devices*
In these systems, drug diffusion from the device is controlled by the presence of a membrane. Mathematically, drug diffusion from the core of the device is defined by the following equations:

$$\frac{dM_t}{dt} = \frac{DAKC_s}{l} \quad \text{for a slab geometry} \quad (3)$$

$$\frac{dM_t}{dt} = \frac{2\pi hDKC_s}{\ln\left(\frac{r_0}{r_1}\right)} \quad \text{for a cylinder geometry} \quad (4)$$

$$\frac{dM_t}{dt} = \frac{4\pi hDKC_s r_0 r_1}{r_0 - r_1} \quad \text{for a sphere geometry} \quad (5)$$

where: D is the diffusion coefficient
l is the thickness of the slab geometry
M_t is the mass of drug released at time t
h is the length of the cylinder
r_0 and r_1 are the outside and inside radii of the cylinder/sphere
A is the area of the device
K is the partition coefficient of the drug between the core and membrane

Under the above circumstances it may be observed that the mass of drug released is directly proportional to time.

More recently, Peppas (1985) described the use of a generic equation to model and characterise drug release from pharmaceutical platforms, as follows:

$$\frac{M_t}{M_\infty} = kt^n \tag{6}$$

where:
k is the release constant
$\frac{M_t}{M_\infty}$ is the fractional drug release
n is the release exponent.

In this approach, the equation encompasses the previous mathematical model, the value of the release exponent being used to define whether the mechanism of drug release from slab systems is:

(a) Fickian ($n = 0.5$)
(b) Reservoir controlled ($n = 1$)
(c) Anomalous ($0.5 < n < 1$)

Defining the Statistical Problem

Whilst the mathematical approaches described above seem quite straightforward, there is an ongoing issue with the application of these models within a statistical framework. There are several issues, which may be defined as follows:

(1) *Use of the incorrect mathematical model*

The choice of the correct mathematical model should be performed following consideration of the design of the dosage form and also the experimental conditions. In many situations, the limitations of the models are overlooked to render the mathematical analysis more straightforward. For example, in Fickian diffusion controlled systems, the mathematical model may only be used whenever there is no swelling of the pharmaceutical device. Furthermore, as highlighted in one of the examples above, the geometry of the device will affect the choice of equation. However, whilst the above concerns may seem obvious to those experienced in the pharmaceutical sciences, one common concern regards the modelling process. Typically the Peppas model is used to model release data however, in the early stages the model may yield an exponent of unity which may not be a true reflection of the release kinetics of the system as both diffusion controlled release and anomalous release will also yield similar exponents over this period of testing.

(2) *Choice of Statistical Tests*

Having acquired drug diffusion/dissolution data, the next challenge to the pharmaceutical scientist concerns the choice of the correct statistical method. One test that is recommended by the FDA is the f_2 test, which is used to compare the dissolution of two products, typically a test product (e.g., a generic product) and a reference product. The f_2 value is calculated using the following equation (Bolton and Bon 2004):

$$f_2 = 50 \log\left(\left[1 + \frac{1}{N}\right]\sum(R_t - T_t)^2 \times 100\right), \tag{7}$$

where: R_t and T_t are the % dissolution of the reference and test product at time t.

In this test an f_2 value >50 illustrates similarity of dissolution profiles. However, it should be noted that this test has several limitations; most notably individual differences at early time points may render the dissolution of two formulations different whenever the overall profiles are similar. The f_2 test has been principally used in the pharmaceutical industry to compare the dissolution of two dosage forms however; it is not commonly used within pharmaceutical research due to its relative inflexibility. The question may then be asked, "How are the drug release profiles of two, or more than two dosage forms compared?" Examples of the strategies that may be used are provided below.

(a) *Comparison of the release rates of the different formulations*

Mathematically the release of a drug from a dosage form is frequently described using the release rate, i.e., the slope of the plot of cumulative drug release against time^n. To use this method it must initially be *correctly* proven that the mechanisms of drug release from the different formulations are similar, a point often overlooked within the scientific literature. In light of the potential similarities of the kinetics of drug release for diffusion controlled, anomalous and zero order systems at early time points, it is essential to statistically establish similarity. Therefore, drug release should be allowed to progress to ensure that up to 60% release has occurred. To establish similarity of release mechanisms, it is appropriate to model drug release using the Peppas model and to then compare the release exponent values. For this purpose the Peppas model is transformed logarithmically, the release exponent (n) being the

resultant slope of the line following linear regression.

$$\ln \frac{M_t}{M_\infty} = \ln k + n \ln t. \quad (8)$$

The underlying prerequisite of this approach is the requirement for linearity. Typically linearity should be proven using both an ►Analysis of Variance and reference to Pearson's correlation coefficient (this should be greater than 0.99 [Jones 2002]). To facilitate meaningful statistical analysis of the data, it is suggested that approximately six replicate measurements should be performed as this increase the likelihood of the use of parametric tests for subsequent comparisons of the release exponents. Following the acquisition of this information the following points should be considered:

- To establish the release mechanism of the drugs from the pharmaceutical systems, the calculated release exponent should be statistically compared to 0.5 and also to 1.0. This is typically performed using a one sample *t* test. Retaining of the null hypothesis in these tests confirms that the release is either zero-order or diffusion controlled. Rejection of the null hypothesis verifies that the release mechanism is anomalous, i.e., $0.5 < n < 1.0$. The reader should note that the values of *n* representative of diffusion controlled and zero-order release are dependent on the geometry of the system. For a cylindrical system the release exponents are 0.45 and 0.89 for Fickian controlled and zero-order systems, respectively whereas for spherical systems these values become 0.43 and 0.85.
- Assuming that the release mechanism of all formulations under examination is similar, it is therefore appropriate to statistically compare the drug release kinetics from the various formulations. Therefore, for reservoir systems (in which the mechanism of release is zero-order), the plot of cumulative drug release against time is linear whereas in Fickian diffusion, the plot of cumulative drug release against $\sqrt{}$time is linear. Using linear regression analysis (and remembering not to include the point 0,0 in the analysis), the slope of the plot may be statistically determined for each individual replicate, which for diffusion controlled release and reservoir (zero-order) controlled release have the units of $(\text{concentration})(\text{time})^{-0.5}$ and $(\text{concentration})(\text{time})^{-1}$. Replication of these analyses (e.g., $n = 6$) enables calculation of the mean ± standard deviation or the median and ranges of the rates of release. Finally comparison of the rates of release may be easily performed using either the Analysis of Variance or the Kruskal-Wallis test if more than two samples/formulations require to be compared or, alternatively, the unpaired *t* test or the Mann Whitney *U* test, if the number of formulations under comparison is two. The choice of parametric or non-parametric tests to analyse the data is performed according to conventional statistical theory, the former tests being used if the populations from which the data were sampled were normally distributed (commonly tested using, e.g., the ►Kolmogorov-Smirnov test or the Shapiro-Wilk test) and if the variances of the populations from which the data were samples were statistically similar (commonly tested using e.g., Leveine's test or ►Bartlett's test). It should be noted that this approach is employed if the release mechanisms of different formulations are statistically similar, independent of the mechanism of drug release. Accordingly, the release exponent of different formulations may be identical within the range of $0.5 < n < 1.0$.

(b) *Comparing drug release from pharmaceutical systems that exhibit different release mechanisms*

In the above scenarios, the release rate of the drug from the pharmaceutical platform was obtained from linear regression of the associated cumulative drug release plot, i.e., cumulative drug release against time for the zero-order system and cumulative drug release against the square root of time for diffusion control systems. The above approach is predicated on the identical mechanisms of drug release; however, this requirement does raise a statistical dilemma. Consequently if the release mechanisms (and hence measured units) are different, therefore it is impossible to generate a single parameter that may be used as the basis for comparisons of the various formulations.

Under these conditions there are two approaches that may be employed to generate meaningful comparisons of drug release from different formulations.

(1) *Analysis of the data sets using a repeated measures Analysis of Variance*

This approach uses a repeated measures experimental design to compare drug release from different formulations. In this the repeated measure is time (which should be identical for each formulation) and the factor is formulation type. Individual differences between the various formulations may then be identified using an appropriate post hoc test. It is essential to ensure that the experimental design does not become overly complicated and that the demands of the ANOVA (with respect to homogeneity of population variances and the use of normally-distributed populations) hold.

(2) *Analysis of data at single time points*

The main requirements for the use of the repeated measures Analysis of Variance are, firstly that the requirements for the use of this test are met and secondly, that the times at which the data were collected (sampled) are identical for each formulation. In practice these problems are straightforward to overcome at the experimental design stage however, there may be issues concerning the ability to perform the required number of replicates (typically ≥5) to allow a parametric test is suitable to use for the data analysis. For example, experiments in which the release is relatively rapid (<48 h) may be easier to perform with many replicates whereas the converse is true for experiments in which the release is protracted. In such circumstances (e.g., whenever there are few replicates, typically $n \leq 3$), one method that may be employed to compare the drug release profiles of different formulation involves the comparison of the formulations at each sampling point using a multiple hypothesis test, e.g., the Kruskal-Wallis test. In a similar fashion, individual differences between formulations may be identified by the application of an appropriate post hoc test, e.g., Dunn's test, Nemenyi's test.

In an alternative approach, typically encountered whenever the sampling periods differ, comparison of the drug release kinetics of candidate formulations may be performed by ascertaining the time required for a defined fraction of the initial drug loading to be released. A regression of the release profile (using the Peppas model) is performed and, using the output from this model, the times required for each formulation to release a defined fraction is obtained and statistically compared using the appropriate statistical test (Jones et al. 1999; Jones et al. 2000). The choice if test to perform the analysis is important and the reader should be reminded that the use of parametric statistical tests (the unpaired *t* test and the ANOVA) should be validated.

Conclusions

Analysing release data is an essential component in the development and assessment of the performance of pharmaceutical systems. In spite of this, suitable methods to analyse release data are not clearly defined. In this monograph strategies for the statistical comparisons of release data are defined.

About the Author

David Jones is Professor of Biomaterial Science at the Queen's University of Belfast. Professor Jones is a Chartered Engineer, Chartered Chemist and holds Fellowships of the Royal Statistical Society and the Institute of Materials, Minerals and Mining and is a Member of the Royal Society of Chemistry, the Institute of Engineers in Ireland and the Pharmaceutical Society of Northern Ireland. He is the Editor of the *Journal of Pharmacy and Pharmacology* and has been the Statistical Advisor to the International Journal of Pharmacy Practice. Professor Jones is a former winner of the Eli Lilly Award and the British Pharmaceutical Conference Science Medal.

Cross References

►Analysis of Variance
►Biopharmaceutical Research, Statistics in
►Medical Research, Statistics in
►Parametric Versus Nonparametric Tests
►Pharmaceutical Statistics: Bioequivalence
►Repeated Measures
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Baker RW (1987) Controlled release of biologically active agents. Wiley-Interscience, New York

Bolton S, Bon C (2004) Pharmaceutical statistics: practical and clinical applications, vol 135. Marcel Dekker, New York, p 755

Chien YW (1992) Novel drug delivery systems, 2nd edn. vol 50. Marcel Dekker, New York

Jones DS (2002) Pharmaceutical statistics. Pharmaceutical Press, London, p 608

Jones DS, Irwin CR, Woolfson AD, Djokic J, Adams V (1999) Physicochemical characterization and preliminary in vivo efficacy of bioadhesive, semisolid formulations containing flurbiprofen for the treatment of gingivitis. J Pharm Sci 88(6):592–598

Jones DS, Woolfson AD, Brown AF, Coulter WA, McClelland C, Irwin CR (2000) Design, characterisation and preliminary clinical evaluation of a novel mucoadhesive topical formulation containing tetracycline for the treatment of periodontal disease. J Cont Rel 67(2–3):357–368

Peppas NA (1985) Analysis of Fickian and Non-Fickian drug release from polymers. Pharm Acta Helvet 60(4):110–111

Statistical Analysis of Longitudinal and Correlated Data

David Todem
Michigan State University, East Lansing, MI, USA

Introduction

Correlated data are typically generated from studies where the outcomes under investigation are collected on clustered units. Specific examples include; (1) longitudinal data where outcomes are collected on the same experimental unit (for instance, the same person) at two or more different points in time; and (2) studies where outcomes

are recorded at one single point in time on clustered units. Such studies have one major attraction, the ability to control for unobserved variables in making inferences. Sampled units serve as controls for other units in the same cluster. As an example, in a longitudinal study, each subject serves as his or her own control in the study of change across time. Therefore, these studies allow the researcher to eliminate a number of competing explanations for observed effects. The determination of causal ordering in making solid inferences contitutes another attraction for longitudinal studies.

Despite these advantages, statistical analysis of correlated data raises a number of challenging issues. It is well known, for example, that the multiplicity of outcomes recorded over time on the same unit necessitates the use of methods for correlated data. This entry reviews some of the common statistical techniques to analyze such data. A focus is on longitudinal data as statistical models for clustered data are typically simple versions of techniques for longitudinal data. In longitudinal data analysis, the response $y(t)$ is a time-varying variable and the covariate can be a baseline vector x, a time-varying covariate vector $x(t)$, or a combination of both. A key issue for such data is to relate the longitudinal mean responses to covariates and draw related inferences while accounting for the within-subject association. In essence, two classes of models exist for modeling the mean outcomes and covariates relationship; (1) the parametric models and; (2) the semiparametric and nonparametric models. This entry examines each of these models in some detail, with an eye to discerning their relative advantages and disadvantages. A discussion on emerging issues in analyzing longitudinal data is also given but touched on briefly.

Parametric Models

Parametric models are the predominant approaches for longitudinal data. They make parametric assumptions about the relationship between the mean of a longitudinal response to covariates. They are known as growth curve models and include the popular mixed-effects models (Laird and Ware 1982) and generalized estimating equations models (Liang and Zeger 1986). Verbeke and Molenberghs (2000) and Diggle et al. (2002) provide an extensive review of this literature.

Mixed-Effects Models

Mixed-effects models are a useful tool to analyze repeated measurements recorded on the same subject. They were primarily developed for continuous outcomes in time (Laird and Ware 1982) and were later extended to categorical and discrete data (Breslow and Clayton 1993). For continuous outcomes with an identity link, they are known as linear mixed-effects models. Generalized linear mixed-effects models constitute the broader class of mixed-effects models for correlated continuous, binary, multinomial, ordinal and count data (Breslow and Clayton 1993). They are likelihood-based and often are formulated as hierarchical models. At the first stage, a conditional distribution of the responses given random effects is specified, usually assumed to be a member of the exponential family. At the second stage, a prior distribution is imposed on the random effects. The conditional expectations (given random effects) are made of two components, a fixed-effects and a random-effects term. The fixed-effects term represents covariate effects that do not change with the subject. Random effects represent a deviation of a subject's profile from the average profile. Most importantly, they account for the within-subject correlation across time under the conditional independence assumption. For continuous outcomes with an identity link function, these models have an appealing feature in that the fixed-effects parameters have a subject-specific as well as a population-averaged interpretation (Verbeke and Molenberghs 2000). For non continuous data and nonlinear relationships, this elegant property is lost. The fixed-effects parameters, with the exception of few link functions, only have a subject-specific interpretation, conditional on random effects. This interpretation is only meaningful for covariates that change within a subject such as time-varying covariates. These effects capture the change occurring within an individual profile. To assess changes for time-independent covariates, the modeler is then required to integrate out the random effects from the quantities of interest.

Mixed-effects models are likelihood-based and therefore can be highly sensitive to any distribution misspecification. But they are known to be robust against less restrictive missing data mechanisms. There exist other likelihood-based methods for analyzing correlated data. Before the advent of ▶linear mixed models, longitudinal continuous data were analyzed using techniques such as repeated measures analysis of variance (ANOVA). This approach has a number of disadvantages and has generally been superseded by linear mixed-effects models, which can easily be fit in mainstream statistical software. For example, repeated measures ANOVA models require a balanced design in that measurements should be recorded at the same time points for all subjects, a condition not required by linear mixed models.

Generalized Estimating Equations Models

Although there is a variety of standard likelihood-based models available to analyze data when the outcome is approximately normal, models for discrete outcomes (such

as binary outcomes) generally require a different methodology. Liang and Zeger (1986) have proposed the so-called Generalized Estimating Equations-GEE model, which is an extension of ▶generalized linear models to correlated data. The basic idea of this family of models is to specify a function that links the linear predictor to the mean response, and use a set of estimating functions with any working correlation model for parameter estimation. A sandwich estimator that corrects for any misspecification of the working correlation model is then used to compute the parameters' standard errors. GEE-based models are very popular as an all-round technique to analyze correlated data when the exact likelihood is difficult to specify. One of the strong points of this methodology is that the full joint distribution of the data does not need to be specified to guarantee asymptotically consistent and normal parameter estimates. Instead, a working correlation model between the clustered observations is required for estimation. GEE regression parameter estimates have a population-averaged interpretation, analogous to those obtained from a cross-sectional data analysis. This property makes GEE-based models desirable in population-based studies, where the focus is on average affects accounting for the within-subject association viewed as a nuisance term.

The GEE approach has several advantages over a likelihood-based model. It is computationally tractable in applications where the parametric approaches are computationally very demanding, if not impossible. It is also less sensitive to distribution misspecification as compared to full likelihood-based models. A major limitation of GEE-based models at least in their 1986 original formulation is that they require a more stringent missing data mechanism (missing data completely at random) to produce valid inferences. Weighted GEE-based models have been proposed to accommodate a less stringent missing data mechanism, the missing data at random process (Robins et al. 1995).

Semiparametric and Nonparametric Models

A major limitation of parametric models is that the relationship of the mean of a longitudinal response to covariates is assumed fully parametric. Although such parametric mean models enjoy simplicity and ease of interpretation, they often suffered from inflexibility in modeling complicated relationships between the response and covariates in various longitudinal studies. Specific examples include modeling of; (1) longitudinal *CD4+* counts as function of time in HIV/AIDS research; and (2) trajectories of angiogenic and antiogenic factors in maternal plasma concentrations (s-eng, sVEGFR-1 and PlGF) in perinatal research. Parametric models typically require higher degree polynomials to capture the relationship between these mean responses and covariates. This has been seen as an indication of poor fit and has motivated the development of more complex and flexible approaches to model these data. Semiparametric and nonparametric regression models, well known to be more data adaptive, have emerged as promising alternative to parametric models in these settings. Nonparametric models make no parametric assumption about the relationship between the mean response and covariates. Semiparametric models assume a parametric relationship between some covariates and the mean response while maintaining a nonparametric relationship between other covariates and the mean response. These methods are well developed for independent data, but their extensions to longitudinal data remain an active area of research. A major difficulty often cited in the literature for this extension is the inherent within-subject correlation in longitudinal studies. This correlation presents significant challenges in the development of kernel and spline smoothing methods for longitudinal data. Specifically, as reported by many researchers in the field (see for example, Lin and Carroll 2000; Lin et al. 2004), local likelihood-based kernel methods are not able to effectively account for the within-subject correlation in longitudinal data.

Discussion

This entry has reviewed some of the common techniques to model longitudinal data. A focus was on parametric models. Nonparametric and semiparametric approaches based on smoothing techniques have emerged as a flexible way to model longitudinal data. Other approaches that do not require smoothing have recently been proposed (Lin and Ying 2001). But much research, especially from a theoretical standpoint, is needed to understand these methods. Moreover, statistical software to fit these models routinely in real time is much needed. This is in contrast to parametric models which can be fit using mainstream statistical software such as SAS, Stata, R, Splus and SPSS. There are emerging areas in connection to longitudinal data analysis that need further research such as; (1) the joint modeling of longitudinal and ▶survival data, (2) missing data and (3) causal inference. These areas have enjoyed some significant developments in the past several years. But there are numerous open questions that remain unanswered and are the subject of future research.

About the Author

Dr. David Todem is a Biostatistics Associate Professor in the Division of Biostatistics of the Department of Epidemiology at Michigan State University, USA. He has authored

and co-authored more than 30 papers and 2 entries in encyclopedic publications. Dr Todem is an Editorial Board member for *The Open Statistics and Probability Journal*.

Cross References

►Data Analysis
►Exponential Family Models
►Linear Mixed Models
►Medical Statistics
►Multilevel Analysis
►Nonlinear Mixed Effects Models
►Nonparametric Regression Using Kernel and Spline Methods
►Panel Data
►Random Coefficient Models
►Repeated Measures
►Semiparametric Regression Models
►Statistical Software: An Overview

References and Further Reading

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25
Diggle PJ, Heagerty PJ, Liang K-Y, Zeger S (2002) Analysis of longitudinal data. Oxford University Press, Oxford
Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974
Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Lin D, Ying Z (2001) Semiparametric and nonparametric regression analysis of longitudinal data. J Am Stat Assoc 96:103–126
Lin X, Carroll RJ (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. J Am Stat Assoc 95:520–534
Lin X, Wang N, Welsh A, Carroll RJ (2004) Equivalent kernels of smoothing splines in nonparametric regression for clustered data. Biometrika 91:177–193
Robins J, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. J Am Stat Assoc 90:106–121
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

Statistical Approaches to Protecting Confidentiality in Public Use Data

Jerome P. Reiter
Associate Professor
Duke University, Durham, NC, USA

Many national statistical agencies, survey organizations, and researchers – henceforth all called agencies – collect data that they intend to share with others. Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities. Often, however, agencies cannot release data as collected, because doing so could reveal data subjects' identities or values of sensitive attributes. Failure to protect confidentiality can have serious consequences for agencies, since they may be violating laws or institutional rules enacted to protect confidentiality. Additionally, when confidentiality is compromised, the agencies may lose the trust of the public, so that potential respondents are less willing to give accurate answers, or even to participate, in future studies (Reiter 2004).

At first glance, sharing safe data with others seems a straightforward task: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data. However, these actions alone may not suffice when quasi-identifiers, such as demographic variables, employment/education histories, or establishment sizes, remain on the file. These quasi-identifiers can be used to match units in the released data to other databases. For example, Sweeney (1997) showed that 97% of the records in a medical database for Cambridge, MA, could be identified using only birth date and nine-digit ZIP code by linking them to a publicly available voter registration list.

Agencies therefore further limit what they release, typically by altering the collected data (Willenborg and de Waal 2001). Common strategies include those listed below. Most public use data sets released by national statistical agencies have undergone at least one of these methods of statistical disclosure limitation.

Aggregation. Aggregation reduces disclosure risks by turning atypical records – which generally are most at risk – into typical records. For example, there may be only one person with a particular combination of demographic characteristics in a city, but many people with those characteristics in a state. Releasing data for this person with geography at the city level might have a high disclosure risk, whereas releasing the data at the state level might not. Unfortunately, aggregation makes analysis at finer levels difficult and often impossible, and it creates problems of ecological inferences.

Top coding. Agencies can report sensitive values exactly only when they are above or below certain thresholds, for example reporting all incomes above $200,000 as "$200,000 or more." Monetary variables and ages are frequently reported with top codes, and sometimes with bottom codes as well. Top or bottom coding by definition eliminates detailed inferences about the distribution

beyond the thresholds. Chopping off tails also negatively impacts estimation of whole-data quantities.

Suppression. Agencies can delete sensitive values from the released data. They might suppress entire variables or just at-risk data values. Suppression of particular data values generally creates data that are not missing at random, which are difficult to analyze properly.

Data swapping. Agencies can swap data values for selected records – for example, switch values of age, race, and sex for at-risk records with those for other records – to discourage users from matching, since matches may be based on incorrect data (Dalenius and Reiss 1982). Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low – agencies do not reveal the rates to the public – because swapping at high levels destroys relationships involving the swapped and unswapped variables.

Adding random noise. Agencies can protect numerical data by adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero (Fuller 1993). This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables. The degree of confidentiality protection depends on the nature of the noise distribution; for example, using a large variance provides greater protection. However, adding noise with large variance introduces measurement error that stretches marginal distributions and attenuates regression coefficients (Yancey et al. 2002).

Synthetic data. The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions (Rubin 1993). These distributions are specified to reproduce as many of the relationships in the original data as possible. Synthetic data approaches come in two flavors: partial and full synthesis (Reiter and Raghunathan 2007). Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values. For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables. Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file. In both types, the agency generates and releases multiple versions of the data (as in multiple imputation for missing data, see ►Multiple Imputation). Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models, but they may not give good results for other analyses.

Statisticians play an important role in determining agencies' data sharing strategies. First, they measure the risks of disclosures of confidential information in the data, both before and after application of data protection methods. Assessing disclosure risks is a challenging task involving modeling of data snoopers' behavior and resources; see Reiter (2005) and Elamir and Skinner (2006) for examples. Second, they advise agencies on which protection methods to apply and with what level of intensity. Generally, increasing the amount of data alteration decreases the risks of disclosures; but, it also decreases the accuracy of inferences obtained from the released data, since these methods distort relationships among the variables. Statisticians quantify the disclosure risks and data quality of competing protection methods to select ones with acceptable properties. Third, they develop new approaches to sharing confidential data (*see* ►Data Privacy and Confidentiality). Currently, for example, there do not exist statistical approaches for safe and useful sharing of network and relational data, remote sensing data, and genomic data. As complex new data types become readily available, there will be an increased need for statisticians to develop new protection methods that facilitate data sharing.

About the Author

Jerry Reiter is currently an Associate Editor for the *Journal of the American Statistical Association*, *Survey Methodology*, the *Journal of Privacy and Confidentiality*, and the *Journal of Statistical Theory and Practice*. He is the current Chair of the Committee on Privacy and Confidentiality of the American Statistical Association (2009–2012). He is an Elected member of the International Statistical Institute. He has authored more than 60 papers, including foundational works on the use of multiple imputation for confidentiality protection. He was awarded the Alumni Distinguished Undergraduate Teaching Award at Duke University.

Cross References

►Census
►Data Analysis
►Data Privacy and Confidentiality
►Federal Statistics in the United States, Some Challenges
►Multi-Party Inference and Uncongeniality

References and Further Reading

Dalenius T, Reiss SP (1982) Data-swapping: a technique for disclosure control. J Stat Plan Infer 6:73–85

Elamir E, Skinner CJ (2006) Record level measures of disclosure risk for survey microdata. J Off Stat 22:525–539

Fuller WA (1993) Masking procedures for microdata disclosure limitation. J Off Stat 9:383–406

Reiter JP (2004) New approaches to data dissemintation: a glimpse into the future (?). Chance 17(3):12–16

Reiter JP (2005) Estimating identification risks in microdata. J Am Stat Assoc 100:1103–1113
Reiter JP, Raghunathan TE (2007) The multiple adaptations of multiple imputation. J Am Stat Assoc 102:1462–1471
Rubin DB (1993) Discussion: statistical disclosure limitation. J Off Stat 9:462–468
Sweeney L (1997) Computational disclosure control for medical microdata: the Datafly system. In: Proceedings of an international workshop and exposition, pp 442–453
Willenborg L, de Waal T (2001) Elements of statistical disclosure control. Springer, New York
Yancey WE, Winkler WE, Creecy RH (2002) Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer J (ed) Inference control in statistical databases. Springer, Berlin, pp 135–152

Statistical Aspects of Hurricane Modeling and Forecasting

Mark E. Johnson[1], Charles C. Watson[2]
[1]Professor
University of Central Florida, Orlando, FL, USA
[2]Watson Technical Consulting, Savannah, GA, USA

Hurricanes are complex, natural phenomena that can cause property damage on a catastrophic scale. The human toll depends on the preparedness of the population – historical events with thousands of casualties are rare but do occur (e.g., the 1900 Galveston storm – Larson 1999). Depending on where hurricanes form and traverse, they have other names such as typhoons (western Pacific) and cyclones (Indian Ocean and Australia). Officially, a hurricane is defined as a closed circulation, warm core, and convective weather system with maximum 10-min average winds of 33 m/s or higher, measured at 10 m above ground level (WMO 2007). This precise and technical definition is important since insurance payouts for losses often depend on the declaration of a hurricane event. The definition also provides a threshold for establishing the event frequency at specific locations, a criterion especially important for climate change studies. For planning purposes, the return period of hurricanes of various intensities is needed – i.e., what is the probability that 100 mph winds will strike a specific location this season or what wind speed corresponds to the 100 year worst event? Fortunately, hurricanes are relatively rare events (as compared to thunderstorms or tornadoes) and thus, extreme value methods are used to assess their frequencies (Embrechts et al. 1997). An excellent introduction to hurricanes is given by Emanuel (2005) while a more technical treatise is available by Anthes (1982).

Iman et al. (2006) reviewed many aspects of statistical forecasting and planning in the premier Interdisciplinary Section of The American Statistician. The invitation to prepare this article was motivated in part by the hyperactive 2004 and 2005 Atlantic hurricane seasons which stunned the American public following relatively minor hurricane activity in the United States since Hurricane Andrew in 1992. Various researchers took these two seasons as the onset of sustained, increased activity, only to witness the four subsequent years of little hurricane activity impacting Florida (O'Hagan et al. 2008). This perspective illustrates a United States-centric perspective regarding hurricane activity. The 2007 season endured two very strong events (Hurricanes Dean and Erin) which pummeled the Mexican Yucatan and the Gulf of Campeche, causing massive havoc with their oil and gas industry. Similarly, in 2009, the Philippines experienced multiple typhoons left nearly 1,000 dead, thousands homeless, and widespread agricultural devastation, yet received little media attention.

Forecasting hurricane track and intensity are key problems that must be addressed in real time for actual events under a harsh public and media spotlight as hurricane watches and warnings go into effect. The "obvious" forecast is to extrapolate the current track with a linear trend in intensity. A more sophisticated version of this forecast is to draw upon the historical record to develop a regression model using comparable information on the movement of storms getting to the current position of the storm (CLIPER and CLIPER5 in use by the National Hurricane Center). More advanced models take into account current and forecast upper level winds ("steering currents"), while the most advanced include fluid dynamics calculations of mesoscale storm structure. In addition to the many individual forecast models, ensemble models are also in use (for a technical summary, see www.nhc.noaa.gov/modelsummary.shtml). The increase in skill (accuracy of prediction) of the more sophisticated models is offset by data input needs and computational run times. Forecasts must be timely – a 6 h forecast that takes 5 h to produce may be inferior to a much simpler forecast that can be formulated in a matter of minutes. For a further discussion of the many pitfalls associated with forecasts, especially the problems encountered with Hurricane Charley in 2004, see the aforementioned article by Iman et al. (2006).

In determining hurricane impacts for insurance purposes, a more leisurely time frame for computation is available. The computational burden is severe in that a probabilistic assessment of hurricane losses is necessary.

Most approaches have proceeded by choosing specific, individual models of hurricane frequency, wind field, track, friction impacts, wind field decay, damage, and actuarial summaries. Given the approximately 150 year Atlantic storm history, less in other regions, practitioners have tended to fit probability distributions to key characteristics and then proceed to simulate 50–300,000 years of future hurricane seasons, accumulating losses for each generated event. To assess the uncertainty and sensitivity of the parameter specifications for these models, the Florida Commission on Hurricane Loss Methodology has prescribed the use of Latin hypercube sampling (McKay et al. 1979). One specific implementation pertinent to hurricane modeling is described by Iman et al. (2005a, b). The latest research focuses on the use of climate models to provide track and intensity guidance (Watson and Johnson 2008).

A basic issue with evaluating hurricane modeling efforts is that every hurricane is somewhat different and any model that "fine tunes" its modeling approach to a specific event will ultimately suffer for it (not all future events are just like the particular event. For some historical events, a very simple hurricane windfield model can do extremely well with respect to matching modeled to actual losses. An approach used by the Florida Commission to address this difficulty follows the contextual analysis developed by Watson and Johnson (2004) and expounded from an actuarial perspective by Watson, Johnson and Simons. In brief, a factorial combination of model components are considered (nine wind fields, four friction models, nine damage functions and three frequency approaches) and the loss costs for specific models are placed in the context of 972 model combination results. ►Outliers with respect to the range of the factorial models generate relevant probing questions of specific models.

Nelder (2010) noted the importance of learning another jargon for statisticians doing interdisciplinary research. The effort is well-rewarded for statisticians dealing with the topic of hurricanes which will likely entail collaborations with meteorologists, atmospheric scientists, geophysicists, and wind engineers.

About the Authors

Dr. Mark E. Johnson is professor, Department of Statistics, University of Central Florida. He was Department Chair (1990–1996). Dr Johnson is a Fellow of the American Statistical Association (1988), Elected Member, International Statistical Institutue (1994), Chartered Statistician, Royal Statistical Society (1993). He has (co-)authored more than 60 refereed papers and is author of *Multivariate Statistical Simulation* (Wiley 1987). Professor Johnson was awarded the Jack Youden Prize (1981), ASQC Shewell Award (1985), Thomas L. Saaty Prize (1984, 1989 and 1997), and ASQC Brumbaugh Award (1991), He was Associate Editor of *Technometrics* (1984–1991), *Journal of Quality Technology*, *American Journal of Management and Mathematical Sciences*, and *Journal of Statistical Computation and Simulation*.

Mr. Charles C. Watson Jr. is the founder and Director of Research and Development of Kinetic Analysis Corporation. He has authored or co-authored more than 70 papers and book contributions in the fields of satellite remote sensing, geophysics, and meteorology, in such diverse publications such as *Bulletin of the American Meteorological Society*, *Photogrammetric Engineering & Remote Sensing*, the *Journal of Insurance Regulation*, and *The American Statistician*. Mr. Watson has served or is active as a scientific consultant on hazard planning and remote sensing to numerous national and international projects and agencies such as the Caribbean Catastrophe Risk Insurance Facility, the Intergovernmental Panel on Climate Change, UN Agencies such as the World Meteorological Organization, UN Environment Program, and World Food Program, as well as US agencies such as National Aeronautics and Space Administration.

Cross References

►Actuarial Methods
►Forecasting: An Overview
►Statistics and Climate Change
►Statistics of Extremes
►Stochastic Difference Equations and Applications
►Time Series

References and Further Reading

Anthes RA (1982) Tropical cyclones, their evolution, structure, and effects, American Meteorological Society meteorological monographs, vol 19(41). AMS, Boston

Emanuel K (2005) Divine wind: the history of science and hurricanes. Oxford University Press, New York

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events. Springer, Berlin

Iman RL, Johnson ME, Watson C Jr (2005a) Sensitivity analysis for computer model projections of hurricane losses. Risk Anal 25:1277–1297

Iman RL, Johnson ME, Watson C Jr (2005b) Uncertainty analysis for computer model projections of hurricane losses. Risk Anal 25:1299–1312

Iman RL, Johnson ME, Watson C Jr (2006) Statistical aspects of forecasting and planning for hurricanes. Am Stat 60(2):105–121

Larson E (1999) Isaac's Storm. A man, a time, and the deadliest hurricane in history. Crown Publishers, New York

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245

Nelder J (2010) "Statistics: Nelder's view," International encyclopedia of statistical science. Springer, New York
O'Hagan T, Ward B, Coughlin K (2008) How many Katrinas? Predicting the number of hurricanes striking the USA. Significance 5(4):163–167
Watson C Jr, Johnson ME (2008) Integrating hurricane loss models with climate models. In: Murnane R, Diaz H (eds) Climate extremes and society. Cambridge University Press, Cambridge, pp 209–224
Watson C Jr, Johnson ME (2004) Hurricane loss estimation models: opportunities for improving the state of the art. B Am Meteorol Soc 84:1713–1726
Watson C Jr, Johnson ME, Simons M (2004) Insurance rate filings and hurricane loss estimation models. J Insur Regul 22(3):39–64
World Meteorological Organization (WMO) (2007) Global guide to tropical cyclone forecasting, WMO/TC-No. 560, Report No. TCO-31, World Meteorological Organization, Geneva, Switzerland

Statistical Consulting

Rolf Sundberg
Professor of Mathematical Statistics
Stockholm University, Stockholm, Sweden

What Is Statistical Consulting?

Here is a sketch of a normal consultation in the consulting unit of my department, in a faculty of sciences. One or a couple of researchers/Ph.D-students from a biology/geology/... department contact us asking for help with the analysis of data from a study they are carrying out. At the meeting the client first describes the background, the set-up, and (some of) the data of the study. The aims of the study are often in a general, vague form that needs specification and statistical reformulation in quantifiable units. What is the client's problem, really, and what kind of questions can possibly be answered from that kind of data? Often the clients will be forced to think about their problems in fresh ways. The consultant will also ask a lot of questions in order to make clear how the data were collected. What populations do the data represent? Was there ▶randomization, stratification, censoring, etc? On what parts of the data should the focus be? Explore the data! What is the structure of these data? This can lead up to a tentative statistical model, and later to parameter estimation procedures and hypothesis tests, etc.

The first meeting hopefully ends at a stage where the client and the statistician have agreed about what questions should be addressed statistically, and how this might be attempted on the data. Either this appears so simple and clear that the clients want and can do this themselves, or else a time plan and a work plan for the contribution by the statistician is agreed on. After a week or two, with some e-mail correspondence in between, client and consultant meet again to discuss the results so far and what kind of report from the statistician that the clients might want. Often also the answer to one question triggers new questions.

Another statistical consultation type of work could be more of a collaborative/partnership character, where the statistician is a member of a team, and the aims are more far-reaching. The statistician then invests a lot of time and effort, to become knowledgeable in the subject matter area and expert in the applications of statistical methods in that area, but can therefore also expect more influence and credit, and is a natural coauthor of the project publications.

Also a consultation where the client is seen only once or twice is rewarding for the statistician, but in a more indirect way. Hopefully it will be an intellectually stimulating challenge that together with other such experiences can have a profound influence on our personal development as statisticians. And it might still lead to a joint publication.

Consultation work is typically done under time pressure from one or both parties. Too often the client has unrealistic expectations in this respect. On the other hand, the clients usually do not need or want a perfect model for data (remember the George Box phrase: "All models are wrong, but some are useful") or the most sophisticated method of analysis. A solution that is approximately right is much better than one that is precisely wrong. The consultant should think of the acronym KISS, here read out as "Keep It Simple, but Scientific," or rephrased as another quotation: "as simple as possible, but no simpler." "Errors of the third kind" (testing the wrong hypothesis) are most dangerous, Common sense and a critical mind are important. As statistical consultants we must beware of falling in the traps of being a More Data Yeller or a Nit Picker, or any other of the consultant stereotypes coined by Hyams (1971).

Desirable Qualities for a Statistical Consultant

Among the desirable qualities to be possessed by an ideal consultant are:

- Interest in the statistical problems of others (Derr: "Regard each client as a potential collaborator"), and a general interest in science, technology, nature, society.

- Sound basis in theoretical and applied statistics. As a start it should certainly include linear and loglinear models (►generalized linear models), some experimental design (and sampling), and some multivariate analysis, but also experience from a few courses in methods for particular fields of application, and experience from applying such methods to data.
- Eagerness to extend and improve one's statistical knowledge.
- Computer skills in at least one (preferably more) statistical packages.
- Good ability to communicate with clients (includes understanding and adjusting to the client's statistical level).
- Skills in report writing (using a word processor).
- Efficiency under time restrictions and time pressure.
- Awareness of ethical dilemmas that can appear, and an ability to deal with problematic clients.

Teaching Statistical Consulting

Nowadays a large number of universities provide education in statistical consulting, in one form or the other. At my department, as an example, this is a master level course for mathematical statistics students, involving real clients, and real problems in real time. Much of the training in the course is orientated towards three aspects:

- The first meeting with a client (in particular asking questions to find out about the problem)
- Statistical thinking
- Structuring problems and seeing the structure in data

The students are also provided some extended knowledge of statistical methods and models, and they are in a concrete way involved in one consulting project, ending with the writing of a project report.

Some Suggested Reading

The entry by Stinnet et al. (2009) in Encyclopedia of Biostatistics describes the roles of biostatisticians in a variety of medical/biological environments (medical school, pharmaceutical industry, governmental agency, etc.), and discusses some of the special challenges in consulting with physicians, as well as the training of consultants in biostatistics. Joiner's (1982) older entry in Encyclopedia of Statistical Sciences also exemplifies what consulting statisticians might do, before it sets up and discusses a list of desirable skills. The discussion of computers and literature is a bit out-of-date, for natural reasons.

Mallows (1998) discusses "statistical thinking" and the question "how do the data relate to the problem?", in an attempt to formulate a "theory of applied statistics." Cox (2007) provides a review of applied statistics in his typical style, while Chatfield's (1995) nicely written book provides more concrete advice.

Efficient communication is a key element in statistical consultation, and it is the topic of Derr's (2000) book, with an accompanying CD-ROM showing illustrative short movies of positive and negative examples. Communication is the main topic also of Boen and Zahn (1982), who provide much discussion of how to deal with clients, not least with difficult clients, cf. Hyams (1971).

Cabrera and McDougall (2002) is written as a textbook on the whole topic. The first half is on consulting, communication, and statistical methods. I do not agree fully with the statistical methods chapter, but who would expect two statisticians to agree fully? The second half consists of case studies. Such a mix also characterizes Chatfield's (1995) book, and the older book by Cox and Snell (1981), that can be recommended in this context for a section on strategy and for its many case studies. More case studies are found in Hand and Everitt (1987) and in Tweedie et al. (1998). Greenfield's contribution to the former is an entertaining chapter on the encounters he has had with some difficult client characters (cf. Hyams 1971, again).

To finish, here is a quote from one of Terry Speed's columns in the IMS Bulletin (2005), entitled "How to do Statistical Research." Former IMS President Speed explains his research strategy to be that of doing

- *Consulting*: a very large amount
- *Collaboration*: quite a bit
- *Research*: some

"Why? A very large amount of consulting means meeting many people and many problems, learning a lot, including finding out where we are ignorant. Then we might spot some low-hanging fruit."

About the Author

For biography *see* the entry ►Chemometrics.

Cross References

►Careers in Statistics
►Data Analysis
►Generalized Linear Models
►Model Selection
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Research Designs
►Sample Survey Methods
►Statistical Design of Experiments (DOE)

►Statistical Literacy, Reasoning, and Thinking
►Statistical Software: An Overview
►Statistics: Nelder's view

References and Further Reading

ASA Section on Statistical Consultation (2003) When you consult a statistician ... what to expect? Downloadable from www.amstat.org/sections/cnsl/brochures/SCSbrochure.pdf
Boen JR, Zahn DA (1982) The human side of statistical consulting. Wadsworth, Belmont, CA
Cabrera J, McDougall A (2002) Statistical consulting. Springer, New York
Chatfield C (1995) Problem solving. A statistician's guide, 2nd edn. Chapman & Hall, London
Cox DR (2007) Applied statistics: a review. Ann Appl Stat 1:1–16
Cox DR, Snell EJ (1981) Applied statistics. Principles and examples. Chapman & Hall, London
Derr J (2000) Statistical consulting. A guide to effective communication. Duxbury Press, Pacific Grove, CA
Hand DJ, Everitt BS (1987) The statistical consultant in action. Cambridge University Press, Cambridge
Hyams L (1971) The practical psychology of biostatistical consultation. Biometrics 27:201–211
Joiner BL (1982) Consulting, statistical. In: Encyclopaedia of statistical sciences. Wiley, New York, pp 147–155
Mallows C (1998) The zeroth problem. Am Stat 52:1–9
Speed TP (2005) Terence's stuff: How to do statistical research. IMS Bull 1:6. http://bulletin.imstat.org/archive/34/1
Stinnet SS, Derr JA, Gehan EA (2009) Statistical consulting. In: Encyclopedia of biostatistics, 2nd edn. Wiley, Chichester, UK
Tweedie R et al (1998) Consulting: real problems, real interactions, real outcomes. Stat Sci 13:1–29

Statistical Design of Experiments (DOE)

Jürgen Pilz
Professor, Head of the Institute of Statistics of the University of Klagenfurt
University of Klagenfurt, Klagenfurt, Austria

Model and Denotations

As in regression analysis, DoE is concerned with modelling the dependence of a random target variable Y in dependence of a number of controllable deterministic variables $x_1, \ldots, x_k$ (called *factors*). The major goal of DoE is to find configurations for $\boldsymbol{x} = (x_1, \ldots, x_k)$ out of a given region $V \subset R^k$ which lead to "optimal" results for the target variable under consideration. The different configurations $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$ for the factors are summarized in a statistical design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}) \in V^n$ of *size* n. The optimality criterion is usually defined through some objective function, e.g., the information or ►entropy associated with an experiment, the variance of some predictor $\hat{Y}(\boldsymbol{x}^*)$ for an unobserved configuration $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ etc. The main areas of concern in DoE are:

(a) statistical design in regression analysis and analysis of variance
(b) factorial designs
(c) identification and elimination of disturbing influences (blocking)

This often includes, as a first step, the design of the size of the experiment; i.e., the number of observations n to be taken in order to achieve a predefined goal, see e.g., Rasch et al. (2010). The mean function of $Y = Y(\boldsymbol{x})$ given $\boldsymbol{x} = (x_1, \ldots, x_k) \in V$ is called the *response surface*, usually denoted by $\eta(\boldsymbol{x}) = EY(\boldsymbol{x})$, and the model becomes

$$Y(\boldsymbol{x}) = \eta(\boldsymbol{x}) + \varepsilon, \ \boldsymbol{x} \in V \tag{1}$$

where the random error term is assumed to be independent of $\boldsymbol{x}$ and such that $E(\varepsilon) = Var(\varepsilon) = \sigma^2$. Interpreting $\boldsymbol{x}$ as realisation of a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$, the response function is simply the regression function of Y w.r.t. $\boldsymbol{X}$. The unknown response surface is often modelled through a linear setup

$$\eta(\boldsymbol{x}) = \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \ldots + \beta_r f_r(\boldsymbol{x}) \tag{2}$$

with given functions $f_1, \ldots, f_r$. For example, $\eta(\boldsymbol{x})$ could be a second order polynomial setup

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i \le j=1}^{k} \beta_{ij} x_i x_j \tag{3}$$

arising from a second order Taylor expansion of η. Here, the first sum contains all *main effects* $x_1, \ldots, x_k$ and the second sum contains the (second order) interactions $x_i x_j$.

Optimal Designs

For any given concrete design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$ of size n; where $\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ik})$; $i = 1 \ldots n$ are not necessarily distinct from each other, it is well-known that the estimated response surface yields the best linear unbiased estimate (BLUE)

$$\hat{\eta}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 f_1(\boldsymbol{x}) + \ldots + \hat{\beta}_r f_r(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^T \hat{\boldsymbol{\beta}}$$

where $\boldsymbol{f}(\boldsymbol{x}) = (1, f_1(\boldsymbol{x}), \ldots, f_r(\boldsymbol{x}))^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r)^T$ provided the parameters are estimated by the method of ►least squares (LS); i.e., $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$.
Here $\boldsymbol{Y} = (Y(\boldsymbol{x}_{(1)}), \ldots, Y(\boldsymbol{x}_{(n)}))$ stands for the vector of observations taken at the design points and X for the so-called *design matrix*

$$X = (f_j(\boldsymbol{x}_{(i)})) = \begin{pmatrix} 1 & f_1(\boldsymbol{x}_{(1)}) & \ldots & f_r(\boldsymbol{x}_{(1)}) \\ \vdots & \vdots & & \vdots \\ 1 & f_1(\boldsymbol{x}_{(n)}) & & f_r(\boldsymbol{x}_{(n)}) \end{pmatrix} \tag{4}$$

which is of type $n \times (r+1)$. For a first order regression setup $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ we have $r = k$ and the design matrix has the simple form

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ & \vdots & & \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{x}_{(1)}^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{(n)}^T \end{pmatrix} \tag{5}$$

Criteria for the optimal choice of a design, as e.g., minimum prediction variance, are based on the covariance matrix

$$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

of the LSE $\hat{\beta}$. For i.i.d. normally distributed observations this matrix is proportional to the Fisher information matrix, therefore

$$M(d_n) = \frac{1}{n} X^T X \tag{6}$$

is called the *information matrix* of the design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$. Thus it makes sense to base optimality criteria for designs on functionals of (the inverse of) this matrix.

Definition The design d_n^* is called

(a) *L-optimal* w.r.t. some positive definite matrix U if

$$tr\left(UM\left(d_n^*\right)^{-1}\right) = \min_{d_n} tr(UM(d_n)^{-1})$$

(b) *G-optimal* if it minimizes the maximum variance of $\hat{\eta}(x) = \boldsymbol{f}(\boldsymbol{x})^T \hat{\boldsymbol{\beta}}$ over some region $H \subset R^k$, i.e., $\max_{x \in H} f(x)^T M\left(d_n^*\right)^{-1} f(x) = \min_{d_n} \max_{x \in H} f(x)^T M(d_n)^{-1} f(x)$

(c) *D-optimal* if it minimizes the determinant:

$$\det\left(M\left(d_n^*\right)^{-1}\right) = \min_{d_n} \det\left(M(d_n)^{-1}\right)$$

Important special cases of L-optimality include A-optimality and $\boldsymbol{c}$-optimality, where $U = I_{r+1}$ and $U = \boldsymbol{c}\boldsymbol{c}^T$ for a given vector $\boldsymbol{c} \in R^{r+1}$, respectively. An A-optimal design minimizes the sum of the variances $Var(\hat{\beta}_0) + \ldots + Var(\hat{\beta}_r)$ and thus the average variance of the regression coefficients, and a **c**-optimal design minimizes the variance of the linear combination $Var(\boldsymbol{c}^T \hat{\boldsymbol{\beta}}) = Var(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_r\hat{\beta}_r)$. A D-optimal design minimizes the volume of the dispersion (confidence) ellipsoid for $\hat{\boldsymbol{\beta}}$.

Further criteria and numerical procedures for the construction of optimal designs may be found in Pukelsheim (1993), Atkinson et al. (2001), and Fedorov and Hackl (1997) on the basis of fundamental results by *Kiefer* and Wolfowitz in the late 1950s and early 1960s. Bayesian extensions of this theory are given in Pilz (1991) and Chaloner and Verdinelli (1995). An extensive theory of optimal designs for correlated errors in a spatial setting can be found in Müller (2007), Pilz and Spöck (2008) and Spöck and Pilz (2010) develop a theory of optimal spatial design for the construction of environmental monitoring networks using spectral theory for random fields. Optimal designs for higher-dimensional random fields are considered in Santner et al. (2003), with applications in the area of the design of computer experiments, see also Fang et al. (2005). Here, *Kriging* approximation models are constructed and then used as surrogates for the computer model. The design problem then refers to the optimal choice of the inputs at which to evaluate the computer model. Several software toolboxes are available for constructing optimal designs, see, e.g., Santner et al. (2003), DACE (http://www.2.imm.dtu.dk/hbn/dace) and the R-toolbox DoE (see Rasch et al. 2010).

Factorial Designs

Contrary to the mathematically well-defined optimality criteria considered in the last section, it is also customary to consider heuristically motivated and "practically useful" criteria for the construction of designs. Briefly, the first branch is called the "Kiefer design theory" and the latter branch is referred to as "Box design theory, " in honour of their pioneers.

We assume that the response surface can be sufficiently well described by a polynomial of degree $g \geq 1$ in $k \geq 2$ factors $x_1, \ldots, x_k$. In order to guarantee the non-singularity of the information matrix it is necessary that each factor can take at least $g + 1$ different values, the latter are called the *levels* of the factors. A *factorial design* then means a design which defines a subset of all possible combinations of the levels of the k factors. It is said to be a *full factorial design* if it contains all of the $(g + 1)^k$ combinations of the levels, otherwise it is said to be a *fractional factorial* design. In most applications the response surface is investigated in a sequential manner. In a first step, a screening of the essential factors has to be made, using tools from regression analysis or from multivariate analysis (e.g., ▶principal component analysis). Hereafter, a first order polynomial in the remaining (essential) factors is formed to study the response surface and quantities of interest (e.g., extrema). If this setup is insufficient then a second or third order polynomial setup is chosen and the factor levels are updated until no further significant improvements are obtained. A formal way for proceeding in this manner had already been developed by Box and Wilson in 1951, with the aim of finding factor configurations leading to optimum experimental results.

Full Factorial Designs of the Type 2^k

Usually, one starts with a full factorial design, where all factors are controlled at two levels, "high" and "low," say. Such a design contains 2^k configurations (design points). By an appropriate scaling the design region can be transformed to the k-dimensional cube $V = \{\boldsymbol{x} = (x_1, \ldots, x_k) : -1 \le x_i \le +1,\ i = 1, \ldots, k\}$ and the design points are just the vertices of the cube. The full factorial design of size $n = 2^k$, $d_n = FF(2^k)$ for short, allows the estimation of all 2^k parameters of the model

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}} \beta_{ij} x_i x_{j+\ldots+} \beta_{12\ldots k} x_1 x_2 \ldots x_k \tag{7}$$

As an example, consider a full factorial 2^3 design with factors x_1, x_2, and x_3 which can be adjusted at two levels -1 ("low") and $+1$ ("high"), respectively. The design has $n = 8$ points and allows the estimation of all parameters of the model $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$. The basic structure of this design is displayed in the following table:

Trial no.	Coding	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	(1)	–	–	–	+	+	+	–
2	a	+	–	–	–	–	+	+
3	b	–	+	–	–	+	–	+
4	c	–	–	+	+	–	–	+
5	ab	+	+	–	+	–	–	–
6	ac	+	–	+	–	+	–	–
7	bc	–	+	+	–	–	+	–
8	abc	+	+	+	+	+	+	+

The coding follows the usual standard in the literature; the letters $a, b, c, \ldots$ represent the factors $x_1, x_2, x_3, \ldots$ and are used to indicate that the corresponding factor is adjusted at the level $+1$.

It is easily seen that $M(d_n) = \frac{1}{n} X^T X = I_n$ for a full factorial $d_n = FF(2^k)$ and the estimated regression coefficients are uncorrelated, in case of normally distributed observations they are even independent, and have a simple structure: $\hat{\beta} = \frac{1}{n} X^T \boldsymbol{Y}$, $Cov(\hat{\beta}) = \frac{\sigma^2}{n} I_n$.

Such designs are called *orthogonal*, they can easily be constructed using Hadamard matrices. When restricting attention to first order polynomials $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots +$ $\beta_k x_k$ then an $FF(2^k)$ design leads to minimum variance estimates with $Var(\hat{\beta}_i) = \sigma^2/n$, moreover these full factorial designs turn out to be A-, D- and G-optimal. Finally, the estimated response surface has variance $Var(\hat{\eta}(\boldsymbol{x})) = \frac{\sigma^2}{n}(1 + \boldsymbol{x}^T \boldsymbol{x})$ which only depends on the distance of $\boldsymbol{x} = (x_1, \ldots, x_k)$ from the center point $\boldsymbol{0} = (0, \ldots, 0)^T$ of the design region V. Such designs are called *rotatable*, i.e., for first order polynomial setups full factorial designs of the type 2^k are rotatable.

Fractional Factorial Designs of the Type 2^{k-p}

If the number of factors is getting large, then one is interested in having less than 2^k observations to reduce the experimental efforts. On the other hand, such a reduction is justified if it is clear that there are no higher-order interactions between all or some of the factors. In practical applications it is very common that only the main effects and second-order interaction effects matter. To illustrate this: a full factorial 2^6 design requires $n = 64$ observations, but only 6 degrees of freedom are needed to estimate the main effects and another 15 are needed for the extimation of the two-factorial interchanging effects. Thus, only one third of the 64 observations would be needed for parameter estimation if third- and higher-order interactions were negligible. Therefore, fractional (incomplete) factorial designs are widely used in practice. They had first been introduced by Finney in 1945.

We call a design d_n of size $n = 2^{k-p}$, $1 \le p < k$, a *fractional factorial design* of the type 2^{k-p} if it forms the 2^{-p}-th part of a full factorial design of type 2^k. Such designs are constructed algorithmically by means of p defining relations. To illustrate the ideas, let $k = 4$ and $p = 1$, i.e., we construct a half replication of the $FF(2^k)$ using the defining relation $x_4 = x_1 x_2 x_3$ or, equivalently, multiplying by $x_4, 1 = x_1 x_2 x_3 x_4$.

Using the coding of the previous full factorial $FF(2^3)$ for the new $FF(2^4)$ and observing the defining relation $1 = x_1 x_2 x_3 x_4$ we arrive at the coding for the required fractional factorial 2^{4-1} design: $(1), ab, ac, ad, bc, bd, cd, abcd$. Finally, using the alternative defining relation $1 = -x_1 x_2 x_3 x_4$ we arrive at the alternative 2^{4-1} design: $a, b, c, d, abc, abd, acd, bcd$. The union of both half replicates results in the full factorial $FF(2^4)$ design.

The reduction of the number of observations achieved with fractional factorial designs, however, comes at the price of confounded parameter estimates. In our example, multiplying the defining relation $1 = x_1 x_2 x_3 x_4$ by x_1, x_2, x_3, and x_4, respectively, we obtain $x_1 = x_2 x_3 x_4$, $x_2 = x_1 x_3 x_4$, $x_3 = x_1 x_2 x_4$, $x_4 = x_1 x_2 x_3$, which implies that the main effects parameters $\beta_1, \beta_2, \beta_3$, and β_4 are confounded with

the third-order interaction parameters $\beta_{234}, \beta_{134}, \beta_{124}$, and β_{123}, respectively. From the defining relation $1 = x_1x_2x_3x_4$ itself follows that the intercept term β_0 is confounded with the fourth-order interaction parameter β_{1234}. However, there is no confounding of main effects with low-order interaction (second order interaction) parameters $\beta_{12}, \ldots, \beta_{34}$. Designs d_n for which $n = 2^s$ for some integer $s \geq 2$ are called *regular*, designs for which $n = r + 1$ (= number of unknown regressin parameters) are called *saturated*. Clearly, full factorial as well as fractional factorial designs are regular; full factorial designs $FF(2^k)$ are saturated for the linear regression setup (7) including all possible interactions between the main factors. The construction of saturated orthogonal designs for the hypercube region $V = \{\boldsymbol{x} = (x_1, \ldots, x_k) : -1 \leq x_i \leq +1, i = 1, \ldots, k\}$ is only possible for sizes n which are multiples of 4, such designs had already been constructed by Plackett and Burman in 1946.

Blocking in Factorial Designs

Random disturbances in the experimental conditions lead to an increased variance of the experimental error. In order to reduce this variance it is necessary to randomize the sequence of level combinations of a given design. If the number of factors k is getting larger (which usually implies an increased duration of experimentation in time) then systematic changes in the experimental conditions can occur (e.g., changing weather conditions in agricultural experiments). In this case, reductions in the variance of the experimental error can be achieved by *blocking*. *Blocks* are subsets of an experimental design which are constructed such that they guarantee the homogeneity of experimental conditions within the corresponding subsets. Such blocks can be formed, e.g., from subsets of full or fractional factorial designs, the sequence of trials within the blocks again chosen at random. For example, having k factors $x_1, \ldots, x_k$ and assuming that only the main effects and two-factorial interaction effects are significant, then the response surface takes the form

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}}^{k} \beta_{ij} x_i x_j$$

For an unconfounded estimation of the effects a full factorial $FF(2^k)$ may be chosen, or, for $k \geq 6$, some fractional factorial 2^{k-p} with small $p \geq 1$. In order to take account of the block effect a block factor x_B is introduced, adjusted to the levels of the product $x_1x_2 \ldots x_k$ (or some other generator when starting with a fractional factorial). The block factor x_B can then be interpreted as an indicator variable taking values +1 and −1, and the resulting design can be interpreted as a fractional factorial design of type $2^{(k+1)-1}$ wtih the defining relation $1 = x_1x_2 \ldots x_kx_B$. Assuming the interaction effects $\beta_{1B}, \ldots, \beta_{kB}, \beta_{12B}, \beta_{13B}, \ldots, \beta_{12\ldots kB}$ to be negligible, the main effects and two-factorial interaction effects can be estimated without confounding. Moreover, since the design is orthogonal, blocking has no influence on these estimates.

For further results on fractional factorial designs, blocking, multilevel designs and other topics relevant in the vast field of statistical (optimum) experimental design we refer to the extensive monograph by Wu and Hamada (2009).

About the Author

Dr. Jürgen Pilz is a Professor and Head, Department of Statistics, University of Klagenfurt, Austria. He is the Head of the Department of Statistics since 2007. He is also Director of the Ph.D study program in Mathematics and Statistics at the University of Klagenfurt. He is an Elected member of the International Statistical Institute (1996). He has authored and co-authored more than 100 papers and 6 books, including *Bayesian Estimation and Experimental Design in Linear Regression Models* (Wiley 1991) and *Interfacing Geostatistics and GIS* (Springer, 2009). Professor Pilz was Associate Editor of the following international journals: *Journal of Statistical Planning and Inference* (1992–1999) and Metrika (2004–2009). Currently, he is an Associate editor of *Stochastic Environmental Research and Risk Assessment*. He has supervised more than 25 Ph.D dissertations.

Cross References

►Clinical Trials: An Overview
►Design of Experiments: A Pattern of Progress
►Factorial Experiments
►Optimum Experimental Design
►Randomization
►Uniform Experimental Design

References and Further Reading

Atkinson AC, Bogacka B, Zhigljavsky A (eds) (2001) Optimum design – 2000. Kluwer, Dordrecht, The Netherlands

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:237–304

Fang KT, Fang K, Runze L (2005) Design and modeling for computer experiments. Chapman & Hall/CRC Press, Boca Raton, FL

Fedorov VV, Hackl P (1997) Model-oriented design of experiments. Lecture notes in statistics 125. Springer, Berlin

Müller WG (2007) Collecting spatial data: optimum design of experiments for random fields. Springer, Berlin

Pilz J (1991) Bayesian estimation and experimental design in linear regression models. Wiley, Chichester, UK

Pilz J, Spöck G (2008) Bayesian spatial sampling design. In: Ortiz JM, Emery X (eds) Proceedings of 8th international geostatistics congress Gecamin Ltd., Santiago de Chile, pp 21–30
Pukelsheim F (1993) Optimal design of experiments. Wiley, New York
Rasch D, Pilz J, Verdooren R, Gebhardt A (2011) Optimal Experimental Design with R. Chapman & Hall/CRC Press, Boca Raton, FL
Santner Th, Williams BJ, Notz W (2003) The design and analysis of computer experiments. Springer, Berlin
Spöck G, Pilz J (2010) Spatial sampling design and covariance-robust minimax prediction based on convex design ideas. Stoch Environ Res Risk Assess 24(3):463–482
Wu CFJ, Hamada M (2009) Experiments: planning, analysis and parameter design optimization, 2nd edn. Wiley, New York

Statistical Distributions: An Overview

Kalimuthu Krishnamoorthy
Philip and Jean Piccione Professor of Statistics
University of Louisiana at Lafayette, Lafayette, LA, USA

Introduction

Statistical distributions are used to model sample data that were collected from a population or to model the outcomes of a *random* experiment. The statistical distribution is simply the probability distribution of a random variable. These probability models are commonly used in many applied areas such as economics, education, engineering, social, health, and biological sciences. The distributions of discrete random variables (whose possible values are countable) are referred to as the discrete distribution while those of continuous random variables are called continuous distribution. To begin with an example, let X denote the number of heads that can be observed by flipping a fair coin three times. The sample space of X includes eight outcomes, namely, HHH, HTH, THH, TTH, HHT, HTT, THT, TTT, where H denotes the head and T denotes the tail. The probability that X equals one is the probability of observing any one of the mutually exclusive outcomes TTH, HTT and THT. As all eight outcomes are equally likely, $P(X = 1) = \frac{3}{8}$. Proceeding this way, we obtain the probability distribution of X as

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The above distribution is a member of the family of binomial distributions indexed by n and p, where n is the number of independent Bernoulli trials (each trial results into either "success" or "failure") and p is the probability of observing a success in each trial. The function that gives the probability that a discrete random variable takes a specified value is referred to as the probability mass function (pmf). For example, the pmf of a binomial random variable is given by

$$P(X = x|n,p) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0,1,2,\ldots,n.$$

For a continuous random variable X, $P(X = x) = 0$ for any fixed x, and so we consider only $P(X \in A)$ for any given interval $A \in R$, and this probability can be evaluated as $P(X \in A) = \int_A f(x;\theta)dx$, where $f(x;\theta)$ is called the probability density function (pdf), and θ is a parameter vector. The pdf $f(x)$ should satisfy two conditions: $f(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} f(x;\theta)dx = 1$.

In the following we shall list some commonly used discrete and continuous distributions, their physical significance, relations among them and some measures that describe features of a distribution.

Discrete Distributions

Most commonly used discrete distributions are the binomial, Poisson, hyper geometric, negative binomial and logarithmic series distributions. The first four distributions are closely related. The ▶binomial distribution is used to estimate the proportion of individual with an attribute of interest in a population. In particular, the number of individuals with an attribute of interest in a random sample from a large population (e.g., proportion of defective items in a large shipment) is a binomial random variable with the sample size as the value of n, and the true proportion (usually unknown) in the sampled population is the parameter p. On the other hand, if the sample is drawn (without replacement) from a finite population, then the number of units in the sample with the characteristic of interest is a hypergeometric random variable with the size of the population N (usually known) as the "lot size," the true number of units M (usually unknown) with the attribute in the population as the parameter, and the sample size n as another (known) parameter. The pmf of a hypergeometric random variable is given by $P(X = x|n,M,N) = \binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}$, $L \leq x \leq U$, where $L = \max\{0, M-N+n\}$ and $U = \min\{n, M\}$. If the population is reasonably large, then one could use the binomial model instead of the hypergeometric.

The Poisson distribution (see ▶Poisson Distribution and Its Application in Statistics) is postulated to model the probability distribution of rare events. Specifically, if

Statistical Distributions: An Overview. Table 1 Some discrete distributions

Distribution	Probability mass function	Description
Uniform	$f(x;N) = \frac{1}{N}, \quad k = 1, \ldots, N.$	Positive integer N
Binomial	$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	n = No. of trials p = Success probability
Hypergeometric	$f(x;n,M,N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$, $\max\{0, M-N+n\} \le x \le U = \min\{n, M\}$	n = Sample size; M = No. of defects N = Lot size
Poisson	$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	λ = Mean
Geometric	$f(x;p) = (1-p)^x p$, $x = 0, 1, 2, \ldots$	p = Success probability x = No. of failures until the first success
Negative binomial	$f(x;r,p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots$	p = Success probability x = Number of failures until the rth success
Logarithmic series	$f(x;\theta) = -\frac{\theta^x}{x \ln(1-\theta)}$	$0 < \theta < 1$

an event is almost unlikely to occur in a moment of time, but the number of occurrences over a long period of time could be very large, then a Poisson model is appropriate to describe the frequency distribution of the event. This description implies that the binomial distribution with large n and small p can be approximated by a Poisson distribution with mean $\lambda = np$. More specifically, for a binomial(n,p) random variable with large n and small p, $P(X \le x|n,p) \simeq \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$, $x = 0, 1, \ldots, n$, where $\lambda = np$ and $e^{-\lambda}\lambda^x/x!$ is the pmf of a Poisson random variable with mean λ.

The geometric distribution arises as the probability distribution of number of trials in a sequence of independent Bernoulli trials needed to get the first success. The negative-binomial distribution is a generalization of the geometric distribution where we consider the number of trials required to get r successes. Note that in the binomial distribution, the number of successes in a fixed number of independent Bernoulli trials is a random variable where as in the case of negative-binomial the number of trials is a random variable. The number of failures K in a sequence of independent Bernoulli trials that can be observed before observing exactly r successes is also referred to as the negative-binomial random variable. In the former case, n takes on values $r, r+1, r+2, \ldots$. whereas in the latter case K takes on values 0, 1, 2,.... Both binomial and negative-binomial distributions are related to the beta distribution: If X is a binomial(n, p) random variable then, for $x \neq 0$, $P(X \ge x|n,p) = P(Y \le p)$, where Y is a beta$(x, n-x+1)$ random variable. Also, for $x \neq n$ $P(X \le x|n,p) = P(W \ge p)$, where W is a beta$(x+1, n-x)$ random variable. If X is the number of failures before the rth success (in a sequence of independent Bernoulli trials), then $P(X \le x|r,p) = P(W \le p)$, where W is a beta$(r, x+1)$ random variable. Similar relation exists between the Poisson and the chi-square distributions. Specifically, $P\left(\chi^2_n > x\right) = P(Y \le n/2-1)$, where Y is a Poisson random variable with mean $x/2$.

The probability mass function of a logarithmic series distribution with parameter θ is given by $P(X = k) = \frac{a\theta^k}{k}$, $0 < \theta < 1$, $k = 1, 2, \ldots$, where $a = -1/[\ln(1-\theta)]$. The logarithmic series distribution is useful to describe a variety of biological and ecological data. It is often used to model the number of individuals per species. This distribution is also used to fit the number of products requested per order from a retailer.

Some popular discrete distributions are listed in Table 1. For detailed descriptions, properties and applications of various discrete distributions, see the books by Johnson et al. (1992), Evans et al. (2000), and Krishnamoorthy (2006).

Continuous Distributions

Continuous distributions are grouped into a few families based on the form of pdfs: location family, scale family, location-scale family and exponential family, etc. In the following we shall describe some of these families.

Location-Scale Family: The pdf of a location-scale distribution can be expressed as $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where μ is the location parameter, $\sigma > 0$ is the scale parameter and f is any

pdf that does not depend on any parameter. As an example, the pdf of a normal distribution can be expressed as

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \text{ with}$$
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The two-parameter exponential distribution, normal, Cauchy, double exponential (Laplace), extreme-value and logistic are popular location-scale distributions. The cumulative distribution function (cdf) of a location-scale random variable can be computed using its standard form as $P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right)$. For a location-scale family, $\frac{\hat{\mu}-\mu}{\sigma}$ and $\frac{\hat{\sigma}}{\sigma}$ are pivotal quantities provided $\hat{\mu}$ and $\hat{\sigma}$ are equivariant estimators. These pivotal quantities are useful to find inferential procedures for μ, σ or for any invariant function of (μ, σ).

The normal distribution is the most popular among the location-scale families. In fact there is nothing inherently normal about the normal distribution, and its common use in applications is due its simplicity. Distributions of many commonly used statistics can be approximated by the standard normal distribution via the central limit theorem (see ►Central Limit Theorems). Furthermore, the asymptotic distribution of a maximum likelihood estimator is normal with the variance determined by the Fisher information matrix.

Exponential Family: A family of distributions whose pdf or pmf can be written in the form $f(x;\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^{k} q_i(\theta) w_i(x)\right)$ is called an exponential family. As an example, the binomial family is an exponential family because the pmf $f(x;p) = h(x)c(p) \exp(q_1(p)w_1(x))$, with $h(x) = \binom{n}{x}$, $c(p) = (1-p)^n$, $q_1(p) = \ln(p/(1-p))$ and $w_1(x) = x$. The normal distribution and lognormal distribution are members of exponential families. A statistical model from an exponential family is easy to work with because exponential families have some nice mathematical properties. For instance, it is easier to find sufficient statistics for an exponential family. In fact, for a sample $X_1, \ldots, X_n$ from an exponential family, $\left(\sum_{i=1}^{n} w_1(X_i), \ldots, \sum_{i=1}^{n} w_k(X_i)\right)$ is a sufficient statistic for θ.

Some distributions are routinely used to model lifetime data, and they are referred to as lifetimes (or failure times) distributions. The ►Weibull distribution is one of the most widely used lifetime distributions in reliability and survival analysis. It is a versatile distribution that can take on the characteristics of other types of distributions, based on the value of the shape parameter. If X follows a Weibull distribution with shape parameter c and the scale parameter b, then $\ln(X)$ has the extreme-value distribution with the location parameter $\mu = \ln(b)$ and the scale parameter $\sigma = 1/c$. This one–one relation allows us to transform the results based on a Weibull model to an extreme-value distribution (see ►Weibull distribution). Other lifetime distributions include exponential, two-parameter exponential, lognormal, and gamma distributions. Some popular continuous distributions are listed in Table 2.

Relations Among Distributions: Many of the continuous distributions have one–one relation with others. For example, normal and lognormal (via logarithmic transformation of lognormal random variable), two-parameter exponential and Pareto (via logarithmic transformation of Pareto random variable), two-parameter exponential and power distribution (via negative log transformation of power random variable). This one–one relation enables us to transform some invariant inferential procedures for one distribution to another. Another important distribution that has relation with the t, F, binomial and negative binomial distributions is the beta distribution. An efficient program that evaluates the beta distribution can be used to computes the cumulative distribution functions (cdfs) of other related random variables just cited. The gamma distribution with the shape parameter $\alpha = n/2$ and the scale parameter $\beta = 2$ specializes to the ►chi-square distribution with n degrees of freedom; when $\alpha = 1$, it simplifies to the exponential distribution with mean β. A diagram that describes relations among various distributions is given in Casella and Berger (2002, p. 627).

Moments and Other Measures

Moments are set of measures that are useful to judge some important properties of a probability distribution. Mean and median are commonly used measure of location or center of the distribution. Range and variance are used to quantify the variability of a random variable. We shall now overview some of these measures that describe important characteristics of a distribution.

The mean of a random variable is usually denoted by μ, which is expectation of the random variable. For a discrete random variable X, $\mu = E(X) = \sum_k kP(X = k)$, where the sum is over all possible values of X. If X is continuous, then $\mu = \int_{-\infty}^{\infty} xf(x)dx$, where $f(x)$ is the pdf of X. The expectation $E(X^k)$, $k = 1, 2, \ldots$, is referred to as the kth moment about the origin, while $E(X-\mu)^k$ is called the kth moment about the mean or the kth *central moment*. The second moment about the mean is the variance (denoted by σ^2), and its positive square root is called the standard deviation. The absolute moment $E(|X-\mu|)$ is referred to as the *mean deviation*. The mean deviation and variance

Statistical Distributions: An Overview. Table 2 Some continuous distributions

Distribution	Probability density function	Description of parameters
Uniform	$f(x;a,b) = \frac{1}{b-a}, \quad a \le x \le b$	$a < b$; known or unknown
Normal	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$	$-\infty < \mu < \infty, \sigma > 0$ Mean μ Standard deviation σ
Chi-square	$f(x;n) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1}, \quad x > 0$	Degrees of freedom (df) $n > 0$
F-distribution	$f(x;m,n) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\left(\frac{m}{2}\right)^{m/2} x^{m/2-1}}{\left(\frac{n}{2}\right)^{m/2}\left[1+\frac{mx}{n}\right]^{m/2+n/2}}, x > 0$	m = Numerator df n = Denominator df
Student's-*t*	$f(x;n) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}} \frac{1}{(1+x^2/n)^{(n+1)/2}}, \quad -\infty < x < \infty$	df $n \ge 1$
Exponential	$f(x;\mu,\sigma) = \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)}{\sigma}\right), \quad x > \mu$	Location μ Scale $\sigma > 0$
Gamma	$f(x;a,b) = \frac{1}{\Gamma(a)b^a} e^{-x/b} x^{a-1}, \quad x > 0$	Shape $a > 0$ Scale $b > 0$
Beta	$f(x;a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$	Shape $a > 0$ Scale $b > 0$
Noncentral Chi-square	$f(x;n,\delta) = \sum_{k=0}^{\infty} \frac{\exp\left(-\frac{\delta}{2}\right)\left(\frac{\delta}{2}\right)^k}{k!} \frac{\exp\left(-\frac{x}{2}\right) x^{\frac{n+2k}{2}-1}}{2^{\frac{n+2k}{2}}\Gamma\left(\frac{n+2k}{2}\right)}$	df $n > 0$ δ = Noncentrality parameter > 0
Noncentral *F*	$\text{cdf} = \sum_{k=0}^{\infty} \frac{\exp\left(-\frac{\delta}{2}\right)\left(\frac{\delta}{2}\right)^k}{k!} P\left(F_{m+2k,n} \le \frac{mx}{m+2k}\right)$	Numerator df $m > 0$ Denominator df $n > 0$ Noncentrality parameter $\delta > 0$
Noncentral *t*	$f(x;n,\delta) = \frac{n^{n/2}\exp(-\delta^2/2)}{\sqrt{\pi}\,\Gamma(n/2)(n+x^2)^{(n+1)/2}} \sum_{i=0}^{\infty} \frac{\Gamma[(n+i+1)/2]}{i!} \left(\frac{x\delta\sqrt{2}}{\sqrt{n+x^2}}\right)^i$	df $n \ge 1$ $-\infty < \delta < \infty$
Laplace (Double exponential)	$f(x;a,b) = \frac{1}{2b} \exp\left[-\frac{\lvert x-a\rvert}{b}\right], -\infty < x < \infty$	$-\infty < a < \infty, b > 0$ Location a, scale $b > 0$
Logistic	$f(x;a,b) = \frac{1}{b} \frac{\exp\left\{-\left(\frac{x-a}{b}\right)\right\}}{\left[1+\exp\left\{-\left(\frac{x-a}{b}\right)\right\}\right]^2}, -\infty < x < \infty$	Location a, scale $b > 0$
Lognormal	$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \quad x > 0$	$\sigma > 0, -\infty < \mu < \infty$
Pareto	$f(x;a,b) = \frac{ba^b}{x^{b+1}}, x \ge a$	$a > 0; b > 0$
Weibull	$f(x;b,c,m) = \frac{c}{b}\left(\frac{x-m}{b}\right)^{c-1} \exp\left\{-\left[\frac{x-m}{b}\right]^c\right\}, x > m$	Scale $b > 0$ Shape $c > 0$ Location m
Extreme-value	$f(x;a,b) = \frac{1}{b} \exp\left[-\frac{x-a}{b}\right] \exp\left\{-\exp\left[-\frac{x-a}{b}\right]\right\}$	Location a Scale $b > 0$
Cauchy	$f(x;a,b) = \frac{1}{\pi\, b[1+((x-a)/b)^2]}, \quad -\infty < x < \infty$	Location a, scale $b > 0$
Inverse Gaussian	$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right), x > 0$	$\lambda > 0, \mu > 0$

are used to judge the spread of a distribution. The measure of variability that is independent of the units of measurements is called *coefficient of variation*, and is defined as (standard deviation/mean = σ/μ).

The measures that are used to judge the shape of a distribution are the *coefficient of* ▶*skewness* and the *coefficient of kurtosis* (see ▶Kurtosis: An Overview). The coefficient of skewness is defined as (the third moment about the

mean)/(variance)$^{3/2}$. The skewness measures the lack of symmetry. A negative coefficient of skewness indicates that the distribution is left-skewed (larger proportion of the population is below the mean) while a positive value indicates that the distribution is right-skewed. The *coefficient of kurtosis*, defined as γ = (the fourth moment about the mean)/(variance)^2, is a measure of peakedness or flatness of the probability density curve. As an example, for the normal distribution, the coefficient of skewness is zero (symmetric about the mean), and the coefficient of kurtosis is three. For a Student t distribution with n degrees of freedom, the coefficient of skewness is zero and the coefficient of kurtosis is $3(n-2)/(n-4)$, which approaches 3 as $n \to \infty$.

The ►moment generating function for a random variable is defined as $M_X(t) = E[e^{tX}]$ provided the expectation exists for t in some neighborhood of zero. Note that the kth derivative of $M_X(t)$ evaluated at $t = 0$ is $E(X^k)$, the kth moment about the origin. The logarithm of moment generating function, $G_X(t) = \ln(M_X(t))$, is called the cumulant generating function. The kth derivative of $G_X(t)$ evaluated at $t = 0$ is the kth moment about the mean. Thus, $G'(t)|_{t=0} = \mu$, $G''(t)|_{t=0} = \sigma^2$, and so on.

Fitting a Probability Model

There are several methods available to fit a probability distribution for a given sample data. A popular simple method is quantile–quantile plot (Q-Q plot) which is the plot of the sample quantiles (percentiles) and the corresponding population quantiles. The population quantiles are usually unknown, and they are obtained using the estimates of the model parameters. If the Q–Q plot exhibits a linear pattern, then the data can be regarded as a sample from the postulated probability distribution. There are other rigorous approaches available to check if the sample is from a specific family of distributions. For instance, the Wilks–Shapiro test and the Anderson–Darling test (see ►Anderson-Darling Tests of Goodness-of-Fit) are popular tests to determine if the sample is from a normal population. Another well-known nonparametric test is the ►Kolmogorov–Smirnov test which is based on the difference between the empirical distribution of the sample and the cumulative distribution function of the hypothesized probability model.

Multivariate Distributions

The probability distribution of a random vector is called multivariate distribution. In general, it is assumed that all the components of the random vector are continuous or all of them are discrete. The most popular continuous multivariate distribution is the multivariate normal (see ►Multivariate Normal Distributions). A random vector $\boldsymbol{X}$ is multivariate normally distributed with mean vector $\boldsymbol{\mu}$ and the variance–covariance matrix $\boldsymbol{\Sigma}$ if and only if $\boldsymbol{\alpha X} \sim N(\boldsymbol{\alpha'\mu}, \boldsymbol{\alpha'\Sigma\alpha})$ for every non-zero $\boldsymbol{\alpha'} \in R^p$. Many results and properties of the univariate normal can be extended to the multivariate normal distribution (see ►Multivariate Normal Distributions) using this definition. Even though there are other multivariate distributions, such as multivariate gamma and multivariate beta, are available in literature, their practical applications are not well-known. One of the most popular books in the area of multivariate analysis is Anderson (2003) and its earlier editions.

A popular multivariate discrete distribution is the ►multinomial distribution, which is a generalization of the ►binomial distribution. This distribution is routinely used to analyze the categorical data in the form of contingency table. Another distribution to model a sample of categorical vector observations from a finite population is the multivariate hypergeometric distribution. A useful reference for multivariate discrete distributions is the book by Johnson et al. (1997).

About the Author

Dr. Kalimuthu Krishnamoorthy is Professor, Department of Mathematics, University of Louisiana at Lafayette, Louisiana, USA. He is holder of Philip and Jean Piccione Professor of statistics. He has authored and co-authored more than 75 papers and 2 books, *Handbook of Statistical Distributions with Applications* (Chapman & Hall/CRC, 2006), and *Statistical Tolerance Regions: Theory, Applications and Computation* (Wiley 2009). He is currently an Associate editor for *Communications in Statistics*.

Cross References

►Approximations to Distributions
►Beta Distribution
►Binomial Distribution
►Bivariate Distributions
►Chi-Square Distribution
►Contagious Distributions
►Distributions of Order K
►Exponential Family Models
►Extreme Value Distributions
►F Distribution
►Financial Return Distributions
►Gamma Distribution
►Generalized Extreme Value Family of Probability Distributions
►Generalized Hyperbolic Distributions
►Generalized Rayleigh Distribution
►Generalized Weibull Distributions

►Geometric and Negative Binomial Distributions
►Heavy-Tailed Distributions
►Hyperbolic Secant Distributions and Generalizations
►Hypergeometric Distribution and Its Application in Statistics
►Inverse Gaussian Distribution
►Location-Scale Distributions
►Logistic Distribution
►Logistic Normal Distribution
►Multinomial Distribution
►Multivariate Normal Distributions
►Multivariate Statistical Distributions
►Normal Distribution, Univariate
►Poisson Distribution and Its Application in Statistics
►Relationships Among Univariate Statistical Distributions
►Skew-Normal Distribution
►Skew-Symmetric Families of Distributions
►Student's t-Distribution
►Testing Exponentiality of Distribution
►Uniform Distribution in Statistics
►Univariate Discrete Distributions: An Overview
►Weibull Distribution

References and Further Reading

Anderson TW (2003) An introduction to multivariate statistical analysis. Wiley, New York
Casella G, Berger RL (2002) Statistical inference. Duxbury, Pacific Grove, CA
Evans M, Hastings N, Peacock B (2000) Statistical distributions. Wiley, New York
Johnson NL, Kotz S, Kemp AW (1992) Univariate discrete distributions. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York
Krishnamoorthy K (2006) Handbook of statistical distributions with applications. Chapman & Hall/CRC Press, Boca Raton, FL
Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York

Statistical Ecology

David Fletcher
Associate Professor
University of Otago, Dunedin, New Zealand

Ecologists study complex systems, and often need to use non-standard methods of sampling and data analysis. The data might be collected over a long-time scale, involve little spatial replication, or be highly aggregated in space. There have been many fruitful collaborations between ecologists and statisticians, often leading to the development of new statistical methods. In this brief overview of the subject, I will focus on three areas that have been of particular interest in the management of animal populations. I will also discuss the use of statistical methods in other areas of ecology, the aim being to highlight interesting areas of development rather than a comprehensive review.

Mark-Recapture Methods

Mark-recapture methods are commonly used to estimate abundance and survival rates of animal populations (Lebreton et al. 1992; Williams et al. 2002). Typically, a number of individuals are physically captured, marked and released. The information obtained from successive capture occasions is summarized in a "capture history," which indicates whether or not an individual was captured on the different occasions. The likelihood is specified in terms of demographic parameters of interest, such as annual survival probabilities, and nuisance parameters that model the capture process. A range of goodness-of-fit diagnostics have been developed, including estimation of overdispersion (Anderson et al. 1994). Overdispersion usually arises as a consequence of heterogeneity, or lack of independence, amongst individuals in the survival and/or capture probabilities; attempts have also been made to model such heterogeneity directly (Pledger et al. 2003). ►Model selection often involves use of ►Akaike's information criterion (AIC), and model-averaging is also commonly used (Johnson and Omland 2004). Bayesian methods are becoming popular, particularly as means of fitting hierarchical models (Brooks et al. 2000). Recent developments include the use of genotyping of fecal, hair or skin samples to identify individuals (Lukacs and Burnham 2005; Wright et al. 2009), and spatially-explicit models that allow estimation of population density (Borchers and Efford 2008). A related area of recent interest has been the estimation of the occupancy rate, i.e., the proportion of a set of geographical locations that are occupied by a species (MacKenzie et al. 2006). This can be of interest in large-scale monitoring programs, for which estimation of abundance is too costly, and in understanding metapopulation dynamics. In this setting, the "individuals" are locations and the "capture history" records whether or not a species was observed at that location, on each of several occasions.

Distance Sampling

A common alternative method for estimating population abundance or density is distance sampling. This involves recording the distance of each observed individual from

a transect line or a point. The analysis then involves estimation of the probability of detection of an individual as a function of distance (Buckland et al. 2004), thereby allowing estimation of the number of individuals that have not been detected. Two important assumptions in using this method is that detection is certain for an individual on the line or point and that individuals do not move during the observation process, although modifications have been suggested for situations in which these assumptions are not met (Borchers et al. 1998; Buckland and Turnock 1992). Compared to the use of mark-recapture methods for estimating abundance, distance sampling typically provides savings in terms of field effort, and will usually be more appropriate when the population is widely dispersed. A useful discussion of the theory underlying use of distance sampling is given by Fewster and Buckland (2004), while Schwarz and Seber (1999) provide an extensive review of methods for estimating abundance.

Population Modeling

Population projection models have long been used as a tool in the process of managing animal and plant populations, most often as means of assessing the impact of management on the population growth rate or on the probability of quasi-extinction (Caswell 2001; Burgman et al. 1993). A population model will typically involve one or more demographic parameters, such as annual survival probabilities and annual reproductive rates, for individuals in different ages or stages. In the past, estimation of the parameters has been performed by separately fitting statistical models to the different sets of data; recent work in this area has focussed on regarding the population model as a statistical model that can be fitted to all the available data (Buckland et al. 2007). The benefit of this approach is that all the uncertainty can be allowed for, and that estimation of the parameters can be improved by including data that provide a direct indication of the population growth rate (Besbeas et al. 2002). This development has the potential to allow ecologists to fit a broad range of population models to their data, including ones that allow for immigration (cf., Nichols and Hines 2002; Peery et al. 2006).

Other Developments

A key aspect of studying many plant and animal populations is their aggregated spatial distribution. This distribution might be of interest in itself, or be something that needs to be allowed for in the sampling and data analysis. There is a long tradition of the analysis of spatial pattern in ecology, involving a range of statistical techniques, including distance-based methods and spatial ►point processes (Fortin and Dale 2005). Various statistical distributions have been suggested as a means of allowing for the fact that aggregation often leads to zero-inflated and/or positively skewed data. These include the negative binomial, lognormal and gamma distributions, plus zero-inflated versions of these (Dennis and Patil 1984; Martin et al. 2005; Fletcher 2008). Likewise, methods have been developed for fitting models that incorporate spatial autocorrelation (Legendre 1993; Fortin and Dale 2005).

►Adaptive sampling is a modification of classical sampling that aims to allow for spatial aggregation by adaptively increasing the sample size in those locations where the highest abundances have been found in an initial sample (Thompson and Seber 1996; Brown and Manly 1998). Information on the number and relative abundance of individual species in one or more geographical areas has been of interest to many ecologists, leading to the use of species abundance models (Hughes 1986; Hill and Hamer 1998), estimation of species richness (Chao 2005), modeling species-area relationships (Connor and McCoy 2001), and the analysis of species co-occurrence (Mackenzie et al. 2004; Navarro-Alberto and Manly 2009).

In studying ecological communities, it is often natural to consider the use of multivariate methods. There is a large literature in this area, primarily focussing on classification and ordination techniques for providing informative summaries of the data (McGarigal et al. 2000). Likewise, multivariate analysis of variance (see ►Multivariate Analysis of Variance (MANOVA)) has been used to assess the ecological impact of human disturbance on a range of species (Anderson and Ter Braak 2003).

In order to study processes operating at large spatial scales, it is useful to carry out studies at those scales. In doing so, there is a tension between satisfying the statistical requirements of replication and keeping the study at a scale that is large enough to provide meaningful results (Schindler 1998; Hewitt et al. 2007). There has been some discussion in the ecological literature regarding appropriate statistical methods for such studies (Cottenie and De Meester 2003). One approach is to consider a single large-scale study as insufficient to provide the level of evidence that is usually required of a small-scale experiment, with the hope that information from a number of studies can eventually be combined, either informally of using meta analysis (Gurevitch and Hedges 1999).

Future

It is clear that the increasing popularity of computationally-intensive Bayesian methods of analysis will lead to ecologists being able to fit statistical models that provide them

with a better understanding of the spatial and temporal processes operating in their study populations (Clark 2007). Likewise, recently-developed techniques such as ►neural networks (Lek et al. 1996) and boosted trees (Elith et al. 2008), are likely to appear more frequently in the ecological literature. In tandem with the development of new techniques, there will always be a need to balance complexity and simplicity in the analysis of ecological data (Murtaugh 2007).

About the Author

David Fletcher is regarded as one of New Zealand's top ecological statisticians. He has worked in statistical ecology since arriving in New Zealand 20 years ago, both in academia and as a private consultant. He is the author of 70 refereed papers and numerous technical reports for government agencies.

Cross References

►Adaptive Sampling
►Akaike's Information Criterion
►Analysis of Areal and Spatial Interaction Data
►Distance Sampling
►Non-probability Sampling Survey Methods
►Spatial Point Pattern
►Statistical Inference in Ecology

References and Further Reading

Anderson MJ, Ter Braak CJF (2003) Permutation tests for multi-factorial analysis of variance. J Stat Comput Sim 73:85–113

Anderson DR, Burnham KP, White GC (1994) AIC model selection in overdispersed capture-recapture data. Ecology 75:1780–1793

Besbeas P, Freeman SN, Morgan BJT, Catchpole EA (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics 58:540–547

Borchers DL, Efford MG (2008) Spatially explicit maximum likelihood methods for capture-recapture studies. Biometrics 64:377–385

Borchers DL, Zucchini W, Fewster RM (1998) Mark-recapture models for line transect surveys. Biometrics 54:1207–1220

Brooks SP, Catchpole EA, Morgan BJT (2000) Bayesian annual survival estimation. Stat Sci 15:357–376

Brown JA, Manly BJF (1998) Restricted adaptive cluster sampling. Environ Ecol Stat 5:49–63

Buckland ST, Turnock BJ (1992) A robust line transect method. Biometrics 48:901–909

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) (2004) Advanced distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford

Buckland ST, Newman KB, Fernández C, Thomas L, Harwood J (2007) Embedding population dynamics models in inference. Stat Sci 22:44–58

Burgman MA, Ferson S, Akcakaya HR (1993) Risk assessment in conservation biology. Chapman and Hall, London

Caswell H (2001) Matrix population models, 2nd edn. Sinauer Associates, Massachusetts

Chao A (2005) Species richness estimation. In: Encyclopedia of statistical sciences, 2nd edn. Wiley, New York

Clark JS (2007) Models for ecological data: an introduction. Princeton University Press, Princeton, NJ

Connor EF, McCoy ED (2001) Species-area relationships. In: Encyclopedia of biodiversity, vol 5. Academic, New York, pp 397–411

Cottenie K, De Meester L (2003) Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). Oikos 100:394–396

Dennis B, Patil GP (1984) The gamma distribution and weighted multimodal gamma distributions as models of population abundance. Math Biosci 68:187–212

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Fewster RM, Buckland ST (2004) Chapter 10 of advanced distance sampling: estimating abundance of biological populations. In: Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) Advanced distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford

Fletcher DJ (2008) Confidence intervals for the mean of the delta-lognormal distribution. Environ Ecol Stat 15:175–189

Fortin M-J, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge

Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. Ecology 80:1142–1149

Hewitt JE, Thrush SF, Dayton PK, Bonsdorff E (2007) The effect of spatial and temporal heterogeneity on the design and analysis of empirical studies of scale-dependent systems. Am Nat 169:398–408

Hill JK, Hamer KC (1998) Using species abundance models as indicators of habitat disturbance in tropical forests. J Appl Ecol 35:458–460

Hughes RG (1986) Theories and models of species abundance. Am Nat 128:879–899

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Lebreton J-D, Burnham KP, Clobert J, Anderson DR (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecol Monogr 62:67–118

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659–1673

Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S (1996) Application of neural networks to modelling nonlinear relationships in ecology. Ecol Model 90:39–52

Lukacs PM, Burnham KP (2005) Review of capture-recapture methods applicable to noninvasive genetic sampling. Mol Ecol 14:3909–3919

McArdle BH (1996) Levels of evidence in studies of competition, predation, and disease. New Zeal J Ecol 20:7–15

Mackenzie DI, Bailey LL, Nichols JD (2004) Investigating species co-occurrence patterns when species are detected imperfectly. J Anim Ecol 73:546–555

MacKenzie D, Nichols J, Royle J, Pollock K, Bailey L, Hines J (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic
McGarigal K, Cushman S, Stafford S (2000) Multivariate statistics for wildlife and ecology research. Springer, New York
Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8:1235–1246
Murtaugh PA (2007) Simplicity and complexity in ecological data analysis. Ecology 88:56–62
Navarro-Alberto JA, Manly BFJ (2009) Null model analyses of presence–absence matrices need a definition of independence. Popul Ecol 51:505–512
Nichols JD, Hines JE (2002) Approaches for the direct estimation of λ, and demographic contributions to λ, using capture-recapture data. J Appl Stat 29:539–568
Peery MZ, Becker BH, Beissinger SR (2006) Combining demographic and count-based approaches to identify source–sink dynamics of a threatened seabird. Ecol Appl 16:1516–1528
Pledger S, Pollock KH, Norris JL (2003) Open capture-recapture models with heterogeneity: I Cormack-Jolly-Seber model. Biometrics 59:786–794
Schindler DW (1998) Replication versus realism: the need for ecosystem-scale experiments. Ecosystems 1:323–334
Schwarz CJ, Seber GAF (1999) Estimating animal abundance: review III. Stat Sci 14:427–456
Taylor LR (1961) Aggregation, variance and the mean. Nature 189:732–735
Thompson SK, Seber GAF (1996) Adaptive sampling. Wiley, New York
Williams BK, Conroy MJ, Nichols JD (2002) Analysis and management of animal populations. Academic, San Diego, CA
Wright JA, Barker RJ, Schofield MR, Frantz AC, Byrom AE, Gleeson DM (2009) Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65:833–840

Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

Abdelhakim Necir
Professor
Mohamed Khider University of Biskra, Biskra, Algeria

Introduction

Risk measures are used to quantify insurance losses and measure financial risk assessments. Several risk measures have been proposed in actuarial science literature, namely, the value at risk, the expected shortfall or the conditional tail expectation, and the distorted risk measures (DRM). Let X be a nonnegative random variable (rv) representing losses of an insurance company with a continuous distribution function (df) F. The DRM of X is defined by

$$\Pi_g = \int_0^\infty g\left(1 - F\left(x\right)\right) dx,$$

where the distortion function g is an increasing function such that $g\left(0\right) = 0$ and $g\left(1\right) = 1$ (see, Wang 1996). In terms of the generalized inverse function $Q(s) := \inf\left\{x : F(x) \geq s\right\}$, the DRM may rewritten as

$$\Pi_g = \int_0^1 g'\left(s\right) Q\left(1 - s\right) ds,$$

provided that g is differentiable. In this entry, we consider the DRM corresponding to the distortion function $g\left(s\right) = s^{1/\rho}$, $\rho \geq 1$ called the proportional hazard transform (PHT) risk measure. In this case we write

$$\Pi_\rho = \rho^{-1} \int_0^1 s^{1/\rho - 1} Q\left(1 - s\right) ds.$$

Empirical Estimation of Π_ρ

Suppose we have independent random variables $X_1, X_2, \ldots,$ each with the cdf F, and let $X_{1:n} < \ldots < X_{n:n}$ be the ▶order statistics corresponding to $X_1, \ldots, X_n$. It is most natural to define an empirical estimator of Π_ρ as follows

$$\widehat{\Pi}_\rho := \rho^{-1} \int_0^1 s^{1/\rho - 1} Q_n\left(1 - s\right) ds,\ \rho \geq 1, \tag{1}$$

where $Q_n(s)$ is the empirical quantile function, which is equal to the ith order statistic $X_{i:n}$ when $s \in \left((i-1)/n, i/n\right]$, $i = 1, \ldots, n$. We note that $\widehat{\Pi}_\rho$ is a linear combinations order statistics, that is, $\widehat{\Pi}_\rho = \sum_{i=1}^n a_{i,n} X_{n-i+1,n}$, with $a_{i,n} := \rho^{-1} \int_{(i-1)/n}^{i/n} s^{1/\rho - 1} ds$, $i = 1, \ldots, n$, and $n \in \mathbb{N}$. A statistic having the form (1) is an L-statistic (see, for instance, Shorack and Wellner 1986, p. 260). The ▶asymptotic normality of the estimator $\widehat{\Pi}_\rho$ is discussed in Jones and Zitikis (2003).

Theorem 1 (Jones and Zitikis, 2003). *For any* $1 < \rho < 2$, *we have*

$$n^{1/2}\left(\widehat{\Pi}_\rho - \Pi_\rho\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_\rho^2\right),\ as\ n \to \infty,$$

where

$$\sigma_\rho^2 := \rho^{-2} \int_0^1 \int_0^1 (min(s,t) - st) s^{1/\rho - 1} t^{1/\rho - 1} dQ\left(1 - s\right) dQ\left(1 - t\right),$$

provided that $\mathbb{E}\left[X^\eta\right] < \infty$ *for some* $\eta > 2\rho/\left(2 - \rho\right)$.

The premium, which is greater than or equal to the mean risk, must be finite for any $\rho \geq 1$. That is, we have $1 \leq \rho < 1/\gamma$. For $\gamma > 1/2$, the second-order moment $\mathbb{E}\left[X^2\right]$ is infinite and $1 \leq \rho < 2$. In this case, we have $2\rho/\left(2 - \rho\right) > 2$ that implies that $\mathbb{E}\left[|X|^\eta\right]$ is infinite for any $\eta > 2\rho/\left(2 - \rho\right)$. Therefore, Theorem 1 does not hold for regularly varying

distributions with tail indices $-1/\gamma > -1/2$. To solve this problem, we propose an alternative estimator for Π_ρ with normal asymptotic distribution for any $-1/\gamma > -1/2$. To get into a more general setting, assume that F is heavy-tailed, which means that $\lim_{x\to\infty} e^{\lambda x}(1-F(x)) = \infty$ for every $\lambda > 0$. The class of regularly varying cdfs is a good example for heavy-tailed models: The cdf F is said to be regularly varying at infinity with index $(-1/\gamma) < 0$ if the condition

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}, \tag{2}$$

is satisfied for every $x > 0$. This class includes a number of popular distributions such as Pareto, Generalized Pareto, Burr, Fréchet, Student, …, which are known to be appropriate models for fitting large insurance claims, large fluctuations of prices, log-returns, etc. (see, e.g., Beirlant et al. 2001). In the remainder of this entry, we therefore restrict ourselves to this class of distributions, and for more information on them we refer to, for example, de Haan and Ferreira (2006).

New Estimator for Π_ρ : Extreme Values Based Estimation

We have already noted that the estimator $\widehat{\Pi}_\rho$ does not yield asymptotic normality beyond the condition $\mathbf{E}[X^2] < \infty$. For this reason, Necir and Meraghni (2009) proposed an alternative of PHT estimator, which would take into account differences between moderate and high quantiles, that is

$$\widetilde{\Pi}_\rho := \sum_{i=k+1}^{n} a_{i,n} X_{n-i+1,n} + (k/n)^{1/\rho} \frac{X_{n-k,n}}{1-\rho\widehat{\gamma}_n},$$

where we assume that the tail index $\gamma \in [1/2, 1)$ and estimate it using the Hill (1975) estimator $\widehat{\gamma}_n := k^{-1}\sum_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n}$. Here, let $k = k_n$ be a sequence such that $k \to \infty$, and $k/n \to 0$ as $n \to \infty$. The construction of this estimator is inspired from the work of Necir et al. (2007) and Necir and Boukhetala (2004).

Asymptotic Normality of $\widetilde{\Pi}_\rho$

The main theoretical result of this entry is Theorem 2, below, in which we establish weak approximations for $\widetilde{\Pi}_\rho$ by functional of Brownian bridges and therefore asymptotic confidence bounds for Π_ρ. To formulate it, we need to introduce an assumption that ensures the weak approximation of Hill's estimator $\widehat{\gamma}_n$. The assumption is equivalent to the following second-order condition (see Geluk et al. 1997). Namely, it said that the cdf F satisfies the generalized second-order regular variation condition with second-order parameter $\beta \le 0$ (see de Haan and Stadtmüller 1996) if there exists a function $a(s)$, which does not change its sign in a neighborhood of infinity and is such that, for every $x > 0$,

$$\lim_{s\to\infty} (a(s))^{-1}\left\{\frac{1-F(sx)}{1-F(s)} - x^{-1/\gamma}\right\} = x^{-1/\gamma}\frac{x^{\rho/\gamma}-1}{\rho/\gamma}, \tag{3}$$

where $\rho \le 0$ is the so-called second-order parameter; when $\rho = 0$, then the ratio on the right-hand side of Eq. (3) should be interpreted as $\log x$. In the formulation of Theorem 2, we shall use $A(z) := \gamma^2 a(\mathbb{U}(z))$ with $a(s)$ as above and $\mathbb{U}(z) := Q(1-1/z)$.

Theorem 2 (Necir and Meraghni 2009). *Let F be a df satisfying (2) with $\gamma > 1/2$ and suppose that $Q(\cdot)$ is continuously differentiable on $[0,1)$. Let $k = k_n$ be such that $k \to \infty$, $k/n \to 0$ and $k^{1/2}A(n/k) \to 0$ as $n \to \infty$. For any $1 \le \rho < 1/\gamma$, there exists a sequence of independent Brownian bridges (B_n) such that*

$$\frac{n^{1/2}\left(\widetilde{\Pi}_\rho - \Pi_\rho\right)}{(k/n)^{1/\rho-1/2}\,Q(1-k/n)} =_d \mathcal{L}_1(B_n,\rho,\gamma) + o_p(1),$$

where

$$\mathcal{L}_1(B_n,\rho,\gamma) := \delta(\rho,\gamma)(n/k)^{1/2}B_n(1-k/n) - \lambda_{\rho,\gamma}(n/k)^{1/2}\int_{1-k/n}^{1}\frac{B_n(s)}{1-s}ds - \frac{\rho^{-1}\int_{k/n}^{1} s^{1/\rho-1}B_n(1-s)\,Q'(1-s)\,ds}{(k/n)^{1/\rho-1/2}\,Q(1-k/n)},$$

with $\delta(\rho,\gamma) := \lambda_{\rho,\gamma}\left(\rho\gamma^2 - \gamma + 1 - \gamma\lambda_{\rho,\gamma}^{-1}\right)$, and $\lambda_{\rho,\gamma} := \frac{\rho\gamma}{(1-\rho\gamma)^2}$.

corollary 1 Under the assumptions of Theorem 2, we have

$$\frac{n^{1/2}\left(\widetilde{\Pi}_{\rho,n} - \Pi_\rho\right)}{(k/n)^{1/\rho-1/2}\,X_{n-k:n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\sigma_{\rho,\gamma}^2\right), \text{ as } n\to\infty,$$

where the asymptotic variance $\sigma_{\rho,\gamma}^2$ is given by the sum of the following terms

$$\kappa_1 = \frac{\left(\gamma\rho - \gamma + \gamma^2\rho\right)^2}{(1-\rho\gamma)^4}, \kappa_2 = \frac{2\rho^2\gamma^2}{(1-\rho\gamma)^4}$$

$$\kappa_3 = \frac{2\gamma^2}{(1-\rho-\rho\gamma)(2-\rho-2\rho\gamma)}, \kappa_4 = \frac{2\rho\gamma\left(\gamma-\gamma\rho-\gamma^2\rho\right)}{(1-\rho\gamma)^4}$$

and $\kappa_5 = -\frac{2\rho\gamma^3}{(1-\rho\gamma)^2}$.

Cross References

►Actuarial Methods
►Asymptotic Normality

►Estimation: An Overview
►Heavy-Tailed Distributions
►Insurance, Statistics in
►Risk Analysis

References and Further Reading

Artzner Ph, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. Math Financ 9:203–228
Beirlant J, Matthys G, Dierckx G (2001) Heavy-tailed distributions and rating. Astin Bull 31:37–58
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
de Haan L, Stadtmüller U (1996) Generalized regular variation of second order. J Aust Math Soc A 61:381–395
Geluk J, de Haan L, Resnick S, Starica C (1997) Second order regular variation, convolution and the central limit theorem. Stoch Proc Appl 69:139–135
Hill BM (1975) A simple approach to inference about the tail of a distribution. Ann Stat 3:1136–1174
Jones BL, Zitikis R (2003) Empirical estimation of risk measures and related quantities. N Am Actuarial J 7:44–54
Necir A, Boukhetala K (2004) Estimating the risk adjusted premium of the largest reinsurance covers. In: Antoch J (ed) Proceeding of computational statistics, Physica-Verlag, pp 1577–1584. http://www.springer.com/statistics/computanional+statistics/book/978-3-7908-1554-2
Necir A, Meraghni D (2009) Empirical estimation of the proportional hazard premium for heavy-tailed claim amounts. Insur Math Econ 45:49–58
Necir A, Meraghni D, Meddi F (2007) Statistical estimate of the proportional hazard premium of loss. Scand Actuarial J 3:147–161
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley, New York
Wang SS (1996) Premium calculation by transforming the layer premium density. Astin Bull 26:71–92

Statistical Evidence

Subhash R. Lele[1], Mark L. Taper[2]
[1]Professor
University of Alberta, Edmonton, AB, Canada
[2]Research Scientist
Montana State University, Bozeman, MT, USA

Scientists want to know how nature works. Different scientists have different ideas or hypotheses about the mechanisms that underlie a phenomenon. To test the validity of these ideas about mechanisms, they need to be translated into quantitative form in a mathematical model that is capable of predicting the possible outcomes from such mechanisms. Observations of real outcomes, whether obtained by designed experiment or observational study, are used to help discriminate between different mechanisms. The classical approach of hypothesis refutation depends on showing that the data are impossible under a specific hypothesis. However, because of the intrinsic stochasticity in nature, appropriate mathematical models tend to be statistical rather than deterministic. No data are impossible under a statistical model and hence this classic approach cannot be used to falsify a statistical model. On the other hand, although not impossible, data could be more improbable under one statistical model than a competing one. Quantifying evidence for one statistical model vis-à-vis a competing one is one of the major tasks of statistics. The evidential paradigm in statistics addresses the fundamental question: How should we interpret the observed data as evidence for one hypothesis over the other? Various researchers have tried to formulate ways of quantifying evidence, most notably Barnard (1949) and Edwards (1992). The monograph by Hacking (Hacking 1965) explicitly stated the problem and its solution in terms of the law of the likelihood:

► *If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$ and the likelihood ratio $p_A(x) > p_B(x)$, measures the strength of that evidence.*

Royall (1997) developed this simple yet powerful idea and turned it into something that is applicable in practice. He emphasized that the commonly used approaches in statistics are either decision-theoretic (Neyman-Pearson-Wald) that address the question "given these data, what should I do?" or, are belief based (Bayesian) that address the question "given these data, how do I change my beliefs about the two hypotheses?" He suggested that statisticians should first address the more fundamental question "how should we interpret the observed data as evidence for one hypothesis over the other?", and only then think about how the beliefs should be changed or decisions should be made in the light of this evidence. Royall also pointed out that evidence is a strictly comparative concept. We need two competing hypotheses before we can compare the evidence for one over the other. His critique of the commonly used evidence measures showed that the practice of using Fisherian p-value as a measure of evidence is incorrect because it is not a comparative measure, while the Bayesian posterior probability, aside from being dependent on the prior beliefs and not solely on the observed data, is also an incorrect measure of evidence because it is not invariant to the choice of the parameterization.

One of the reasons, the Neyman-Pearson ideas are prominent in science is that they accept the fact that decisions can go wrong. Hence in scientific practice, one quantifies and controls the probabilities of such wrong decisions. Royall (1997) introduced concepts of error probabilities that are similar to the Type-I and Type-II error probabilities in the Neyman-Pearson formulation, but relevant to the evidential paradigm. He realized, evidence, properly interpreted, can be misleading and asked how often would be we be misled by strong evidence (see below) if we use the law of the likelihood and how often would we be in a situation that neither hypothesis is supported to the threshold of strong evidence.

Three concepts answer those questions. Suppose we say that hypothesis A has strong evidence supporting it over hypothesis B if the likelihood ratio is greater than K, for some a priori fixed $K > 0$. Then:

(a) The probability of misleading strong evidence: $M(K) = P_A\left(x : \frac{p_B(x)}{p_A(x)} > K\right)$,
(b) The probability of weak evidence: $W(K) = P_A\left(x : \frac{1}{K} < \frac{p_B(x)}{p_A(x)} < K\right)$,
(c) The probability of strong evidence for the correct model: $S(K) = P_A\left(x : \frac{p_A(x)}{p_B(x)} > K\right)$.

A remarkable result that follows is that there exists a universal upper bound on the probability of misleading evidence under any model, namely $M(K) \leq 1/K$. Furthermore, as one increases the sample size, both $M(K)$ and $W(K)$ converge to 0 and $S(K) \to 1$. Thus, with enough observations we are sure to reach the right conclusion without any error. This is in stark contrast with the Neyman-Pearson Type-I error that remains fixed, no matter how large the sample size. In the Neyman-Pearson formulation, as sample size increases, K increases while error probability is held constant. Thus, as one increases the sample size, the criterion for rejection changes so that it is harder and harder to distinguish the hypotheses. This seems quite counter-intuitive and makes it difficult to compare tests of different sample size.

The concepts of misleading and weak evidence have implications in the sample size calculations and optimal experimental designs. For example, the experimenter should make sure the minimal sample size is such that probability of weak evidence is quite small and at the end of the experiment one can reach a conclusion. Furthermore, by controlling the probability of misleading evidence through sample size, experimental/sampling design and evidence threshold one can also make sure that the conclusions reached are likely to be correct. Besides these *a priori* uses, the probability of misleading evidence can be calculated as a post data error statistic reminiscent of a p-value, but explicitly constructed for the comparison of two hypotheses (Taper and Lele 2010).

There are, however, limitations to the evidential ideas developed by Royall and described above. One major limitation is that the law of likelihood can only quantify evidence when the hypotheses are simple, but most scientific problems involve comparing composite hypotheses. This may arise because the scientist may be interested in testing only some feature of the model without restrictions on the rest of the features. Similarly, a proper statistical model might involve infinitely many nuisance parameters in order to model the underlying mechanism realistically but the parameters of interest may be finite. Such cases arise in many practically situations, for example, the longitudinal data analysis or random effects models among others. Aside from raising the need to consider composite hypothesis, in these situations, the full likelihood function may be difficult to write down. One may want to specify only a few features of the model such as the mean or the variance, leading to the use of quasi-likelihood, estimating functions and such other modifications. The question of ►model selection where one is selecting between families of models instead of a specific element of a given family is important in scientific practice. For example, whether to use a linear regression model (see ►Linear Regression Models) or a non-linear regression model (see ►Nonlinear Regression) is critical for forecasting.

Can we generalize the law of likelihood and concepts of error probabilities to make it applicable in such situations? An initial attempt is described in Lele (2004), Taper and Lele (2004, 2010). The key observation in such a generalization is that quantifying the strength of evidence is the same as comparing distances between the truth and the competing models that are estimated from data. The likelihood ratio simply compares an estimate of the ►Kullback-Leibler divergence.

One can consider many different kinds of divergences, each leading to different desirable properties. For example, if one uses Hellinger distance to quantify strength of evidence, one gets a measure that is robust against ►outliers. If one uses Jeffrey's divergence, one needs to specify only the mean and variance function, similar to the quasi-likelihood formulation, to quantify strength of evidence. One can use profile likelihood or integrated likelihood or conditional likelihood to compare evidence about a parameter of interest in the presence of nuisance parameters. These simply correspond to different divergence measures and hence have different properties. Lele (2004) terms these as "evidence functions". They may be

compared in terms of how fast the probability of strong evidence for the correct model converges to 1. Not surprisingly, for simple versus simple hypothesis comparison, it turns out that the Kullback-Leibler divergence or the likelihood ratio is the best evidence function, provided the model is correctly specified. Other evidence functions, however, might be more robust against outliers or may need less specification; and hence may be more desirable in practice.

Error probabilities can be calculated for general evidence functions using bootstrapping (Taper and Lele 2010). When the data are independent and identically distributed one can circumvent the conceptual constraint that the true model is in one of the alternative hypotheses by using a non-parametric bootstrap. We briefly describe this in the likelihood ratio context. Notice that the likelihood ratio is simply a statistic, a function of the data. One can generate a ►simple random sample with replacement from the original data and compute the strength of evidence based on this new sample. By repeating this procedure large number of times, one obtains the bootstrap estimate of the distribution of the strength of evidence. The percentile-based confidence interval tells us the smallest level of strength of evidence one is likely to obtain if the experiment is repeated. One of the vexing questions in evidential paradigm is how to relate evidence to decision making without invoking beliefs. It may be possible to use the bootstrap distribution of the strength of evidence, in conjunction with the ►loss function, for decision-making. Because this distribution is obtained empirically from the observations, such decisions will be robust against model specifications.

The evidential paradigm is still in its adolescence, with much scope for innovation. Nevertheless the paradigm is sufficiently developed to make immediate contributions; in fact, information criterion comparisons, which are evidence functions, have already revolutionized the practice of many sciences. The references below will be useful to further widen the reader's knowledge and understanding beyond just our views.

About the Authors

Dr. Subhash Lele is a professor of statistics in the Department of Mathematical and Statistical Sciences, University of Alberta, Canada. He has published over 70 papers in statistical and scientific journals on various topics such as morphometrics, quantitative ecology, hierarchical models, estimating functions and philosophy of statistics. He has served as the President of the Biostatistics section of the Statistical Society of Canada and Secretary of the Statistical Ecology section of the Ecological Society of America. He is an Elected member of the International Statistical Institute. He has served on the editorial boards of the Journal of the American Statistical Association, Ecology and Ecological Monographs and Ecological and Environmental Statistics. He has co-authored (with Dr. J.T. Richtsmeier) a book *An invariant approach to statistical analysis of shapes* (Chapman and Hall, 2000) and co-edited a book (with Dr. M.L. Taper) on *The nature of scientific evidence: Empirical, philosophical and statistical considerations* (University of Chicago press, 2004). He has served on three U.S. National Academy of Sciences committees on climate change, public health and other issues.

Dr. Mark L. Taper (Department of Ecology, Montana State University, USA) is a statistical and quantitative ecologist who uses analytic and computational modeling to answer questions in conservation biology, spatial ecology, population dynamics, macro ecology, and evolutionary ecology. He also has a deep interest in the epistemological foundations of both statistics and science. Dr. Taper has chaired the Ecological Society of America's statistics section and is the founding Director of the Montana State University interdisciplinary program in Ecological and Environmental Statistics. Dr. Taper has published over 90 scientific, statistical, and philosophical articles and co-edited with Subhash Lele a volume titled *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations* published by The University of Chicago Press in 2004. He has served on the editorial boards of *Frontiers in Ecology and the Environment* and of *Ecological and Environmental Statistics*.

Cross References

►Bootstrap Methods
►Frequentist Hypothesis Testing: A Defense
►Kullback-Leibler Divergence
►Likelihood
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Model Selection
►Most Powerful Test
►Neyman-Pearson Lemma
►Presentation of Statistical Testimony
►P-Values
►Sample Size Determination
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Fallacies

►Statistical Inference: An Overview
►Statistics and the Law

References and Further Reading

Research Papers

Barnard GA (1949) Statistical Inference. J Roy Stat Soc B 11:115–149

Blume JD (2002) Likelihood methods for measuring statistical evidence. Stat Med 21:2563–2599

Lele SR (2004) Evidence functions and the optimality of the law of likelihood. In: Taper ML, Lele SR (eds) The nature of scientific evidence: statistical, philosophical and empirical considerations. University of Chicago Press, Chicago

Strug LJ, Hodge SE (2006a) An alternative foundation for the planning and valuation of linkage analysis I. Decoupling "error probabilities" from "measures of evidence". Hum Hered 61:166–188

Strug LJ, Hodge SE (2006b) An alternative foundation for the planning and evaluation of linkage analysis II. Implications for multiple test adjustments. Hum Hered 61:200–209

Strug LJ, Rohde C, Corey PN (2007) An introduction to evidential sample size. Am Stat 61(3):1–5

Taper ML, Lele SR (2010) Evidence, evidence functions and error probabilities. In: Forster MR, Bandyophadhyay PS (eds) Handbook for philosophy of statistics. Elsevier

Books

Edwards AWF (1992) Likelihood, Expanded edn. Johns Hopkins University Press, Baltimore

Forster MR, Bandopadhyay PS (2010 expected) Handbook for philosophy of statistics. Elsevier

Hacking I (1965) Logic of statistical inference. Cambridge University Press, Cambridge

Mayo DG (1996) Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago

Royall R (1997) Statistical Evidence: a likelihood paradigm. Chapman & Hall, London

Taper ML, Lele SR (eds) (2004) The nature of scientific evidence: statistical, philosophical and empirical considerations. University of Chicago Press, Chicago

Taper ML, Lele SR (2010) Evidence, evidence functions and error probabilities. In: Forster MR, Bandyophadhyay PS (eds) Handbook for philosophy of statistics. Elsevier

Statistical Fallacies

Watter Krämer
Professor and Chairman
Technische Universität Dortmund, Dortmund, Germany

The range of possible fallacies in statistics is as wide as the range of statistics itself (see Cohen 1938; Good 1962, 1978; Moran 1973 for convenient overviews); there is probably no application and no theory where one does not find examples of intentional or unintentional misuse of statistical facts and theories (which of course is not unique to statistics – there is probably no science or social science whatsoever which is immune to such abuse). When collecting data, there is the well known problem of biased or self-selected samples, or ill-phrased questionnaires where answers are already imbedded in the questions. A nice example is provided by two surveys on workers' attitude towards working on Saturdays which were conducted in Germany in the same months of the same year (Krämer 2008, p. 121). The first survey produced a rejection rate of 95% whereas in the second survey, 80% of workers who were asked were happy to work on Saturdays if only they could. After inspection of the questionnaires it was clear how these results came about: The first survey was sponsored by a trade union and started with reminding the audience of the hard work it had taken to push through the five day work week, ending with the question (I exaggerate slightly): Are you really prepared to sacrifice all of what your fellow workers have fought about so hard? The second survey started with a comment on fierce competition for German industry from Asia which in the end let to the final question of whether workers were prepared to work on Saturdays if otherwise their employer went bankrupt.

Such extreme examples are of course quite rare, but it is rather easy to lead people in any direction which is convenient from the researcher's point of view.

In the area of biased and self-selected samples, the best known example is of course the historical disaster of the Literary Digest magazine back in 1936. The magazine had asked well above ten million Americans, a record sample by any standards, whom they intended to vote for in the upcoming presidential election. According to this survey, the republican candidate was going to win handsomely whereas in reality Roosevelt, the incumbent, won by a landslide. The Digest's sample was drawn from lists of automobile and telephone owners (likely to vote republican) and among those asked, less than a quarter actually replied (presumably voters with an axe to grind with the incumbent; see Bryson 1976).

Other fallacies arise in the context of interpreting or presenting the results of statistical analyses. There is the obvious area of confusing correlation and causation or of misreading the meaning of statistical tests of significance, where even professional statisticians have a hard time to correctly interpret a positive test result at – say – a 5% level of significance (there are even textbooks which state that this means: "The null hypothesis is wrong with 95% probability"). Another problem here is that true significance levels are in many applications much higher then nominal ones due to the fact that only "significant" outcomes are reported.

Such problems with interpreting statistical tests are tightly connected with the misuse of conditional probabilities, which is probably the both most widespread and most dangerous way that one can misread statistical evidence (Krämer and Gigerenzer 2005). One of these is to infer, from a conditional probability $P(A|B)$ that is seen as "large," that the conditional event A is "favorable" to the conditioning event B, in the sense that $P(B|A) > P(B)$.

This confusion occurs in various contexts and is possibly the most frequent logical error that is made in the interpretation of statistical information. Here are some examples from the German press (with the headlines translated into English):

- "Beware of German tourists" (According to *Der Spiegel* magazine, most skiers involved in accidents in a Swiss skiing resort came from Germany).
- "Boys more at risk on bicycles" (the newspaper *Hannoversche Allgemeine Zeitung* reported that among children involved in bicycle accidents, the majority were boys).
- "Soccer most dangerous sport" (the weekly magazine *Stern* commenting on a survey of accidents in sports).
- "Private homes as danger spots" (the newspaper *Die Welt* musing about the fact that a third of all fatal accidents in Germany occur in private homes).
- "German shepherd most dangerous dog around" (The newspaper *Ruhr-Nachrichten* on a statistic according to which German shepherds account for a record 31% of all reported attacks by dogs).
- "Women more disoriented drivers" (The newspaper *Bild* commenting on the fact that among cars that were found entering a one-way street in the wrong direction, most were driven by women).

These examples can easily be extended. Most of them result from unintentionally misreading the statistical evidence. When there are cherished stereotypes to conserve, such as the German tourist bullying his fellow vacationers, or women somehow lost in space, perhaps some intentional neglect of logic may have played a role as well. Also, not all of the above statements are necessarily false. It might, for instance, well be true that when 1,000 men and 1,000 women drivers are given a chance to enter a one-way street the wrong way, more women than men will actually do so, but the survey by *Bild* simply counted wrongly entering cars and this is certainly no proof of their claim. For example, what if there were no men on the street at that time of the day? And in the case of the Swiss skiing resort, where almost all foreign tourists came from Germany, the attribution of abnormally dangerous behavior to this class of visitors is clearly wrong.

In terms of favorable events, *Der Spiegel*, on observing that P(German tourist | skiing accident) was "large," concluded that the reverse conditional probability was also large, in particular, that being a German tourist increases the chances of being involved in a skiing accident:

$$P(\text{skiing accident}|\text{German tourist}) > P(\text{skiing accident}).$$

Similarly, *Hannoversche Allgemeine Zeitung* concluded from P(boy | bicycle accident) = large that P(bicycle accident | boy) > P(bicycle accident) and so on. In all these examples, the point of departure was always a large value of $P(A|B)$, which then led to the – possibly unwarranted – conclusion that $P(B|A) > P(B)$. From the symmetry

$$P(B|A) > P(B) \Longleftrightarrow P(A|B) > P(A)$$

it is clear, however, that one cannot infer anything regarding A's favorableness for B from $P(A|B)$ alone, and that one needs information on $P(A)$ as well.

Another avenue through which the attribute of favorableness can be incorrectly attached to certain events is ▶Simpson's paradox, which in our context asserts that it is possible that B is favorable to A when C holds, B is also favorable to A when C does not hold, yet overall, B is unfavorable to A. Formally, one has

$$\begin{aligned} &P(A|B \cap C) > P(A) && \text{and} \\ &P(A|B \cap \overline{C}) > P(A) && \text{yet} \\ &P(A|B) < P(A). \end{aligned}$$

This paradox also extends to situations where $C_1 \cup \ldots \cup C_n = \Omega$, $C_i \cap C = \emptyset\,(i \neq j)$.

One instance where Simpson's paradox (to be precise: the refusal to take account of Simpson's paradox) has been deliberately used to mislead the public is the debate on the causes of cancer in Germany. The official and fiercely defended credo of the Green movement has it that the increase in cancer deaths from well below 20% of all deaths after the war to almost 30% today, is mostly due to industrial pollution and chemical waste of all sorts. However, as Table 1 shows, among women, the probability of dying from cancer has actually *decreased* for young and old alike! Similar results hold for men.

A final and more trivial example for faulty inferences from conditional probabilities concerns the inequality

$$P(A|B \cap D) > P(A|C \cap D).$$

Plainly, this does not imply

$$P(A|B) > P(A|C),$$

yet this conclusion is still sometimes drawn. A German newspaper once claimed that people get happier as they

Statistical Fallacies. Table 1 Probability of dying from cancer Number of women (among 100,000 in the respective age groups) who died from cancer in Germany

Age	1970	2001
0–4	7	3
5–9	6	2
10–14	4	2
15–19	6	2
20–24	8	4
25–29	12	6
30–34	21	13
35–39	45	25
40–44	84	51
45–49	144	98
50–54	214	161
55–59	305	240
60–64	415	321
65–69	601	468
70–74	850	656
75–79	1183	924
80–84	1644	1587

(Statistisches Jahrbuch für die Bundesrepublik Deutschland)

grow older. The paper's "proof" runs as follows: Among people who die at age 20–25, about 25% commit suicide. This percentage then decreases with advancing age; thus, for instance, among people who die over the age of 70, only 2% commit suicide. Formally, one can put these observations as

$$P(\text{suicide} \mid \text{age } 20-25 \text{ and death}) > P(\text{suicide} \mid \text{age} > 70 \text{ and death}),$$

and while this is true, it certainly does not imply

$$P(\text{suicide} \mid \text{age } 20-25) > P(\text{suicide} \mid \text{age} > 70).$$

In fact, a glance at any statistical almanac shows that quite the opposite is true.

Here is a more recent example from the US, where likewise $P(A|B)$ is confused with $P(A|B \cap D)$. This time the confusion is spread by renowned Harvard Law professor who advised the O. J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflects a motive to kill, advancing the premise that "a slap is a prelude to homicide." The defence – in the end successfully – argued that the probability of the event K that a husband killed his wife if he battered her was rather small, so battering showed not be viewed as evidence of murder.

$$P(K \mid \text{battered}) = 1/2,500.$$

The relevant probability, however, is not this one. It is that of a man murdering his partner given that he battered her *and* that she was murdered:

$$P(K \mid \text{battered and murdered}).$$

This probability is about 8/9 (Good 1996). It must not of course be confused with the probability that O. J. Simpson is guilty. But it shows that battering is a fairly good predictor of guilt for murder.

About the Author

Dr. Walter Krämer is a Professor and Chairman of Department of Statistics, University of Dortmund. He is Editor of *Statistical Papers* and *German Economic Review*. He is a member of ISI and NRW Akademie der Wissenschaften. Professor Krämer is author of more than 100 articles and 30 books. His book *So lügt man mit Statistik* (in German), modelled after Durrel Huffs classic "How to lie with Statics" has been translated into several languages.

Cross References

►Fraud in Statistics
►Misuse of Statistics
►Questionnaire
►Simpson's Paradox
►Statistical Evidence
►Statistical Fallacies: Misconceptions, and Myths
►Telephone Sampling: Frames and Selection Techniques

References and Further Reading

Bryson MC (1976) The literary digest poll: making of a statistical myth. Am Stat 30:184–185
Cohen JB (1938) The misuse of statistics. J Am Stat Soc 33:657–674
Good IJ (1962) A classification of fallacious arguments and interpretations. Technometrics 4:125–132
Good IJ (1978) Fallacies, statistical. In: Kruskal WH, Tanar JM (eds) International encyclopedia of statistics, vol 1. pp 337–349
Good IJ (1996) When batterer becomes murderer. Nature 381:481
Krämer W (2008) So lügt man mit Statistik, 11th paperback edn. Piper-Verlag, München

Krämer W, Gigerenzer G (2005) How to confuse with statistics: the use and misuse of conditional probabilities. Stat Sci 20:223–230
Moran PAP (1973) Problems and mistakes in statistical analysis. Commun Stat 2:245–257

Statistical Fallacies: Misconceptions, and Myths

Shlomo Sawilowsky
Professor
Wayne State University, Detroit, MI, USA

Compilations and illustrations of statistical fallacies, misconceptions, and myths abound (e.g., Brewer 1985; Huck 2008; Huff 1954; Hunter and May 1993; King 1986; Sawilowsky 1993, 2003a, b, c, d, 2005, 2007a, b; Vandenberg 2006). The statistical faux pas is appealing, intuitive, logical, and persuasive, but demonstrably false. They are uniformly presented based on authority and supported based on assertion. Unfortunately, these errors spontaneously regenerate every few years, propagating in peer reviewed journal articles; popular college textbooks; and most prominently, in the alternate (e.g., qualitative), nonprofessional (e.g., Wikipedia), and dissident literature. Some of the most egregious and grievous are noted below.

1. *Law of Large Numbers, Central Limit Theorem (CLT), population normality, and asymptotic theory*. This quartet is asserted to inform the statistical properties (i.e., Type I and II errors, comparative statistical power) of parametric tests for small samples (e.g., n≤50 or so). In fact, much of what was asserted regarding small samples based on these eighteenth to nineteenth century theorems was wrong. Most of what is correctly known about the properties of parametric statistics has been learned through Monte Carlo studies and related methods conducted in the last quarter of the twentieth century to the present.

   Examples of wrong statements include (a) random selection is mooted by drawing a sufficiently large sample, (b) the CLT guarantees $\overline{X}$ is normally distributed, (c) the CLT safeguards parametric tests as long as n ≥ 30, and (d) asymptotic relative efficiency is a meaningful predictor of small sample power. A corollary that is particularly destructive is journal editor and reviewer bias in favor of this quartet over Monte Carlo evidence, relegating the inelegance of the latter to be a function of "anyone who has a personal computer and knowledge of Algebra I."

   (e) Perhaps the most pervasive myth is that real variables are normally distributed. Micceri (1989) canvassed authors of psychology and education research over a number of years and determined that less than 3% of their data sets (even those where n > 5,000) could be considered even remotely bell-shaped (e.g., symmetric with light tails). Not a single data set was able to pass any known statistical test of normality. Similar studies have been conducted in other disciplines with the same result. Population normality is not the norm.

   (f) Journal editors and reviewers mistakenly attach more importance to lemmas, theorems, and corollaries from this quartet than on evidence from small samples Monte Carlo studies and related methods.
2. *Random assignment*. It is commonly asserted that the lack of random assignment can be rehabilitated via matching, ANCOVA, regression, econometric simultaneous modeling, latent-variable modeling, etc. In truth, "*there is no substitute for randomization*" (Sawilowsky 2007b, p 214.)
3. *Control group*. It is frequently asserted by journal editors and referees, and funding agency reviewers, that science and rigorous experimental design demand the use of a control, comparison, or second treatment group. Actually, there are many designs that do not require this, such as factorial ANOVA, times series, and single subject repeated measures layouts.
4. *Data transformations*. (a) One reason for transforming data is to better meet a parametric test's underlying assumptions. The inexplicable pressure to shoehorn a parametric test into a situation where doesn't fit has prompted textbook authors to recommend transforming data to better meet underlying assumptions. For example, if the data are skewed then the square root transformation is recommended. The debate on the utility of transforming for this purpose is known as the Games-Levine controversy that was waged in the early 1980s, primarily recorded in *Psychological Bulletin*.

   There is a misguided presumption that the statistician has a priori knowledge of when or how best to transform. Also, it is a fallacy to interpret results from a transformation in the original metric. What does it mean to conclude that the arcsin of children's weight in the intervention group was statistically significantly higher than the arcsin of children's weight in the comparison group? When was the last time a patient chal-

lenged the physician's recommended medication by demanding to know the logarithm of the expected reduction in weight as predicted from the clinical trial?

(b) Another reason for transforming the data is to convert a parametric procedure into a nonparametric procedure. The rank transformation is the prime example. Based on asymptotic theory published in very prestigious journals, and subsequent recommendations from high profile statistical software companies, data analysts were encouraged to routinely run their data through a ranking procedure, and follow with the standard parametric test on those ranks.

Careful data analysts have shown through Monte Carlo studies that good results may be obtained for the two independent samples, one-way independent ANOVA, and two independent samples multivariate layouts. The myth persists, however, that this procedure is a panacea. Those same careful data analyst have also shown the rank transformation does not work in the context of two dependent samples, factorial ANOVA, factorial ANCOVA, MANOVA, or MANCOVA layouts, yielding Type I error rates as high as 1, and greatly suppressed power (e.g., Sawilowsky 1985a; Sawilowsky et al. 1989; Blair et al. 1987). Yet, software vendors continue to promote this procedure.

(c) It is also a myth that secondary transformations resolve this problem. The original data are transformed into ranks, and the ranks are in turn transformed into expected normal scores, random normal scores, or some other type of score. However, careful data analysts have also shown that secondary transformations fare no better than the rank transformation in terms of displaying poor Type I error control and severely depressed power (Sawilowsky 1985b).

5. *p values*. (a) Significance testing, as opposed to hypothesis testing, is mistakenly asserted to be scientific. Whereas hypothesis testing is objective due to the a priori stated threshold of what constitutes a rare event, significance testing is not objective. With the advent of easily obtained (and even exact) p values through statistical software, significance testing permits citing the resulting p value and letting the reader decide a posteriori if it is significant. Unfortunately, post and ad hoc significance testing obviates objectivity in interpreting the results, which is a fatal violation of a cornerstone of science. (b) Obtained p values are asserted to be transitory. For example, a p value that is close to nominal alpha (e.g., $\alpha = 0.05$ and $p = 0.06$) is incorrectly claimed to be approaching statistical significance, when in fact the result of the experiment is quite stationary. (c) The magnitude of the p value is asserted to inform the magnitude of the treatment effect. A p value of 0.0001 is erroneously claimed to mean the effect is of great practical importance. Although that may be true, it is not because of any evidence based on the magnitude of p.
6. *Effect Size*. Statistical philosophers stipulate that the null hypothesis can never literally be true. By virtue of all phenomena existing in a closed universe, at some part of the mantissa the population values must diverge from zero. Thus, it is claimed that effect sizes should be reported even if a hypothesis test was not conducted, or even if the result of a hypothesis test is not statistically significant.

   This viewpoint is presaged on an imputed meta-analytic intent that will arise in the future even if there is no such intent at the time the experiment was conducted. This fallacy arises, as do many errors in interpretation of statistics, by ignoring the null hypothesis being tested. Under the truth of the null hypothesis observed results for the sample are not statistically significantly different from zero, and thus the magnitude of the observed result is meaningless. Hence, effect sizes are only meaningfully reported in conjunction with a statistically significant hypothesis test.
7. *Experiment-wise Type I error*. It is universally recommended that prudent statisticians should conduct preliminary tests of underlying assumptions (e.g., homoscedasticity, normality) prior to testing for effects. It is asserted that this does no harm to the experiment-wise Type I error rate. However, Monte Carlo evidence demonstrates that the experiment-wise Type I error rate will inflate if preliminary tests are conducted without statistical adjustment for multiple testing. Moreover, there will be a Type I inflation even if the decision to proceed is based on eye-balling the data.
8. *Confidence Intervals*. Confidence intervals have recently been promoted over the use of hypothesis tests for a litany of unsupported reasons. (a) Among its supposed benefits is the assertion that confidence intervals provide more confidence than do hypothesis tests. This is based on the fallacy that confidence intervals are based on some system of probability theory other than that of hypothesis tests, when in fact they are the same. (b) Another prevalent misconception is confidence intervals must be symmetric.
9. ►*Robust statistics*. Typically, proposed expansions of descriptive robust statistics into inferential procedures are substantiated via comparisons with para-

metric methods. It is rare to find direct comparisons of inferential robust statistics with nonparametric procedures. (a) It is asserted that robust descriptive statistics maintain their robustness when evolved into inferential counterparts. This is a fallacy, however, because robust descriptive statistics were derived under parametric models confronted with perturbations. Therefore, Monte Carlo studies show they exhibit inflated Type I errors in many layouts. (b) It is similarly asserted that robust inferential statistics are high in comparative statistical power, but they are generally less powerful than rank based nonparametric methods when testing hypotheses for which the latter are intended.

10. ►*Permutation tests*. Permutation analogs to parametric tests are correctly stated to have equal power, and indeed can rehabilitate parametric tests' poor Type I error properties. However, it is incorrectly asserted that they are more powerful than nonparametric methods when testing for shift in location, when in fact the power spectrum of permutation tests generally follows (albeit somewhat higher) the power spectrum of their parametric counterparts, which is considerably less powerful than nonparametric procedures.
11. *Exact statistics*. Exact statistics, recently prevalent due to the advent of statistical software, are often advertised by software venders as being the most powerful procedure available to the statistician for the analysis of small samples. Actually, the advantage of exact statistics is that the p values are correct, but as often as not a smaller p value will result from the use of tabled asymptotic p values.
12. *Parametric tests*. The t and F tests are asserted to be (a) completely robust to Type I errors with respect to departures from population normality, (b) generally robust with respect to departures from population homoscedasticity, and (c) at least somewhat robust with respect to departures from independence. All three of these assertions are patently false. (d) Parametric tests are incorrectly asserted to trump the need for random selection or assignment of data, particularly due to Sir Ronald Fisher's paradigm of analysis on the data at hand.

    (e) Parametric tests (e.g., t, F) are asserted to be more powerful than nonparametric tests (e.g., Wilcoxon Rank Sum (see ►Wilcoxon–Mann–Whitney Test), Wilcoxon Signed Ranks (see ►Wilcoxon-signed-rank test)) when testing for shift in location. In fact, for skewed distributions, the nonparametric tests are often three to four times more powerful than their parametric counterparts. (f) As sample size increases, these parametric tests are asserted to increase their power advantages over nonparametric tests. In fact, the opposite is true until the upper part of the power spectrum is reached (e.g., the ceiling is 1) when the parametric tests eventually converge with the nonparametric test's statistical power.
13. *Nonparametric rank tests*. The assertions denigrating the Wilcoxon tests are so pervasive (to the extent that the two independent samples case is more frequently attributed as the Mann Whitney U, even though Wilcoxon had priority by 2 years) that the reader is referred to Sawilowsky (2005) for a listing of 22 frequently cited fallacies, misconceptions, and myths. Among the highlights are the incorrect beliefs that (a) the uniformly most powerful unbiased moniker follows the usage of the parametric t test for data sampled from nonnormally distributed populations, (b) the Wilcoxon tests should only be used with small data sets, (c) the Wilcoxon tests should only be used with ordinal scaled data, and (d) the Wilcoxon tests' power properties are oblivious to ►outliers.
14. χ^2. (a) We live in a χ^2 society due to political correctness that dictates equality of outcome instead of equality of opportunity. The test of independence version of this statistic is accepted *sans voire dire* by many legal systems as the single most important arbiter of truth, justice, and salvation. It has been asserted that any statistical difference between (often even nonrandomly selected) samples of ethnicity, gender, or other demographic as compared with (often even inaccurate, incomplete, and outdated) census data is *primae faciea* evidence of institutional racism, sexism, or other ism. A plaintiff allegation that is supportable by a significant χ^2 is often accepted by the court (judges and juries) *praesumptio iuris et de iure*. Similarly, the goodness of fit version of this statistic is also placed on an unwarranted pedestal.

    In fact, χ^2 is super powered for any arbitrary large number of observations. For example, in the goodness of fit application where the number of observed data points is very large and the obtained χ^2 can be of an order of magnitude greater than three, there is the custom not to even bother with the divisor E_i, and instead to proclaim a good fit if the new empirical process results in a reduced obtained value of the numerator. The converse is true where the number of observed data points are small (e.g., $N < 20$ or 30), in which case the χ^2 test of independence is among the least powerful methods available in a statistician's repertoire.

15. *Stepwise regression*. Stepwise (or "unwise", Leamer 1985) regression and replicability are two mutually exclusive concepts. It is asserted to be an appropriate data mining technique (see ►Data Mining). However, it is analogous to talking a walk in the dark in the park, tripping over a duffle bag, inspecting the bag and finding data sheets crumpled together, transcribing and entering the data into a statistical software program, having the software command the CPU to regress all possible combinations of independent variables on the dependent variable until the probability to enter has been met, reporting the results, and eyeballing the results to construct an explanation or prediction about an as yet unstated research hypothesis. There is nothing scientifically rigorous about Stepwise regression, even when it is adorned with the appellation of nonmodel-based regression. It is tantamount to a search for Type I errors.
16. *ANOVA main and interaction effects*. (a) It is asserted that because certain transformations can be invoked to make interaction effects apparently vanish, main effects are real and interaction effects are illusionary. Actually, it is easily demonstrated through symbolic modeling that main effects in the presence of interactions are spurious.

    (b) It is a misguided tendency to interpret significant main effects first and significant interaction effects second. The correct interpreting and stopping rules (see Sawilowsky 2007a) are to begin with the highest order effect, and cease with the highest order statistically significant effect(s) on that level.

    For example, in a 2 × 2 × 2 ANOVA layout, meaningful interpretation begins with the a × b × c interaction. Analysis should cease if it is statistically significant. If it is not, then the focus of analysis descends to the a × b, a × c, and b × c lower order interactions. If none are statistically significant, it is then appropriate to give attention to the a, b, and c main effects. (c) It is true that MANOVA is useful even when there are only univariate hypotheses, because the sole reason for invoking it is to provide increased statistical power. Thus, it is meaningful to follow with univartiate tests to provide further insight after a statistically significant MANOVA result. However, it is a misconception that so-called step-down univariate tests are necessary, or meaningful, to interpret a statistically significant MANOVA that was conducted to examine a multivariate hypothesis, which by definition is multivariate because it consists of hopelessly intertwined dependent variables (see Sawilowsky 2007a).
17. *ANCOVA*. (a) This procedure is the Catch-22 of statistical methods. Because it is erroneously assumed to correct for baseline differences, and baseline differences are concomitant with the lack of ►randomization, the myth has arisen that using ANCOVA rehabilitates the lack of randomization. Unfortunately, to be a legitimate test ANCOVA requires randomization, only after which it serves to decrease the error term in the denominator of the F ratio, and hence increase statistical power.

    (b) ANCOVA, even when legitimately applicable due to randomization, is used to control for unwanted effects. The logic of partitioning and then removing sums of squares of an effect known to be significant is nearly meritless. It is by far more realistic to retain and model the unwanted effects by entering it (by some technique other than dummy coding) into a general linear model (i.e., regression) than it is to remove it from consideration.

    Consider a hypothetical treatment for the fresh water fish disease ichthyophthirius multifilis (ich). Suppose to determine its effectiveness the following veterinarian prescribed treatment protocol must be followed: (1) Remove the water while the fish remain in the aquarium. (2) Wait ten days until all moisture is guaranteed to have evaporated from the fish. (3) Apply Sawilowsky's miracle *ich-b-gone*™®© salve to the fish. (4) Wait an additional ten days for the salve to completely dry. (5) Refill the aquarium with water. Results of the experiment show no evidence of ich. Hence, the salve is marketable as a cure for ich, controlling for water.

    (c) There is a propensity, especially among doctoral dissertation proposals, and proposals submitted to funding agencies, to invoke as many covariates into ANCOVA as possible, under the mistaken impression that any covariate will reduce the error term and result in a more powerful test. In fact, a covariate must be carefully chosen. If it is not highly correlated with the dependent variable the trivial sum of squares that it may remove from the residual in the denominator will not overcome the impact of the loss of the *df*, resulting in a less powerful test. See Sawilowsky (2007b) for other myths regarding ANCOVA.
18. *Readership's view on publication differs from retraction and errata*. One of the most unfortunate, and sometimes insidious, characteristics of peer reviewed statistical outlets is the propensity to publish new and exciting statistical procedures that were derived via elegant squiggles, but were never subjected to Monte Carlo or other real data analysis methodologies to

determine their small samples Type I error and power properties. It appears that the more prestigious the outlet, the greater is the reluctance in publishing subsequent notices to the readership that the statistic or procedure fails, is severely limited, or has no practical value. If an editor imagines an article is so important to the readership that it is publishable, it is a misconception for editors to presume that the same readership would be uninterested in subsequently learning that the article was erroneous.

Some editors and reviewers, in an effort to protect the prestige of the outlet, create great barriers to correcting previously published erroneous work, such as demanding that the critical manuscript also solve the original problem in order to be worthy of publication (e.g., Hyman 1995). For example, this removes oversight if an ineffective or counter-productive cure for cancer was published by demanding the rebuttal author first cure cancer in order to demonstrate the published cure was vacuous.

19. *Mathematical and applied statistics/data analysis*. It is a myth that mathematical statistics and applied statistics/data analysis share a common mission and toolkit. The former is a branch of mathematics, whereas the latter are not. The consumer of real world statistics rejoices over an innovation that increases the ability to analyze data to draw a practical conclusion that will improve the quality of life, even if the memoir in which it was enshrined will never appear in the American Mathematical Society's *Mathematical Reviews* and its *MathSciNet* online database.
20. *Statisticians, authors of statistical textbooks, and statistics*. The following are myths: (a) Statisticians are subject matter experts in all disciplines. (b) Statisticians are mathematician wannabes. (c) Anyone who has a cookbook of statistical procedures is a qualified statistician. Corollary: Only the British need to certify statisticians. (d) Anyone who has taken an undergraduate course in statistics is qualified to teach statistics or serve as an expert witness in court. (e) Statistics textbooks are free from computational errors. (f) Statistics textbook authors are consistent in their use of symbols. (g) If three randomly selected statistics textbook authors opine the same view it must be true. Corollary: It is a myth that if a statistical topic is examined in three randomly selected statistics textbooks the explanations will be *i.i.d.* (h) t, F, regression, etc., aren't statistics – they are data analysis. (i) It is a myth that statistics can be used to perform miracles.

About the Author

Biography of Shlomo Sawilowsky is in ►Frequentist Hypothesis Testing: A Defense

Cross References

►Analysis of Covariance
►Asymptotic Relative Efficiency in Estimation
►Box–Cox Transformation
►Confidence Interval
►Data Analysis
►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Interaction
►Misuse of Statistics
►Monte Carlo Methods in Statistics
►Multivariate Analysis of Variance (MANOVA)
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Normal Scores
►Null-Hypothesis Significance Testing: Misconceptions
►Permutation Tests
►Power Analysis
►P-Values
►Randomization
►Rank Transformations
►Robust Statistics
►Scales of Measurement and Choice of Statistical Methods
►Statistical Fallacies
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Blair RC, Sawilowsky SS, Higgins JJ (1987) Limitations of the rank transform in factorial ANOVA. Communications in Statistics-Computations and Simulations B16:1133–1145

Brewer JK (1985) Behavioral statistics textbooks: Source of myths and misconceptions? J Educ Stat 10:252–268

Huck SW (2008) Statistical Misconceptions. Psychology Press, London

Huff D (1954) How to lie with statistics. Norton, New York

Hunter MA, May RB (1993) Some myths concerning parametric and nonparametric tests. Can Psychol 34(4):365–469

Hyman R (1995) How to critique a published article. Psychol Bull 118(2):178–182

King G (1986) How not to lie with statistics: avoiding common mistakes in quantitative political science. Am J Polit Sci 30(3):666–687

Leamer E (1985) Sensitivity analyses would help. Am Econ Rev 75:308–313

Micceri T (1989) The Unicorn, the normal curve, and other improbable creatures. Psychol Bull 105(1):156–166

Sawilowsky S (1985) Robust and power analysis of the 2 × 2 × 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida

Sawilowsky S (1985b) A comparison of random normal scores test under the F and Chi-square distributions to the 2x2x2 ANOVA test. Florida J Educ Res 27:83–97
Sawilowsky S (1990) Nonparametric tests of interaction in experimental design. Rev Educ Res 60(1):91–126
Sawilowsky SS (1993) Comments on using alternatives to normal theory statistics in social and behavioral sciences. Can Psychol 34(4):432–439
Sawilowsky S (2003a) A different future for social and behavioral science research. J Mod Appl Stat Meth 2(1):128–132
Sawilowsky SS (2003b) You think you've got trivials? J Mod Appl Stat Meth 2(1):218–225
Sawilowsky SS (2003c) Trivials: The birth, sale, and final production of meta-analysis. J Mod Appl Stat Meth 2(1):242–246
Sawilowsky S (2003d) Deconstructing arguments from the case against hypothesis testing. J Mod Appl Stat Meth 2(2):467–474
Sawilowsky S (2005) Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney U test for shift in location parameter. J Mod Appl Stat Meth 4(2):598–600
Sawilowsky S (2007a) ANOVA: effect sizes, simulation interaction vs. main effects, and a modified ANOVA table. In: Sawilowsky S (ed) Real data analysis, Ch. 14, Information Age Publishing, Charlotte, NC
Sawilowsky S (2007b) ANCOVA and quasi-experimental design: the legacy of campbell and stanley. In: Sawilowsky S (ed) Real data analysis, Ch. 15, Information Age Publishing, Charlotte, NC
Sawilowsky S, Blair RC, Higgins JJ (1989) An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. J Educ Stat 14:255–267
Thompson B (1995) Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educ Psychol Meas 55(4):525–534
Vandenberg RJ (2006) Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea? Organ Res Meth 9:194–201

Statistical Genetics

Susan R. Wilson
Professor, Faculty of Medicine and Faculty of Science
University of New South Wales, Sydney, NSW, Australia

Statistical genetics broadly refers to the development and application of statistical methods to problems arising in genetics. Genetic data analysis covers a broad range of topics, from the search for the genetic background affecting manifestation of human diseases to understanding genetic traits of economic importance in domestic plants and animals. The nature of genetic data has been evolving rapidly, particularly in the past decade, due mainly to ongoing advancements in technology.

The work over a century ago of Gregor Mendel, using inbred pea lines that differed in easily scored characteristics, marks the start of collecting and analysing genetic data. Today we can easily, and relatively inexpensively, obtain many thousands, even millions or more, of genetic and phenotypic, as well as environmental, observations on each individual. Such data include high-throughput gene expression data, single nucleotide polymorphism (SNP) data and high-throughput functional genomic data, such as those that examine genome copy number variations, chromatin structure, methylation status and transcription factor binding. The data are being generated using technologies like microarrays, and very recently, next-generation sequencing. In the next few years, it is anticipated that it will be possible to sequence an entire human genome for $100, in a matter of days or even hours. The sheer size and wealth of these new data are posing many, ongoing, challenges.

Traditionally there have been close links between developments in genetics and in statistics. For example Sir RA Fisher's proposal of ▶analysis of variance (ANOVA) can be traced back to the genetic problems in which he was interested. It is not widely known that probabilistic graphical models have their origins at about the same time in S Wright's genetic path analysis. A current thrust of modern statistical science concerns research into methods for dealing with data in very high dimensional space, such as is being generated today in molecular biology laboratories. New opportunities abound for analysing extremely complex biological data structures.

Basic analyses of genetic data include estimation of allele and haplotype frequencies, determining if Hardy-Weinberg equilibrium holds, and evaluating linkage disequilibrium. Statistical analyses of sequence, structure and expression data cover a range of different types of data and questions, from mapping, to finding sequence homologies and gene prediction, and to finding protein structure. Although many tools appear ad hoc, often it is found that there are some solid, statistical underpinnings. For example, the very widely used heuristic computational biology tool, Basic Local Alignment Sequence Tool (BLAST) is based on random walk theory (see ▶Random Walk).

In animal and plant breeding, there are a range of approaches to finding and mapping quantitative trait loci, in both inbred lines and outbred pedigrees. Population genetics is a large topic in its own right, and is concerned with the analysis of factors affecting the genetic composition of a population. Hence it is centrally concerned with evolutionary questions, namely the change in the genetic composition of a population over time due to

natural selection, mutation, migration, and other factors. The knowledge of the structure of genes as DNA sequences has completely changed population genetics, including retrospective theory, in which a sample of genes is taken, DNA sequence determined, and the questions relate to the way in which, through evolution, the population has arrived at its presently observed state. For intrapopulation genetic inferences, coalescent theory (whereby from a sample of genes one traces ancestry back to the common ancestor) is fundamental. Evolutionary genetics is another, huge, topic. Many approaches have been developed for phylogenetic analyses, from applying likelihood methods, to use of parsimony and distance methods. In forensics, the use of DNA profiles for human identification often requires statistical genetic calculations. The probabilities for a matching DNA profile can be evaluated under alternative hypotheses about the contributor(s) to the profile, and presented as likelihood ratios. Conditional probabilities are needed, namely the probabilities of the profiles given that they have already been seen, and these depend on the relationships between known and unknown people.

Genetic epidemiology is a growing area, especially with current research to find the genes underpinning complex genetic diseases. "Methodological research in genetic epidemiology (is developing) at an ever-accelerating pace, and such work currently comprises one of the most active areas of methodological research in both ►biostatistics and epidemiology. Through an understanding of the underlying genetic architecture of common, complex diseases modern medicine has the potential to revolutionize approaches to treatment and prevention of disease" (Elston et al. 2002). Pharmacogenetics research is concerned with the identification and characterization of genes that influence individual responses to drug treatments and other exogenous stimuli. Modern pharmacogenetics involves the evaluation of associations between genetic polymorphisms and outcomes in large-scale clinical trials typically undertaken to evaluate the efficacy of a particular drug in the population at large. Meta-analysis methods (see ►Meta-Analysis) are an increasingly important tool for modern genetic analysis.

A starting point for the whole area of statistical genetics is the "Handbook" (Balding et al. 2004) that is also available online. Interestingly, the final chapter addresses ethics in the use of statistics in genetics. An encyclopaedic approach is used in the reference text of Elston et al. (2002). Software also is proliferating, and a good starting point is the suite of R packages in the Comprehensive R Archive Network (CRAN) Task View: Statistical Genetics (http://cran.r-project.org/web/views/Genetics.html) and in Bioconductor (http://www.bioconductor.org), an open source and open development software project for the analysis of genomic data.

About the Author

For biography *see* the entry ►Biostatistics.

Cross References

►Analysis of Variance
►Bioinformatics
►Biostatistics
►Forensic DNA: Statistics in
►Medical Statistics

References and Further Reading

Balding DJ, Bishop M, Cannings C (2004) Handbook of statistical genetics, 2nd edn. Wiley

Elston RC, Olson JM, Palmer L (2002) Biostatistical genetics and genetic epidemiology. Wiley, New York

Weir BS (1996) Genetic data analysis II. Sinaur Assocs, Sunderland, MA

Statistical Inference

Richard A. Johnson
Professor Emeritus
University of Wisconsin, Madison, WI, USA

At the heart of statistics lie the ideas of statistical inference. Methods of statistical inference enable the investigator to argue from the particular observations in a sample to the general case. In contrast to logical deductions from the general case to the specific case, a statistical inference can sometimes be incorrect. Nevertheless, one of the great intellectual advances of the twentieth century is the realization that strong scientific evidence can be developed on the basis of many, highly variable, observations.

The subject of statistical inference extends well beyond statistics' historical purposes of describing and displaying data. It deals with collecting informative data, interpreting these data, and drawing conclusions. Statistical inference includes all processes of acquiring knowledge that involve fact finding through the collection and examination of data. These processes are as diverse as opinion polls, agricultural field trials, clinical trials of new medicines, and the studying of properties of exotic new materials. As a consequence, statistical inference has permeated all fields of human endeavor in which the evaluation of information must be grounded in data-based evidence.

A few characteristics are common to all studies involving fact finding through the collection and interpretation of data. First, in order to acquire new knowledge, relevant data must be collected. Second, some variability is unavoidable even when observations are made under the same or very similar conditions. The third, which sets the stage for statistical inference, is that access to a complete set of data is either not feasible from a practical standpoint or is physically impossible to obtain.

To more fully describe statistical inference, it is necessary to introduce several key terminologies and concepts. The first step in making a statistical inference is to model the population(s) by a *probability distribution* which has a numerical feature of interest called a *parameter*. The problem of statistical inference arises once we want to make generalizations about the *population* when only a *sample* is available.

A *statistic*, based on a sample, must serve as the source of information about a parameter. Three salient points guide the development of procedures for statistical inference

1. Because a sample is only part of the population, the numerical value of the statistic will not be the exact value of the parameter.
2. The observed value of the statistic depends on the particular sample selected.
3. Some variability in the values of a statistic, over different samples, is unavoidable.

The two main classes of inference problems are *estimation* of parameter(s) and *testing hypotheses* about the value of the parameter(s). The first class consists of point estimators, a single number estimate of the value of the parameter, and interval estimates. Typically, the interval estimate specifies an interval of plausible values for the parameter but the subclass also includes prediction intervals for future observations. A test of hypotheses provides a yes/no answer as to whether the parameter lies in a specified region of values.

Because statistical inferences are based on a sample, they will sometimes be in error. Because the actual value of the parameter is unknown, a test of hypotheses may yield the wrong yes/no answer and the interval of plausible values may not contain the true value of the parameter.

Statistical inferences, or generalizations from the sample to the population, are founded on an understanding of the manner in which variation in the population is transmitted, via sampling, to variation in a statistic. Most introductory texts (see Johnson and Bhattacharyya 2010; Johnson, Freund, and Miller 2011) give expanded discussions of these topics.

There are two primary approaches, *frequentist* and *Bayesian*, for making statistical inferences. Both are based on the *likelihood* but their frameworks are entirely different.

The frequentist treats parameters as fixed but unknown quantities in the distribution which governs variation in the sample. Then, the frequentist tries to protect against errors in inference by controlling the probabilities of errors. The long-run relative frequency interpretation of probability then guarantees that if the experiment is repeated many times only a small proportion of times will produce incorrect inferences. Most importantly, using this approach in many different problems keeps the overall proportion of errors small.

Frequentists are divided on the problem of testing hypotheses. Some statisticians (Cox 2006) follow R. A. Fisher and perform *significance tests* where the decision to reject a *null hypothesis* is based on values of the statistic that are extreme in directions considered important by subject matter interest. It is more common to take a *Neyman–Pearson* approach where an *alternative hypothesis* is clearly specified together with the corresponding distributions for the statistic. *Power*, the probability of rejecting the null hypothesis when it is false, can then be optimized. A definitive account of Neyman–Pearson theory is given in Lehmann and Casella (2003) and Lehmann and Romano (2008).

In contrast, Bayesians consider unknown parameters to be random variables and, prior to sampling, assign a *prior distribution* for the parameters. After the data are obtained, the Bayesian takes the product prior times likelihood and obtains the *posterior distribution* of the parameter after a suitable normalization. Depending on the goal of the investigation, a pertinent feature or features of the posterior distribution are used to make inferences. The mean is often a suitable point estimator and a suitable region of highest posterior density gives an interval of plausible values. See Box and Tiao (1973) and Gelman et al. (2004) for discussions of Bayesian approaches.

A second phase of statistical inference, *model checking*, is required for both frequentist and Bayesian approaches. Are the data consonant with the model or must the model be modified in some way? Checks on the model are often subjective and rely on graphical diagnostics.

D. R. Cox, gives an excellent introduction to statistical inference in Cox (2006) where he compares Bayesian and frequentist approaches and highlights many of the important issues.

Statistical inferences have been extended to semiparametric and fully nonparametric models where functions are the infinite dimension parameters.

About the Author

Richard A. Johnson is Professor Emeritus at the University of Wisconsin following 42 years on the regular faculty of the Department of Statistics (1966–2008). He served as Chairman (1981–1984). Professor Johnson has coauthored six books including (a) *Applied Multivariate Statistical Analysis* (1982, 6th edition 2007), Prentice-Hall with D. W. Wichern, (b) *Probability and Statistics for Engineers* (1990, 8th edition 2011), Prentice-Hall, with I. Miller and J. E. Freund, and (c) *Statistics-Principles and Methods* (1985, 6th edition 2010), J. Wiley and Sons, with G. K. Bhattacharyya. Richard is a recipient of the *Technometrics* Frank Wilcoxon Prize (1991), the Institute of Mathematical Statistics Carver Award (2008), and the American Statistical Association, San Antonio Chapter, Don Owen Award (2009). He is founding editor of *Statistics and Probability Letters* and served as editor for the first 25 years (1992–2007). Professor Johnson is a Fellow of the American Statistical Association, Fellow of the Institute of Mathematical Statistics, Fellow of the Royal Statistical Society and an Elected member of the International Statistical Institute. He has published 125 research papers and has lectured in more than 22 countries. His research interests include multivariate analysis, reliability and life testing, and large sample theory.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Confidence Interval
►Estimation
►Estimation: An Overview
►Likelihood
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Robust Inference
►Significance Testing: An Overview
►Statistical Inference: An Overview

References and Further Reading

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley
Cox DR (2006) Principles of statistical inference, Cambridge University Press, Cambridge
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC Press, Boca Raton, FL
Johnson R, Freund J, Miller I (2011) Miller and Freund's probability and statistics for engineers, 8th edn. Prentice-Hall, Upper Saddle River
Johnson R, Bhattacharyya GK (2010) Statistics – principles and methods, 6th edn. Wiley, Hoboken, NJ
Lehmann EL, Casella GC (2003) Theory of point estimation, 2nd edn. Springer, New York
Lehmann EL, Romano JP (2008) Testing of statistical hypotheses, 3rd edn. Springer, New York

Statistical Inference for Quantum Systems

Alexander S. Helevo
Professor
Steklov Mathematical Institute, Moscow, Russia

With the advent of lasers and optical communication it was realized that specific restrictions on the fidelity of information transmission due to quantum-mechanical nature of a communication channel need be taken into account and require a special approach. In the 1960–1970s this led to creation of a consistent quantum statistical decision theory which gave the framework for investigation of fundamental limits for detection and estimation of the states of quantum systems (Helstrom; Holevo 1976; 1982). In this theory statistical uncertainty is described by using mathematical apparatus of quantum mechanics – operator theory in a Hilbert space. Thus, the quantum statistical decision theory is a "noncommutative" counterpart of the classical one which was based on the Kolmogorov probability model and both of them can be embedded into a general framework (Holevo 1976). The interest to quantum statistical inference got the new impetus at the turn of the century (Barndorff-Nielsen et al. 2003). In high precision and quantum optics experiments researchers became able to operate with elementary quantum systems such as single ions, atoms and photons leading to potentially important applications such as quantum cryptography and novel communication protocols. In currently discussed proposals for quantum computing, the information is written into states of elementary quantum cells – qubits, and is read off via quantum measurements. Therefore the issue of extracting the maximum statistical information from the state of a given quantum system becomes important. On the other hand, building a consistent statistical theory of quantum measurement has significant impact onto foundations of quantum mechanics resulting in clarification of several subtle points. Last but not the least, quantum statistical inference has a number of appealing specifically

noncommutative features which open new perspectives for avantgarde research in the mathematical statistics.

As in the classical statistical decision theory, there is a set Θ of values of an unknown parameter θ, a set $\mathcal{X}$ of decisions x and a loss function $L_\theta(x)$, defining the quality of the decision x for a given value of parameter θ. The difference comes with the description of statistical uncertainty: here to each θ corresponds a density operator ρ_θ in the separable Hilbert space $\mathcal{H}$ of the system. *Density operator* ρ is a positive operator in $\mathcal{H}$ with unit trace, describing *state* of the quantum system. In physical problems the quantum system is the information carrier such as coherent electromagnetic field, prepared by transmitter in a state which depends on the signal θ.

A *decision rule* is defined by a quantum *measurement* with outcomes $x \in \mathcal{X}$. In the case of finite set $\mathcal{X}$ corresponding to hypotheses testing (detection), decision rule is described mathematically by a *resolution of the identity* in $\mathcal{H}$, i.e., the family of operators $M = \{M_x; x \in \mathcal{X}\}$ satisfying

$$M_x \geq 0, \quad \sum_{x\in\mathcal{X}} M_x = I, \tag{1}$$

where I is the identity operator. The probability of making decision x in the state ρ_θ is defined by the basic formula generalizing the Born-von Neumann statistical postulate

$$P_M(x|\theta) = \mathrm{Tr}\rho_\theta M_x.$$

Decision rule is implemented by a receiver making a quantum measurement and the problem is to find the optimal measurement performance.

The mean risk corresponding to the decision rule M is given by the usual formula

$$R_\theta\{M\} = \sum_{x\in\mathcal{X}} L_\theta(x) P_M(x|\theta). \tag{2}$$

In this way one has a family $\{R_\theta\{M\}, \theta \in \Theta\}$ of affine functionals defined on the convex set $\mathfrak{M}(\mathcal{X})$ of decision rules (1). The notions of admissible, minimax, Bayes decision rule are then defined as in the classical Wald's theory. The profound difference lies in the much more complicated convex structure of the sets of quantum states and decision rules.

The *Bayes risk* corresponding to a priori distribution π on Θ is

$$R_\pi\{M\} = \int_{\theta\in\Theta} R_\theta\{M\} d\pi(\theta) = \mathrm{Tr} \sum_{x\in\mathcal{X}} \hat{L}(x) M_x, \tag{3}$$

where

$$\hat{L}(x) = \int_{\theta\in\Theta} \rho_\theta L_\theta(x) d\pi(\theta) \tag{4}$$

is the operator-valued posterior loss function. Bayes decision rule minimizing $R_\pi\{M\}$ always exists and can be found among extreme points of the convex set $\mathfrak{M}(\mathcal{X})$. An illustration of the effect of noncommutativity is the following analog of the classical rule saying that Bayes procedure minimizes posterior loss: M is Bayes if and only if there exists Hermitian trace-class operator Λ such that

$$\Lambda \leq \hat{L}(x), \quad (\hat{L}(x) - \Lambda) M_x = 0, \quad x \in \mathcal{X}. \tag{5}$$

The operator Λ plays here the role of the minimized posterior loss.

The Bayes problem can be solved explicitly in a number of important cases, notably in the case of two hypotheses and for the families of states with certain symmetry. In general, symmetry and invariance play in quantum statistical inference much greater role; on the other hand, the concept of sufficiency has less applicability because of the severe restrictions onto existence of conditional expectations in the noncommutative probability theory (Petz 2008).

The optimum is found among the extreme points of the convex set of decision rules which therefore play a central role. In the classical case the extreme points are precisely deterministic decision rules. Their quantum analog are *orthogonal resolutions of the identity* satisfying $M_x M_y = \delta_{xy} M_x$ in addition to (1). However in the noncommutative case these form only a subset of all extreme decision rules. According to a classical result of Naimark, any resolution of the identity can be extended to an orthogonal one in a larger Hilbert space. In statistical terms, such an extension amounts to an outer quantum randomization. Consequently, there are quantum Bayes problems in which the optimal rule is inherently "randomized" (Holevo 1982). This paradoxical fact has a profound physical background, namely, the measurement *entanglement* between the system and the outer randomizer, which is a kind of intrinsically quantum correlation due to tensor product structure of the composite systems in quantum theory. Notably, in standard approach to quantum mechanics only orthogonal resolutions of the identity (namely, spectral measures of self-adjoint operators) were considered as representing *observables* (i.e., random variables). Thus, quantum statistical decision theory gives a strong argument in favor of the substantial generalization of the fundamental notion of quantum observable.

As in the classics, the case of two simple hypotheses ρ_0, ρ_1 is the most tractable one: there are quantum counterparts of the Neumann-Pearson criterion and of the asymptotics for the error probability and for the Bayes risk (the quantum Chernoff bound). However the derivation of these asymptotics is much more involved due to possible noncommutativity of the density operators ρ_0, ρ_1 (Hayashi 2006).

In estimation problems Θ and $\mathcal{X}$ are parametric varieties (typically $\mathcal{X} = \Theta \subset \mathbb{R}^s$) and the decision rules are given by *positive operator-valued measures* on Θ which are (generalized) spectral measures for operators representing the estimates. Solution of the Bayes estimation problem can be obtained by generalizing results for finite $\mathcal{X}$ with appropriate integration technique (Holevo 1976). Explicit solutions are obtained for problems with symmetry and for estimation of the mean value of Bosonic Gaussian states. The last is quantum analog of the classical "signal+noise" problem, however with the noise having quantum-mechanical origin and satisfying the canonical commutation relations (Holevo 1982).

Quantum statistical treatment of models with the shift or rotation parameter provides a consistent approach to the issue of canonical conjugacy and nonstandard uncertainty relations in quantum mechanics, such as time-energy, phase-number of quanta, as well as to approximate joint measurability of incompatible observables. In the quantum case estimation problems with multidimensional parameter are inherently more complex than those with one-dimensional parameter. This is due to the possible non-commutativity of the components reflecting existence of *incompatible* quantities that in principle cannot be measured exactly in one experiment. This sets new statistical limitations to the components of multidimensional estimates, absent in the classical case, and results in essential non-uniqueness of logarithmic derivatives and of the corresponding quantum Cramér–Rao inequalities (Helstrom 1976; Holevo 1982).

Another special feature of quantum statistical inference appears when considering series of i.i.d. quantum systems: the statistical information in quantum models with independent observations can be strictly superadditive. This means that the value of a measure of statistical information for a quantum system consisting of independent components can be strictly greater than the sum of its values for the individual systems. The property of strict superadditivity is again due to the existence of entangled (collective) measurements over the composite system (Hayashi 2005).

One of the most important quantum estimation models is the *full model*, in which the state is assumed completely unknown. In the case of finite dimensionality d this is a parametric model with a specific group of symmetries (the unitary group), in particular, for $d = 2$ it is the model of unknown qubit state (i.e., 2×2-density matrix), with the three-dimensional Stokes parameter varying inside the Bloch sphere. The most advanced results here concern the asymptotic estimation theory for the i.i.d. observations, culminating in the noncommutative analog of Le Cam's local asymptotic normality for estimation of an arbitrary mixed state of a finite dimensional quantum system (Guta and Kahn 2009; Hayashi 2005). The full model in infinite dimensions belongs to nonparametric quantum mathematical statistics, which is at present in a stage of development. In this connection the method of *homodyne tomography* of a density operator widely used in quantum optics is particularly important (Artiles et al. 2005).

Quantum statistical decision theory provides powerful general methods for computing fundamental limits to accuracy of physical measurements, which serve as benchmarks for evaluating the quality of existing physical measurement procedures. It also gives the mathematical description of the optimal decision rule; however the quantum theory in principle provides no universal recipe for constructing a measuring device from the corresponding resolution of the identity and such kind of problems have to be treated separately in each concrete situation. Still, in several cases methods of quantum statistical inference give important hints towards the realization (based, e.g., on covariance with respect to the relevant symmetries) and can provide an applicable description of the required (sub)optimal measurement procedure (Artiles et al. 2005; Hayashi 2005; Helstrom 1976) .

Acknowledgment

Supported in part by RFBR grant 09-01-00424 and the program "Mathematical control theory" of Russian Academy of Sciences.

About the Author

Alexander S. Holevo (Kholevo) is Professor at the Steklov Mathematical Institute. He is also a Professor at the Moscow State University and Moscow Institute for Physics and Technology. A. S. Holevo has been awarded the Markov Prize of Russian Academy of Sciences (1997) for his work in the noncommutative probability, the International Quantum Communication Award (1996) and A. von Humboldt Research Award (1999) for the development of mathematical theory of quantum information systems. He is the author of five monographs and more than 160 research articles in classical and quantum probability, statistics and information theory and in the mathematical foundations of quantum mechanics. He is currently Co-editor-in-chief of the journal *Theory of Probability and Its Applications*.

Cross References

►Astrostatistics
►Bayesian Statistics
►Chernoff Bound

►Decision Theory: An Introduction
►Decision Theory: An Overview
►Loss Function
►Markov Chain Monte Carlo
►Random Matrix Theory
►Statistical Inference: An Overview
►Stochastic Processes

References and Further Reading

Artiles L, Gill RD, Guta M (2005) An invitation to quantum tomography. J Roy Stat Soc B 67:109–134
Barndorff-Nielsen OE, Gill RD, Jupp PE (2003) On quantum statistical inference. J Roy Stat Soc B 65:775–816
Guta M, Kahn J (2009) Local asymptotic normality for finite dimensional quantum systems. Commun Math Phys 289(2):597–652
Hayashi M (ed) (2005) Asymptotic theory of quantum statistical inference. Selected papers. World Scientific, New York
Hayashi M (2006) Quantum information: an introduction, Springer, New York
Helstrom CW (1976) Quantum detection and estimation theory. Academic, New York
Holevo AS (1976) Investigations in the general theory of statistical decisions. Proc Steklov Math Inst 124:1–140 (AMS Translation, 1978, Issue 3)
Holevo AS (1982) Probabilistic and statistical aspects of quantum theory. North-Holland, Amsterdam
Petz D (2008) Quantum information theory and quantum statistics. Springer, Berlin

Statistical Inference for Stochastic Processes

M. B. Rajarshi
President of the International Indian Statistician Association (Indian Chapter)
Professor
University of Pune, Pune, India

Statistical inference for ►stochastic processes deals with dependent observations made at time points in $\{0, 1, 2, \cdots\}$ or $[0, \infty)$. Thus, the time parameter can be either discrete or continuous in nature.

Markov Chains and Sequences

Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a time-homogeneous L-order Markov sequence with the state-space S. Let $p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots x_{t-L})$ be the conditional probability mass function(p.m.f.) or probability density function(p.d.f.) of X_t given $X_{t-1}, X_{t-2}, \cdots X_{t-L}$, θ being an unknown parameter in Θ, an open set in the K-dimensional Euclidean space. The (conditional) log-likelihood (given $(X_{(1)}, X_{(2)}, \cdots, X_{(L)})$) is given by $\ln(L_T(\theta)) = \sum_{t=L,T} \ln[p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots, x_{t-L})]$, $T > L$. We assume that the conditional p.m.f./p.d.f. satisfies the Cramer regularity conditions and that $\{X_t, t = 0, 1, 2, \ldots\}$ is a strictly stationary and ergodic sequence. The Fisher Information matrix is defined by

$$I(\theta) = ((\, -E[\partial^2 \ln(p_\theta(X_t|X_{t-1}, X_{t-2}, \cdots X_{t-L}))/\partial\theta_i \partial\theta_j]\,))$$

and is assumed to be positive definite (the expectation is with respect to the joint distribution of $(X_t, X_{t-1}, \cdots, X_{t-L})$ and is computed under the assumption of stationarity). Under these conditions, it can be shown that there exists a consistent solution $\widehat{\theta}$ of the likelihood equations, such that $\sqrt{T}(\widehat{\theta} - \theta) \to N_K(0, [I(\theta)]^{-1})$ in distribution (Billingsley 1961). We apply the ►martingale central limit theorem to the score function (i.e., the vector of $\partial \ln(L_T(\theta))/\partial\theta_i$, $i = 1, 2, \cdots, K$) (Billingsley 1961; Hall and Heyde 1980) and the Strong Law of Large numbers for various sample averages of stationary and ergodic sequences to prove this result. The large-sample distribution theory of Likelihood Ratio Tests (LRTs) and confidence sets follows in a manner similar to the case of independently and identically distributed (i.i.d.) observations.

Some of the assumptions made above can be relaxed, cf. Basawa and Prakasa Rao (1980), Chap. 7. The LRT can be used for selecting the order of a model by testing a model against the alternatives of a higher order model. However, the ►Akaike's Information Criterion (AIC) and Bayes criterion (BIC), respectively given by $AIC = -2 \ln L_T(\hat{\theta}) + K$ and $BIC = -2 \ln L_T(\hat{\theta}) + K \ln(T)$ are more appropriate for selection of a model and an order. The model with the least AIC/BIC is selected. When S is finite, the procedure based on BIC yields a consistent estimator of the true order, cf. Katz (1981). The AIC is an inconsistent procedure, cf. Davison (2003), Sect. 4.7. For finite Markov chains, Pearson's χ^2-statistic can be used in place of the LRT for various hypotheses of interest. In moderate samples, the chi-square approximation to Pearson's χ^2-statistic is better than the same to LRT.

First order Markov models offer a satisfactory fit to observations somewhat infrequently. Lindsey (2004, p. 113) discusses approaches based on ►logistic regression and log-linear models (contingency table analysis) for higher order finite ►Markov chains. A distinct advantage of such a modeling is that both time-dependent and time-independent covariates can be incorporated, see discussion of Generalized Auto-Regressive Moving Average (GARMA) models below. A limitation of such models is that the conditional probabilities depend upon the numerical values (coding) assigned to the states, which is not suitable for models for data without any numerical structure, such as linguistic classes.

Higher order Markov chains and sequences can be handicapped by a large number of parameters. An important Markov model of order L with a substantially small number of parameters is due to Raftery (1985) and it is given by

$$p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots x_{t-L}) = \sum_{l=1,L} \lambda_l q_{x_{t-l},x_t}, \quad \lambda_l \geq 0, \quad \sum_l \lambda_l = 1.$$

Here, $q_{x,y}$ is a transition probability matrix (t.p.m.) or a transition density. The model is known as Mixture Transition Density (MTD) model. For an M-state chain, the number of parameters of the MTD model is $M(M-1) + L - 1$, far less (particularly for $M > 2$) than $(M^L)(M-1)$, the number of parameters in the corresponding saturated Markov chain. In the MTD models, like the Auto-Regressive (AR) time series models, we need to add only a single parameter to the r-order model to get the $(r+1)$-order model. We may note that if the state-space is continuous or countably infinite, the transition density $q_{x,y}$ is a specified function of K unknown parameters.

Non-Markovian Models

Hidden Markov Model (HMM). HMM was introduced in speech recognition studies. It has a very wide range of applications. Let $\{Y_t, t = 0, 1, 2, \ldots\}$ be a first-order Markov chain with the state-space $S_y = \{1, 2, \cdots M\}$ and the one-step t.p.m. P. The Markov chain $\{Y_t, t = 0, 1, 2, \ldots\}$ is not observable. Let $\{X_t, t = 0, 1, 2, \ldots\}$ be an observable process taking values in S_x with M_1 elements such that $P[X_t = j|Y_t = i, Y_{t-1}, \cdots Y_0, X_{t-1} \cdots X_{t-2}, \cdots, X_0] = q_{ij}$, $i \in S_x$, $j \in S_y$. Thus, if $M_1 = M$, the number of parameters of a Hidden Markov chain is $2M(M-1)$ which is considerably smaller than a higher order Markov chain. For estimation of unobserved states $\{Y_t, t = 0, 1, 2, \ldots T\}$ and estimation of parameters, the Baum-Welsch algorithm is widely used, which is an early instance of the Expectation-Maximization (EM) algorithm.

For a discussion of Hidden Markov chains, we refer to MacDonald and Zucchini (1997) and Elliot et al. (1995). Cappe et al. (2005) give a thorough and more recent account of a general state-space HMM.

ARMA Models for integer valued random variables. A non-negative Integer-valued ARMA (INARMA) sequence is defined as follows. The binomial operator $\gamma \circ W$ is defined by a binomial random variable with W as the number of trials and γ as the success probability (if $W = 0$, $\gamma \circ W = 0$). Let $\{Z_t, t = 0, \pm 1, \pm 2, \cdots\}$ be a sequence of i.i.d. non-negative integer valued random variables with a finite variance. Then, the $INARMA(p, q)$ process is defined by $X_t = \sum_{i=1,p} \alpha_i \circ X_{t-i} + \sum_{j=1,q} \beta_j \circ Z_{t-j} + Z_t$. All the binomial experiments required in the definition of the process are independent. The process $\{Z_t\}$ is not observable. The process $\{X_t\}$ is (second order) stationary if $\sum \alpha_i < 1$ and is invertible if $\sum \beta_j < 1$. An excellent review of such processes has been given in McKenzie (2003). Interesting special cases such as AR, MA and Poisson, Binomial, Negative Binomial as the stationary distributions are reported therein.

GARMA models. These are extensions of the ▶Generalized Linear Models based on an exponential family of distributions and can incorporate vector of time-dependent covariates z_t along with past observations. The conditional mean of X_t given the past is given by $h(\eta_t)$ where $h^{-1} = g$ (say) is the link function of the chosen exponential family and $\eta_t = z_t'\gamma + \sum_{i=1,p} \phi_i \left[g(x_{t-i}) - z_{t-i}'\gamma\right] + \sum_{j=1,q} \theta_j [g(x_{t-j}) - \eta_{t-j}]$. The parameters $\{\phi_i\}$ and $\{\theta_j\}$ denote the auto-regressive and moving average parameters respectively. The parameter γ explains the effect of covariates. A modification of the mean function is required to take care of the range of the observations. A limitation of this class of models is that in the absence of regressors or when the vector γ is null, it may not be possible to have a stationary series. We refer to Benjamin et al. (2003) and Fahrmeir and Tutz (2004), Chap. 6 for more details.

Bienayme-Galton-Watson Branching Process

Billingsley's work based on martingale methods for deriving asymptotic properties of the maximum likelihood estimator paved the way for many interesting theoretical developments for non-ergodic models such as a Bienayme-Galton-Watson (BGW) branching process.

Let $\{X_t, t = 0, 1, \cdots\}$ be a BGW Branching process with the state-space $S = \{0, 1, \cdots\}$ and the off-spring distribution $p_k, k = 0, 1, \cdots$. Parameters of interest are the offspring distribution and its functions such as the mean μ and the variance σ^2. A number of estimators for μ have been suggested: Lotka's estimator X_T/X_{T-1} (taken to be 1 if $X_{T-1} = 0$), Heyde's estimator $(X_T)^{1/T}$ and the nonparametric maximum likelihood estimator $\widehat{\mu}_T = (Y_T - X_0)/Y_{T-1}$ with $Y_t = X_0 + X_1 + \ldots + X_t$. The maximum likelihood estimator has a natural interpretation that it is the ratio of the total number of off-springs (in the realization) born to the total number of parents. By using the Scott central limit theorem for martingales (Scott 1978), it can be shown that, on the non-extinction path, $\sqrt{Y_{T-1}}(\hat{\mu}_T - \mu)/\sigma$ is asymptotically standard Normal. A natural estimator of σ^2, resulting from regression considerations, is given by $(1/T)\sum_t X_{t-1}(X_t/X_{t-1} - \hat{\mu}_T)^2$. This can be shown to be consistent and asymptotic normal with $\sqrt{T}$-norming, if

the fourth moment of the offspring distribution is finite. These results are useful to construct tests and confidence intervals for μ.

Based on a single realization, only μ and σ^2 are estimable on the non-extinction path of the process (i.e., consistent estimators exist for these parameters), if no parametric form of the offspring distribution is assumed. A good account of inference for branching processes, their extensions and related population processes along with applications can be found in Guttorp (1991).

Non-parametric Modeling Based on Functional Estimation

For a stationary process, where every finite dimensional distribution is absolutely continuous, we may opt for a non-parametric approach. We estimate the conditional density of X_t given $X_{t-1}, X_{t-2}, \ldots, X_{t-L}$ by the ratio of estimators of appropriate joint densities. The joint density of p consecutive random variables is estimated by a kernel-based estimator as follows. Let $K_p(x)$ be a probability density function, where $x \in R^p$, the p-dimensional Euclidean space. Let h_T be a sequence of positive constants such that $h_T \to 0$ and $Th_T^p \to \infty$ as $T \to \infty$. The estimator of joint density of consecutive p observations at $(x_1, x_2, \cdots, x_p)$ is then given by $\widehat{f}(x_1, x_2, \cdots, x_p) = \left[1/\left(Th_T^p\right)\right] \sum_{j=1,T-p} K((x_1 - X_j)/h_T, (x_2 - X_{j+1})/h_T, \cdots, (x_p - X_{j+p})/h_T)$. Based on the estimator of the conditional p.d.f., one can estimate the conditional mean (or other parameters such as conditional median or mode).

Properties of conditional density estimators are established assuming that the random sequence $\{X_t, t = 0, 1, 2, \cdots\}$ satisfies certain mixing conditions. We discuss strong or α-mixing, since most of the other forms of mixing imply the strong mixing. Let $F_{0,s}$ be the σ-field generated by the random variables $(X_0, X_1, \cdots, X_s)$ and let $F_{s+t,\infty}$ be the σ-field generated by the collection of random variables $\{X_{s+t}, X_{s+t+1}, \cdots\}$. The stationary sequence $\{X_t, t = 0, 1, 2, \cdots\}$ is said to be strong mixing if $sup_{A \in F_{0,s}, B \in F_{s+t,\infty}} \{|P(A \cap B) - P(A)P(B)|\} \leq \alpha(t)$ and $\alpha(t) \to 0$ as $t \to \infty$. For most of the results, we need faster rates of decay of $\alpha(t)$. Asymptotic properties of the kernel-based estimator have been established in Robinson (1983) who also illustrates how plots of conditional means can be helpful in bringing out nonlinear relationships. Prakasa Rao (1996) discusses, in detail, non-parametric analysis of time series based on functional estimation.

Non-parametric inference. Tests for median or tests and estimation procedures based on order or rank statistics, like the widely used tests in the case of i.i.d. observations can be suggested. However, the exact distribution is neither free from the unknown parameters, nor it is known, except in some special cases. Thus, such procedures for stationary observations lack simplicity and elegance of the rank-based tests. Further, robustness of an estimator is much more complex for dependent observations, since the effect of a spurious observation or an outlier (which can be an innovation outlier in an ARMA model) spreads over a number of succeeding observations. In an important paper, Martin and Yohai (1986) discuss influence functions of estimators obtained from ARMA Time Series model.

Bootstrap. Efron's Bootstrap (see ▶Bootstrap Methods) for i.i.d. samples is now routinely used to estimate the variance or the sampling distributions of estimators, test statistics and approximate pivotals. In most of the situations of practical interests, it gives a more accurate estimator of the sampling distribution than the one obtained by the traditional methods based on the Central Limit Theorem. In the i.i.d. case, we obtain B bootstrap samples, each sample being a Simple Random Sample With Replacement (SRSWR) of size T from the observed sample. This generates B values of a statistic or pivotal of interest.

For a stationary AR model of order L, the first L values of a bootstrap series may be the same as those of the observed time series. We take a SRSWR sample of size $T-L$ from residuals. The randomly selected residuals are then successively used to generate a bootstrap time series. We then have B time series, each of length T. For stationary and invertible MA or ARMA models, a bootstrap series is constructed from a SRSWR sample of the residuals. Rest of the methodology is the same as the usual bootstrap procedure. Bose (1988) (AR models) and (1990) (MA models) has shown that such a bootstrap approximation to the sampling distribution of the least square estimators is superior to the traditional normal approximation.

Bootstrap procedures for (strictly) stationary and ergodic sequence are based on blocks of consecutive observations. Bootstrap procedure is a boon for stochastic models, since in most of the cases, working out the variance of a statistic or its sampling distribution is very complex. By and large, it is beyond the reach of an end-user of statistics. (Consider, for example, computing the variance of a 10 per cent trimmed mean computed from stationary observations.) In a Moving Blocks Bootstrap (MBB)(Kunsch 1989; Liu and Singh 1992), we form K blocks of L consecutive observations to capture the dependence structure of the process. There are $N = T - L + 1$ blocks of L consecutive observations. We obtain a SRSWR of size K from these N *blocks* to get a bootstrap sample of size $T^* = KL$. If T is divisible by L, $K = T/L$, otherwise, it can be taken to be the

integer nearest to T/L. Let F_T be the empirical distribution function of T observations and let H be a functional on the space of distribution functions, computed at F_T (such as the trimmed mean or a percentile). A bootstrap statistic H^* is computed from the empirical distribution function of T^* bootstrap observations. Other procedures are NBB (Non-overlapping Blocks Bootstrap) and CBB (Circular Blocks Bootstrap), cf. Lahiri (2003), Chap. 2. Carlstein (1986) considers non-overlapping subseries of size L.

Let us assume that $L \to \infty$, $T \to \infty$ such that $T/L \to \infty$. Kunsch has shown that the bootstrap estimator of the variance of the normalized sample mean ($\sqrt{T}\overline{X}$) is consistent. (He further discusses jackknife procedures wherein we delete a block at a time.) The MBB procedure correctly estimates the sampling distribution of the sample mean. This property holds for a large number of mean-like statistics and smooth (continuously differentiable) functions of the mean vector, see Lahiri (2003 p. 177). Statistics based on averages of consecutive observations or their smooth functions (such as serial correlation coefficients) can be similarly bootstrapped. Second-order properties of the bootstrap estimator of the sampling distribution of the normalized/Studentized smooth functions of the sample mean (vector) have been obtained by Lahiri (1991) and Gotze and Kunsch (1996). Let $G(\mu)$, a third order differentiable function of the population mean vector μ, be the parameter of interest. While constructing the bootstrap version of the pivotal, we need to consider $G(\overline{X}^*) - G(\hat{\mu}_T)$, where $\hat{\mu}_T = E^*(\overline{X}^*)$. If the block length L is of the order $T^{1/4}$, the best possible error rate of the MBB approximation for estimation of the distribution function is $O(T^{-3/4})$. Though it is not as good as the accuracy that we have in the case of i.i.d. or residual based ARMA bootstrap, it is still better than the normal approximation to an asymptotic pivotal. Optimal block lengths for estimator of variance and the sampling distribution of a smooth statistics have been discussed in Chap. 7 of Lahiri (2003).

Under certain conditions, it is possible to bootstrap the empirical process, cf. Radulovic (2002). Such results as well as those discussed above for block based bootstrap, assume that the underlying process is strong mixing with a specified rate of decay of the mixing coefficients along with the block lengths L. We can construct confidence bands for the distribution function, by using the bootstrap distribution of the empirical process. Further, a number of statistics such as natural estimators of a compactly differentiable functional of the distribution function can be bootstrapped. Such a class of estimators include most of the estimators that we use in practice.

Kulperger and Prakasa Rao (1989) discuss bootstrap estimation of the sampling distribution of the estimator of a suitable function of P, the one-step t.p.m. of a finite ergodic irreducible Markov chain. They consider the expected value of time taken to reach a state from another state of a Markov chain, as a parametric function P. Computing the variance of such an estimator is very tedious. Bootstrap samples are generated by regarding the maximum likelihood estimate of the t.p.m. P as the underlying parameter.

State-space models (Doubly stochastic processes/Randomly driven stochastic processes). Let $\{X_t, t = 0, 1, \cdots\}$ be an unobservable process. Let $\{Y_t, t = 0, 1, \cdots\}$ be an observable process with the conditional p.m.f. or p.d.f. $f(y_0, y_1, \cdots, y_t \mid x_0, x_1, \cdots, x_t)$. In practice, often the process $\{X_t, t = 0, 1, \cdots\}$ is a Markov sequence and the conditional distribution of Y_t given $(y_0, y_1, \cdots, y_{t-1}, x_0, x_1, \cdots, x_t)$ depends upon x_t and y_{t-1} only. Such models are useful in situations where parameters vary slowly over time. It may be noted that models such as HMM, MTD or ARMA among others can be conveniently viewed as state-space models. Varying parameters can be modeled by a random process, see Guttorp (1995, p. 111) for an example involving a two state Markov chain.

Counting and Pure Jump Markov Processes

Let $\{X(t), t \geq 0\}$ be a counting process with $X(0) = 0$. Let $F(t_-)$ be the complete history up to t but not including t (technically the σ-field generated by the collection of random variables $\{X(u), u < t\}$). The intensity function $\lambda(t)$ can be stochastic (a random variable with respect to $F(t_-)$). It is characterized by the properties that $P[X(t + dt) - X(t) = 1 | F(t_-)] = \lambda(t)dt + o(dt)$, $P[X(t + dt) - X(t) = 0 | F(t_-)] = 1 - \lambda(t)dt + o(dt)$ and $P[X(t + dt) - X(t) \geq 1 | F(t_-)] = o(dt)$ for small dt. We assume that $E(X(t)) < \infty$ for every t. Let $M(t) = X(t) - E[X(t)|F(t_-)]$. It can be shown that $\{M(t), t > 0\}$ is a continuous time martingale with respect to $F(t_-)$, i.e., $E[M(t + s) - M(t) \mid F(t_-)] = 0$ for every $s > 0$. Time-dependent or time independent regressors can be included in the intensity function $\lambda(t)$.

Let the intensity $\lambda(t)$ be $\lambda(t, \theta)$, a specified function of the time and the parameters θ. In practice, to informally compute the likelihood, a partition $t_0 = 0, t_1, t_2, \ldots, t_N = T$ of $[0, T]$ is selected and the likelihood for such a partition is computed first. One then allows the norm of this partition to converge to 0. It turns out that the likelihood is given by $\ln(L(\theta) = \int \ln(\lambda(u, \theta))dX(u) - \int \lambda(u, \theta)I(u)du$, where $I(t) = 1$, if there is a jump at t and 0, otherwise. Such a general formulation linking counting processes inference with martingales in continuous time is due to Aalen (1978).

Important special cases include (a) Poisson process (see ▶Poisson process) with $\lambda(t) = \lambda$ for all t; (b) a Non-homogeneous Poisson Process where $\lambda(t)$ is a deterministic function, (c) Pure birth process $\lambda(t) = \lambda X(t_-)$, and (d) Renewal process (see ▶Renewal Processes) $\lambda(t) = h[t - t(x(t))]$ where $h(t)$ is the failure rate or hazard function of the absolutely continuous lifetime distribution of the underlying i.i.d. lifetimes and $t(x(t))$ is the time epoch at which the last failure before t takes place. (d) Semi-Markov or Markov renewal process. Here the intensity function depends on the state of the process at $t(x(t))$ and the state observed at t (assuming that there is an event at t).

Inference for counting processes and asymptotic properties of the maximum likelihood estimators have been discussed in Karr (1986) and Andersen et al. (1993).

Likelihood of a time-homogeneous continuous time *Pure Jump Markov process* follows similarly. Let, for $i \neq j$, $P[X(t+dt) = j|X(t) = i] = \lambda_{ij}dt + o(dt)$ and let $P[X(t+dt) = i|X(t) = i] = 1 - Q_{ii}dt + o(dt)$. The probability of other events is $o(dt)$. Here, $Q_{ii} = -\sum_{j\neq i}\lambda_{ij}$. If the state space is finite, each of the row-sums of the matrix $Q = ((Q_{ij}))$ is 0. The transition function $P[X(t) = j|X(0) = i]$ of the process is assumed to be differentiable in t for every i, j. The log-likelihood, conditional on $X(0) = x(0)$, is given by $\ln L = \sum_{i\neq j} N_{ij} \ln Q_{ij} - \sum_i Q_{ii}\tau_i$, where N_{ij} is the number of direct transitions from i to j and τ_i is the time spent in the state i, both during $[0, T]$. If the number of states is finite, the non-parametric maximum likelihood estimator of Q_{ij} is given by Nij/τ_i. Properties of maximum likelihood estimators have been discussed in Adke and Manjunath (1984) and Guttorp (1995, Chap. 3). Important cases include (Linear or Non-linear) Birth-Death-Immigration-Emigration processes and Markovian Queuing models.

Goodness of fit procedures are both graphical and formal. The Q-Q plot of the times spent in a state i scaled by the maximum likelihood estimates of their expected values, reveals departures from the exponential distribution. Since N_{ij}'s form transition counts of the embedded Markov chain, one can check whether such transitions have any memory. If the model under study has a stationary distribution, the observed frequencies of the test can be compared with the fitted stationary distribution, see Keiding (1975) who analyzes a Birth-Death-Immigration process model.

Diffusion Processes

Let $\{X(t), t \geq 0\}$ be a diffusion process with $\mu(x, \theta)$ and $\sigma^2(x)$ as the trend and diffusion functions respectively. The likelihood for the observed path $\{X(t), 0 \leq t \leq T\}$ is the Radon-Nikodym derivative of the probability measure of $\{X(t), 0 \leq t \leq T\}$ under the assumed diffusion process with respect to the probability measure of $\{X(t), 0 \leq t \leq T\}$ under the assumption of a diffusion process with the mean function equal to 0 for all x and the variance function $\sigma^2(x)$. It is assumed that $\sigma^2(x)$ is a known function. The log-likelihood is given by

$$\ln(L(\theta)) = \int_{0,T} \mu(x(t), \theta)/(\sigma(x(t))dx(t) - 1/2 \int_{0,T} \mu^2(x(t), \theta)/(\sigma(x(t))dt.$$

(If the variance functions is unknown, a time transformation is used to reduce the process with a known variance function.) Some special cases are (a) Brownian motion, (b) Geometric Brownian Motion, and (c) Ornstein-Uhlenbeck process. $\sqrt{T}-$ consistency and ▶asymptotic normality of the estimator of the mean of the process can be shown under the assumption that the process is non-null persistent (i.e., the process almost surely returns to any bounded set and the corresponding mean return time is finite). In this case, we can obtain non-parametric estimators of the common distribution function and the probability density function of $X(t)$. We refer to Prakasa Rao (1999a) and Kutoyants (2004) for details. Kutoyants (2004) also discusses asymptotic distributions of the estimator of the mean of the process in the null persistent case.

Observing a continuous time process may not be always feasible. We choose a partition of $[0, T]$, write the likelihood of such a partially observed process and then take the limit as the norm of the partition tends to 0. Validity of such operations has been established in Kutoyants (2004). Sorensen (2004) gives an extensive review for inference for stationary and ergodic diffusion processes observed at discrete points. The following techniques are discussed therein: (a) estimating functions with special emphasis on martingale estimating functions and so-called simple estimating functions, (b) analytical and numerical approximations of the likelihood function which can, in principle, be made arbitrarily accurate, (c) Bayesian analysis and MCMC methods, and (d) indirect inference and Generalized Method of Moments which both introduce auxiliary (but wrong) models and correct for the implied bias by simulation.

Statistical analysis and theoretical derivation of diffusion processes (as well as counting processes) is based on the theory of semimartingales. A semimartingale is a sum of a local martingale and a function of bounded variation. A class of diffusion processes and counting processes form

a subclass of the family of submartingales. A unified theory of statistical inference for semimartingales is presented in Prakasa Rao (1999b).

A fractional diffusion process is driven by the fractional Brownian motion (see ►Brownian Motion and Diffusions), which is not a semimartingale. Such processes can be useful in modeling phenomena with long range dependence, but the earlier techniques based on the theory of semimartingales are not applicable. Statistical inference for fractional diffusion processes has been discussed in Prakasa Rao (2010).

Concluding Remarks

Computational aspects. Computation of likelihood and its subsequent maximization are involved for most of the stochastic models. There are many procedures such as Kalman Filter, EM algorithm and Monte Carlo EM algorithm (which is based on Markov Chain Monte Carlo methods, see ►Markov Chain Monte Carlo), to compute the likelihood and the maximum likelihood estimator. From a computer programming view-point, implementation of the EM algorithm and its stochastic versions, require a special routine for each model. The conditional expectation step may require extensive simulations from a joint density, the constant of integration of which is not known. For state-space models, one needs to carry out a T-tuple integral (or a sum) to compute the likelihood. It seems that various methods based on numerical analysis to get a good approximation to the likelihood, its maximization and derivatives (which are needed to compute standard error of the maximum likelihood estimator), are preferred to other procedures. Possibly this is due to a very slow rate for convergence of the EM algorithm (and its stochastic versions) and yet another round of computations required to compute the estimator of the variance of the maximum likelihood estimator.

Efficiency of Estimators

(a) *Finite sample optimality.* Godambe's criterion (Godambe 1985) of a finite sample optimality of an estimator is based on optimality of the estimating equation it solves. Under the usual differentiability-based regularity conditions, an estimating function g^* is said to be optimal in G, if it minimizes $E(g(A)^2)/(E(\partial g(A)/\partial\theta))^2$. Let F_t be the $\sigma-$field generated by the collection of random variables $\{X_s, s = 0, 1, \cdots, t\}$. Let $g(t,\theta)$ be an F_t measurable random variable involving θ, a real parameter, such that $E[g(t,\theta) \mid F_{t-1}] = 0$ and $Var[g(t,\theta) \mid F_{t-1}] = V(t)$. Let $g(A) = \sum_t A(t)g(t,\theta)$, where $A(t)$ is an F_{t-1} measurable random variable, $t \geq 1$. Let $G = \{g(A)\}$ be the class of estimating functions $g(A)$ which satisfy the regularity conditions together with the assumptions that $E(g(A)^2) < \infty$ and $E(\partial g(A)/\partial\theta) \neq 0$. Godambe proves that the optimal choice of $A(t)$ is given by $E[\partial g(t,\theta)/\partial\theta \mid F_{t-1}]/V(t)$. In practice, we need to assume that such optimal weights do not involve other (incidental or nuisance) parameters.

A number of widely used estimators turn out to be solutions of such an optimal estimating equations $g^* = 0$. Further, Godambe's result justifies the estimator for each finite sample size and in addition, it broadens the class of parametric models to a larger class of semi-parametric models, for which the estimating function is optimal. The score function is optimal in a class of regular estimating functions, justifying use of the maximum likelihood estimator in finite samples. Continuous time analogues of these results with applications to counting processes have been discussed in a number of papers in a volume edited by Godambe (1991) and Prakasa Rao and Bhat (1996).

Optimality of an estimating function in a class is also equivalent to an optimal property of confidence intervals based on it. In large samples, the optimal g^* leads to a shortest confidence interval for θ at a given confidence coefficient. In a number of situations, the confidence interval, obtained from a Studentized estimating function, is typically better than the approximate pivotal obtained by Studentizing the corresponding estimator, in the sense that the true coverage rate of the procedure based on estimating function is closer to the nominal confidence coefficient. Bootstrapping the Studentized estimating function further improves performance of the corresponding confidence interval.

(b) *Asymptotic efficiency*. In non-ergodic models such as a BGW process, large-sample efficiency issues are rather complex. Though the random norming is convenient from an application view-point, the non-random norming is more appropriate and meaningful for efficiency issues. Further, notions of asymptotic efficiency based on variance of an estimator are no more applicable, since the variance of the asymptotic distribution for a large number of estimators does not exist. The W-efficiency of the maximum likelihood estimator, under certain regularity conditions, has been established by Hall and Heyde (1980) and Basawa and Scott (1983). Estimators based on other criteria can also be W-efficient. The Bayes estimator, under certain conditions, is asymptotically distributed like the maximum likelihood estimator. This result is known as the Bernstein-von Mises theorem and for its proof in the case of stochastic processes, we refer to Chap. 10 of Basawa and Prakasa Rao (1980).

Inference problems in stochastic processes have enriched both theoretical investigations and applied statistics.

Theoretical research in bootstrap, estimating functions, functional estimation and non-Gaussian non-Markov processes has widened scope of stochastic models. Use of fast and cheap computing has been helpful in computing likelihood, maximum likelihood estimators and Bayes estimators in very complicated stochastic models.

Acknowledgment

I am thankful to Professor B.L.S. Prakasa Rao and two other reviewers for a number of helpful suggestions.

About the Author

Dr. M. B. Rajarshi retired in 2009, as a Professor of statistics from the University of Pune. His areas of of interests are inference for stochastic processes, applied probability and stochastic modeling. He has published about 35 papers, some of which have appeared in *Annals of Statistics*, *Journal of Applied Probability*, *Journal of the American Statistical Association*, *Annals of the Institute of Statistical Mathematics*, *Naval Logistic Quarterly*, *Statistics and Probability Letters*, *Communications in Statistics*, *Ecology and Theoretical Population Biology*. He has held visiting appointments at Penn State University, University of Waterloo and Memorial University of Newfoundland (Canada). He was elected as Member of the International Statistical Institute (1998). He was the Chief Editor of the *Journal of the Indian Statistical Association* (2000–2006). At present, Dr. Rajarshi is the President of the International Indian Statistician Association-Indian Chapter and the Vice-President of the Indian Statistical Association.

Cross References

►Akaike's Information Criterion
►Asymptotic Relative Efficiency in Estimation
►Bootstrap Methods
►Brownian Motion and Diffusions
►Central Limit Theorems
►Generalized Linear Models
►Kalman Filtering
►Likelihood
►Markov Chain Monte Carlo
►Markov Chains
►Markov Processes
►Martingale Central Limit Theorem
►Martingales
►Methods of Moments Estimation
►Nonparametric Estimation
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Poisson Processes
►Renewal Processes
►Statistical Inference: An Overview
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726
Adke SR, Manjunath SM (1984) An introduction to finite Markov processes. Wiley Eastern, New Delhi
Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical Models based on counting processes. Springer, New York
Basawa IV, Prakasa Rao BLS (1980) Statistical inference for stochastic processes. Academic, London
Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York
Benjamin M, Rigby R, Stasinopoulos M (2003) Generalized autoregressive moving average models. J Am Stat Assoc 461:214–223
Billingsely P (1961) Statistical inference for Markov processes. Chicago University Press, Chicago, IL
Bose A (1988) Edgeworth correction by bootstrap in autoregressions. Ann Stat 1709–1722
Bose A (1990) Bootstrap in moving average models. Ann Inst Stat Math 42:753–768
Cappe O, Moulines E, Ryden T (2005) Inference in Hidden Markov models. Springer, New York
Carlstein E (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann Stat 14:1172–1179
Davison AC (2003) Statistical models. Cambridge University Press, Cambridge
Elliot RJ, Aggaoun L, Moore JB (1995) Hidden Markov models: estimation and control. Springer, New York
Fahrmeir L, Tutz G (2004) Multivariate statistical modelling based on generalized linear models. Springer, New York
Godambe VP (1985) The foundation of finite sample estimation in stochastic processes. Biometrika 72:419–428
Godambe VP (ed) (1991) Estimating Functions. Oxford Science Publications, New York
Gotze F, Kunsch HR (1996) Second-order correctness of the block-wise bootstrap for stationary observations. Ann Stat 24: 1914–1933
Guttorp P (1991) Statistical inference for branching processes. Wiley, New York
Guttorp P (1995) Stochastic modeling of scientific data. Chapman & Hall, London
Hall P, Heyde CC (1980) Martingale limit theory and its applications. Academic, New York
Karr AF (1986) Point processes and their statistical inference. Marcel Dekker, New York
Katz RW (1981) On some criteria for estimating the order of a Markov chain. Technometrics 23:243–249
Keiding N (1975) Maximum likelihood estimation in birth and death process. Ann Stat 3:363–372
Kulperger RJ, Prakasa Rao BLS (1989) Bootstrapping a finite Markov chain. Sankhya A 51:178–191
Kunsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1261
Kutoyants YA (2004) Statistical inference for ergodic diffusion processes. Springer, New York

Lahiri SN (1991) Second order optimality of stationary bootstrap. Statist Probab Lett 11:335–341
Lahiri SN (2003) Resampling methods for dependent data. Springer, New York
Lindsey JK (2004) Statistical analysis of stochastic processes in time. Cambridge University Press, Cambridge
Liu RY, Singh K (1992) Moving blocks jackknife and bootstrap capture weak dependence. In: Lepage R, Billard L (eds) Exploring the limits of bootstrap. Wiley, New York, pp 225–248
MacDonald I, Zucchini W (1997) Hidden Markov and other models for discrete valued time series. Chapman & Hall, London
Martin RD, Yohai VJ (1986) Influence functionals for time series. Ann Stat 14:781–818
McKenzie E (2003) Discrete variate time series. In: Shanbhag DN, Rao CR (eds) Stochastic processes: modeling and simulation. Handbook of statistics. North-Holland, Amsterdam, pp 573–606
Prakasa Rao BLS (1996) Nonparametric Approach in Time Series Analysis. In: Prakasa Rao BLS, Bhat BR (eds) Stochastic processes and statistical inference. New Age International New Delhi, pp 73–89
Prakasa Rao BLS (1999a) Statistical inference for diffusion type processes. Arnold, London
Prakasa Rao BLS (1999b) Semimartingales and their statistical inference. CRC Press, Boca Raton, FL
Prakasa Rao BLS (2010) Statistical inference for fractional diffusion processes. Wiley, New York
Prakasa Rao BLS, Bhat BR (eds) (1996) Stochastic processes and statistical inference. New Age International, New Delhi
Radulovic D (2002) On the bootstrap and the empirical processes for dependent sequences. In: Dehling H, Mikosch T, Sorensen M (eds) Empirical process techniques for dependent data, Birkhauser, Boston, pp 345–364
Raftery AE (1985) A model for high order Markov chains. J Roy Stat Soc B 47:528–539
Robinson PM (1983) Nonparametric estimators for time series. J Time Ser Anal 4:185–207
Scott DJ (1978) A central limit theorem for martingales and an application to branching processes. Stoch Processes Appl 6:241–252
Sorensen H (2004) parametric inference for diffusion processes observed at discrete points in time: a survey. Internat Statist Rev 72:337–354

Statistical Inference in Ecology

Subhash R. Lele[1], Mark L. Taper[2]
[1]Professor
University of Alberta, Edmonton, AB, Canada
[2]Research Scientist
Montana State University, Bozeman, MT, USA

Researchers in ecology and evolution have long recognized the importance of understanding randomness in nature in order to distinguish the underlying pattern. Sir Francis Galton developed regression analysis to answer questions about heredity; Karl Pearson's systems of distributions were motivated by the desire to fit evolutionary data on the size of crab claws. Fisher's contributions from the fundamental theorem of evolution to fields of quantitative genetics, species abundance distributions and measurement of diversity are legendary. Studies on the geographic distribution of species led to the study of spatial statistics in ecology in the early part of the 20th century. The classification and discrimination methods developed by Fisher and others for numerical taxonomy and community ecology are still commonly used in ecology.

Unfortunately, Karl Pearson believed that causation was an illusion of scientific perception, stating in the introduction to the 1911 3rd edition of *The Grammar of Science*, "Nobody believes now that science explains anything; we all look upon it as a shorthand description, as an economy of thought." Under Pearson's influence, statistical techniques in ecology tended, until recently, to be more descriptive than predictive with a major early exception of path analysis developed by Sewall Wright in the first decades of the 20th century.

In curious contradiction, mathematical models used by ecologists to model population dynamics and related processes were highly sophisticated and predictive in nature. For example, Lotka–Volterra models were developed in the 1930s. Generalization of these models to multispecies cases such as the Predator-Prey, Host-Parasitoid and other systems of models were available soon after that. Skellam (1951) pioneered the use of spatial diffusion processes to model spread of invasive species.

Gause's work (Gause 1934) was unique in that he tried to validate the mathematical models using experimental data. He used non-linear regression to fit Logistic growth model to the population growth series for paramecia. Most of this work was based on the assumption that error comes into the process only through observational inaccuracies, and thus he missed the modern nuance of inherent randomness or process variation.

Statistical ecology received a large impetus in the 1970s after the publication of Professor E.C. Pielou's numerous classic books (e.g., Pielou 1977) and number of conferences and the resultant edited volumes by Professor G.P. Patil (e.g., Patil et al. 1971). These provided nice summaries of what was known then and also indicated future directions. Driven by the passage of the 1973 Endangered Species Act (ESA) and the dozens of other environmental laws passed in the United States during the 1970's the field of ecology gained substantial prominence in the context of managing and not simply describing ecosystems. This necessitated the development of models that were predictive and not simply descriptive.

Population Viability Analysis (PVA) where one uses stochastic models to predict the distribution of extinction times for a population or species of concern became an important tool for studying the effect of various human activities on nature. Political decisions regarding the conservation of species are often legally required by the ESA to consider the results of a PVA. The importance of demographic and environmental stochasticity as well as the measurement error in forecasting became apparent. Expanding beyond a single population focus, the development of meta-population theory was based on probabilistic models for spatial dispersal and growth. Ecologists became more familiar and comfortable with the idea of modeling randomness and studying its impact on prediction. While much of what is modeled as random in ecology undoubtedly represents unrecognized deterministic influences, it seems likely that true stochasticity is as much a fundamental part of ecology as it is in physics. For example, demographic events such as the sex of offspring are truly random, and not simply the consequence unrecognized deterministic influences. Such demographic stochasticity strongly influences population dynamics when population size is low.

Although stochastic models became prominent in the 1970s and 80s, statistical inference, the methods that connect theoretical models to data, or inductive inference, was still limited. Most of the statistical techniques used were based on linear regression and its derivatives such as the ►Analysis of Variance. The main hurdles were limited data, limited computational power and mathematical nature of the statistical inferential tools. Dennis et al. (1991) and Dennis and Taper (1994) made a major advance by incorporating stochastic population dynamic models as the skeleton for a full likelihood based inference in ecological time series.

The rapid rise in computational power available to ecologists, coupled with the development of computational statistical techniques especially the bootstrap (see ►Bootstrap Methods) and Monte-Carlo approaches have reduced the threshold of mathematical expertise necessary to apply sophisticated statistical inference techniques making the analysis of complex ecological models feasible. This has provided significant impetus for developing strong inferential tools in ecology.

Following are some of the important examples of the application of statistical thinking in ecology.

1. *Sampling methods for estimation of population abundances and occurrences*: Mark-Capture-Recapture (Seber 2002) methods have formed an important tool in the statistical ecology toolbox, but have also led to development of new statistical methods that have found applications in epidemiology and other sciences. Capture probabilities may change temporally or spatially. ►Generalized Linear models and mixed models have proved their usefulness in these situations. Biases due to visibility are adjusted using distance based sampling methods. In many instances, it is too expensive to conduct abundance estimation and one has to settle for site occupancy models based on presence-absence data. Site occupancy data and methods have made a broader range of ecologists aware of the ubiquitous nature of measurement error. Although a species may be present, it may not be detected because of various factors such as lack of visibility, time of the day when birds may not be singing etc. (MacKenzie et al. 2006). This is an active area of research.
2. *Resource selection by animals*: Ecologists need to know what resources animals select and how does this selection affect their fitness and survival. Human developments such as dams or a gas pipe line across a habitat that might be critical to the animals can doom their survival. Recent technological advances such as GPS collars and DNA analysis help in collecting information on where animals spend their time and what they eat. The resource selection probability function (RSPF) (Manly et al. 2002; Lele and Allen 2006) and habitat suitability maps (Hirzel et al. 2006) have been essential tools for environmental impact assessments (EIA) for studying impact of various developments.
3. *Model identification and selection*: The statistical models used for prediction can be either process driven or phenomenological, "black box", models (Breiman 2001). Predictions from ecological models are often made for the distant and not the immediate future. This extrapolation makes it essential that ecological models be process driven. The use of powerful likelihood methods for analyzing population time series models is a relatively new development. The predictions are strongly affected by the particular process based model chosen. This has forced ecologist to consider many models simultaneously and to search for good methods for ►model selection. Information based model selection (Burnham and Anderson 2002) has received considerable attention in this context. Although alternative methods and modifications are constantly being suggested and tested (Taper et al. 2008).
4. *Hierarchical models*: This is one of the most exciting developments in statistical ecology. General hierarchical models are also known as latent variable models, random effects models, mixed models and ►mixture models. These models are natural models to account

for the hierarchical structure inherent in many ecological processes. They also simplify statistical analysis in the presence of missing data, sampling variability, covariates measured with error and other problems commonly faced by ecologists. Reviews of the use of hierarchical models in ecology are available in Royle and Dorazio (2009), Cressie et al. (2009) or Clark and Gelfand (2006). Survival analysis methods and random effects models have found important applications in avian nest survival studies (Natarajan and McCulloch 1999). Linear mixed effects models have been used in evolution and animal breeding since the 1940's. However, generalization of those ideas to more complex models was not possible until recently. Writing down the likelihood function for general hierarchical models is difficult (Lele et al. 2007) and hence use of standard likelihood based inference is not popular. On the other hand, non-informative Bayesian inference using Markov Chain Monte Carlo algorithm (see ►Markov Chain Monte Carlo) is computationally feasible. These calculations are simulation based and replicate the causal processes that ecologists seek to understand. Due to its simplicity, the non-informative Bayesian approach has become quite popular in ecology. However, there are important philosophical and pragmatic issues that should be considered before using this approach (Lele and Allen 2006, Lele and Dennis 2009). Moreover, the recent development of the data-cloning algorithm (Lele et al. 2007; Ponciano et al. 2009) has removed the computational obstacle to likelihood inference for general hierarchical models.

Powerful statistical methods are being developed for ecology, generally coupled with software. The development of accessible tools has greatly facilitated the application of complex statistical analysis to ecological problems. These advances have come at a cost. Researchers are under pressure to be cutting edge and consequently tend to use techniques because they are convenient and fashionable not necessarily because they are appropriate.

Ecological statistics is vibrant and contributing greatly to the advancement of the science, but what are the future directions? One clear recommendation that can be made is in the realm of teaching. Education in ecological statistics has not kept pace with statistical practice in ecology, and improvements are desperately needed (Lele and Taper 2002, Dennis 2004). While methods instruction will always be essential, what is needed most by young ecologists is the development of strong foundational thinking about the role of statistical inference in ecological research.

On the other hand, recommendations regarding the development of new statistics are less clear. Techniques generally follow the questions that need to be answered. However, we are confident that while descriptive statistics and black box prediction will have their place, the greatest advances to knowledge in ecology will come from challenging the probabilistic predictions from explicit models of ecological process with data from well-designed experiments and surveys.

About the Authors

For biographies of both authors *see* the entry ►Statistical Evidence.

Cross References

►Bayesian Statistics
►Distance Sampling
►Factor Analysis and Latent Variable Modelling
►Linear Mixed Models
►Marine Research, Statistics in
►Modeling Survival Data
►Multilevel Analysis
►Nonlinear Mixed Effects Models
►Non-probability Sampling Survey Methods
►Proportions, Inferences, and Comparisons
►Statistical Ecology

References and Further Reading

Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16:199–215

Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag, New York

Clark JS, Gelfand A (eds) (2006) Hierarchical modelling for the environmental sciences: statistical methods and applications. Oxford university press, Oxford, U.K

Cressie N, Calder CA, Clark JS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–570

Dennis B (2004) Statistics and the scientific method in ecology. In: Taper ML, Lele SR (eds) The nature of scientific evidence: statistical, empirical and philosophical considerations. University of Chicago Press, USA, pp 327–378

Dennis B, Taper ML (1994) Density dependence in time series observations of natural populations: estimation and testing. Ecol Monogr 64:205–224

Dennis B, Munholland PL, Scott JM (1991) Estimation of growth and extinction parameters for endangered species. Ecol Monogr 61:115–143

Gause GF (1934) The struggle for existence. Williams and Wilkins, Baltimore, MD, USA

Hirzel AH, LeLay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presence. Ecol Model 199:142–152

Lele SR, Allen KL (2006) On using expert opinion in ecological analyses: a frequentist approach. Environmetrics 17:683–704
Lele SR, Dennis B (2009) Bayesian methods for hierarchical model: are ecologists making a Faustian bargain? Ecol Appl 19:581–584
Lele SR, Keim JL (2006) Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021–3028
Lele SR, Taper ML (2002) What shall we teach in environmental statistics? Discussion. Environ Ecol Stat 9(2):145–146
Lele S, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Lett 10:551–563
MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, NY
Manly BFJ, McDonald LL, Thomas DL, McDonald TL, Erickson WP (2002) Resource selection by animals: statistical analysis and design for field studies, 2nd edn. Kluwer Press, Boston, USA
Natarajan R, McCulloch CE (1999) Modeling heterogeneity in nest survival data. Biometrics 55:553–559
Patil GP, Pielou EC, Waters WE (1971) Statistical ecology: proceedings of the 1969 International Conference on Statistical Ecology, New Haven. Pennsylvania State University Press, PA, USA
Pielou EC (1977) Mathematical ecology. Wiley, New York
Ponciano J, Taper ML, Dennis B, Lele SR (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing and model selection using data cloning. Ecology 90:356–362
Royle A, Dorazio R (2009) Hierarchical models and inference in ecology: the analysis of data from populations, metapopulations and communities. Elsevier Inc., UK
Seber GAF (2002) The estimation of animal abundance and related parameters, 2nd edn. Blackburn Press, Caldwel, NJ
Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38:196–218
Taper ML, Staples DF, Shepard BB (2008) Model structure adequacy analysis: selecting models on the basis of their ability to answer scientific questions. Synthese 163:357–370

Statistical Inference: An Overview

Aris Spanos
Wilson Schmidt Professor
Virginia Tech, Blacksburg, VA, USA

Introduction

Statistical inference concerns the application and appraisal of methods and procedures with a view to *learn from data* about observable stochastic phenomena of interest using probabilistic constructs known as *statistical models*. The basic idea is to construct statistical models using probabilistic assumptions that "capture" the chance regularities in the data with a view to adequately account for the underlying data-generating mechanism; see **?** (**?**). The discussion that follows focuses primarily on frequentist inference, and to a lesser extent on Bayesian inference.

The perspective on statistical inference adopted here is broader than earlier accounts, such as: "making inferences about a population from a random sample drawn from it" (Dodge 2003), in so far as it extends its intended scope beyond *random samples* and static *populations*, to include dynamic phenomena giving rise to observational (non-experimental) data. In addition, the discussion takes into account the fact that the demarcation of the intended scope of statistical inference is intrinsically challenging because it is commonly part of broader scientific inquiries; see Lehmann (1990). In such a broader context statistical inference is often *preceded* with substantive questions of interest, combined with the selection of data pertaining to the phenomenon being studied, and *succeeded* with the desideratum to relate the inference results to the original substantive questions.

This special placing of statistical inference raises a number of crucial methodological problems pertaining to the adequateness of the statistical model to provide a well-grounded link between the phenomenon of interest, at one end of the process, and furnishing evidence for or against the substantive hypotheses of interest, at the other. The link between the phenomenon of interest and the statistical model – thru the data – raises several methodological issues including: the role of substantive and statistical information (Lehmann 1990), as well as the criteria for selecting a statistical model and establishing its adequacy Spanos (2007). The link between the data – construed in the context of a statistical model – and evidence for or against particular substantive claims also raises a number of difficult problems including the fact that "accept" or "reject" the null hypothesis (or a small p-value) does not mean that there is evidence for the null or the alternative, respectively. Indeed, one can make a case that most of the foundational problems bedeviling statistical inference since the 1930s stem from its special place in this broader scientific inquiry; see Mayo (2006).

Frequentist Statistical Inference

Modern statistical inference was founded by Fisher (1922) who initiated a change of paradigms in statistics by recasting the then dominating *Bayesian-oriented induction*, relying on large sample size (n) approximations (Pearson 1920), into a frequentist *statistical model-based induction*, relying on *finite sampling distributions*, inspired by Gosset's (1908) derivation of the Student's t distribution for any sample size $n > 1$. Before Fisher, the notion of a statistical model was implicit, and its role was primarily confined to the *description* of the distributional features

of *the data in hand* using the histogram and the first few sample moments. Unlike Karl Pearson who would commence with data $\mathbf{x}_0$: = $(x_1, \ldots, x_n)$ in search of a frequency curve to describe the histogram of $\mathbf{x}_0$, he proposed to begin with (a) a prespecified model (a hypothetical infinite population), and (b) view $\mathbf{x}_0$ as a realization thereof. Indeed, he made the initial choice (specification) of the prespecified statistical model a response to the question: "Of what population is this a random sample?" (Fisher 1922, p. 313), emphasizing that: "the adequacy of our choice may be tested a posteriori" (ibid., p. 314).

The Notion of a Statistical Model

Fisher's notion of a prespecified statistical model can be formalized in terms of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$, underlying data $\mathbf{x}_0$. This takes the form of parameterizing the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ to specify a *statistical model*:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathbb{R}_X^n,\ \text{for } \Theta\subset\mathbb{R}^m,\ m<n. \tag{1}$$

$f(\mathbf{x};\theta)$ denotes the joint *distribution of the sample* $\mathbf{X}$: = $(X_1, \ldots, X_n)$ that encapsulates the whole of the probabilistic information in $\mathcal{M}_\theta(\mathbf{x})$, by giving a general description of the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ (Doob 1953). $\mathcal{M}_\theta(\mathbf{x})$ is chosen to provide an idealized description of the mechanism that generated data $\mathbf{x}_0$ with a view to appraise and address the substantive questions of interest.

The quintessential example of a statistical model is *the simple Normal model*:

$$\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \backsim \text{NIID}(\mu,\sigma^2),\ \theta\text{: } = (\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ k=1,2,\ldots,n,\ldots, \tag{2}$$

where "$\backsim$ NIID(μ,σ^2)" stands for "distributed as Normal, Independent and Identically Distributed, with mean μ and variance σ^2".

The statistical model $\mathcal{M}_\theta(\mathbf{x})$ plays a pivotal role in statistical inference in so far as it determines what constitutes *a legitimate*:

(a) Event — any well-behaved (Borel) functions of the sample $\mathbf{X}$ —
(b) Assignment of probabilities to legitimate events via $f(\mathbf{x};\theta)$
(c) Data $\mathbf{x}_0$ for inference purposes
(d) Hypothesis or inferential claim
(e) Optimal inference procedure and the associated error probabilities

Formally an event is legitimate when it belongs to the σ-field generated by $\mathbf{X}$ (Billingsley 1995). Legitimate data come in the form of data $\mathbf{x}_0$ that can be realistically viewed as a truly typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$, as specified by $\mathcal{M}_\theta(\mathbf{x})$. Legitimate hypotheses and inferential claims are invariably about the data-generating mechanism and framed in terms of the unknown parameters θ. Moreover, the optimality (effectiveness) of the various inference procedures depends on the validity of the probabilistic assumptions constituting $\mathcal{M}_\theta(\mathbf{x})$; see Spanos (1999).

The interpretation of probability underlying frequentist inference associates probability with the *limit* of relative frequencies anchored on the Strong Law of Large Numbers (SLLN). "Stable relative frequencies" (Neyman 1952), i.e., one's that satisfy the SLLN, constitute a crucial feature of real-world phenomena we call stochastic. The *long-run* metaphor associated with this interpretation enables one to conceptualize probability in terms of viewing $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}_X^n$ as an idealized description of the data-generating mechanism. The appropriateness of this interpretation stems primarily from its capacity to facilitate the task of bridging the gap between stochastic phenomena and the mathematical underpinnings of $\mathcal{M}_\theta(\mathbf{x})$, as well as elucidate a number of issues pertaining to modeling and inference; see Spanos (2009).

Different Forms of Statistical Inference

Fisher (1925), almost single-handedly, put forward a frequentist theory of *optimal estimation,* and Neyman and Pearson (1933) modified Fisher's significance testing to propose an analogous theory for *optimal testing*; see Cox and Hinkley (1974). Optimality of inference in frequentist statistics is defined in terms of the capacity of different procedures to give rise to valid inferences, evaluated in terms of the associated *error probabilities*: how often these procedures lead to erroneous inferences.

The main forms of statistical inference in frequentist statistics are: (a) point estimation, (b) interval estimation, (c) hypothesis testing, and (d) prediction.
All these forms share the following features:

(a) Assume that the prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$ is valid vis-à-vis data $\mathbf{x}_0$.
(b) The objective of inference is always to learn about the underlying data-generating mechanism, and it is framed in terms of the unknown parameter(s) θ.
(c) An inference procedure is based on a *statistic* (estimator, test statistic, predictor), say $Y_n = g(X_1, X_2, \ldots, X_n)$,

whose sampling distribution provides the relevant error probabilities that calibrate its reliability. In principle, the sampling distribution of Y_n is derived via:

$$P(Y_n \leq y) = \underbrace{\int\int \cdots \int}_{\{\mathbf{x}:\, g(x_1,\ldots,x_n)\leq y\}} f(\mathbf{x};\theta)dx_1dx_2\cdots dx_n. \quad (3)$$

Point estimation centers on a mapping: $h(.): \mathbb{R}_X^n \to \Theta$, say $\widehat{\theta}_n(\mathbf{X}) = h(X_1, X_2, \ldots, X_n)$, known as an estimator of θ. The idea underlying optimal estimation is to select a mapping $h(.)$ that locates, as closely as possible, the *true* value of θ; whatever that happens to be. The qualification "as closely as possible" is quantified in terms of certain features of the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$, known as estimation properties: unbiasedness, efficiency, sufficiency, consistency, etc.; see Cox and Hinkley (1974).

A *key concept* in Fisher's approach to inference is the *likelihood function*:

$$L(\theta;\mathbf{x}) = \ell(\mathbf{x}) \cdot f(\mathbf{x};\theta),\ \theta \in \Theta, \quad (4)$$

where $\ell(\mathbf{x}) > 0$ denotes a proportionality constant. Fisher (1922) defined the *Maximum Likelihood* (ML) estimator $\widehat{\theta}_{ML}(\mathbf{X})$ of θ to be the one that maximizes $L(\theta;\mathbf{x})$. He was also the first to draw a sharp distinction between the *estimator* $\widehat{\theta}(\mathbf{X})$ and the *estimate* $\widehat{\theta}(\mathbf{x}_0)$, and emphasized the importance of using the sampling distribution of $\widehat{\theta}(\mathbf{X})$ to evaluate the reliability of inference in terms of the relevant error probabilities.

Example In the case of the simple Normal model, the statistics:

$$\overline{X}_n = \frac{1}{n}\Sigma_{k=1}^n X_k \backsim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right),$$
$$s^2 = \frac{1}{(n-1)}\Sigma_{k=1}^n (X_k - \overline{X}_n)^2 \backsim \left(\frac{\sigma^2}{n-1}\right)\chi^2(n-1), \quad (5)$$

where $\mathsf{N}(.,.)$ and $\chi^2(.)$ denote the Normal and chis-square distributions, constitute "good" estimators of (μ, σ^2) is terms of satisfying most of the above properties.

Point estimation is often considered *inadequate* for the purposes of scientific inquiry because a "good" point estimator $\widehat{\theta}_n(\mathbf{X})$, by itself, does not provide any measure of the reliability and precision associated with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$. This is the reason why $\widehat{\theta}_n(\mathbf{x}_0)$ is often accompanied by some significance test result (e.g., p-value) associated with the *generic* hypothesis $\theta = 0$.

Interval estimation rectifies this crucial weakness of point estimation by providing the relevant error probabilities associated with inferences pertaining to "covering" the true value of θ. This comes in the form of the Confidence Interval (CI):

$$\mathbb{P}(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha, \quad (6)$$

where the statistics $L(\mathbf{X})$ and $U(\mathbf{X})$ denote the lower and upper (random) bounds that "covers" the true value θ^* with probability $(1-\alpha)$, or equivalently, the "coverage error" probability is α.

Example In the case of the simple Normal model:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) = 1 - \alpha, \quad (7)$$

provides a $(1-\alpha)$ Confidence Interval (CI) for μ. The evaluation of the coverage probability $(1-\alpha)$ is based on the following sampling distribution result:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \backsim \mathsf{St}(n-1), \quad (8)$$

where $\mathsf{St}(n-1)$ denotes the Student's t distribution with $(n-1)$ degrees of freedom, attributed to Gosset (1908).

What is often not appreciated sufficiently about estimation in general, and CIs in particular, is the underlying reasoning that gives rise to sampling distribution results such as (5) and (8). The reasoning that underlies estimation is *factual,* based on evaluating the relevant sampling distributions "under the True State of Nature" (TSN), i.e., the *true* data-generating mechanism: $\mathcal{M}^*(\mathbf{x}) = \{f(\mathbf{x};\theta^*)\}$, $\mathbf{x} \in \mathbb{R}_X^n$, where θ^* denotes the true value of the unknown parameter(s) θ. Hence, the generic CI in (6) is more accurately stated as:

$$\mathbb{P}\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X});\ \theta = \theta^*\right) = 1 - \alpha, \quad (9)$$

where $\theta = \theta^*$ denotes 'evaluated under the TSN'. The remarkable thing about factual reasoning is that one can make probabilistic statements like (9), with a precise error probability (α), *without* knowing the true θ^*.

Example In the case of the simple Normal model, the distributional results (5) and (8) are more accurately stated as:

$$\overline{X}_n \overset{\text{TSN}}{\backsim} \mathsf{N}\left(\mu_*, \frac{\sigma_*^2}{n}\right), \quad \frac{(n-1)s^2}{\sigma_*^2} \overset{\text{TSN}}{\backsim} \chi^2(n-1),$$
$$\frac{\sqrt{n}(\overline{X}_n - \mu^*)}{s} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1), \quad (10)$$

where $\theta^*{:=} \left(\mu_*, \sigma_*^2\right)$ denote the "true" values of the unknown parameters $\theta{:=}(\mu, \sigma^2)$.

Prediction is similar to estimation in terms of its underlying factual reasoning, but it differs from it in so far as it is concerned with finding the most representative

value of X_k beyond the observed data, say X_{n+1}. An optimal predictor of X_{n+1} is given by:

$$\widehat{X}_{n+1} = \overline{X}_n, \tag{11}$$

whose reliability can be calibrated using the sampling distribution of the prediction error:

$$\widehat{u}_{n+1} = (X_{n+1} - \overline{X}_n) \overset{\text{TSN}}{\backsim} \mathsf{N}\left(0, \sigma_*^2\left(1+\frac{1}{n}\right)\right), \tag{12}$$

to construct a $(1-\alpha)$ prediction interval:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}\left(s\sqrt{\left(1+\frac{1}{n}\right)}\right) \leq X_{n+1} \leq \overline{X}_n + c_{\frac{\alpha}{2}}\left(s\sqrt{\left(1+\frac{1}{n}\right)}\right); \theta = \theta^*\right) = 1-\alpha. \tag{13}$$

Hypothesis testing. In contrast to estimation, the reasoning underlying hypothesis testing is *hypothetical*. The sampling distribution of a test statistic is evaluated under several hypothetical scenarios concerning the statistical model $\mathcal{M}_\theta(\mathbf{x})$, referred to as "under the null" and "under the alternative" hypotheses of interest.

Example Consider testing the hypotheses in the context of (2):

$$H_0\text{: } \mu \leq \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0. \tag{14}$$

What renders the hypotheses in (14) legitimate is that: (a) they pose questions concerning the underlying data-generating mechanism, (b) they are framed in terms of the unknown parameter θ, and (c) in a way that partitions $\mathcal{M}_\theta(\mathbf{x})$. In relation to (c), it is important to stress that even in cases where substantive information excludes or focuses exclusively on certain subsets (or values) of the parameter space, the entire Θ is relevant for statistical inference purposes. Ignoring this, and focusing only on the substantively relevant subsets of Θ, gives rise to fallacious results.

The N-P test for the hypotheses (14) $T_\alpha := \{\tau(\mathbf{X}), C_1(\alpha)\}$, where:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s},\ C_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\}, \tag{15}$$

can be shown to be Uniformly Most Powerful (UMP) in the sense that, its type I error probability (significance level) is:

$$\begin{aligned}(\text{a})\ \ \alpha &= \max_{\mu \leq \mu_0} \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; H_0)\\ &= \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; \mu = \mu_0),\end{aligned} \tag{16}$$

and among all the α-level tests T_α has highest *power* (Lehmann 1986):

$$\begin{aligned}&(\text{b})\ \mathbb{P}(\mathbf{x}: \tau(\mathbf{X}) > c_\alpha; \mu = \mu_1), \textit{ for all } \mu_1 > \mu_0,\\ &\mu_1 = \mu_0 + \gamma,\ \gamma \geq 0;\end{aligned} \tag{17}$$

In this sense, a UMP test provides the most effective α-level probing procedure for detecting any discrepancy ($\gamma \geq 0$) of interest from the null.

To evaluate the error probabilities in (16) and (17) one needs to derive the sampling distribution of $\tau(\mathbf{X})$ under several *hypothetical* values of μ relating to (14):

$$(\text{a})\ \tau(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{St}(n-1), \quad (\text{b})\ \tau(\mathbf{X}) \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta(\mu_1); n-1), \quad \text{for any } \mu_1 > \mu_0, \tag{18}$$

where $\delta(\mu_1) = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$ is known as the non-centrality parameter. The sampling distribution in (18a) is also used to evaluate Fisher's (1935) p-value:

$$p(\mathbf{x}_0) = \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_0), \tag{19}$$

where a small enough $p(\mathbf{x}_0)$ can be interpreted as indicating discordance with H_0.

Remark It is unfortunate that most statistics books use the vertical bar (|) instead of the semi-colon (;) in formulae (16)–(17) to denote the evaluation *under* H_0 or H_1, as it relates to (18), encouraging practitioners to misinterpret error probabilities as being *conditional* on H_0 or H_1; see Cohen (1994). It is worth emphasizing these error probabilities are: (1) never conditional, (2) always assigned to inference procedures (never to hypotheses), and (3) invariably depend on the sample size $n > 1$.

Comparing the sampling distributions in (18) with those in (10) brings out the key difference between hypothetical and factual reasoning: in the latter case there is only one unique scenario, but in hypothetical reasoning there is usually an infinity of scenarios. The remarkable thing about hypothetical reasoning is that one can pose sharp questions by comparing $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$, for different hypothetical values of θ, with $\mathcal{M}^*(\mathbf{x}_0)$, to learn about $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$. This often elicits more informative answers from $\mathbf{x}_0$ than factual reasoning. This difference is important in understanding the nature of the error probabilities associated with each type of inference as well as in interpreting the results of these procedures.

In particular, factual reasoning can only be used pre-data to generate the relevant error probabilities, because when data $\mathbf{x}_0$ is observed (i.e., post-data) the unique factual scenario has been realized and the sampling distribution in question becomes degenerate. This is the reason why the p-value in (19) is a well-defined post-data error probability, but one cannot attach error probabilities to an observed CI: $(L(\mathbf{x}_0) \leq \theta \leq U(\mathbf{x}_0))$; see the exchange between Fisher (1955) and Neyman (1956). In contrast, the scenarios in hypothetical reasoning are equally relevant to both pre-data and post-data assessments. Indeed, one can go a long

way towards delineating some of the confusions surrounding frequentist testing, as well as addressing some of the criticisms leveled against it – statistical vs. substantive significance, with a large enough n one can reject any null hypothesis, no evidence against the null is *not* evidence for it – using post-data error probabilities to provide an evidential interpretation of frequentist testing based on the severity rationale; see Mayo and Spanos (2006) for further discussion.

Bayesian Inference

Bayesian inference also begins with a prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$, as specified in (1), but modifies it in three crucial respects:

(1) Probability is now interpreted as (subjective or rational) *degrees of belief* (not as the limit of relative frequencies).
(2) The unknown parameter(s) θ are now viewed as *random variables* (not as constants) with their own distribution $\pi(\theta)$, known as the *prior distribution*.
(3) The distribution of the sample is now viewed as *conditional* on θ, and denoted by $f(\mathbf{x} \mid \theta)$ instead of $f(\mathbf{x};\theta)$.

All three of these modifications have been questioned in the statistics literature, but the most prominent controversies concern the nature and choice of the prior distribution. There are ongoing disputes concerning subjective vs. default (reference) priors, informative vs. non-informative (invariant) priors, proper vs. improper priors, conjugate vs. non-conjugate, matching vs. non-matching priors, and how should these choices be made in practice; see Kass and Wasserman (1996) and Roberts (2007).

In light of these modifications, one can use the definition of conditional probability distribution between two jointly distributed random vectors, say (Z,W):

$$f(\mathbf{z} \mid \mathbf{w}) = \frac{f(\mathbf{z},\mathbf{w})}{f(\mathbf{w})} = \frac{f(\mathbf{z},\mathbf{w})}{\int_{\mathbf{z}} f(\mathbf{z},\mathbf{w})d\mathbf{z}} = \frac{f(\mathbf{w} \mid \mathbf{z})f(\mathbf{z})}{\int_{\mathbf{z}} f(\mathbf{w} \mid \mathbf{z})f(\mathbf{z})d\mathbf{z}},$$

to define *Bayes formula* that determines the *posterior* distribution of θ:

$$\pi(\theta \mid \mathbf{x}_0) = \frac{f(\mathbf{x}_0|\theta)\cdot\pi(\theta)}{\int_\theta f(\mathbf{x}_0|\theta)\cdot\pi(\theta)d\theta} \propto \pi(\theta)\cdot L(\theta \mid \mathbf{x}_0),\ \theta \in \Theta, \tag{20}$$

where $L(\theta \mid \mathbf{x}_0)$ denotes the *reinterpreted* likelihood function, not (4).

Bayesian inference is based exclusively on the posterior distribution $\pi(\theta \mid \mathbf{x}_0)$ which is viewed as the *revised* (from the initial $\pi(\theta)$) degrees of belief for different values of θ in light of the summary of the data by $L(\theta \mid x_0)$. A Bayesian point estimate of θ specified by selecting the *mean* ($\widehat{\theta}_B(\mathbf{x}_0) = E(\pi(\theta \mid \mathbf{x}_0))$) or the *mode* of the posterior. A Bayesian interval estimate for θ is given by finding two values $a < b$ such that:

$$\int_a^b \pi(\theta \mid \mathbf{x}_0)d\theta = 1-\alpha, \tag{21}$$

known as a $(1-\alpha)$ *posterior* (or credible) *interval*.

Bayesian *testing of hypotheses* is more difficult to handle in terms of the posterior distribution, especially for point hypotheses, because of the technical difficulty in attaching probabilities to particular values of θ, since the parameter space Θ is usually *uncountable*. There have been numerous attempts to address this difficulty, but no agreement seems to have emerged; see Roberts (2007). Assuming that one adopts his/her preferred way to sidestep this difficulty, Bayesian testing for H_0: $\theta = \theta_0$ vs. H_1: $\theta = \theta_1$ relies on comparing their respective degrees of belief using the *posterior ratio*:

$$\frac{\pi(\theta_0|\mathbf{x}_0)}{\pi(\theta_1|\mathbf{x}_0)} = \frac{L(\theta_0|\mathbf{x}_0)\cdot\pi(\theta_0)}{L(\theta_1|\mathbf{x}_0)\cdot\pi(\theta_1)}, \tag{22}$$

or, its more widely used modification in the form of the *Bayes Factor* (BF):

$$BF(\mathbf{x}_0) = \left(\frac{\pi(\theta_0|x_0)}{\pi(\theta_1|\mathbf{x}_0)}\right) / \left(\frac{\pi(\theta_0)}{\pi(\theta_1)}\right) = \frac{L(\theta_0|\mathbf{x}_0)}{L(\theta_1|\mathbf{x}_0)}, \tag{23}$$

together with certain rules of thumb, concerning the *strength* of the degrees of belief *against* H_0 based on the magnitude of $\ln BF(\mathbf{x}_0)$: for $0 \leq \ln BF(\mathbf{x}_0) \leq .5$, $.5 < \ln BF(\mathbf{x}_0) \leq 1$, $1 < \ln BF(\mathbf{x}_0) \leq 2$ and $\ln BF(\mathbf{x}_0) > 2$, the degree of belief against H_0 is *poor*, *substantial*, *strong* and *decisive*, respectively; see Roberts (2007). Despite their intuitive appeal, these rules of thumb have been questioned by Kass and Raftery (1995) inter alia.

The question that naturally arises at this stage concerns the nature of the reasoning underlying *Bayesian inference*. In Bayesian inference *learning* is about revising one's degrees of belief pertaining to $\theta \in \Theta$, from $\pi(\theta)$ (pre-data) to $\pi(\theta \mid \mathbf{x}_0)$ (post-data). In contrast to frequentist inference — which pertains to the *true* data-generating mechanism $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$ — Bayesian inference is concerned with more or less appropriate (in terms of $\pi(\theta \mid \mathbf{x}_0)$) models within $\mathcal{M}_\theta(\mathbf{x}_0)$, $\theta \in \Theta$. In terms of the underlying reasoning the Bayesian is similar to the decision theoretic inference which is also about selecting among more or less cost (or utility)-appropriate models. This questions attempts to present N-P testing as naturally belonging to the decision theoretic approach.

The problem with the inference not pertaining to the underlying data-generating mechanism can be brought out more clearly when Bayesian inference is viewed in

the context of the broader scientific inquiry. In that context, one begins with substantive questions pertaining to the phenomenon of interest, and the objective is to learn about the phenomenon itself. Contrasting frequentist with Bayesian inference, using interval estimation as an example, Wasserman (2008) argued: "Frequentist methods have coverage guarantees; Bayesian methods don't. In science, coverage matters" (p. 463).

About the Author

Dr. Aris Spanos is Professor of Economics and former Chair of the Department of Economics (2001–2006) at Virginia Tech, USA. Previously he has taught at London University (England), Cambridge University (England), University of California (USA) and the University of Cyprus. He is the author of two textbooks entitled: *Statistical Foundations of Econometric Modelling* (Cambridge University Press, 1986), and *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data* (Cambridge University Press, 1999). He has published over 70 papers in leading econometric, economic, philosophical and statistical journals.

Cross References

►Bayes' Theorem
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Nonparametric Statistics
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Causation and Causal Inference
►Confidence Interval
►Degrees of Freedom in Statistical Inference
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Estimation
►Estimation: An Overview
►Exact Inference for Categorical Data
►Fiducial Inference
►Frequentist Hypothesis Testing: A Defense
►Generalized Quasi-Likelihood (GQL) Inferences
►Inference Under Informative Probability Sampling
►Likelihood
►Multi-Party Inference and Uncongeniality
►Neyman-Pearson Lemma
►Nonparametric Predictive Inference
►Nonparametric Statistical Inference
►Null-Hypothesis Significance Testing: Misconceptions
►Optimal Statistical Inference in Financial Engineering
►Parametric Versus Nonparametric Tests
►Philosophical Foundations of Statistics
►Proportions, Inferences, and Comparisons
►P-Values
►Ranking and Selection Procedures and Related Inference Problems
►Robust Inference
►Sampling Distribution
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Inference
►Statistical Inference for Quantum Systems
►Statistical Inference for Stochastic Processes
►Statistical Inference in Ecology

References and Further Reading

Billingsley P (1995) Probability and measure, 4th edn. Wiley, New York
Cohen J (1994) The earth is round ($p < .05$). Am Psychol 49:997–1003
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman & Hall, London
Dodge Y (ed) (2003) The Oxford dictionary of statistical terms. The International Statistical Institute, Oxford University Press, Oxford
Doob JL (1953) Stochastic processes. Wiley, New York
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans Roy Soc A 222:309–368
Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Philos Soc 22:700–725
Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh
Fisher RA (1955) Statistical methods and scientific induction. J Roy Stat Soc B 17:69–78
Gosset WS (1908) The probable error of the mean, Biometrika 6:1–25
Kass RE, Raftery AE (1995) Bayes factor and model uncertainty. J Am Stat Assoc 90:773–795
Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91:1343–1370
Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York
Lehmann EL (1990) Model specification: the views of Fisher and Neyman, and later developments. Stat Sci 5:160–168
Mayo DG (1996) Error and the growth of experimental knowledge, The University of Chicago Press, Chicago
Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. Br J Philos Sci 57:323–357
Neyman J (1952) Lectures and conferences on mathematical statistics and probability, 2nd edn. U.S. Department of Agriculture, Washington, DC
Neyman J (1956) Note on an article by Sir Ronald Fisher. J Roy Stat Soc B 18:288–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans Roy Soc A 231: 289–337
Pearson K (1920) The fundamental problem of practical statistics. Biometrika XIII:1–16

Roberts CP (2007) The Bayesian choice, 2nd edn. Springer, New York
Spanos A (1999) Probability theory and statistical inference: econometric modeling with observational data. Cambridge University Press, Cambridge
Spanos A (2007) Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach, Philosophy of Science, 74:1046–1066
Spanos A (2009) Model-based inference and the frequentist interpretation of probability. Working Paper, Virginia Tech
Student (pseudonym for Gosset W) (1908) The probable error of the mean. Biometrika 6:1–25
Wasserman L (2008) Comment on article by Gelman. Bayesian Anal 3:463–466

Statistical Literacy, Reasoning, and Thinking

Joan Garfield
Professor
University of Minnesota, Minneapolis, MN, USA

Statistics educators often talk about their desired learning goals for students, and invariably, refer to outcomes such as being statistically literate, thinking statistically, and using good statistical reasoning. Despite the frequent reference to these outcomes and terms, there have been no agreed upon definitions or distinctions. Therefore, the following definitions were proposed by Garfield (2005 and have been elaborated in Garfield and Ben-Zvi (2008).

Statistical literacy is regarded as a key ability expected of citizens in information-laden societies, and is often touted as an expected outcome of schooling and as a necessary component of adults' numeracy and literacy. Statistical literacy involves understanding and using the basic language and tools of statistics: knowing what basic statistical terms mean, understanding the use of simple statistical symbols, and recognizing and being able to interpret different representations of data (Garfield 1999; Rumsey 2002; Snell 1999).

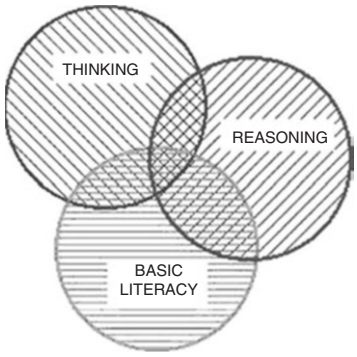
There are other views of statistical literacy such as Gal's (2000, 2002), whose focus is on the data consumer: Statistical literacy is portrayed as the ability to interpret, critically evaluate, and communicate about statistical information and messages. Gal (2002) argues that statistically literate behavior is predicated on the joint activation of five interrelated knowledge bases (literacy, statistical, mathematical, context, and critical), together with a cluster of supporting dispositions and enabling beliefs. Watson and Callingham (2003) proposed and validated a model of three levels of statistical literacy (knowledge of terms, understanding of terms in context, and critiquing claims in the media).

Statistical reasoning is the way people reason with statistical ideas and make sense of statistical information. Statistical reasoning may involve connecting one concept to another (e.g., understanding the relationship between the mean and standard deviation in a distribution) or may combine ideas about data and chance (e.g., understanding the idea of confidence when making an estimate about a population mean based on a sample of data). Statistical reasoning also means understanding and being able to explain statistical processes, and being able to interpret statistical results (Garfield 2002). For example, being able to explain the process of creating a sampling distribution for a statsistics and why this distribution has particular features. Statistical reasoning invovles the mental representations and connections that students have regarding statistical concepts. Another examples is being able to see how and why an outlier makes the mean and standard deviation larger than when that outlier is removed, or reasoning about the effect of an influential data value on the correlation coefficient.

Statistical thinking involves a higher order of thinking than statistical reasoning. Statistical thinking is the way professional statisticians think (Wild and Pfannkuch 1999). It includes knowing how and why to use a particular method, measure, design or statistical model; deep understanding of the theories underlying statistical processes and methods; as well as understanding the constraints and limitations of statistics and statistical inference. Statistical thinking is also about understanding how statistical models are used to simulate random phenomena, understanding how data are produced to estimate probabilities, recognizing how, when, and why existing inferential tools can be used, and being able to understand and utilize the context of a problem to plan and evaluate investigations and to draw conclusions (Chance 2002). Finally, statistical thinking is the normative use of statistical models, methods, and applications in considering or solving statistical problems.

Statistical literacy, reasoning, and thinking are unique learning outcomes, but there is some overlap as well as a type of hierarchy, where statistical literacy provides the foundation for reasoning and thinking (see Fig. 1). A summary of additional models of statistical reasoning and thinking can be found in Jones et al. (2004).

There is a growing network of researchers who are interested in studying the development of students' statistical literacy, reasoning, and thinking (e.g., SRTL – The



Statistical Literacy, Reasoning, and Thinking. Fig. 1 The overlap and hierarchy of statistical literacy, reasoning, and thinking (Artist Website, https://app.gen.umn.edu/artist)

International Statistical Reasoning, Thinking, and Literacy Research Forums, http://srtl.stat.auckland.ac.nz/). The topics of the research studies conducted by members of this community reflect a shift in emphasis in statistics instruction, from developing procedural understanding, i.e., statistical techniques, formulas, computations and procedures; to developing conceptual understanding and statistical literacy, reasoning, and thinking.

Words That Characterize Assessment Items for Statistical Literacy, Reasoning, and Thinking

One way to distinguish between these related outcomes is by examining the types of words used in assessment of each outcome. Table 1 (modified from delMas (2002)) lists words associated with different assessment items or tasks.

Statistical Literacy, Reasoning, and Thinking. Table. 1 Typical words associated with different assessment items or tasks

Basic Literacy	Reasoning	Thinking
Identify	Explain why	Apply
Describe	Explain how	Critique
Translate		Evaluate
Interpret		Generalize
Read		
Compute		

The following three examples (from Garfield and Ben-Zvi 2008) illustrate how statistical literacy, reasoning, and thinking may be assessed.

Example of an Item Designed to Measure Statistical Literacy

A random sample of 30 first-year students was selected at a public university to estimate the average score on a mathematics placement test that the state mandates for all freshmen. The average score for the sample was found to be 81.7 with a sample standard deviation of 11.45. Describe to someone who has not studied statistics what the standard deviation tells you about the variability of placement scores for this sample.

This item assesses statistical literacy because it focuses on understanding (knowing) what the term "standard deviation" means.

Example of an Item Designed to Measure Statistical Reasoning

The following stem plot displays the average annual snowfall amounts (in inches, with the stems being tens and leaves being ones) for a random sample of 25 American cities:

0	000000024
1	028
2	00228
3	8
4	2248
5	48
6	0

Without doing any calculations, would you expect the mean of the snowfall amounts to be larger, smaller, or about the same as the median? Why?

This item assess statistical reasoning because students need to connect and reason about how shape of a distribution affects the relative locations of measures of center, in

this case, reasoning that the mean would be larger than the mean because of the positive skew.

Example of an Item Designed to Assess Statistical Thinking

A random sample of 30 first year students was selected at a public university to estimate the average score on a mathematics placement test that the state mandates for all freshmen. The average score for the sample was found to be 81.7 with a sample standard deviation of 11.45.

A psychology professor at a state college has read the results of the university study. The professor wants to know if students at his college are similar to students at the university with respect to their mathematics placement exam scores. This professor collects information for all 53 first year students enrolled this semester in a large section (321 students) of his "Introduction to Psychology" course. Based on this sample, he calculates a 95% confidence interval for the average mathematics placement scores exam to be 69.47 to 75.72. Below are two possible conclusions that the psychology professor might draw. For each conclusion, state whether it is valid or invalid. Explain your choice for both statements. Note that it is possible that neither conclusion is valid.

(a) The average mathematics placement exam score for first year students at the state college is lower than the average mathematics placement exam score of first year students at the university.
(b) The average mathematics placement exam score for the 53 students in this section is lower than the average mathematics placement exam score of first year students at the university.

This item assesses statistical thinking because it asks students to think about the entire process involved in this research study in critiquing and justifying different possible conclusions.

Comparing Statistical Literacy, Reasoning, and Thinking to Bloom's Taxonomy

These three statistics learning outcomes also seem to coincide somewhat with Bloom's more general categories of learning outcomes (1956). In particular, some current measurement experts feel that Bloom's taxonomy is best used if it is collapsed into three general levels (knowing, comprehending, and applying). Statistical literacy may be viewed as consistent with the "knowing" category, statistical reasoning as consistent with the "comprehending" category (with perhaps some aspects of application and analysis) and statistical thinking as encompassing many elements of the top three levels of Bloom's taxonomy (application, analysis, and synthesis).

About the Author

Dr. Joan Garfield is Professor of Educational Psychology and Head of a unique graduate program in Statistics Education at the University of Minnesota, USA. She is Associate Director for Research and co-founder of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), Past Chair of the American Statistical Association Section on Statistical Education (ASA), and past Vice President of the International Association for Statistical Education. She has co-authored or co-edited 5 books including *Developing Students' Statistical Reasoning: Connecting Research and Teaching practice* (Garfield and Ben-Zvi, Springer, 2008) as well as numerous journal articles. Professor Garfield has received the ASA Founders' Award, the CAUSE Lifetime Achievement Award, is a fellow of ASA and AERA, and has received both Post- baccalaureate and Undergraduate Outstanding Teaching awards given by the University of Minnesota. She has helped found three journals in statistics Education (JSE, SERJ and TISE) and currently serves as Associate Editor for SERJ and TISE. Finally, she is co-founder and co-chair of the biennial International Research Forum on Statistical Reasoning, Thinking, and Literacy.

Cross References

►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics in Advancing Quantitative Education
►Statistical Consulting
►Statistics Education

References and Further Reading

Bloom BS (ed) (1956) Taxonomy of educational objectives: the classification of educational goals: handbook I, cognitive domain. Longmans, Green, New York

Chance BL (2002) Components of statistical thinking and implications for instruction and assessment. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/chance.html Retrieved 15 July2007

delMas RC (2002) Statistical literacy, reasoning, and learning: a commentary. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/delmas_intro.html. Retrieved 6 November 2006

Gal I (ed) (2000) Adult numeracy development: theory, research, practice. Hampton Press, Cresskill, NJ
Gal I (2002) Adults' statistical literacy: meaning, components, responsibilities. Int Stat Rev 70(1):1–25
Garfield J (1999) Thinking about statistical reasoning, thinking, and literacy. Paper presented at the first international research forum on statistical reasoning, thinking, and literacy (STRL-1), Kibbutz Be'eri, Israel
Garfield J (2002) The challenge of developing statistical reasoning. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/garfield.html Retrieved 15 July 2007
Garfield J, delMas R, Chance B (2005) The Web-Based Assessment Resource for Improving Statistics Thinking (ARTIST) Project. Project funded by the National Science Foundation. Accessed 1 Aug 2010
Garfield J, Ben-Zvi D (2008) Developing students' statistical reasoning: connecting research and teaching practice. Springer, Dordrecht, The Netherlands
Jones GA, Langrall CW, Mooney ES, Thornton CA (2004) Models of development in statistical reasoning. In: Ben-Zvi D, Garfield J (eds) The challenge of developing statistical literacy, reasoning, and thinking. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 97–117
Rumsey DJ (2002) Statistical literacy as a goal for introductory statistics courses. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/rumsey2.html Retrieved 15 July 2007
Snell L (1999) Using chance media to promote statistical literacy. Paper presented at the 1999 Joint Statistical Meetings, Dallas, TX
Watson JM, Callingham R (2003) Statistical literacy: a complex hierarchical construct. Stat Educ Res J 2:3–46, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf Retrieved 26 April 2008
Wild CJ, Pfannkuch M (1999) Statistical thinking in empirical enquiry. Int Stat Rev 67(3):223–265

Statistical Methods for Non-Precise Data

Reinhard Viertl
Professor and Head
Vienna University of Technology, Vienna, Austria

Non-Precise Data

Real data obtained from measurement processes are not precise numbers or vectors, but more or less non-precise, also called fuzzy. This uncertainty is different from measurement errors and has to be described formally in order to obtain realistic results from data analysis. A real life example is the water level of a river at a fixed time. It is typically not a precise multiple of the scale unit for height measurements. In the past this kind of uncertainty was mostly neglected in describing such data. The reason for that is the idea of the existence of a "true" water level which is identified with a real number times the measurement unit. But this is not realistic. The formal description of such non-precise water levels can be given using the intensity of the wetness of the gauge to obtain the so called *characterizing functions* from the next section. Further examples of non-precise data are readings on digital measurement equipments, readings of pointers on scales, color intensity pictures, and light points on screens.

Remark 1 Non-precise data are different from measurement errors because in error models the observed values y_i are considered to be numbers, i.e., $y_i = x_i + \epsilon_i$, where ϵ_i denotes the error of the i-th observation.

Historically non-precise data were not studied sufficiently. Some earlier work was done in interval arithmetics. General non-precise data in form of so called fuzzy numbers were considered in the 1980s and first publications combining fuzzy imprecision and stochastic uncertainty came up, see Kacprzyk and Fedrizzi (1988). Some of these approaches are more theoretically oriented. An applicable approach for statistical analysis of non-precise data is given in Viertl (1996).

Characterizing Functions of Non-Precise Data

In case of measurements of one-dimensional quantities non-precise observations can be reasonably described by so-called *fuzzy numbers* $x^\star$. Fuzzy numbers are generalizations of real numbers in the following sense. Each real number $x \in I\!R$ is characterized by its indicator function $I_{\{x\}}(\cdot)$. A fuzzy number is characterized by its so-called *characterizing function* $\xi(\cdot)$ which is a generalization of an indicator function. A characterizing function is a real function of a real variable obeying the following:

1. $\xi : I\!R \longrightarrow [0,1]$
2. $\forall\, \delta \in (0,1]$ the so called δ-cut $C_\delta(x^\star) := \{x \in I\!R : \xi(x) \geq \delta\}$ is a non-empty and closed bounded interval

Remark 2 A characterizing function is describing the imprecision of *one* observation. It should not be confused with a probability density which is describing the stochastic variation of a random quantity X.

A fundamental problem is how to obtain the characterizing function of a non-precise observation. This depends on the area of application. Some examples can be given.

Example 1 For data in form of gray intensities in one dimension as boundaries of regions the gray intensity $g(x)$ as an increasing function of one real variable x can be used to obtain the characterizing function $\xi(\cdot)$ in the following way. Take the derivative $\frac{d}{dx}g(x)$ and divide it by its maximum then the resulting function or its convex hull can be used as characterizing function of the non-precise observation.

Non-Precise Samples

Taking observations of a one-dimensional continuous quantity X in order to estimate the distribution of X usually a finite sequence $x_1^\star, \cdots, x_n^\star$ of non-precise numbers is obtained. These non-precise data are given in form of n characterizing functions $\xi_1(\cdot), \cdots, \xi_n(\cdot)$ corresponding to $x_1^\star, \cdots, x_n^\star$. Facing this kind of samples even the most simple concepts like *histograms* have to be modified. This is necessary by the fact that for a given class K_j of a histogram in case of a non-precise observation $x_i^\star$ with characterizing function $\xi_i(\cdot)$ obeying $\xi_i(x) > 0$ for an element $x \in K_j$ and $\xi_i(y) > 0$ for an element $y \in K_j^c$ it is not possible to decide if $x_i^\star$ is an element of K_j or not.

A generalization of the concept of histograms is possible by so-called *fuzzy histograms*. For those histograms the height of the histogram over a fixed class K_j is a fuzzy number $h_j^\star$. For the definition of the characterizing function of $h_j^\star$ compare Viertl (2006). For other concepts of statistics in case of non-precise data compare Viertl (2006).

Fuzzy Vectors

In case of multivariate continuous data $\boldsymbol{x} = (x_1, \cdots, x_n)$, for example the position of an object on a radar screen, the observations are non-precise vectors $\boldsymbol{x}^\star$. Such non-precise vectors are characterized by so called *vector-characterizing functions* $\zeta_{\boldsymbol{x}^\star}(\cdot, \cdots, \cdot)$. These vector-characterizing functions are real functions of n real variables $x_1, \cdots, x_n$ obeying the following:

(1) $\zeta_{\boldsymbol{x}^\star} : I\!R^n \longrightarrow [0,1]$

(2) $\forall\, \delta \in (0,1]$ the δ-cut $C_\delta(\boldsymbol{x}^\star) := \{\boldsymbol{x} \in I\!R^n : \zeta_{\boldsymbol{x}^\star}(\boldsymbol{x}) \geq \delta\}$ is a non-empty, closed and star shaped subset of $I\!R^n$ with finite n-dimensional content

In order to generalize statistics $t(x_1, \cdots, x_n)$ to the situation of fuzzy data the fuzzy sample has to be combinded into a fuzzy vector called *fuzzy combined sample*.

Generalized Classical Inference

Based on combined fuzzy samples point estimators for parameters can be generalized using the so-called *extension principle* from fuzzy set theory. If $\vartheta(x_1, \cdots, x_n)$ is a classical point estimator for θ, then $\vartheta\left(x_1^\star, \cdots, x_n^\star\right) = \vartheta(\boldsymbol{x}^\star)$ yields a fuzzy element $\hat{\theta}^\star$ of the parameter space Θ.

Generalized confidence regions for θ can be constructed in the following way. Let $\kappa(x_1, \cdots, x_n)$ be a classical confidence function for θ with coverage probability $1 - \alpha$, i.e., $\Theta_{1-\alpha}$ is the corresponding confidence set. For fuzzy data $x_1^\star, \cdots, x_n^\star$ a generalized confidence set $\Theta_{1-\alpha}^\star$ is defined as the fuzzy subset of Θ whose membership function $\varphi(\cdot)$ is given by its values

$$\varphi(\theta) = \left\{ \begin{array}{ll} \sup\left\{\zeta(\boldsymbol{x}) : \boldsymbol{x} \in M_X^n, \theta \in \kappa(\boldsymbol{x})\right\} & \text{if } \exists\, \boldsymbol{x} : \theta \in \kappa(\boldsymbol{x}) \\ 0 & \text{if } \nexists\, \boldsymbol{x} : \theta \in \kappa(\boldsymbol{x}) \end{array} \right\} \quad \forall\, \theta \in \Theta.$$

Statistical tests are mostly based on so-called *test statistics* $t(x_1, \cdots, x_n)$. For non-precise data the values $t\left(x_1^\star, \cdots, x_n^\star\right)$ become non-precise numbers. Therefore test decisions are not as simple as in the classical (frequently artificial) situation. There are different generalizations possible. Also in case of non-precise values of the test statistic it is possible to find ▶*p*-values and the test decision is possible similar to the classical case. Another possibility is to define fuzzy *p*-values which seems to be more problem adequate. For details see Viertl (2006).

There are other approaches for the generalization of classical inference procedures to the situation of fuzzy data. References for that are Gil et al. (1988) and Näther (1997).

Generalized Bayesian Inference

In Bayesian inference for non-precise data, besides the imprecision of data there is also imprecision of the a-priori distribution. So ▶Bayes' theorem is generalized in order to take care of this. The result of this generalized Bayes' theorem is a so-called *fuzzy a-posteriori distribution* $\pi^\star\left(\cdot \mid x_1^\star, \cdots, x_n^\star\right)$ which is given by its so-called δ-*level functions* $\underline{\pi}_\delta(\cdot \mid \boldsymbol{x}^\star)$ and $\overline{\pi}_\delta(\cdot \mid \boldsymbol{x}^\star)$ respectively.

From the fuzzy a-posteriori distributions generalized Bayesian confidence regions, fuzzy highest a-posteriori density regions, and fuzzy predictive distributions can be constructed. Moreover also decision analysis can be generalized to the situation of fuzzy utilities and non-precise data.

Applications

Whenever measurements of continuous quantities have to be modeled non-precise data appear. This is the case with initial conditions for differential equations, time dependent description of quantities, as well as in statistical analysis of environmental data.

About the Author

Professor Reinhard Viertl is Past Head of the Austrian Statistical Society, 1987 and 1991. He had founded the Austrian Bayesian Society in 1981. Dr. Viertl organized an International Symposium on Statistics with Non-precise Data at Innsbruck, 1993. He is an Elected Member of the New York Academy of Science (1997). He is author or co-author of

more than 100 papers and 10 books, including *Statistical Methods for Non-Precise Data* (CRC Press, Boca Raton, Florida, 1996)

Cross References

►Bayesian Statistics
►Fuzzy Logic in Statistical Data Analysis
►Fuzzy Sets: An Introduction

References and Further Reading

Bandemer H (1993) Modelling uncertain data. Akademie Verlag, Berlin
Bandemer H (2006) Mathematics of uncertainty. Springer, Berlin
Dubois D, Lubiano M, Prade H, Gil M, Grzegorzewski P, Hryniewicz O (eds) (2008) Soft methods for handling variability and imprecision. Springer, Berlin
Gil M, Corral N, Gil P (1988) The minimum inaccuracy estimates in χ^2-tests for goodness of fit with fuzzy observations. J Stat Plan Infer 19:95–115
Kacprzyk J, Fedrizzi M (eds) (1988) Combining fuzzy imprecision with probabilistic uncertainty in decision making. Lecture notes in economics and mathematical systems, vol 310, Springer, Berlin
Näther W (1997) Linear statistical inference for random fuzzy data. Statistics 29(3):221–240
Ross T, Booker J, Parkinson W (eds) (2002) Fuzzy logic and probability applications – bridging the gap. SIAM, Philadephia, PA
Viertl R (1996) Statistical methods for non-precise data. CRC Press, Boca Raton, FL
Viertl R (2006) Univariate statistical analysis with fuzzy data. Comput Stat Data Anal 51:133–147
Viertl R (2008) Foundations of fuzzy Bayesian inference. J Uncertain Syst 2:3
Viertl R, Hareter D (2006) Beschreibung und Analyse unscharfer Information – statistische Methoden für unscharfe Daten. Springer, Wien

Statistical Methods in Epidemiology

Giovanni Filardo[1], John Adams[2], Hon Keung Tony Ng[3]
[1]Director of Epidemiology
Baylor Health Care System, Dallas, TX, USA
[2]Epidemiologist
Baylor Health Care System, Dallas, TX, USA
[3]Associate Professor
Southern Methodist University, Dallas, TX, USA

Introduction

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations and the translation of study results to control health problems at the group level. The major objectives of epidemiologic studies are to describe the extent of disease in the community, to identify risk factors (factors that influence a persons risk of acquiring a disease), to determine etiology, to evaluate both existing and new preventive and therapeutic measures (including health care delivery), and to provide the foundation for developing public policy and regulatory decisions regarding public health practice. Epidemiologic studies provide research strategies for investigating public health questions in a systematic fashion relating a given health outcome to the factors that might cause and/or prevent this outcome in human populations. Statistics informs many decisions in epidemiologic study design and statistical tools are used extensively to study the association between risk factors and health outcomes.

When analyzing data for epidemiologic research, the intent is usually to extrapolate the findings from a sample of individuals to the population of all similar individuals to draw generalizable conclusions. Despite the enormous variety of epidemiologic problems and statistical solutions, there are two basic approaches to statistical analysis: regression and non-regression methods.

Types of Epidemiologic Studies and Related Risk Measures

Epidemiologist, in conceptualizing basic types of epidemiologic studies, often group them as experimental (e.g., randomized control trials) and observational (cohort, case-control, and cross-sectional) studies. This manuscript will focus on cohort and ►case-control studies. The study design determines how risk is measured (e.g., person-time at risk, absolute risk, odds) and which probability model should be employed.

Cohort Studies

In a cohort study, a group of persons are followed over a period of time to determine if an exposure of interest is associated with an outcome of interest. The key factor identifying a cohort study is that the exposure of interest precedes the outcome of interest. Depending on the exposure, different levels of exposure are identified for each subject and the subjects are subsequently followed over a period of time to determine if they experienced the outcome of interest (usually, health-related). Cohort studies are also called prospective studies, retrospective cohort studies, follow-up studies or longitudinal studies. Among all the observational studies (which includes cohort, case-control, and cross-section studies), cohort studies are the "gold standard." However, the major limitation of cohort studies is that they may require a large number of study

participants and usually many years of follow-up (which can be expensive). Loss to follow-up is another concern for cohort studies. Disease prevalence in the population under study may also determine the practicality of conducting a cohort study. Should the prevalence of an outcome be very low, the number of subjects needed to determine if there is an association between an exposure and outcome may be prohibitive within that population.

Cohort studies may result in counts, incidence (cumulative incidence or incidence proportion), or incidence rate of the outcome of interest. Suppose each subject in a large population-based cohort study is classified as exposed or unexposed to a certain risk factor and positive (case) or negative (noncase) for some disease state. Due to the loss-to-follow-up or late entry in the study, the data are usually presented in terms of number of diseases developed per person-years at risk.

The incidence rate in the exposed group and unexposed groups are then expressed as $\pi_1 = y_1/t_1$ per person-year and $\pi_2 = y_2/t_2$ per person-year, respectively (Table 1). In this situation, the numbers of disease developed in exposed and unexposed groups are usually modeled assuming a Poisson distribution when the event is relatively rare (see, Haight 1967; Johnson et al. 2005).

If there is no loss-to-follow-up or late entry in the study (closed cohort in which all participants contribute equal follow-up time), it may be convenient to present the data in terms of proportion experiencing the outcome (i.e., cumulative incidence or incidence proportion). A 2 × 2 table of sample person-count data in a cohort study is presented in Table 2.

Let p_1 and p_2 be the probabilities denoting risks for developing cases in the population for exposed and unexposed groups, respectively. The most commonly used sample estimates for p_1 and p_2 are obtained as

$$\pi_1 = \frac{x_{11}}{n_1} \text{ and } \pi_2 = \frac{x_{12}}{n_2}.$$

Statistical Methods in Epidemiology. Table 1 Data presented in terms of person-year at risk and the number of diseases developed

	Exposed	Unexposed
Disease develops	y_1	y_2
Person-year at risk	t_1	t_2
Incidence rate	y_1/t_1	y_2/t_2

Statistical Methods in Epidemiology. Table 2 2 × 2 table of sample person-count data

	Exposed	Unexposed	Total
Cases	x_{11}	x_{12}	m_1
Noncases	x_{21}	x_{22}	m_2
Total	n_1	n_2	N

Note that p_1 and p_2 are the incidence proportion in the exposed and unexposed groups, respectively. In this situation, the probability of disease in exposed and unexposed groups are usually modeled assuming a ▶binomial distribution. Statistical estimation and related inference for incidence can be found in Lui (2004) and Sahai and Khurshid (1995).

It is oftentimes the goal in epidemiologic studies to measure the association between an exposure and an outcome. Depending upon how subjects are followed, in regard to time, different measures of risk are used. Relative risk (RR) is defined as

$$RR = \frac{\text{incidence proportion (or rate) in exposed group}}{\text{incidence proportion (or rate) in unexposed group}} = \frac{\pi_1}{\pi_2}.$$

The relative risk is a ratio, therefore, it is dimensionless and without unit. It is a measure of the strength of an association between an exposure and a disease, and is the measure used in etiologic studies. In most real-world situations, subjects enter the study at different times and they are follow for variable lengths of time. In this situation, we should consider the number of cases per the total person-time contributed and the relative rate that approximates the RR defined as

$$\text{Relative rate} = \frac{\text{incidence rate in exposed group}}{\text{incidence rate in unexposed group}} = \frac{\pi_1}{\pi_2}.$$

Note that the units for π_1 and π_2 are per person-year. As it is a ratio, it is also unitless. Another measure of risk is the attributable risk (AR) which is defined as:

$$AR = \text{incidence rate in exposed group} - \text{incidence rate in unexposed group} = \pi_1 - \pi_2.$$

In the rare event of a closed cohort study framework, π_1 and π_2 can be replaced by p_1 and p_2. Attributable risk is the magnitude of disease incidence attributable to a specific exposure. It tells us the most we can hope to accomplish in reducing the risk of disease among the exposed if we totally eliminated the exposure. In other words, AR is a measure of how much of the disease incidence is attributable to the exposure. It is useful in assessing the exposures public health importance. Attributable risk

percent (ARP) in exposed group, the percent of disease incidence attributable to a specific exposure, is also used to measure the risk of disease

$$ARP = \frac{(RR-1)}{RR} \times 100.$$

ARP tells us what percent of disease in the exposed population is due to the exposure. The statistical inference on these measures of risk is discussed extensively in the literature, see, for example, Lui (2004) and Sahai and Khurshid (1995).

Case-Control Studies

Case-control studies (*see* also ►Case-Control Studies) compare a group of persons with a disease (cases) with a group of persons without the disease (controls) with respect to history of past exposures of interest. In contrast to a cohort study where an exposure of interest is determined preceding the development of future outcome, in a case-control, the disease status is known a priori while the exposure of interest is subsequently assessed among cases and controls.

Although the underlying concept of case-control studies is different from cohort study, the data for case-control study can be summarized as in a 2 × 2 table in Table 2. We can calculate the probability that cases were exposed as

$$\Pr(\text{exposed}|\text{case}) = \frac{x_{11}}{m_1}$$

and the probability that cases were not exposed as

$$\Pr(\text{unexposed}|\text{case}) = \frac{x_{12}}{m_1}.$$

We can also calculate the odds of a case being exposed as

$$\frac{x_{11}/m_1}{x_{12}/m_1} = \frac{x_{11}}{x_{12}}$$

and the odds of a case not being exposed as x_{21}/x_{22}. In case-control studies, although risk factors might contribute to the development of the disease, we cannot distinguish between risk factors for the development of the disease and risk factors for cure or survival. A major weakness in case control studies is that they are inherently unable to discern whether the exposure of interest precedes the outcome (with few exceptions). Additionally, there is some difficulty in the selection of controls. It is often the case that selected controls are not necessarily from the source population that gave rise to the cases. Therefore, measurement of association can be problematic. We cannot measure incidence rate (or proportion) in the exposed and non-exposed groups, and therefore cannot calculate rate ratios or relative risk directly. Because direct measures of risk are not applicable here, it is necessary to describe the relationship between an exposure and outcome using odds of exposure. The odds ratio (*OR*), ratio of the odds of exposure in cases and the odds of exposure in controls, is

$$OR = \frac{x_{11}/x_{12}}{x_{21}/x_{22}} = \frac{x_{11}x_{22}}{x_{12}x_{21}}.$$

The odds ratio is the cross-product ratio in the 2 × 2 table. The odds ratio is a good approximation of the relative risk when the disease being studied occurs infrequently in the population under study (case-control studies are conducted most frequently in this situation). An $OR = 1$ indicates that there is no association between exposure and outcome. When $OR > 1$ ($OR < 1$), it indicates a positive (negative) association between the exposure and disease and the larger (smaller) the *OR*, the stronger the association. An example of the calculation and interpretation of the odds ratio is given by Bland and Altman (2000).

Note that there are other variations in case-control studies and related statistical techniques which are applicable in particular situations. For instance, McNemar's test is used in matched case-control studies. For an extensive review on major development on statistical analysis of case-control studies, one can refer to Breslow (1996).

Regression vs. Non-Regression Methods

In analyzing data from epidemiologic studies, non-regression and regression methods are often used to study the relationship between an outcome and exposure. Non-regression methods of analysis control for differences in the distribution of covariates among subjects in exposure groups of interest by stratifying, while regression methods control for covariates by including possible confounders (*see* ►Confounding and Confounder Control) of the association of interest in a regression model. In some situations, regardless of whether regression techniques are used, stratification may still be necessary.

Statistical techniques used in epidemiologic studies are determined by the study design and data type. For cohort or case-control studies dealing with proportions, non-regression statistical methods based on binomial or negative binomial distribution could be applied, depending on the sampling method used (if any). Mantel-Haenszel procedures and ►Chi-square tests are the common approaches to access the association between the disease and risk factor with or without stratification. **Logistic regression** and **generalized linear models** are other possible regression methods that can be used for observational studies (see, for example, Harrell 2001). For stud-

ies with count data, statistical methods based on Poisson distribution could be applied (Cameron and Trivedi 1998).

Study designs that employ matched pairs or one-to-one matching are often approached by methods that assume a certain uniqueness of each member of the pair. The rationale for matching resembles that of blocking in statistical design, in that each stratum formed by the matching strategy is essentially the same with respect to the factors being controlled. When matching in cohort or case-control studies, McNemar's test, Mantel-Haenszel test and conditional logistic regression are normally used for analysis.

When the outcome variable is time-to-event, non-regression statistical estimation techniques for survival curves and log-rank tests can be applied, for example, the well-known **Kaplan-Meier estimator** can be used to estimate the survival curve. Lifetime parametric or **semiparametric regression models**, such as the Weibull regression model and Cox proportional hazard model (*see* ►Hazard Regression Models), can be used to model time-to-event data while controlling for possible confounders.

Cross References

►Binomial Distribution
►Biostatistics
►Case-Control Studies
►Confounding and Confounder Control
►Geometric and Negative Binomial Distributions
►Hazard Regression Models
►Incomplete Data in Clinical and Epidemiological Studies
►Medical Statistics
►Modeling Count Data
►Poisson Regression
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Bland JM, Altman DG (2000) Statistics notes: the odds. BMJ 320:1468
Breslow NE (1996) Statistics in epidemiology: the case-control study. J Am Stat Assoc 91:14–28
Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, New York
Haight FA (1967) Handbook of the Poisson distribution. Wiley, New York
Harrell FE (2001) Regression modeling strategies. Springer, New York
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, New York
Lui KJ (2004) Statistical estimation of epidemiological risk. Wiley, New York
Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott Williams & Wilkins, Philadelphia, PA
Sahai H, Khurshid A (1995) Statistics in epidemiology: methods, techniques, and applications. CRC Press, Boca Raton, FL

Statistical Modeling of Financial Markets

Mhamed-Ali El-Aroui
Associate-Professor of Quantitative Methods
ISG de Tunis, Bardo, Tunisia

Overview

Optimal investment strategies and efficient risk management often need high-performance predictions of market evolutions. These predictions are usually provided by statistical models based on both statistical analyses of financial historical data and theoretical modeling of financial market working.

One of the pioneering works of financial market statistical modeling is the Ph.D. thesis of Bachelier (1900) who was the first to note that financial stock prices have unforecastable and apparently random variations. Bachelier introduced the Brownian process to model the price movements and to assess contingent claims in financial markets. He also introduced the random walk assumption (see ►Random Walk) according to which future stock price movements are generally unforecastable. More precisely, he assumed that the price evolves as a continuous homogeneous Markov process (see ►Markov Processes). Then, by considering the price process as a limit of random walks, he showed that this process satisfies the Chapman–Kolmogorov equation and that the Gaussian distribution with the linearly increasing variance solves this equation.

Between the 1920s and the 1960s, many economists and statisticians (Coles, Working, Kendall, Samuelson, etc.) analyzed several historical stock prices data and supported the random walk assumption.

In the 1960s, Samuelson and Fama gave both theoretical and empirical proofs of the random walk assumption. They introduced the important *efficient market hypothesis* stating that, in efficient markets, price movements should be unforecastable since they should fully incorporate the expectations and informations of all market participants.

Mandelbrot in 1963 criticized the Bachelier Gaussian assumption and stated that "*Despite the fundamental importance of the Brownian motion, (see ►Brownian Motion and Diffusions) it is now obvious that it does not account for the abundant data accumulated since 1900 by empirical economists, simply because the empirical distributions of price changes are usually too peaked to be relative to samples from Gaussian population.*" It is consensually assumed now that financial returns are generally *leptokurtic* and should be modeled by heavy tailed probability distributions. Many mathematical tools were suggested to model

this heavy tailed property: Levy process (see ►Lévy Processes), alpha-stable processes, Pareto-type distributions, Extreme value theory, long memory processes, GARCH time series, etc. Leptokurtosis and heteroskedasticity are stylized facts observed in log-returns of a large variety of financial data (security prices, stock indices, foreign exchange rates, etc.).

In the following, it will be assumed that a market economy contains N financial assets, S_{jt} and R_{jt} will denote, respectively, the daily price and log-return of the j-th asset on day t ($R_{jt} = \log(S_{jt}/S_{jt-1})$). $R_t^{(m)}$ will denote the log-return on day t of the market portfolio. It will also be assumed that there exists a single deterministic lending and borrowing risk-free rate denoted r.

Markowitz in 1952 developed the mean-variance portfolio optimization, where it is assumed that rational investors choose among risky assets purely on the basis of expected return and risk (measured as returns variance). Sharpe in 1964 presented the Capital Asset Pricing Model (CAPM) where the excess return over the risk-free rate r of each asset j is, up to noise, a linear function of the excess return of the market portfolio. In other words, for each asset j: $R_{jt} - r = \alpha_j + \beta_j(R_t^{(m)} - r) + \epsilon_{jt}$; where the noise sequence ϵ_{jt} is uncorrelated with the market portfolio return.

A third major step in the history of statistical modeling of financial markets concerns the problem of pricing derivative securities. Merton, Black, and Scholes introduced a reference paradigm for pricing and hedging derivatives on financial assets. Their paradigm, known as the *Black–Scholes formula*, is based on continuous time modeling of asset price movements. It gave an explicit formula for pricing European options and got tremendous impact on the financial engineering field. Since 1973, the Black–Scholes model was used to develop several extensions combining financial, mathematical, and algorithmic refinements.

Alternative statistical modeling approaches used time series statistical tools. Since the 1980s, time series tools are very frequently used in everyday manipulations and statistical analysis of financial data. Statistical Time series models, such as ARMA, ARIMA, ARCH, GARCH, state space models, and the important Granger cointegration concept, are often used to analyze the statistical internal structure of financial time series. These models, and especially the Engel Auto-Regressed Conditionally Heteroskedastic (ARCH) model, are well suited to the nature of financial markets, they capture time dependencies, volatility clustering, comovements, etc.

In the 1990s, the statistical modeling of financial markets data was linked to the rich literature of Extreme Value Theory (EVT). Many researchers found that EVT is well suited to model maxima and minima of financial returns. This yielded a more efficient assessment of financial market risks. New EVT-based methods were developed to estimate the *Value-at-Risk* (VaR), which is now one of the most used quantitative benchmarks for managing financial risk (recommended by the Basel international committee of banking supervision).

In the last 10 years, *copula functions* (see ►Copulas and ►Copulas: Distribution Functions and Simulation) have been used by many finance researchers to handle observed comovements between markets, risk factors, and other relevant dependent financial variables. The use of copula for modeling multivariate financial series open many challenging methodological questions to statisticians, especially concerning the estimation of copula parameters and the choice of the appropriate copula function.

It is worth noting that many works combining statistical science and market finance were rewarded by Nobel prizes in economics: Samuelson in 1970, Markowitz and Sharpe in 1990, Merton and Scholes in 1997, and Engle and Granger in 2003.

Due to space limitations, only two selected topics will be detailed in the following: Black–Scholes modeling paradigm and the contribution of Extreme Value Theory to the market risk estimation.

Black–Scholes Model

The Black–Scholes model is one of the most used option-pricing models in the trading rooms. For liquid securities, quotations could occur every 30 sec; continuous time models could therefore give good approximations to the variations of asset prices. Price evolution of a single asset is modeled here by a continuous time random process denoted $\{S_t\}_{t\in\mathbb{R}_+}$. Black and Scholes assume that the studied market has some ideal conditions: Market efficiency, no transaction costs in buying or selling the stock, the studied stock pays no dividend, and known and constant risk-free interest-rate r.

The basic modeling equation of Black, Scholes, and Merton, comes from the updating of a risky investment in a continuous time modeling: $(S_{t+dt} - S_t)/S_t = \mu\, dt + \sigma(\mathbb{B}_{t+dt} - \mathbb{B}_t)$, where μ is a constant parameter called *drift* giving the global trend of the stock price; σ is a nonnegative constant called *volatility* giving the magnitude of the price variations and $\mathbb{B}_{t+dt} - \mathbb{B}_t$ are independent increments (the independence results from the market efficiency assumption) from a Brownian motion, i.e., random centered Gaussian variables. So in Black–Scholes dynamics, the stock price $\{S_t\}_{t\in\mathbb{R}_+}$ satisfies the following stochastic differential equation :$dS_t/S_t = \mu\, dt + \sigma d\mathbb{B}_t$.

Using Itô lemma on Black–Scholes equation gives the explicit solution of the previous stochastic differential equation: $S_t = S_0 \exp\left[\left(\mu - \sigma^2/2\right)t + \sigma\mathbb{B}_t\right]$, which is a geometric Brownian motion. The model parameters μ and σ are easily estimated from data.

The Black–Scholes model is still a reference tool for pricing financial derivatives. Its simple formula makes it an everyday benchmark tool in all trading rooms. But its restrictive assumptions contradict many stylized facts recognized by all financial analysts (volatility clustering, leptokurtosis, and left asymmetry of the financial returns).

Many works have extended the Black–Scholes model: in the stochastic volatility extensions, for example, prices are modeled by the two following equations: $dS_t = S_t[\mu\, dt + \sigma_t\, dB_t]$ and $d\sigma_t = \sigma_t[\nu dt + \zeta dW_t]$, where B and W are two correlated Brownian motions having a constant correlation coefficient ρ. Both parametric and nonparametric estimators are available for the parameters μ, ν, ζ, ρ, and σ_0.

Challenging research topics now concern the problem of pricing sophisticated derivative products (American options, Asian or Bermudian options, swaptions, etc.). Longstaff and Schwartz, for example, gave an interesting pricing algorithm for American options, where they combined Monte Carlo simulations with ►least squares to estimate the conditional expected payoff of the optionholder. Monte Carlo simulation is now widely used in financial engineering; for example, Broadie and Glasserman 1996 used simulations to estimate security price derivatives within a modeling framework much more realistically than the simple Black–Scholes paradigm. Monte Carlo simulations are also used in stress testing (which identifies potential losses under simulated extreme market conditions) and in the estimation of nonlinear stochastic volatility models.

EVT and Financial Risks

The Extreme Value theory (EVT) gives interesting tools for modeling and estimating extreme financial risk (see Embrecht et al. 1997 for a general survey). One common use of EVT concerns the estimation of Value-at-Risk (an extreme quantile of the loss distribution). If at day t, $\mathrm{VaR}_t(\alpha)$ denotes the Value-at-Risk of a single asset at confidence level $1 - \alpha$ with a prediction horizon of one day, then VaR writes: $\Pr(R_{t+1} \leq -VaR_t(\alpha)|\mathcal{H}_t) = \alpha$, where R_{t+1} is the return at $t + 1$ and $\mathcal{H}_t$ denotes the σ−algebra modeling all the information available at time t. Many statistical methods were used to estimate the extreme quantile $VaR_t(\alpha)$. McNeil and Frey (2000), for example, combined ARCH and EVT to take into account volatility clustering and leptokurtosis. They used an AR(1) model for the average returns μ_t and a GARCH(1,1) with pseudo-maximum-likelihood estimation for the stochastic volatility dynamics σ_t. McNeil and Frey used the previous AR-GARCH for estimating the parameters of the model $R_t = \mu_t + \sigma_t Z_t$ where $\{Z_t\}_t$ is a strict white noise process. EVT peaks-over-threshold approach is then used on the AR-GARCH-residuals $z_1, \ldots, z_k$ in order to estimate their extreme quantiles. These estimates are plugged in the estimator of the $VaR_t(\alpha)$. The idea behind this method is the elimination of data dependence by the use of time series models and then the use of EVT tools to estimate extreme quantiles of the i.i.d. residuals.

When VaR of a multi-asset portfolio is considered, multivariate statistical tools should be used: variance–covariance, multivariate GARCH, simulation approach, Multivariate Extreme Theory, dynamic copula approach, etc. In the variance–covariance approach, for example, the portfolio returns are modeled as a linear combination of slected market factors. The copula approach gives generally more efficient portfolio VaR estimations since it improves the modeling of the dependence structure between the studied assets and the risk factors.

Conclusions

Statistical science has provided essential tools for market finance. These important contributions concern the problems of portfolio selection and performance analysis, the pricing and hedging of derivative securities, the assessment of financial risks (market risk, operational risk, credit risk), the modeling of crises contagion, etc. Many challenging research topics concern both statistics and finance: the huge amount of data (called high-frequency data) need new statistical modeling approaches. The high complexity of the new financial products and the management of portfolios with high number of assets need more tractable multivariate statistical models. New research challenges are also given by the multivariate extreme value theory where copula functions gave promising results when used to model extreme comovements of asset prices or stock indices. Copula modeling has become an increasingly popular tool in finance, especially for modeling dependency between different assets. However many statistical questions remain open: copula parameter estimations, statistical comparison of competitive copula, etc. Another use of copula functions in market finance concerns the modeling of crises contagion (see, e.g., Rodriguez 2007). Many empirical works proved that dependence structure between international markets during crises is generally nonlinear and therefore better modeled by copula functions.

Cross References

►Banking, Statistics in
►Brownian Motion and Diffusions
►Copulas
►Copulas in Finance
►Financial Return Distributions
►Heavy-Tailed Distributions
►Heteroscedastic Time Series
►Lévy Processes
►Monte Carlo Methods in Statistics
►Nonlinear Time Series Analysis
►Optimal Statistical Inference in Financial Engineering
►Portfolio Theory
►Quantitative Risk Management
►Random Walk
►Statistical Modelling in Market Research

References and Further Reading

Bachelier L (1900) Théorie de la spéculation. Ann Sci École Norm S 81(3):21–86. Available at www.numdam.org
Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654
Broadie M, Glasserman P (1996) Estimating security price derivatives using simulation. Manag Sci 42(2):269–285
Embrecht P, Kluppelberg C, Mikosch T (1997) Modeling extremal events for insurance and finance. Springer, Berlin
Engel RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation and testing. Econometrica 55(2):251–276
Longstaff FA, Schwartz ES (2001) Valuing American options by simulation: a simple least-squares approach. Rev Financ Stud 14(1):113–147
Markowitz H (1952) Portfolio selection. J Financ 7:77–91
McNeil A, Frey R (2000) Estimation of tail related risk measures for heteroscedastic financial time series: an extreme value approach. J Empirical Financ 7:271–300
Rodriguez JC (2007) Measuring financial contagion: a copula approach. J Empirical Financ 14:401–423
Sharpe W (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19:425–442

Statistical Modelling in Market Research

Anatoly Zhigljavsky
Professor, Chair in Statistics
Cardiff University, Cardiff, UK

Mathematical modelling is a key element of quantitative marketing and helps companies around the globe in making important marketing decisions about launching new products and managing existing ones. Most mathematical models used in marketing research are either purely statistical or include elements of statistical models.

An extensive discussion (by the top market research academics) of the state-of-art in the field of marketing modelling and its prospects for the future is contained in Steemkamp (2000), a special issue of the International Journal of Research in Marketing. One can consult Steemkamp (2000) for many references related to the subject; see also recent books (Wierenga 2008; Wittink et al. 2000; Mort 2001; Zikmund and Babin 2009).

We look at the field of market modelling from a viewpoint of a professional statistician with twenty years of experience on designing and using statistical models in market research. We start with distinguishing the following types of statistical models used in market research:

1. Direct simulation models
2. Standard statistical models
3. Models of consumer purchase behaviour
4. Dynamic models for modelling competition, pricing and advertising strategies
5. Statistical components of inventory and other management science models

Let us briefly consider these types of models separately.

1. Direct *simulation models*. These are specialized models based on attempts to directly imitate the market (e.g., via the behaviour of individual customers) using a synergy of stochastic and deterministic rules. These models were popular 20–30 years ago but are less popular now. The reasons are the lack of predictive power, huge number of parameters in the models and impossibility of their validation.

2. *Standard statistical models*. All standard statistical models and methods can be used in market research, see Mort (2001); Zikmund and Babin (2009); Rossi et al. (2005); Hanssens et al. (2003). Most commonly, the following statistical models are used:

- Various types of regression
- ARIMA and other time series models
- Bayesian models
- Models and methods of multivariate statistics; especially, structural equation and multinomial response models, conjoint, factor, and principal component analyses

3. *Models of consumer purchase behaviour*. Several types of statistical models are used for modelling consumer purchase behaviour including brand choice. The following three basic models (and some of their extensions) have

proved to be the most useful: Mixed ►Poisson processes, the Dirichlet model, and Markovian models.

The mixed Poisson process model assumes that a customer makes his/her purchase according to a Poisson process with some intensity λ where λ is random across the population. In the most popular model, called Gamma-Poisson, λ has Gamma distribution (with two unknown parameters); this yields that the number of purchases for a given period is the Negative Binomial Distribution. Typical questions, which the Poisson process model answers, is the forecasting of the behaviour of the market research measures (like penetration, purchase frequency and repeat buying measures) in the form of the so-called growth curves. Extensions of the mixed Poisson models cover the issues like the zero-buyer problem (some zero-buyers do have a positive propensity to buy but some other don't), seasonality of the market and the panel flow-through.

The Dirichlet model is a brand-choice model. It assumes that customers make their brand choice independently with certain propensities; these propensities are different for all customers and are independent realizations from the Dirichlet distribution which parameters are determined by the market shares of the brands. In Markovian brand-choice models, the propensity to buy a given brand for a random customer may vary depending on either the previous purchase or other market variables. These models are more complicated than the mixed Poisson process and Dirichlet models but in some circumstances are easily applicable and sometimes are able to accurately describe some features of the market.

Of course, the models above are unrealistic on the individual level (e.g., few people have the Poisson process pattern as their purchase sequence). However, these models (and especially the mixed Poisson model) often fit data extremely accurately on the aggregated level (when the time period considered and the number of customers are sufficiently large). These models can be classified as descriptive (rather than "prescriptive") and help in explaining different aspects of market research dynamics and some phenomena related to the brand-choice.

4. *Dynamic models for modelling competition, pricing and advertising strategies.* There is extensive literature on this subject, see, e.g., Erickson (1991). The majority of the models are so-called differential games or simpler models still written in terms of differential equations. The models are deterministic and the statistical aspect only arrives through the assumption that the data contain random errors. Statistical modelling part is therefore negligible in these models. Alternatively, in some Markovian brand-choice models mentioned above, there is an option of including the market variables (e.g., promotion) into the updating rule for the buying propensities. These models are proper stochastic models but they are often too complicated (have too many parameters) and therefore difficult to validate.

5. *Statistical components of inventory and other management science models.* Inventory and other management science models applied in market research are typically standard models of Operations Research, see Ingene and Parry (2004) for a recent review of these models. Despite these models often have a large stochastic component, they do not represent anything special from the statistics view-point.

Statistical models are used for the following purposes: (a) forecasting the market behaviour of a new brand to prepare its launch and (b) managing existing brands. In case (a), the models are usually based solely on standard statistical models, type 2 above. Sometimes, other types of models (especially, large simulation models, type 1) are used too. A lot of specific market research data are often collected to feed these models. These data includes market surveys, various types of questionnaires and focus group research in direct contact with customers. All available market data, for example economic trends and specific industry sector reports, is used too. In case (b), the models are used for making decisions about pricing, promotion and advertising strategies, production and inventory management etc. All available statistical models and methods are used to help managers to make their decisions.

While reading academic papers and books on marketing research, one can get an impression that mathematical and statistical modelling in marketing is a mature subject with many models developed and used constantly for helping market research managers in working out their decisions. Indeed, there are many models available (some of them are quite sophisticated). However, only a small number of them are really used in practice: the majority of practical models can be reduced either to a simple regression or sometimes to another standard model among those mentioned above. One of the reasons for this gloomy observation is the fact that managers rarely want a description of the market. Instead, they want 'a prescription'; that is, a number (with a hope that no confidence interval is attached to this number) which would lead them to a right decision. Another reason is the fact that only a very few models used in market research satisfy the following natural requirements for a good statistical model: (a) simplicity, (b) robustness to the deviations from the model assumptions, (c) clear range of applicability, and (d) empirical character, which means that the models have to be built with the data (and data analysis) in view and with the purpose of explaining/fitting/forecasting relevant data.

Despite huge amounts of market data is available to analysts, these data are typically messy, not reliable, badly structured and become outdated very quickly. Development of reliable statistical models dealing with such data is hard. The progress in understanding all these issues and tackling them by means of the development of appropriate models and making them correctly applicable is visible but it is justifiably slow.

Cross References

►Box–Jenkins Time Series Models
►Gamma Distribution
►Model Selection
►Multivariate Statistical Distributions
►Poisson Processes
►Statistical Modeling of Financial Markets

References and Further Reading

Erickson GM (1991) Dynamic models of advertising competition: open- and closed-loop extensions. Springer
Hanssens DM, Parsons LJ, Schultz RL (2003) Market response models: econometric and time series analysis. Springer, Berlin
Ingene CA, Parry ME (2004) Mathematical models of distribution channels. Springer, New York
Mort D (2001) Understanding statistics and market research data. Europa publications, London
Rossi PE, Allenby GM, McCullog R (2005) Bayesian statistics and marketing. Wiley/Blackwell, New York
Steenkamp, J-BEM (ed) (2000) Marketing modeling on the threshold of the 21st century. Int J Res Mark 17(2–3):99–253
Wierenga B (ed) (2008) Handbook of marketing decision models. Springer, New York
Wittink DR, Leeflang PSH, Wedel M, Naert PA (2000) Building models for marketing decisions. Kluwer Academic, Boston, MA
Zikmund WG, Babin BJ (2009) Exploring marketing research. South-western Educational Publishing, Florence, KY

Statistical Natural Language Processing

Florentina T. Hristea
Associate Professor, Faculty of Mathematics and Computer Science
University of Bucharest, Bucharest, Romania

Natural language processing (*NLP*) is a field of artificial intelligence concerned with the interactions between computers and human (natural) languages. It refers to a technology that creates and implements ways of executing various tasks concerning natural language (such as designing natural language based interfaces with databases, machine translation, etc.). NLP applications belong to three main categories:

1. Text-based applications (such as knowledge acquisition, information retrieval, information extraction, text summarization, machine translation, etc.)
2. Dialog-based applications (such as learning systems, question answering systems, etc.)
3. Speech processing (although NLP may refer to both text and speech, work on speech processing has gradually evolved into a separate field)

Natural language engineering deals with the implementation of large-scale natural language–based systems. It refers to the related field of *Human Language Technology* (*HLT*).

NLP represents a difficult and largely unsolved task. This is mainly due to the interdisciplinary nature of the problem that requires interaction between many sciences and fields: linguistics, psycholinguistics, computational linguistics, philosophy, statistics, computer science in general, and artificial intelligence in particular.

Statistical NLP has been the most widely used term to refer to nonsymbolic and nonlogical work on NLP over the past decade. Statistical NLP comprises all *quantitative approaches* to automated language processing, including probabilistic modeling, information theory, and linear algebra (Manning and Schütze 1999).

As computational problems, many problems posed by NLP (such as WSD – word sense disambiguation) were often described as AI-complete, that is, problems whose solutions presuppose a solution to complete natural language understanding or common-sense reasoning. This view originated from the fact that possible statistical approaches to such problems were almost completely ignored in the past. As it is well known, starting with the early 1990s, the artificial intelligence community witnessed a great revival of empirical methods, especially statistical ones. This is due to the success of statistical approaches, as well as of machine learning, in solving problems such as speech recognition or part-of-speech tagging. It was mainly research into speech recognition that inspired the revival of statistical methods within NLP, and many of the techniques used nowadays were developed first for speech and then spread over into NLP (Manning and Schütze 1999). Nowadays statistical methods and machine learning algorithms are used for solving a great number of problems posed by artificial intelligence in general and by NLP in particular. Furthermore, the availability of large

text corpora has changed the scientific approach to language in linguistics and cognitive science, with language and cognition being viewed as probabilistic phenomena.

From the point of view of NLP, the two main components of statistics are:

1. *Descriptive statistics*: methods for summarizing (large) datasets
2. *Inferential statistics*: methods for drawing inferences from (large) datasets

The use of statistics in NLP falls mainly into three categories (Nivre 2002):

1. *Processing*: We may use probabilistic models or algorithms to process natural language input or output.
2. *Learning*: We may use inferential statistics to learn from examples (corpus data). In particular, we may estimate the parameters of probabilistic models that can be used in processing.
3. *Evaluation*: We may use statistics to assess the performance of language processing systems.

As pointed out in Manning and Schütze (1999), "complex probabilistic models can be as explanatory as complex non-probabilistic models – but with the added advantage that they can explain phenomena that involve the type of *uncertainty* and *incompleteness* that is so pervasive in cognition in general and in language in particular."

A practical NLP system must be good at making *disambiguation decisions* of word sense, word category, syntactic structure, and semantic scope. One could say that disambiguation abilities, together with robustness, represent the two main hallmarks of statistical natural language processing models. Again as underlined in Manning and Schütze (1999), "a statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from corpora... The use of statistical models offers a good solution to the ambiguity problem: statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data. Thus statistical NLP methods have led the way in providing successful disambiguation in large scale systems using naturally occurring text. Moreover, the parameters of Statistical NLP models can often be estimated automatically from text corpora, and this possibility of automatic learning not only reduces the human effort in producing NLP systems, but raises interesting scientific issues regarding human language acquisition."

Cross References

►Data Mining
►Distance Measures
►Estimation
►Expert Systems
►Information Theory and Statistics
►Statistical Inference

References and Further Reading

Manning C, Schütze H (1999) Foundations of statistical natural language processing. The MIT Press, Cambridge, MA

Nivre J (2002) On statistical methods in natural language processing. In: Berbenko J, Wangler B (eds) Promote IT. Second Conference for the Promotion of Research in IT at New Universities and University Colleges in Sweden. University of Skovde, Billingehus, Skövde, pp 684–694

Statistical Pattern Recognition Principles

Nicholas A. Nechval[1], Konstantin N. Nechval[2], Maris Purgailis[3]
[1]Professor, Head of the Mathematical Statistics Department
University of Latvia, Riga, Latvia
[2]Assistant Professor
Transport and Telecommunication Institute, Riga, Latvia
[3]Professor, Dean of the Faculty of Economics and Management
University of Latvia , Riga, Latvia

Problem Description

Mathematically, pattern recognition is a classification problem. Consider the recognition of characters. We wish to design a system such that a handwritten symbol will be recognized as an "A," a "B," etc. In other words, the machine we design must classify the observed handwritten character into one of 26 classes. The handwritten characters are often ambiguous, and there will be misclassified characters. The major goal in designing a pattern recognition machine is to have a low probability of misclassification.

There are many problems that can be formulated as pattern classification problems. For example, the weather may be divided into three classes, fair, rain, and possible rain, and the problem is to classify tomorrow's weather into one of these three classes. In the recognition of electrocardiograms, the classes are disease categories plus the class of normal subjects. In binary data transmission, a "one" and a "zero" are represented by signals of amplitudes A_1 and A_0, respectively. The signals are distorted or corrupted by noise when transmitted over communication channels, and the

receiver must classify the received signal into "ones" and "zeros." Hence, many of the ideas and principles in pattern recognition may be applied to the design of communication systems and vice versa (Nechval 1997; Nechval and Nechval 1999).

Pattern recognition theory deals with the mathematical aspects common to all pattern recognition problems. Application of the theory to a specific problem, however, requires a thorough understanding of the problem, including its peculiarities and special difficulties (Bishop 2006).

The input to a pattern recognition machine is a set of p measurements, and the output is the classification. It is convenient to represent the input by a p-dimensional vector $\mathbf{x}$, called a *pattern vector*, with its components being the p measurements. The classification at the output depends on the input vector $\mathbf{x}$, hence we write

$$C = d(x). \tag{1}$$

In other words, the machine must make a decision as to the class to which $\mathbf{x}$ belongs, and $d(\mathbf{x})$ is called a *decision function*.

A pattern recognition machine may be divided into two parts, a feature extractor and a classifier. The classifier performs the classification, while the feature extractor reduces the dimensionality of input vectors to the classifier. Thus, feature extraction is a linear or nonlinear transformation

$$y = Y(x), \tag{2}$$

which transforms a pattern vector $\mathbf{x}$ (in the pattern space Ω_x) into a *feature vector* $\mathbf{y}$ (in a *feature space* Ω_y). The classifier then classifies $\mathbf{x}$ based on $\mathbf{y}$. Since Ω_y is of lower dimensionality than Ω_x, the transformation is singular and some information is lost. The feature extractor should reduce the dimensionality but at the same time maintain a high level of machine performance. A special case of feature extraction is feature selection, which selects as features a subset of the given measurements.

The division of a pattern recognition machine into feature extractor and classifier is done out of convenience rather than necessity. It is conceivable that the two could be designed in an unified manner using a single performance criterion. When the structure of the machine is very complex and the dimensionality p of the pattern space is high, it is more convenient to design the feature extractor and the classifier separately.

The problem of pattern classification may be discussed in the framework of hypothesis testing. Let us consider a simple example. Suppose that we wish to predict a student's success or failure in graduate study based on his GRE (Graduate Record Examination) score. We have two hypotheses – the null hypothesis H_0, that he or she will be successful, and the alternative hypothesis H_1, that he or she will fail. Let x be the GRE score, $f_0(x)$ be the conditional probability density of x, given that the student will be successful, and $f_1(x)$ be the conditional density of x, given that he or she will fail. The density functions $f_0(x)$ and $f_1(x)$ are assumed known from our past experience on this problem. This is a hypothesis testing problem and an obvious decision rule is to retain H_0 and reject H_1 if x is greater than a certain threshold value h, and accept H_1 and reject H_0 if $x \leq h$. A typical example of multiple hypothesis testing is the recognition of English alphabets where we have 26 hypotheses.

Illustrative Examples

Applicant Recognition for Project Realization with Good Contract Risk

One of the most important activities that an employer has to perform is recognition of applicant for realization of project with good contract risk. The employer is defined as a firm or an institution or an individual who is investing in a development. The above problem is a typical example of a pattern classification problem. An applicant for contract can be represented by a random $p \times 1$ vector $\mathbf{X} = (X_1, \ldots, X_p)'$ of features or characteristics. We call this $p \times 1$ vector the applicant's pattern vector. Using historical data and the applicant's pattern vector, a decision-maker must decide whether to accept or reject the contract request. The historical data are summarized in a collection of pattern vectors. There are pattern vectors of former applicants who received contract and proved to be good risks, and there are patterns of former applicants who were accepted and proved to be poor risks. The historical data should include the pattern vectors and eventual contract status of applicants who were rejected. The eventual contract status of rejected applicants is difficult to determine objectively, but without this information, the historical data will contain the basis of former decision rules. The historical data consist of the pattern vectors and eventual contract status of n applicants; $n = n_1 + n_2$: n_1 of the n applicants proved to be good contract risks, and n_2 proved to be poor contract risks. Given this situation and a new applicant's pattern vector, the decision-maker deals with the problem of how to form his or her decision rule in order to accept or reject new applicants. In this entry, we shall restrict attention to the case when $p(\mathbf{X}; H_i)$, $i = 1, 2$, are multivariate normal with unknown parameters. All statistical information is contained in the historical data. In this case, the procedure based on a generalized likelihood ratio test is proposed. This procedure is relatively simple to carry out and can be

recommended in those situations when we deal with small samples of the historical data (Nechval and Nechval 1998).

Generalized Likelihood Ratio Test for Applicant Recognition. Let $\mathbf{X}$ be a random $p\times 1$ vector that is distributed in the population Π_i $(i = 0,1,2)$ according to the p-variate nonsingular normal distribution $N(\mathbf{a}_i, \mathbf{Q}_i)$ $(i = 0,1,2)$. Let $\mathbf{x}_0$ be an observation on $\mathbf{X}$ in Π_0. The n_i independent observations from Π_i will be denoted by $\{\mathbf{x}_{ij}, j = 1,2,\ldots,n_i\}$ distributed with the density $p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)$ for $i = 1,2$ and the density of the unidentified observation $\mathbf{x}_0$ will be taken as $p(\mathbf{x}_0;\mathbf{a}_0,\mathbf{Q}_0)$. The $\mathbf{a}_i$s and $\mathbf{Q}_i$s are unknown and it is assumed that either $(\mathbf{a}_0,\mathbf{Q}_0) = (\mathbf{a}_1,\mathbf{Q}_1)$, or $(\mathbf{a}_0,\mathbf{Q}_0) = (\mathbf{a}_2,\mathbf{Q}_2)$, and $\mathbf{a}_1 \neq \mathbf{a}_2$, $\mathbf{Q}_1 \neq \mathbf{Q}_2$. Assume for the moment that there are prior odds of $\xi/(1-\xi)$ in favor of type 1 for $\mathbf{x}_0$. Then the likelihood ratio statistic for testing the null hypothesis $H_1 : (\mathbf{a}_0 = \mathbf{a}_1,\ \mathbf{Q}_0 = \mathbf{Q}_1)$ versus the alternative hypothesis $H_2 : (\mathbf{a}_0 = \mathbf{a}_2,\ \mathbf{Q}_0 = \mathbf{Q}_2)$ is given by

$$LR = \frac{\xi \max\limits_{H_1} p(\mathbf{x}_0;\mathbf{a}_1,\mathbf{Q}_1)\prod\limits_{i=1}^{2}\prod\limits_{j=1}^{n_i} p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)}{(1-\xi)\max\limits_{H_2} p(\mathbf{x}_0;\mathbf{a}_2,\mathbf{Q}_2)\prod\limits_{i=1}^{2}\prod\limits_{j=1}^{n_i} p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)}, \quad (3)$$

where

$$p(\mathbf{x}_0;\mathbf{a}_0,\mathbf{Q}_0) = (2\pi)^{-P/2}|\mathbf{Q}_0|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_0-\mathbf{a}_0)'\mathbf{Q}_0^{-1}(\mathbf{x}_0-\mathbf{a}_0)\right\}, \quad (4)$$

$$p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i) = (2\pi)^{-P/2}|\mathbf{Q}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_{ij}-\mathbf{a}_i)'\mathbf{Q}_i^{-1}(\mathbf{x}_{ij}-\mathbf{a}_i)\right\}. \quad (5)$$

The maximum likelihood estimators of the unknown parameters under H_1 are

$$\widehat{\mathbf{a}}_1 = \frac{n_1\bar{\mathbf{x}}_1 + \mathbf{x}_0}{n_1+1}, \quad (6)$$

$$\widehat{\mathbf{a}}_2 = \bar{\mathbf{x}}_2, \quad (7)$$

$$\widehat{\mathbf{Q}}_1 = \frac{1}{n_1+1}\left[(n_1-1)\mathbf{S}_1 + \frac{n_1}{n_1+1}(\mathbf{x}_0-\bar{\mathbf{x}}_1)(\mathbf{x}_0-\bar{\mathbf{x}}_1)'\right], \quad (8)$$

$$\widehat{\mathbf{Q}}_2 = \frac{n_2-1}{n_2}\mathbf{S}_2, \quad (9)$$

where

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i}\mathbf{x}_{ij}/n_i, \quad (10)$$

$$S_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'/(n_i-1), \quad i = 1,2, \quad (11)$$

with obvious changes for the corresponding estimators under H_2. Substitution of the estimators in (3) gives, after some simplification,

$$LR = \left[\frac{(n_1+1)(n_2-1)}{(n_2+1)(n_1-1)}\right]^{p/2} = \left[\frac{(n_2/(n_2+1))^{pn_2/2}}{(n_1/(n_1+1))^{pn_1/2}}\left(\frac{|S_2|}{|S_1|}\right)^{1/2} \times \frac{(1+n_2v_2(x_0)/(n_2^2-1))^{(n_2+1)/2}}{(1+n_1v_1(x_0)/(n_1^2-1))^{(n_1+1)/2}}\right]\left(\frac{\xi}{1-\xi}\right), \quad (12)$$

where

$$v_i(\mathbf{x}_0) = (\mathbf{x}_0-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_0-\bar{\mathbf{x}}_i), \quad i = 1,2. \quad (13)$$

For $\mathbf{Q}_1 = \mathbf{Q}_2$, the likelihood ratio statistic simplifies to

$$\mathrm{LR} = \left[\frac{1+\dfrac{n_2v_2(x_0)}{(n_2+1)(n_1+n_2-2)}}{1+\dfrac{n_1v_1(x_0)}{(n_1+1)(n_1+n_2-2)}}\right]^{(n_1+n_2+1)/2}\left(\frac{\xi}{1-\xi}\right), \quad (14)$$

and hypothesis H_1 or H_2 is favoured according to whether LR is greater or less than 1, that is,

$$LR\begin{cases} > 1, & \text{then} \quad H_1 \\ \leq 1, & \text{then} \quad H_2 \end{cases}. \quad (15)$$

Signal Detection in Clutter

The problem of detecting the unknown deterministic signal $\mathbf{s}$ in the presence of a clutter process, which is incompletely specified, can be viewed as a binary hypothesis-testing problem (Nechval 1992; Nechval et al. 2004). The decision is based on a sample of observation vectors $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})'$, $i = 1(1)n$, each of which is composed of clutter $\mathbf{w}_i = (w_{i1},\ldots,w_{ip})'$ under the null hypothesis H_0 and a signal $\mathbf{s} = (s_1,\ldots,s_p)'$ added to clutter $\mathbf{w}_i$ under the alternative H_1, where $n > p$. The two hypotheses that the detector must distinguish are given by

$$H_0 : \mathbf{X} = \mathbf{W} \quad \text{(clutter alone)}, \quad (16)$$

$$H_1 : \mathbf{X} = \mathbf{W} + \mathbf{c}\mathbf{s}' \quad \text{(signal present)}, \quad (17)$$

where

$$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)', \quad (18)$$

$$\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_n)', \quad (19)$$

are $n\times p$ random matrices, and

$$\mathbf{c} = (1,\ldots,1)' \quad (20)$$

is a column vector of n units. It is assumed that $\mathbf{w}_i$, $i = 1(1)n$, are independent and normally distributed with common mean 0 and covariance matrix (positive definite) $\mathbf{Q}$, i.e.,

$$\mathbf{w}_i \sim N_p(0,\mathbf{Q}), \quad \forall i = 1(1)n. \quad (21)$$

Thus, for fixed n, the problem is to construct a test, which consists of testing the null hypothesis

$$H_0 : \mathbf{x}_i \sim N_p(0, \mathbf{Q}), \quad \forall i = 1(1)n, \tag{22}$$

versus the alternative

$$H_1 : \mathbf{x}_i \sim N_p(\mathbf{s}, \mathbf{Q}), \quad \forall i = 1(1)n, \tag{23}$$

where the parameters $\mathbf{Q}$ and $\mathbf{s}$ are unknown.

One of the possible statistics for testing H_0 versus H_1 is given by the generalized maximum likelihood ratio (GMLR)

$$GMLR = \max_{\theta \in \Theta_1} L_{H_1}(\mathbf{X}; \boldsymbol{\theta}) \Big/ \max_{\theta \in \Theta_0} L_{H_0}(\mathbf{X}; \boldsymbol{\theta}), \tag{24}$$

where $\boldsymbol{\theta} = (\mathbf{s}, \mathbf{Q}), \Theta_0 = \{(\mathbf{s}, \mathbf{Q}) : \mathbf{s} = 0, \mathbf{Q} \in Q_p\}$, $\Theta_1 = \Theta - \Theta_0$, $\Theta = \{(\mathbf{s}, \mathbf{Q}) : \mathbf{s} \in \mathbb{R}^p, \mathbf{Q} \in Q_p\}$, Q_p denotes the set of $p \times p$ positive definite matrices. Under H_0, the joint likelihood for $\mathbf{X}$ based on (22) is

$$L_{H_0}(\mathbf{X}; \boldsymbol{\theta}) = (2\pi)^{-np/2} |\mathbf{Q}|^{-n/2} \exp\left(- \sum_{i=1}^{n} \mathbf{x}_i' \mathbf{Q}^{-1} \mathbf{x}_i / 2 \right). \tag{25}$$

Under H_1, the joint likelihood for $\mathbf{X}$ based on (23) is

$$L_{H_1}(\mathbf{X}; \boldsymbol{\theta}) = (2\pi)^{-np/2} |\mathbf{Q}|^{-n/2} \exp\left(- \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{s})' Q^{-1} (\mathbf{x}_i - \mathbf{s}) / 2 \right). \tag{26}$$

It can be shown that

$$GMLR = \left| \widehat{\mathbf{Q}}_0 \right|^{n/2} \left| \widehat{\mathbf{Q}}_1 \right|^{-n/2}, \tag{27}$$

and

$$\widehat{\mathbf{Q}}_0 = \mathbf{X}'\mathbf{X}/n, \tag{28}$$

$$\widehat{\mathbf{Q}}_1 = (\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')(\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')'/n, \tag{29}$$

and

$$\hat{\mathbf{s}} = \mathbf{X}'\mathbf{c}/n \tag{30}$$

are the well-known maximum likelihood estimators of the unknown parameters $\mathbf{Q}$ and $\mathbf{s}$ under the hypotheses H_0 and H_1, respectively. It can be shown, after some algebra, that (27) is equivalent finally to the statistic

$$y = \mathbf{T}_1' \mathbf{T}_2^{-1} \mathbf{T}_1 / n, \tag{31}$$

where $\mathbf{T}_1 = \mathbf{X}'\mathbf{c}$, $\mathbf{T}_2 = \mathbf{X}'\mathbf{X}$. It is known that $(\mathbf{T}_1, \mathbf{T}_2)$ is a complete sufficient statistic for the parameter $\boldsymbol{\theta} = (\mathbf{s}, \mathbf{Q})$. Thus, the problem has been reduced to consideration of the sufficient statistic $(\mathbf{T}_1, \mathbf{T}_2)$. It can be shown that under H_0, the result (31) is a $\mathbf{Q}$-free statistic y, which has the property that its distribution does not depend on the actual covariance matrix $\mathbf{Q}$. It is clear that the statistic y is equivalent to the statistic

$$v = [(n-p)/p]\, y/(1-y) = [n(n-p)/p] \left(\hat{\mathbf{s}}' \left[\widehat{\mathbf{G}}_1 \right]^{-1} \hat{\mathbf{s}} \right), \tag{32}$$

where

$$\widehat{\mathbf{G}}_1 = n\widehat{\mathbf{Q}}_1 = (\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')(\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')' = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{s}})(\mathbf{x}_i - \hat{\mathbf{s}})'. \tag{33}$$

Under H_1, the statistic v is subject to a noncentral F-distribution with p and $n - p$ degrees of freedom, the probability density function of which is (Nechval 1992; Nechval et al. 2004)

$$f_{H_1}(v; n, q) = \left[B\left(\frac{p}{2}, \frac{n-p}{2} \right) \right]^{-1} \frac{\left(\frac{p}{n-p} \right)^{p/2} v^{p/2-1}}{\left(1 + \frac{p}{n-p} v \right)^{n/2}} \times e^{-q/2} {}_1F_1\left(\frac{n}{2}; \frac{p}{2}; \frac{q}{2} \left(\frac{p}{n-p} v \left(1 + \frac{p}{n-p} v \right)^{-1} \right) \right), \quad 0 < v < \infty, \tag{34}$$

where ${}_1F_1(a; b; x)$ is the confluent hypergeometric function (Abramowitz and Stegun 1964),

$$q = n\left(\mathbf{s}' \mathbf{Q}^{-1} \mathbf{s} \right) \tag{35}$$

is a noncentrality parameter representing the generalized signal-to-noise ratio (GSNR). Under H_0, when $q = 0$, (34) reduces to a standard F-distribution with p and $n - p$ degrees of freedom,

$$f_{H_0}(v; n) = \left[B\left(\frac{p}{2}, \frac{n-p}{2} \right) \right]^{-1} \frac{\left(\frac{p}{n-p} \right)^{p/2} v^{p/2-1}}{\left(1 + \frac{p}{n-p} v \right)^{n/2}}, \quad 0 < v < \infty. \tag{36}$$

The test of H_0 versus H_1, based on the GMLR statistic v, is given by

$$v \begin{cases} > h, & \text{then} \quad H_1 \quad \text{(signal present)}, \\ \le h, & \text{then} \quad H_0 \quad \text{(clutter alone)}, \end{cases} \tag{37}$$

and can be written in the form of a decision rule $u(v)$ over $\{v : v \in (0, \infty)\}$,

$$u(v) = \begin{cases} 1, & v > h \quad (H_1), \\ 0, & v \le h \quad (H_0), \end{cases} \tag{38}$$

where $h > 0$ is a threshold of the test that is uniquely determined for a prescribed level of significance so that

$$\sup_{\theta \in \Theta_0} E_\theta \{u(v)\} = \alpha. \tag{39}$$

For fixed n, in terms of the probability density function (36), tables of the central F-distribution permit one to choose h to achieve the desired test size (false alarm probability P_{FA}),

$$P_{FA} = \alpha = \int_h^\infty f_{H_0}(v; n) dv. \tag{40}$$

Furthermore, once h is chosen, tables of the noncentral F-distribution permit one to evaluate, in terms of the probability density function (34), the power (detection probability P_D) of the test,

$$P_D = \gamma = \int_h^\infty f_{H_1}(v; n, q) dv. \tag{41}$$

The probability of a miss is given by

$$\beta = 1 - \gamma. \tag{42}$$

It follows from (36) and (40) that the GMLR test is invariant to intensity changes in the clutter background and achieves a fixed probability of a false alarm, that is, the resulting analyses indicate that the test has the property of a constant false alarm rate (CFAR). Also, no learning process is necessary in order to achieve the CFAR. Thus, operating in accordance to the local clutter situation, the test is adaptive.

About the Authors

Dr. Nicholas A. Nechval is a Professor and Head, Department of Mathematical Statistics, EVF Research Institute, University of Latvia, Riga, Latvia. He is also a Principal Investigator in the Institute of Mathematics and Computer Science at the University of Latvia. Dr. Nechval was a Professor of Mathematics and Computer Science and the Head of the Research Laboratory at the Riga Aviation University (1993–1999). In 1992, Dr. Nechval was awarded a Silver Medal of the Exhibition Committee (Moscow, Russia) for his research on the problem of Prevention of Collisions between Aircraft and Birds. He is a Member of the Russian Academy of Science. Professor Nechval has authored and coauthored more than 350 papers and 9 books, including the book *Aircraft Protection from Birds* (Moscow: Russian Academy of Science, 2007) coauthored with V.D. Illyichev (Academician of the Russian Academy of Science), and the book *Improved Decisions in Statistics* (Riga: SIA "Izglitibas soli", 2004) coauthored with E.K. Vasermanis. This book was awarded the "2004 Best Publication Award" by the Baltic Operations Research Society. Dr. Nechval is also an Associate editor of the following international journals: *Scientific Inquiry* (2005–), *An International Journal of Computing Anticipatory Systems* (2002–), et al.

Dr. Konstantin N. Nechval is an Assistant Professor, Applied Mathematics Department, Transport and Telecommunication Institute, Riga, Latvia. He has authored and co-authored more than 50 papers. Dr. Konstantin N. Nechval was awarded the "CASYS'07 Best Paper Award" for his paper: "Dual Control of Education Process" presented at the Eight International Conference on Computing Anticipatory Systems (Liege, Belgium, August 6-11, 2007) and the "MM2009 Best Paper Award" for his paper: "Optimal Statistical Decisions in a New Product Lifetime Testing" presented at the Fourth International Conference on Maintenance and Facility Management (Rome, Italy, April 22-24, 2009).

Dr. Maris Purgailis is a Professor and Dean, Faculty of Economics and Management, University of Latvia, Riga, Latvia. Professor Purgailis has authored and co-authored more than 120 papers and 6 books.

Cross References

►Data Analysis
►Fuzzy Sets: An Introduction
►Pattern Recognition, Aspects of
►Statistical Signal Processing

References and Further Reading

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions. National Bureau of Standards, New York

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Nechval NA (1992) Radar CFAR thresholding in clutter under detection of airborne birds. In: Proceedings of the 21st meeting of bird strike committee Europe. BSCE, Jerusalem, pp 127–140

Nechval NA (1997) Adaptive CFAR tests for detection of a signal in noise and deflection criterion. In: Wysocki T, Razavi H, Honary B (eds) Digital signal processing for communication systems. Kluwer, Boston, pp 177–186

Nechval NA, Nechval KN (1998) Recognition of applicant for project realization with good contract risk. In: Pranevicius H, Rapp B (eds) Organisational structures, management, simulation of business sectors and systems. Kaunas University of Technology, Lithuania, pp 70–72

Nechval NA, Nechval KN (1999) CFAR test for moving window detection of a signal in noise. In: Proceedings of the 5th international symposium on DSP for communication systems, Curtin University of Technology, Perth-Scarborough, pp 134–141

Nechval NA, Nechval KN, Srelchonok VF, Vasermanis EK (2004) Adaptive CFAR tests for detection and recognition of targets signals in radar clutter. In: Berger-Vachon C, Gil Lafuente AM (eds) The 2004 conferences best of. AMSE Periodicals, Barcelona, pp 62–80

Statistical Publications, History of

Vassiliy Simchera
Director of Rosstat's Statistical Research Institute
Moscow, Russia

Statistical publications are editions that contain summarized numerical data about socio-economic phenomena, usually presented in the form of statistical tables, charts, diagrams, graphs, etc. These statistical publications are an inseparable part of common numerical information concerning the state and development of healthcare, education, science, and culture provided by the statistical authorities.

Depending on the common purpose, one may distinguish their various types. These include *Statistical Yearbook, Annual Statistics; Statistical Abstract; Manual Guide, Handbook of Statistics*; overview of census, and other major surveys. By order of coverage, statistical publications can be common (*National Accounts*), industrial (*Industrial Indicators*), or may deal with other activities of an economy (for example, *Financial Statistics*). By the level of details they can be complete (*Yearbooks, Almanacs*, etc.) or short (*Pocket Book and Statistisches Handbook* are the most common types).

There are also differences by the domain of coverage among one or another statistical publication: an entire country, an administrative territorial part of the country (for example, state, region, land, county, etc.); in international statistical publications, this could be several countries, an entire continent, or the whole world (for example, *UN Statistical Publications*).

The outcomes of large surveys are presented in nonrecurrent statistical publications; among the recurrent statistical publications, the most significant are periodical statistical publications (published annually, quarterly, or monthly), the least significant are non-periodical statistical publications (containing demographic figures, birth and death rates, marriage status, etc.).

The statistical publications cover current and previous years (retrospective statistical publications) with the scope of decades and centuries (*Historical Statistics of the US from* 1789, *Colonial Times to* 1957, 1960, 1975, *and* 2008; *USSR's Economics* 60 *years*, 1987; *Russia*: 100 *Years of Economic Growth* 1900–2008 *Historical Series; Annuaire Statistique de la France, vols*. 1–106, 1878–2003).

Statistical publications have various forms of editions: yearbooks, reports, series of books (for example, a census of the population), bulletins, and journals, "notebooks", which contain statistical reviews (quarterly, monthly, *Bulletin of Statistics, Journal of Statistics, Survey of Statistics and Review of Statistics*), summaries, and reports.

The form and content of statistical publications have been changing along with history.

The first statistical publications (similar to modern ones) appeared in 15^{th} century in Venice and then later on in Holland (a series of 60 small volumes under a common name "Elsevier republics," from 1624). In England numerical statistical figures appeared in the 17^{th} century in works by the founders of "political arithmetic," William Petty and John Graunt, and in the 18^{th} century in the works by Gregory King. In Germany ("The Holy Roman Empire of the German Nation"), the second half of the 17^{th} and 18^{th} centuries were predominated by "descriptive government statistics" (H. Conring, G. Achenwall, A. L. Schlözer); only in the last quarter of the 18^{th} century did a new type of statistical publications appeared, i.e., the works of "linear arithmeticians" tending to represent numerical data about one or several countries in the shape of statistical graphs—diagrams and cartograms (the founder of these statistical publications is August Friedrich Crome, who published "*Producten-Karte von Europa*" (1782) and *Über die Größe und Bevölkerung der sämtlichen europäischen Staaten* (1785)). In Russia, the first statistical publications date back to 1831 (historical, ethnographic, and economic atlases with a statistical description of Russia by I. K. Kirilov). The classified yearbooks (with the scope of data for a period of 100 years and more by various types of figures describing territories, natural resources, population, GDP, standard of living etc.) of the USA have been published in the United States since 1878 (125 yearbooks), in Great Britain since 1850 (150 yearbooks), in France since 1860 (85 yearbooks of old series and 23 of new series), in Germany since 1872, in Canada since 1818, in Sweden since 1915, and in Japan since 1818.

Apart from yearbooks there are also many other specialized statistical publications, the most important among them being "Census of Population," "Census of Manufacturers," etc., annual surveys on separate industries "Annual Survey on Manufacturers," enterprises "Moody's manual" in the U.S. "Compas" in Germany, France, and Belgium, and also personal references such as "Who's Who," "Who's Who in the world," "Poor's Register of Corporations Directors and Executives," "Great Minds of the 21^{st} Century," etc.

The first international statistical dictionary was by Michael G. Mulhall, "The Dictionary of Statistics," which ran into several editions (1884, 1892, 1899, 1909) included figures on 30–50 countries for a period from 1800 to 1900. Augustus D. Webb's "The New Dictionary of Statistics"x covered 1896–1905. From 1916 to 1926, the International

Statistical Institute (ISI) published the "International Statistical Yearbook" (from 1853–1876 there were editions from the International Statistical Congresses). With the establishment of the League of Nations (1919) the number of statistical publications increased. The significant statistical publications by the League of Nations were "Statistical Yearbook of the League of Nations" (11 yearbooks for a period from 1932 to 1945), "Monthly Bulletin of Statistics," "World Economic Surveys" (1933–1945, 11 issues), "World Production and Prices" (1925–1939, 7 issues), "Review of World Trade" (1932–1939, 8 issues), etc. In 1919, the International Labour Organization began publication of the "Yearbook of Labour Statistics," and in 1921 the International Institute of Agriculture started publication of the "International Yearbook of Agriculture Statistics".

In 1949, the United Nations Organization (UN) and its specialized institutions started a new stage of statistical publications subdivided into nine series - A, B, C, D, J, K, M, P, F. The most important of them are: "Statistical Yearbook," "Demographic Yearbook," "Yearbook of National Accounts Statistics," "Yearbook of International Trade Statistics," "Balance of Payments Yearbook," "Annual Epidemiological and Vital Statistics," "United Nations Juridical Yearbook," and "Yearbook of the United Nations."

The Food and Agriculture Organization publishes "Yearbook of Food and Agricultural Statistics," "Yearbook of Fishery Statistics," and "Yearbook of Forest Products."

UNESCO publishes "International Yearbook of Education," "Yearbook of Youth Organizations," and "UNESCO Statistical Yearbook."

EU, OECD, WHO, EuroStat, IMF, and World Bank have their own statistical publications. The most important statistical publications are world economic reviews (published separately by the UN and its commissions for Europe, Asia, Africa and Latin America, on annual basis) and various statistical editions. There are also statistical journals, for example, the UN's "Monthly Bulletin of Statistics" and the UN's reference books, "World Weight and Measures," "Nomenclature of Geographic Areas for Statistical Purposes," "Name's of Countries and Adjectives of Nationality," etc. The international bibliographies, indexes, dictionaries, and encyclopedias are also considered to be statistical publications.

The specialized editions and international statistical classifiers, questionnaires, systems, methods, and standards (there are over 120,000 of titles including 175 standard classifiers in the world) regulate the procedures of the international comparisons, the most recognized standards of which are UN's System of National Accounts, trade, banking and monetary transactions, and standards of EuroStat and IMF on the statistical ethics and assessment of data quality.

About the Author

For Biography *see* the entry ►Actuarial Methods.

Cross References

►Census
►Eurostat
►Statistics, History of

References and Further Reading

Nixon JW (1960) A history of the International Statistical Institute 1855–1960. International Statistical Institute, Hague
Simchera VM, Sokolin VL (2001) Encyclopedia of statistical publications X–XX centuries. Financy i Statistika, Moscow
Simchera VM (2006) Russia: 100 years of economic growth: 1900–2000: historical series, trends of centuries, institutional cycles. Nauka, Moscow

Statistical Quality Control

M. Ivette Gomes
Professor
Universidade de Lisboa, DEIO and CEAUL, Lisboa, Portugal

Quality: A Brief Introduction

The main objective of *statistical quality control* (SQC) is to achieve *quality* in production and service organizations, through the use of adequate statistical techniques. The following survey relates to manufacturing rather than to the service industry, but the principles of SQC can be successfully applied to either. For an example of how SQC applies to a service environment, see Roberts (2005). *Quality* of a product can be defined as its adequacy to be used (Montgomery 2009), which is evaluated by the so-called *quality characteristics*. Those are random variables in a probability language, and are usually classified as: *physical*, like length and weight; *sensorial*, like flavor and color; *temporally oriented*, like the maintenance of a system.

Quality Control (QC) has been an activity of engineers and managers, who have felt the need to work jointly with statisticians. Different quality characteristics are measured and compared with pre-determined specifications, the *quality norms*. QC began a long time ago, when manufacturing began and competition accompanied it,

with consumers comparing and choosing the most attractive product. The *Industrial Revolution*, with a clear distinction between producer and consumer, led producers to the need of developing methods for the control of their manufactured products. On the other hand, SQC is comparatively new, and its greatest developments have taken place during the twentieth century. In 1924, at the Bell Laboratories, Shewhart developed the concept of *control chart* and, more generally, *statistical process control* (SPC), shifting the attention from the product to the production process (Shewhart 1931). Dodge and Romig (1959), also in the Bell Laboratories, developed *sampling inspection*, as an alternative to the 100% inspection.

Among the pioneers in SPC we also distinguish W.E. Deming, J.M. Juran, P.B. Crosby and K. Ishikawa (see other references in Juran and Gryna 1993). But it was during the *Second World War* that there was a generalized use and acceptance of SQC, largely used in USA and considered as primordial for the defeat of Japan. In 1946, the *American Society for Quality Control* was founded, and this enabled a huge push to the generalization and improvement of SQC methods.

After the II World War, Japan was confronted with rare food and lodging, and the factories were in ruin. They evaluated and corrected the causes of such a defeat. The quality of the products was an area where USA had definitely over passed Japan, and this was one of the items they tried to correct, becoming rapidly masters in inspection sampling and SQC, and leaders of quality around 1970. Recently, the quality developments have also been devoted to the motivation of workers, a key element in the expansion of the Japanese industry and economy.

Quality is more and more the prime decision factor in the consumer preferences, and quality is often pointed out as the key factor for the success of organizations. The implementation of a *production* QC clearly leads to a reduction in the manufacturing costs, and the money spent with control is almost irrelevant. At the moment, the quality improvement in all areas of an organization, a philosophy known as *Total Quality Management* (TQM) is considered crucial (see Vardeman and Jobe 1999). The challenges are obviously difficult. But the modern SQC methods surely provide a basis for a positive answer to these challenges. SQC is at this moment much more than a set of *statistical instruments*. It is a global way of thinking of workers in an organization, with the objective of making things *right in the first place*. This is mainly achieved through the systematic *reduction of the variance* of relevant quality characteristics.

Usual Statistical Techniques in SQC

The statistical techniques useful in SQC are quite diverse. In this survey, we shall briefly mention SPC, an on-line control technique of a process production with the use of ►*control charts*. ►*Acceptance sampling*, performed out of the line production (before it, for sentencing incoming batches, and after it, for evaluating the final product), is another important topic in SQC (see Duncan [1986] and Pandey [2007], among others). A similar comment applies to *reliability theory* and *reliability engineering*, off-line techniques performed when the product is complete, in order to detect the resistance to failure of a device or system (see Pandey [2007], also among others).

It is however sensible to mention that, additionally to these techniques, there exist other statistical topics useful in the *improvement* of a process. We mention a few examples: in a line of production, we have the *input variables*, the *manufacturing process* and the *final product* (output). It is thus necessary to model the relationship between input and output. Among the statistical techniques useful in the building of these models, we mention *Regression* and *Time Series Analysis*. The area of *Experimental Design* (see Taguchi et al. 1989) has also proved to be powerful in the detection of the most relevant input variables. Its adequate use enables a reduction of variance and the identification of the controllable variables that enable the optimization of the production process.

Statistical Process Control (SPC). Key monitoring and investigating tools in SPC include *histograms*, *Pareto charts*, *cause and effect diagrams*, *scatter diagrams* and *control charts*. We shall here focus on control chart methodology.

A *control chart* is a popular statistical tool for monitoring and improving quality, and its success is based on the idea that no matter how well the process is designed, there exists a certain amount of nature variability in output measurements. When the variation in process quality is due to random causes alone, the process is said to be *in-control*. If the process variation includes both random and special causes of variation, the process is said to be *out-of-control*. The control chart is supposed to detect the presence of special causes of variation.

Generally speaking, the main steps in the construction of a control chart, performed at a *stable* stage of the process, are the following: determine the process parameter you want to monitor, choose a convenient statistic, say W, and create a *central line* (CL), a *lower control limit* (LCL) and an *upper control limit* (UCL). Then, sample the

production process along time, and group the process measurements into *rational subgroups* of size n, by time period t. For each rational subgroup, compute w_t, the observed value of W_t, and plot it against time t. The majority of measurements should fall in the so-called *continuation interval* $C = [LCL, UCL]$. Data can be collected at *fixed sampling intervals* (FSI), with a size equal to d, or alternatively, at *variable sampling intervals* (VSI), usually with sampling intervals of sizes d_1, d_2 $(0 < d_1 < d_2)$. The region C is then split in two disjoint regions C_1 and C_2, with C_2 around CL. The sampling interval d_1 is used as soon as a measurement falls in C_1; otherwise, it is used the largest sampling interval d_2. If the measurements fall within LCL and UCL no action is taken and the process is considered to be *in-control*. A point w_t that exceeds the control limits signals an alarm, i.e., it indicates that the process is *out of control*, and some action should be taken, ranging from taking a re-check sample to the tracing and elimination of these causes. Of course, there is a slight chance that is a *false alarm*, the so-called α-risk. The design of control charts is a compromise between the risks of not detecting real changes (β-risks) and of α-risks. Other relevant *primary characteristics* of a chart are the *run length* (RL) or *number of samples to signal* (NSS) and the associated mean value, the *average run length*, ARL=$\mathbb{E}(RL) = 1/(1-\beta)$, as well as the *capability indices*, C_k and C_{pk} (see Pearn and Kotz 2006). Essentially, a control chart is a test, performed along time t, of the hypothesis H_0 : the process is in-control *versus* H_1 : the process is out-of-control.

Stated differently, we use historical data to compute the initial control limits. Then the data are compared against these initial limits. Points that fall outside of the limits are investigated and, perhaps, some will later be discarded. If so, the limits need to be recomputed and the process repeated. This is referred to as *Phase* I. Real-time process monitoring, using the limits from the end of Phase I, is *Phase* II. There thus exists a strong link between control charts and hypothesis testing performed along time.

Note that a *preliminary statistical data analysis* (usually *histograms* and *Q-Q plots*) should be performed on the prior collected data. A common assumption in SPC is that quality characteristics are distributed according to a *normal* distribution. However, this is not always the case, and in practice, if data seem very far from meeting this assumption, it is common to transform them through a ▶Box–Cox transformation (Box and Cox 1964). But much more could be said about the case of nonnormal data, like the use of robust control charts (see Figueiredo and Gomes [2004], among others).

With its emphasis on early detection and prevention of problems, SPC has a distinct advantage over quality methods such as inspection, that apply resources to detecting and correcting problems in the final product or service. In addition to reducing waste, SPC can lead to a reduction in the time required to produce the final products. SPC is recognized as a valuable tool from both a cost reduction and a customer satisfaction standpoint. SPC indicates when an action should be taken in a process, but it also indicates when no action should be taken.

Classical Shewhart Control Charts: A Simple Example. In this type of charts, measurements are assumed to be independent and distributed according to a normal distribution. Moreover, the statistics W_t built upon those measurements are also assumed to be independent. The main idea underlying these charts is to find a simple and convenient statistic, W, with a sampling distribution easy to find under the validity of the *in-control* state, so that we can easily construct a confidence interval for a location or spread measure of that statistic. For continuous quality characteristics, the most common Shewhart-charts are the average chart ($\overline{X}$-chart) and the range chart (R-chart), as an alternative to the standard-deviation chart (S-chart). For discrete quality characteristics, the most usual charts are the p-charts and np-charts in a $Binomial(n, p)$ background, and the so-called c-charts and u-charts for $Poisson(c)$ backgrounds.

Example 1 ($\overline{X}$-chart). *Imagine a breakfast cereal packaging line, designed to fill each cereal box with* 500 *grams of product. The production manager wants to monitor on-line the mean weight of the boxes, and it is known that, for a single pack, an estimate of the weight standard-deviation* σ *is* 10 *g. Daily samples of* $n = 5$ *packs are taken during a stable period of the process, the weights* $x_i, 1 \le i \le n$, *are recorded, and their average,* $\overline{x} = \sum_{i=1}^{n} x_i/n$, *is computed. These averages are estimates of the process mean value* μ, *the parameter to be monitored. The center line is* $CL = 500$ *g (the target). If we assume that data are normally distributed, i.e.,* $X \sim N(\mu = 500, \sigma = 10)$, *the control limits can be determined on the basis that* $\overline{X} \sim N(\mu = 500, \sigma/\sqrt{n} = 10/\sqrt{5} = 4.472)$. *In-control, it thus expected that* $100(1-\alpha)\%$ *of the average weights are between* $500 + 4.472\ \xi_{\alpha/2}$ *and* $500 - 4.472\ \xi_{\alpha/2}$ *where* $\xi_{\alpha/2}$ *is the* $(\alpha/2)$*-quantile of a standard normal distribution. For a* α*-risk equal to* 0.002 *(a common value in English literature),* $\xi_{\alpha/2} = -3.09$. *The American Standard is based on "*3 *– sigma" control limits (corresponding to* 0.27% *of false alarms), while the British Standard uses*

"3.09–sigma" limits (corresponding to 0.2% of false alarms). In this case, the 3-sigma *control limits are* $LCL = 500 - 3 \times 10/\sqrt{5} = 486.584$ *and* $UCL = 500 + 3 \times 10/\sqrt{5} = 513.416$.

Other Control Charts. Shewhart-type charts are efficient in detecting medium to large shifts, but are insensitive to small shifts. One attempt to increase the power of these charts is by adding supplementary stopping rules based on runs. The most popular stopping rules, supplementing the ordinary rule, "one point exceeds the control limits," are: two out of three consecutive points fall outside warning (2-sigma) limits; four out of five consecutive points fall beyond 1-sigma limits; eight consecutive points fall on one side of the centerline.

Another possible attempt is to consider some kind of dependency between the statistics computed at the different sampling points. To control the mean value of a process at a target μ_0, one of the most common control charts of this type is the *cumulative sum* (CUSUM) chart, with an associated control statistic given by $S_t := \sum_{j=1}^{t}(x_j - \mu_0) = S_{t-1} + (\bar{x}_t - \mu_0)$, $t = 1, 2, \cdots \quad (S_0 = 0)$. Under the validity of $H_0 : X \frown N(\mu_0, \sigma)$, we thus have a *random walk* with null mean value (see ►Random Walk). It is also common to use the *exponentially weighted moving average* (EWMA) statistic, given by $Z_t := \lambda\bar{x}_t + (1-\lambda)Z_{t-1} = \lambda\sum_{j=0}^{t-1}(1-\lambda)^j\,\bar{x}_{t-j} + (1-\lambda)^t Z_0$, $t = 1, 2, \ldots, \quad Z_0 = \bar{\bar{x}}, \quad 0 < \lambda < 1$, where $\bar{\bar{x}}$ denotes the overall average of a small number of averages collected *a priori*, when the process is considered stable and in-control. Note that it is also possible to replace averages by individual observations (for details, see Montgomery 2009).

ISO 9000, Management and Quality

The main objective of this survey was to speak about statistical instruments useful in the improvement of quality. But these instruments are a small part of the total effort needed to achieve quality. Nowadays, essentially due to an initiative of the International Organization for Standardization (ISO), founded in 1946, all organizations are pushed towards quality. In 1987, ISO published the ISO 9000 series, with general norms for quality management and quality guarantee, and additional norms were established later on diversified topics. The ISO 9000 norms provide a guide for producers, who want to implement efficient quality. They can also be used by consumers, in order to evaluate the producers' quality. In the past, the producers were motivated to the establishment of quality through the increasing satisfaction of consumers. Nowadays, most of the them are motivated by the ISO 9000 certification – if they do not have it, they will lose potential clients.

Regarding management and quality: as managers have a final control of all organization resources, management has a ultimate responsibility in the quality of all products. Management should thus establish a quality policy, making it perfectly clear to all workers (see Burrill and Ledolter 1999, for details).

Acknowledgment

Research partially supported by FCT/OE, POCI 2010 and PTDC/FEDER.

About the Author

Dr. Gomes is Professor of Statistics at the Department of Statistics and Operations Research (DEIO), Faculty of Science, University of Lisbon. She is Past President of Portuguese Statistical Society (1989–1993). She is Founding Editor, *Revstat* (2003–), Associate Editor of *Extremes* (2007–) and Associate Editor of *J. Statistical Planning and Inference* (2007–).

Cross References

►Acceptance Sampling
►Box–Cox Transformation
►Control Charts
►Random Walk
►Rao–Blackwell Theorem
►Relationship Between Statistical and Engineering Process Control
►Statistical Design of Experiments (DOE)
►Statistical Quality Control: Recent Advances

References and Further Reading

Burril CW, Ledolter J (1999) Achieving quality through continual improvement. Wiley, New York
Box GEP, Cox DR (1964) An analisys of transformations. J R Stat Soc B26:211–256
Dodge HF, Romig HG (1959) Sampling inspection tables, single and double sampling, 2nd edn. Wiley
Duncan AJ (1986) Quality control and industrial statistics, 5th edn. Irwin, Homehood
Figueiredo F, Gomes MI (2004) The total median in statistical quality control. Appl Stoch Model Bus 20(4):339–353
Juran JM, Gryna FM (1993) Quality planning and analysis. MacGraw-Hill, New York
Montgomery DC (2009) Statistical quality control: a modern introduction, 6th edn. Wiley, Hoboken, NJ
Pandey BN (2007) Statistical techniques in life-testing, reliability, sampling theory and quality control. Narosa, New Delhi
Pearn WL, Kotz S (2006) Encyclopedia and handbook of process capability indices: a comprehensive exposition of quality control measures. World Scientific, Singapore

Roberts L (2005) SPC for right-brain thinkers: process control for non-statisticians. Quality, Milwaukee
Shewhart WA (1931) Economic control of quality of manufactured product. Van Nostrand, New York
Taguchi G, Elsayed E, Hsiang T (1989) Quality engineering in production systems. Mc-Graw-Hill, New York
Vardeman S, Jobe JM (1999) Statistical quality assurance methods for engineers. Wiley, New York

Statistical Quality Control: Recent Advances

Fugee Tsung[1], Yanfen Shang[2], Xianghui Ning[2]
[1]Professor and Head
Hong Kong University of Science and Technology, Hong Kong, China
[2]Hong Kong University of Science and Technology, Hong Kong, China

Statistical quality control aims to achieve the product or process quality by utilizing statistical techniques, in which statistical process control (SPC) has been demonstrated to be one primary tool for monitoring the process or product quality. Since 1920s, the control chart, as one of the most important SPC techniques, has been widely studied.

Univariate Control Charts Versus Multivariate Control Charts

In terms of the number of variables, ▶control charts can be classified into two types, that is, univariate control charts and multivariate control charts.

The performance of the conventional univariate control charts, including Shewhart control charts, cumulative sum (CUSUM) control charts and exponentially weighted moving average (EWMA) control charts have been extensively reviewed. The research demonstrates that the Shewhart chart is more sensitive to large shifts than the EWMA and CUSUM chart and vice versa. These traditional control charts usually assume that the observations are independent and identically follow the normal distribution. In some practical situations, however, these assumptions are not valid. Therefore, other control charts that are different or extended from the traditional charts are developed for some special cases, such as monitoring autocorrelated processes and/or processes with huge sample data, detecting dynamic mean change and/or a range of mean shifts. See Han and Tsung (2005, 2006, 2007, 2009), Han et al. (2007a, b), Wang and Tsung (2005), Zhao et al. (2005) and Zou et al. (2008c) for detailed discussion.

Although the aforementioned univariate charts perform well in monitoring some process or product qualities, their performance is not satisfactory when the quality of a product or process is characterized by several correlated variables. Therefore, multivariate statistical process control (MSPC) techniques were developed and widely applied. Hotelling's T^2 chart, the traditional multivariate control chart, was proposed in 1947 (Hotelling 1947) to deal with the multivariate monitoring case, which assumed that several variables follow the multivariate normal distribution (see ▶Multivariate Normal Distributions). Following that, a variety of studies extended this research further. Among others, see Tracy et al. (1992), Mason et al. (1995), and Sullivan and Woodall (1996) for discussion concerning the property and performance of the T^2 chart.

Besides the Hotelling's T^2 chart, the other traditional multivariate control charts include the Multivariate cumulative sum (MCUSUM) chart presented by Crosier (1988) and Pignatiello and Runger (1990) and the multivariate exponentially weighted moving average (MEWMA) chart proposed by Lowry et al. (1992). Similarly to Hotelling's T^2, these two charts are sensitive to moderate and small mean shifts. Other extensions of traditional MSPC techniques, i.e., adaptive T^2 chart for dynamic processes (see Wang and Tsung (2007, 2008)), have been analyzed. Besides the multivariate charts for mean shifts, the multivariate charts for monitoring the process variation were also presented recently, such as the multivariate exponentially weighted mean squared deviation (MEWMS) chart and a multivariate exponentially weighted moving variance (MEWMV) chart (Huwang et al. (2007)). The extensive literature reviews were provided by Kourti and MacGregor (1996) and Bersimis et al. (2007), in which other statistical methods applied in MSPC, i.e., ▶principal component analysis (PCA) and partial least square (PLS), are also reviewed.

Most of the mentioned charts have a common assumption that process variables follow normal distributions. When there is no distribution assumption, nonparametric methods, like the depth function (Zuo and Serfling (2000)), can be used, the advantages of which are examined by Chakraborti et al. (2001). However, with the development of technology, a more complicate situation occurs. Numerical process variables may be mixed up with the categorical process variables to represent the real condition of a process. Direct application of the aforementioned methods may lead to inappropriate ARL and unsatisfactory false alarms. An alternative way to solve this problem is to use some distribution-free methods, like the K-chart proposed by Sun and Tsung (2003). More research is needed in this area.

SPC for Profile Monitoring

In most SPC applications, either in the univariate or multivariate cases, it is assumed that the quality of a process or product can be adequately represented by the distribution of a single quality characteristic or by the general multivariate distribution of the several correlated quality characteristics. In some practical situations, however, the quality of a process or product is better characterized and summarized by a relationship between a response variable and one or more explanatory variables (Woodall et al. 2004). Therefore, studies on profile monitoring have been steadily increasing.

The early research on profile monitoring usually assumes that the relationship can be represented by the linear model. There has been extensive existing research on linear profile monitoring in the literature. For example, as early as 2000, Kang and Albin presented two methods in order to monitor the linear profiles. One approach is to monitor the intercept and slope of the linear model by constructing the multivariate chart (T^2 chart). The other is to monitor the average residuals by using the exponential weighted moving average (EWMA) chart and rang (R) chart simultaneously. It can be noted that some different control schemes were also developed for solving different linear profile monitoring problems, i.e., the self-staring control chart for linear profiles with unknown parameters (Zou et al. (2007a)). In addition, Zou et al. (2007b) proposed a multivariate EWMA (MEWMA) scheme for monitoring the general linear profile. Furthermore, recent studies on the nonlinear profile monitoring can be sourced in the relevant literature. Among others, the nonparametric methods are commonly used in monitoring the nonlinear profiles (see Zou et al. 2008b, Jensen et al. 2009). Besides, Woodall et al. (2004) provided an extensive review on profile monitoring. Recent research focused on the control scheme for monitoring profiles with categorical data rather than continuous data (Yeh et al. 2009)), in which a Phase I monitoring scheme for profiles with binary output variables was proposed.

SPC for Processes with Multiple Stages

In modern manufacturing and service environments, it is very common that most manufacturing and/or service processes involve a large number of operating stages rather than one single stage. Many examples of such multistage processes can be found in semiconductor manufacturing, automobile assembly lines and bank services, etc. For instance, the print circuit board (PCB) manufacturing process includes several stages, that is, exposure to black oxide, lay-up, hot press, cutting, drilling, and inspection. However, most of the abovementioned conventional SPC methods focus on single-stage processes without considering the multistage scenario, which do not consider the relationship among different stages. Therefore, the recent research on multistage processes has been widely conducted.

The existing popular SPC methods for multistage processes usually involve three types of approaches, which are the regression adjustment method, the cause-selecting method and methods based on linear state space models. The regression adjustment method was developed by Hawkins (1991, 1993), while Zhang (1984, 1985, 1989, 1992) proposed the cause-selecting method. A review of the cause-selecting method can be found in Wade and Woodall (1993). Recent research on the use of cause-selecting charts for multistage processes can be found in Shu et al. (2003), Shu and Tsung (2003), Shu et al. (2004) and Shu et al. (2005). A variety of current studies on multistage processes also adopt engineering models with a linear state space model structure. This model incorporates physical laws and engineering knowledge in order to describe the quality linkage among multiple stages in a process. Latest works on multistage process monitoring and diagnosis can be referred to Xiang and Tsung (2008), Zou et al. (2008a), Jin and Tsung (2009), and Li and Tsung (2009). With respect to multistage processes with categorical variables, some monitoring schemes were developed recently. For example, Skinner et al. (2003, 2004) proposed the generalized linear model (GLM)-based control chart for the Poisson data obtained from multiple stages.

An extensive review on the quality control of multistage systems including monitoring and diagnosing schemes was presented by Shi and Zhou (2009).

SPC Applications in Service Industries

SPC techniques can be applied in different industries such as manufacturing or service industries, although most of these techniques are originally developed for manufacturing industries, i.e., machining processes, assembly processes, semiconductor processes etc. Because the SPC techniques have been demonstrated to be efficient for manufacturing processes, the application of these techniques in service processes was argued in some papers (see Wyckoff (1984), Palm et al. (1997) and Sulek (2004)). In the existing literature, several control charts have been applied in service processes, i.e., quick service restaurant, the auto loan process that provides better service from the loan company to car dealers and buyers, and invoicing processes. See Apte and Reynolds (1995), Mehring (1995), Cartwright and Hogg (1996) for detailed discussion. In addition, the control charts were also widely applied in health-care and public-health fields (see Wardell and Candia (1999),

Green (1999)). Recently, Woodall (2006) discussed in great detail different control charts that have been proposed in health-care and public-health fields. Both the manufacturing process and the service operation process involve multiple operating stages rather than a single stage. Therefore, Sulek et al. (2005) proposed to use the cause selecting control chart for monitoring the service process with multiple stages in the grocery store and showed that it outperformed the Shewhart chart in monitoring the multistage service process. More recent studies on the application of SPC techniques, especially in service industries, were reviewed by Maccarthy and Wasusri (2002) and Tsung et al. (2008). All these applications showed that SPC techniques were efficient in monitoring and identifying service processes.

Statistical Process Control as one primary tool for quality control is very efficient and important in monitoring the process/product quality. SPC techniques will be applied in more industries with different characteristics. Therefore, more advanced studies on SPC schemes will be widely conducted in order to achieve the quality required for products or processes.

About the Author

Dr. Fugee Tsung is Professor and Head of the Department of Industrial Engineering and Logistics Management (IELM), Director of the Quality Lab, at the Hong Kong University of Science & Technology (HKUST). He is a Fellow of the Institute of Industrial Engineers (IIE), Fellow of the American Society for Quality (ASQ) and Fellow of the Hong Kong Institution of Engineers (HKIE). He received both his MSc and PhD from the University of Michigan, Ann Arbor and his BSc from National Taiwan University. He is currently Department Editor of the IIE Transactions, Associate Editor of *Technometrics*, *Naval Research Logistics*, and on the Editorial Boards for *Quality and Reliability Engineering International* (QREI). He is an ASQ Certified Six Sigma Black Belt, ASQ authorized Six Sigma Master Black Belt Trainer, Co-funder and Chair of the Service Science Section at INFORMS, Regional Vice President (Asia) of IIE. He has authored over 70 refereed journal publications, and is also the winner of the Best Paper Award for the IIE Transactions in 2003 and 2009. His research interests include quality engineering and management to manufacturing and service industries, statistical process control, monitoring and diagnosis.

Cross References

►Control Charts
►Moving Averages
►Multivariate Normal Distributions
►Multivariate Statistical Process Control
►Relationship Between Statistical and Engineering Process Control
►Statistical Quality Control

References and Further Reading

Apte UM, Reynolds CC (1995) Quality management at Kentucky fried chicken. Interfaces 25:6–21
Bersimis S, Psarakis S, Panaretos J (2007) Multivariate statistical process control charts: an overview. Quality Reliab Eng Int 23(5):517–543
Cartwright G, Hogg B (1996) Measuring processes for profit. The TQM Magazine 8(1):26–30
Chakraborti S, Van der Laan P, Bakir ST (2001) Nonparametric control charts: an overview and some results. J Qual Technol 33:304–315
Crosier RB (1988) Multivariate generalizations of cumulative sum quality-control schemes. Technometrics 30(3):291–303
Green RS (1999) The application of statistical process control to manage global client outcomes in behavioral healthcare. Eval Program Plann 22:199–210
Han D, Tsung F (2005) Comparison of the Cuscore. GLRT and CUSUM control charts for detecting a dynamic mean change. Ann Inst Stat Math 57:531–552
Han D, Tsung F (2006) A reference-free cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. J Am Stat Assoc 101:368–386
Han D, Tsung F (2007) Detection and diagnosis of unknown abrupt changes using CUSUM multi-chart schemes. Sequential Anal 26:225–249
Han D, Tsung F (2009) Run length properties of the CUSUM and EWMA control schemes for stationary autocorrelated processes. Statistica Sinica 19:473–490
Han D, Tsung F, Li Y (2007a) A CUSUM chart with local signal amplification for detecting a range of unknown shifts. Int J Reliab Qual Saf Eng 14:81–97
Han D, Tsung F, Hu X, Wang K (2007b) CUSUM and EWMA multi-charts for detecting a range of mean shifts. Statistica Sinica 17:1139–1164
Hawkins DM (1991) Multivariate quality control based on regression-adjusted variables. Technometrics 33:61–75
Hawkins DM (1993) Regression adjustment for variables in multivariate quality control. J Qual Technol 25:170–182
Hotelling H (1947) Multivariate quality control. In: Eisenhart C, Hastay M, Wallis WA (eds) Techniques of statistical analysis. McGraw-Hill, New York
Huwang L, Yeh AB, Wu C (2007) Monitoring multivariate process variability for individual observations. J Qual Technol 39(3):258–278
Jensen WA, Birch JB, Woodall WH (2009) Profile monitoring via nonlinear mixed models. J Qual Technol 41:18–34
Jin M, Tsung F (2009) A chart allocation strategy for multistage processes. IIE Trans 41(9):790–803
Kang L, Albin SL (2000) On-line monitoring when the process yields a linear profile. J Qual Technol 32:418–426
Kourti T, MacGregor JF (1996) Multivariate SPC methods for process and product monitoring. J Qual Technol 28(4):409–428
Li Y, Tsung F (2009) False discovery rate-adjusted charting schemes for multistage process monitoring and fault identification. Technometrics 51:186–205

Lowry A, Woodall WH, Champ CW, Rigdon SE (1992) A multivariate exponentially weighted moving average control chart. Technometrics 34(1):46–53
Maccarthy BL, Wasusri T (2002) A review of non-standard applications of statistical process control (SPC) charts. Int J Qual Reliab Manage 19(3):295–320
Mason RL, Tracy ND, Young JC (1995) Decomposition of T2 for multivariate control chart interpretation. J Qual Technol 27(2): 99–108
Mehring JS (1995) Achieving multiple timeliness goals for auto loans: a case for process control. Interfaces 25:81–91
Palm AC, Rodriguez RN, Spiring FA, Wheeler DJ (1997) Some perspectives and challenges for control chart methods. J Qual Technol 29:122–127
Pignatiello J, Runger GC (1990) Comparison of multivariate CUSUM charts. J Qual Technol 22:173–186
Shi J, Zhou S (2009) Quality control and improvement for multistage systems: a survey. IIE Trans 41:744–753
Shu LJ, Tsung F (2003) On multistage statistical process control. J Chin Inst Ind Eng 20:1–8
Shu LJ, Apley DW, Tsung F (2003) Autocorrelated process monitoring using triggered CUSCORE charts. Qual Reliab Eng Int 18:411–421
Shu LJ, Tsung F, Kapur KC (2004) Design of multiple cause-selecting charts for multistage processes with model uncertainty. Qual Eng 16:437–450
Shu LJ, Tsung F, Tsui KL (2005) Effects of estimation errors on cause-selecting charts. IIE Trans 37(6):559–567
Skinner KR, Montgomery DC, Runger GC (2003) Process monitoring for multiple count data using generalized linear model-based control charts. Int J Prod Res 41(6):1167–1180
Skinner KR, Montgomery DC, Runger GC (2004) Generalized linear model-based control charts for discrete semiconductor process data. Qual Reliab Eng Int 20:777–786
Sulek J (2004) Statistical quality control in services. Int J Serv Tech Manag 5:522–531
Sulek JM, Marucheck A, Lind MR (2005) Measuring performance in multi-stage service operations: an application of cause selecting control charts. J Oper Manag 24:711–727
Sullivan JH, Woodall WH (1996) A comparison of multivariate control charts for individual observations. J Qual Technol 28(4):398–408
Sun R, Tsung F (2003) A kernel-distance-based multivariate control chart using support vector methods. Int J Prod Res 41:2975–2989
Tracy ND, Young JC, Mason RL (1992) Multivariate control Charts for Individual Observations. J Qual Technol 24(2):88–95
Tsung F, Li Y, Jin M (2008) Statistical process control for multistage manufacturing and service operations: a review and some extensions. Int J Serv Oper Inform 3:191–204
Wade MR, Woodall WH (1993) A review and analysis of cause-selecting control charts. J Qual Technol 25(3):161–169
Wang K, Tsung F (2005) Using profile monitoring techniques for a data-rich environment with huge sample size. Qual Reliab Eng Int 21(7):677–688
Wang K, Tsung F (2007) Monitoring feedback-controlled processes using adaptive T2 schemes. Int J Prod Res 45:5601–5619
Wang K, Tsung F (2008) An adaptive T2 chart for monitoring dynamic systems. J Qual Technol 40:109–123
Wardell DG, Candia MR (1999) Statistical process monitoring of customer satisfactior survey data. Qual Manag J 3(4):36–50
Woodall WH (2006) The use of control charts in health-care and public-health surveillance. J Qual Technol 38(2):89–104
Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320
Wyckoff DD (1984) New tools for achieving service quality. Cornell Hotel Rest Admin Quartr 25:78–91
Xiang L, Tsung F (2008) Statistical monitoring of multistage processes based on engineering models. IIE Trans 40(10):957–970
Yeh AB, Huwang L, Li YM (2009) Profile monitoring for binary response. IIE Trans 41(11):931–941
Zhang GX (1984) A new type of control charts and a theory of diagnosis with control charts. World Qual Congr Trans: 175–185
Zhang GX (1985) Cause-selecting control charts - a new type of quality control charts. The QR Journal 12:221–225
Zhang GX (1989) A new diagnosis theory with two kinds of quality. world quality congress transactions. Am Soc Qual Control 00:594–599
Zhang GX (1992) Cause-selecting control chart and diagnosis. Theory and practice. Aarhus School of Business. Department of Total Quality Management. Aarhus, Denmark
Zhao Y, Tsung F, Wang Z (2005) Dual CUSUM control schemes for detecting a range of mean shifts. IIE Trans 37:1047–1057
Zou C, Tsung F, Wang Z (2007a) Monitoring general linear profiles using multivariate EWMA schemes. Technometrics 49: 395–408
Zou C, Zhou C, Wang Z, Tsung F (2007b) A self-starting control chart for linear profiles. J Qual Technol 39:364–375
Zou C, Tsung F, Liu Y (2008a) A change point approach for phase I analysis in multistage processes. Technometrics 50(3): 344–356
Zou C, Tsung F, Wang Z (2008b) Monitoring profiles based on nonparametric regression methods. Technometrics 50: 512–526
Zou C, Wang Z, Tsung F (2008c) Monitoring an autocorrelated processes using variable sampling schemes at fixed-times. Qual Reliab Eng Int 24:55–69
Zuo Y, Serfling R (2000) General notions of statistical depth function. Annal Stat 28(2):461–482

Statistical Signal Processing

Debasis Kundu
Chair Professor
Indian Institute of Technology Kanpur, Kanpur, India

Signal processing may broadly be considered to involve the recovery of information from physical observations. The received signals is usually disturbed by thermal, electrical, atmospheric or intentional interferences. Due to the random nature of the signal, statistical techniques play an important role in signal processing. Statistics is used in the formulation of appropriate models to describe the behavior of the system, the development of appropriate techniques

for estimation of model parameters, and the assessment of model performances. Statistical Signal Processing basically refers to the analysis of random signals using appropriate statistical techniques. The main purpose of this article is to introduce different signal processing models and different statistical and computational issues involved in solving them.

The Multiple Sinusoids Model

The multiple sinusoids model may be expressed as

$$y(t) = \sum_{k=1}^{M} \{A_k \cos(\omega_k t) + B_k \sin \omega_k t)\} + n(t); \quad t = 1, \ldots, N. \tag{1}$$

Here A_k's and B_k's represent the amplitudes of the signal, ω_k's represent the real radian frequencies of the signals, $n(t)$'s are error random variables with mean zero and finite variance. The assumption of independence of the error random variables is not that critical to the development of the inferential procedures. The problem of interest is to estimate the unknown parameters $\{A_k, B_k, \omega_k\}$ for $k = 1, \ldots, M$, given a sample of size N. In practical applications often M is also unknown. Usually, when M is unknown, first estimate M using some model selection criterion, and then it is assumed that M is known, and estimate the amplitudes and frequencies.

The sum of sinusoidal model (1) plays the most important role in the Statistical Signal Processing literature. Most of the periodic signals can be well approximated by the model (1) with the proper choice of M and with the amplitudes and frequencies. For several applications of this model in different fields see Brillinger (1987).

The problem is an extremely challenging problem both from the theoretical and computational points of view. As a statistician Fisher (1929) first considered this problem. It seems that the standard least squares estimators will be the natural choice in this case, but finding the least squares estimators, and establishing their properties are far from trivial issues. Although, the model (1) is a non-linear regression model, but the standard sufficient conditions needed for the least squares estimators to be consistent and asymptotically normal do not hold true in this case. Special care is needed in establishing the consistency and ►asymptotic normality properties of the least squares estimators, see for example Hannan (1973) and Kundu (1997) in this respect. Moreover, for computing the least squares estimators, most of the standard techniques like Newton–Raphson or its variants do not often converge even from good starting values. Even if it converges, it may converge to a local minimum rather than the global minimum due to highly non-linear nature of the least squares surface. Special purpose algorithms have been developed to solve this problem.

Several approximate solutions have been suggested in the literature. Among several approximate estimators, Forward Backward Linear Prediction (FBLP) and modified EquiVariance Linear Prediction (EVLP) work very well. But it should be mentioned that none of these methods behaves uniformly better than the other. More than 200 references on this topic can be found in Stoica (1993), and see also Quinn and Hannan (2001), the only monograph written by statisticians in this topic.

Two-Dimensional Sinusoidal Model

Two dimensional periodic signals are often being analyzed by the two-dimensional sinusoidal model, which can be written as follows:

$$y(s,t) = \sum_{k=1}^{M} \{A_k \cos(\omega_k s + \mu_k t) + B_k \cos(\omega_k s + \mu_k t)\} + n(s,t), \quad s = 1, \ldots S, \; t = \ldots, T. \tag{2}$$

Here A_k's and B_k's are amplitudes and ω_k's and μ_k's are frequencies. The problem once again involves the estimation of the signal parameters namely A_k's, B_k's, ω_k's and μ_k's from the data $\{y(s,t)\}$.

The model (2) has been used very successfully for analyzing two dimensional gray texture data, see for example Zhang and Mandrekar (2001). A three dimensional version of it can be used for analyzing color texture data also, see Prasad (2009) and Prasad and Kundu (2009). Some of the estimation procedures available for the one-dimensional problem may be extended quite easily to two or three dimensions. However, several difficulties arise when dealing with high dimensional data. There are several open problems in multidimensional frequency estimation, and this continues to be an active area of research.

Array Model

The area of array processing has received a considerable attention in the past several decades. The signals recorded at the sensors contain information about the structure of the generating signals including the frequency and amplitude of the underlying sources. Consider an array of P sensors receiving signals from M sources ($P > M$). The array geometry is specified by the applications of interest. In array processing, the signals received at the $i-th$ censor is given by

$$y_i(t) = \sum_{j=1}^{M} a_i(\theta_j) x_j(t) + n_i(t), \quad i = 1, \ldots, P. \tag{3}$$

Here $x_j(t)$ represents the signal emitted by the $j-th$ source, and $n_i(t)$ represents additive noise. The model (3) may be written in the matrix form as;

$$y(t) = [a(\theta_1) : \ldots : a(\theta_M)]\, x(t) + n(t) = A(\theta)x(t) + n(t), \quad t = 1, \ldots, N. \tag{4}$$

The matrix $A(\theta)$ has a Vandermonde structure if the underlying array is assumed to be uniform linear array. The signal vector $x(t)$ and the noise vector $n(t)$ are assumed to be independent and zero mean random processes with covariance matrices Γ and $\sigma^2 I$ respectively. The main problem here is to estimate the signal vector θ, based on the sample $y(1), \ldots, y(N)$, when the structure of A is known.

Interestingly, instead of using the traditional maximum likelihood method, different subspace fitting methods, like MUltipe SIgnal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) and their variants are being used more successfully, see for example the text by Pillai (1989) for detailed descriptions of the different methods.

For basic introduction of the subject the readers are referred to Kay (1987) and Srinath et al. (1996) and for advanced materials see Bose and Rao (1993) and Quinn and Hannan (2001).

Acknowledgments

Part of this work has been supported by a grant from the Department of Science and Technology, Government of India.

About the Author

Professor Debasis Kundu received his Ph.D in Statistics in 1989, Pennsylvania State University. He has (co-)authored about 165 papers. He is an Associate Editor of *Communications in Statistics, Theory and Methods* and an Associate Editor of *Communications in Statistics, Simulation and Computations*. He is an invited member of New York Academy of Sciences and a Fellow of the National Academy of Sciences, India.

Cross References

►Least Squares
►Median Filters and Extensions
►Nonlinear Models
►ROC Curves
►Singular Spectrum Analysis for Time Series
►Statistical Pattern Recognition Principles

References and Further Reading

Bose NK, Rao CR (1993) Signal processing and its applications. Handbook of Statistics, 10, North-Holland, Amsterdam

Brillinger D (1987) Fitting cosines: some procedures and some physical examples. In MacNeill IB, Umphrey GJ (eds) Applied statistics, stochastic processes and sampling theory. Reidel, Dordrecht

Fisher RA (1929) Tests of significance in Harmonic analysis. Proc R Soc London A 125:54–59

Hannan EJ (1973) The estimation of frequencies. J Appl Probab 10:510–519

Kay SM (1987) Modern spectral estimation. Prentice Hall, New York, NY

Kundu D (1997) Asymptotic theory of the least squares estimators of sinusoidal signal. Statistics 30:221–238

Pillai SU (1989) Array signal processing. Springer, New York, NY

Prasad A (2009) Some non-linear regression models and their applications in statistical signal processing. PhD thesis, Indian Institute of Technology Kanpur, India

Prasad A, Kundu D (2009) Modeling and estimation of symmetric color textures. Sankhya Ser B 71(1):30–54

Quinn BG, Hannan EJ (2001) The estimation and tracking of frequency. Cambridge University Press, Cambridge, UK

Srinath MD, Rajasekaran PK, Viswanathan R (1996) Introduction to statistical processing with applications. Prentice-Hall, Englewood Cliffs, NJ

Stoica P (1993) List of references on spectral estimation. Signal Process 31:329–340

Zhang H, Mandrekar V (2001) Estimation of hidden frequencies for 2-D stationary processes. J Time Ser Anal 22:613–629

Statistical Significance

Jan M. Hoem
Professor, Director Emeritus
Max Planck Institute for Demographic Research, Rostock, Germany

Statistical thinking pervades the empirical sciences. It is used to provide principles of initial description, concept formation, model development, observational design, theory development and theory testing, and much more. Some of these activities consist in computing significance tests for statistical hypotheses. Such a hypothesis typically is a statement about a regression coefficient in a linear regression or a relative risk for a chosen life-course event, such as marriage formation or death. The hypothesis can state that the regression coefficient equals zero (or that the relative risk equals 1), implying that the corresponding covariate has no impact on the transition in question and thus does not affect the behavior it represents, or that for all practical purposes the analyst may act as if this

were the case. Alternatively the hypothesis may predict the sign of the coefficient, for example that higher education leads to lower marriage rates, *ceteris paribus*, as argued by some economists. The converse (namely that the sign is zero or positive) would be called the *null hypothesis*. Other hypotheses concern the form of the statistical model for the behavior in question. In such a case the null hypothesis would be that the model specified is correct; this leads to questions of goodness of fit. In any case the statistician's task is to state whether the data at hand justify rejecting whatever null hypothesis has been formulated.

The null hypothesis is typically rejected when a suitable *test statistic* has a value that is unlikely when the null hypothesis is correct; usually the criterion is that the test statistic lies in (say) the upper tail of the probability distribution it has when the hypothesis is correct. An upper bound on the probability of rejecting the null hypothesis when it actually is correct is called *the level of significance* of the test method. It is an important task for the investigator to keep control of this upper bound. A test of significance is supposed to prevent that a conclusion is drawn (about a regression coefficient, say) when the data set is so small that a pattern "detected" can be caused by random variation. Operationally an investigator will often compute the probability (when the null hypothesis is correct) that in a new data set, say, the test statistic would exceed the value actually observed and reject the null hypothesis when this so-called *p-value* is very small, since a small *p*-value is equivalent to a large value of the test statistic.

Ideally, hypotheses should be developed on the basis of pre-existing theory and common sense as well as of empirical features known from the existing literature. Strict protocols should be followed that require any hypothesis experimentation to be made on one part of the current data set, with testing subsequently to be carried out on a virgin part of the same data, or on a new data set. Unfortunately, most empirical scientists in the economic, social, biological, and medical disciplines, say, find such a procedure too confining (assuming that they even know about it). It is common practice to use all available data to develop a model, formulate scientific hypotheses, and to compute test statistics or ►*p*-values from the same data, perhaps using canned computer programs that provide values of test statistics as if scientific statistical protocol could be ignored (Ziliak and McCloskey 2008). The danger of such practices is that the investigator loses control over any significance levels, a fact which has been of concern to professional statisticians for a good while (For some contributions from recent decades see Guttman (1985), Cox (1986), Schweder (1988), and Hurvich and Tsai (1990). Such concerns also extend to many others. For instance, Chow (1996) describes a litany of criticism appearing in the psychological literature in Chapter 1 of a book actually written to *defend* the null-hypothesis significance-test procedure. [See Hoem (2008) for a discussion of further problems connected to common practices of significance testing, namely the need to embed an investigation into a genuine theory of behavior rather than to rely on mechanical significance testing, the avoidance of grouped *p*-values (often using a system of asterisks), the selection of substantively interesting contrasts rather than those thrown up mechanically by standard software, and other issues]). For twenty years and more, remedies have been available to overcome the weaknesses of the procedures just described, including rigorous methods for model development and data snooping. Such methods prevent the usual loss of control over the significance level and also allow the user to handle model misspecification (The latter feature is important because a model invariably is an imperfect representation of reality.). Users of event-history analysis may want to consult Hjort (1988, 1992), Sverdrup (1990), and previous contributions from these authors and their predecessors.

Unfortunately such contributions seem to be little known outside a circle of professional statisticians, a fact which for example led Rothman (1998) to attempt to eradicate significance tests from his own journal (*Epidemiology*). He underlined the need to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, he would prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is significant, as if neither chance nor bias could then account for the findings.

About the Author

For biography *see* the entry ►Demography.

Cross References

►Event History Analysis
►Frequentist Hypothesis Testing: A Defense
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Power Analysis
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Significance Testing: An Overview

►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistics: Nelder's view

References and Further Reading

Chow SL (1996) Statistical significance: rationale validity and utility. Sage Publications, London
Cox DR (1986) Some general aspects of the theory of statistics. J Am Stat Assoc 49:559–575
Guttman L (1985) The illogic of statistical inference for cumulative science. Appl Stoch Model Data Anal 1:3–10
Hjort NL (1988) On large-sample multiple comparison methods. Scand J Stat 15(4):259–271
Hjort NL (1992) On inference in parametric survival data models. Int Stat Rev 60(3):355–387
Hoem JM (2008) The reporting of statistical significance in scientific journals: A reflection. Demographic Res 18(15):437–442
Hurvich CM, Tsai C-L (1990) The impact of model selection on inference in linear regression. The American Statistician 44(3):214–217
Rothman KJ (1998) Special article: writing for epidemiology. Epidemiology 9(3):333–337
Schweder T (1988) A significance version of the basic Neyman–Pearson theory for scientific hypothesis testing. Scand J Stat 15(4):225–235 (with a discussion by Ragnar Norberg, pp 235–242)
Sverdrup E (1990) The delta multiple comparison method: performance and usefulness. Scand J Stat 17(2):115–134
Ziliak ST, McCloskey DN (2008) The cult of statistical significance; how the standard error costs us jobs, justice, and lives. The University of Michigan Press, Ann Arbor

Statistical Software: An Overview

Jan de Leeuw
Distinguished Professor and Chair
University of California-Los Angeles, Los Angeles, CA, USA

Introduction

It is generally acknowledged that the most important changes in statistics in the last 50 years are driven by technology. More specifically, by the development and universal availability of fast computers and of devices to collect and store ever-increasing amounts of data. Satellite remote sensing, large-scale sensor networks, continuous environmental monitoring, medical imaging, micro-arrays, the various genomes, and computerized surveys have not just created a need for new statistical techniques. These new forms of massive data collection also require efficient implementation of these new techniques in software. Thus development of statistical software has become more and more important in the last decades.

Large data sets also create new problems of their own. In the early days, in which the *t*-test reigned, including the data in a published article was easy, and reproducing the results of the analysis did not take much effort. In fact, it was usually enough to provide the values of a small number of sufficient statistics. This is clearly no longer the case. Large data sets require a great deal of manipulation before they are ready for analysis, and the more complicated data analysis techniques often use special-purpose software and some tuning. This makes *reproducibility* a very significant problem. There is no science without replication, and the weakest form of replication is that two scientists analyzing the same data should arrive at the same results.

It is not possible to give a complete overview of all available statistical software. There are older publications, such as Francis (1979), in which detailed feature matrices for the various packages and libraries are given. This does not seem to be a useful approach any more, there simply are too many programs and packages. In fact many statisticians develop ad-hoc software packages for their own projects.

We will give a short historical overview, mentioning the main general purpose packages, and emphasizing the present state of the art. Niche players and special purpose software will be largely ignored. There is a well-known quote from Brian Ripley (2002): "Let's not kid ourselves: the most widely used piece of software for statistics is Excel." This is surely true, but it is equally true that only a tiny minority of statisticians have a degree in statistics. We have to distinguish between "statistical software" and the much wider terrain of "software for statistics." Only the first type is of interest to us here – we will go on kidding ourselves.

BMDP, SAS, SPSS

The original statistical software packages were written for IBM mainframes. BMDP was the first. Its development started in 1957, at the UCLA Health Computing Facility. SPSS arrived second, developed by social scientists at the University of Chicago, starting around 1968. SAS was almost simultaneous with SPSS, developed since 1968 by computational statisticians at North Carolina State University. The three competitors differed mainly in the type of clients they were targeting. And of course health scientists, social scientists, and business clients all needed the standard repertoire of statistical techniques, but in addition some more specialized methods important in their field. Thus the packages diverged somewhat, although their basic components were very much the same.

Around 1985 all three packages added a version for personal computers, eventually developing WIMP (window, icon, menu, pointer) interfaces. Somewhat later they also added matrix languages, thus introducing at least some form of extensibility and code sharing.

As in other branches of industry, there has been some consolidation. In 1996 SPSS bought BMDP, and basically killed it, although BMDP-2009 is still sold in Europe by Statistical Solutions. It is now, however, no longer a serious contender. In 2009 SPSS itself was bought by IBM, where it now continues as PASW (Predictive Analytics Software). As the name change indicates, the emphasis in SPSS has shifted from social science data analysis to business analytics. The same development is going on at SAS, which was originally the Statistical Analysis System. Currently SAS is not an acronym any more. Its main products are SAS Analytics and SAS Business Intelligence, indicating that the main client base is now in the corporate and business community. Both SPSS (now PASW) and SAS continue to have their statistics modules, but the keywords have definitely shifted to analytics, forecasting, decision, and marketing.

Data Desk, JMP, Stata

The second generation of statistics packages started appearing in the 1980's, with the breakthrough of the personal computer. Both Data Desk (1985) and JMP (1989) were, from the start, written for Macintosh, i.e., for the WIMP interface. They had no mainframe heritage and baggage. As a consequence they had a much stronger emphasis on graphics, visualization, and exploratory data analysis.

Data Desk was developed by Paul Velleman, a former student of John Tukey. JMP was the brain child of John Sall, one of the co-founders and owners of SAS, although it existed and developed largely independent of the main SAS products. Both packages featured dynamic graphics, and used graphical widgets to portray and interactively manipulate data sets. There was much emphasis on brushing, zooming, and spinning. Both Data Desk and JMP have their users and admirers, but both packages never became dominant in either statistical research or statistical applications. They were important, precisely because they emphasized graphics and interaction, but they were still too rigid and too difficult to extend.

Stata, another second generation package for the personal computer, was an interesting hybrid of a different kind. It was developed since 1985, like BMDP starting in Los Angeles, near UCLA. Stata had a CLI (command line interface), and did not get a GUI until 2003. It emphasized, from the start, extensibility and user-contributed code. Stata did not get its own matrix language Mata until Stata-9, in 2007.

Much of Stata's popularity is due to its huge archive of contributed code, and a delivery mechanism that uses the Internet to allow for automatic downloads of updates and new submissions. Stata is very popular in the social sciences, where it attracts those users that need to develop and customize techniques, instead of using the more inflexible procedures of SPSS or SAS. For such users a CLI is often preferable to a GUI.

Until Stata developed its contributed code techniques, the main repository had been CMU's statlib, modeled on netlib, which was based on the older network interfaces provided by ftp and email. There were no clear organizing principles, and the code generally was FORTRAN or C, which had to be compiled to be useful. We will see that the graphics from Data Desk and JMP, and the command line and code delivery methods from Stata, were carried over into the next generation.

S, LISP-STAT, R

Work had on the next generation of statistical computing systems had already started before 1980, but it mostly took place in research labs. Bell Laboratories in Murray Hill, N.J., as was to be expected, was the main center for these developments.

At Bell John Chambers and his group started developing the S language in the late seventies. S can be thought of as a statistical version of MATLAB, as a language and an interpreter wrapped around compiled code for numerical analysis and probability. It went through various major upgrades and implementations in the eighties, moving from mainframes to VAX'es and then to PC's. S developed into a general purpose language, with a strong compiled library of linear algebra, probability and optimization, and with implementations of both classical and modern statistical procedures. The first 15 years of S history are ably reviewed by Becker (1994), and there is a 30 year history of the S language in Chambers (2008, Appendix A) . The statistical techniques that were implemented, for example in the *White Book* (Chambers and Hastie 1992), were considerably more up-to-date than techniques typically found in SPSS or SAS. Moreover the S system was build on a rich language, unlike Stata, which until recently just had a fairly large number of isolated data manipulation and analysis commands. Statlib started a valuable code exchange of public domain S programs.

For a long time S was freely available to academic institutions, but it remained a product used only in the higher reaches of academia. AT&T, later Lucent, sold S to

the Insightful corporation, which marketed the product as S-plus, initially quite successfully. Books such as Venables and Ripley; Venables and Ripley (1994; 2000) effectively promoted its use in both applied and theoretical statistics. Its popularity was increasing rapidly, even before the advent of R in the late nineties. S-plus has been quite completely overtaken by R. Insightful was recently acquired by TIBCO, and S-plus is now TIBCO Spotfire S+. We need not longer consider it as a serious contender.

There were two truly exciting developments in the early nineties. Luke Tierney (1990) developed LISP-STAT, a statistics environment embedded in a Lisp interpreter. It provided a good alternative to S, because it was more readily available, more friendly to personal computers, and completely open source. It could, like S, easily be extended with code written in either Lisp or C. This made it suitable as a research tool, because statisticians could rapidly prototype their new techniques, and distribute them along with their articles. LISP-STAT, like Data Desk and JMP, also had interesting dynamic graphics capabilities, but now the graphics could be programmed and extended quite easily. Around 2000 active development of LISP-STAT stopped, and R became available as an alternative (Valero-Mora and Udina 2004).

R was written as an alternative implementation of the S language, using some ideas from the world of Lisp and Scheme (Ihaka and Gentleman 1996). The short history of R is a quite unbelievable success story. It has rapidly taken over the academic world of statistical computation and computational statistics, and to an ever-increasing extend the world of statistics teaching, publishing, and real-world application. SAS and SPSS, which initially tended to ignore and in some cases belittle R, have been forced to include interfaces to R, or even complete R interpreters, in their main products. SPSS has a Python extension, which can run R since SPSS-16. The SAS matrix language SAS/IML, starting at version 3.2. has an interface to an R interpreter.

R is many things to many people: a rapid prototyping environment for statistical techniques, a vehicle for computational statistics, an environment for routine statistical analysis, and a basis for teaching statistics at all levels. Or, going back to the origins of S, a convenient interpreter to wrap existing compiled code. R, like S, was never designed for this all-encompassing role, and the basic engine is straining to support the rate of change in the size and nature of data, and the developments in hardware.

The success of R is both dynamic and liberating. But it remains an open source project, and nobody is really in charge. One can continue to tag on packages extending the basic functionality of R to incorporate XML, multicore processing, cluster and grid computing, web scraping, and so on. But the resulting system is in danger of bursting at the seams. There are now four ways to do (or pretend to do) object-oriented programming, four different systems to do graphics, and four different ways to link in compiled C code. There are thousands of add-on packages, with enormous redundancies, and often with code that is not very good and documentation that is poor. Many statisticians, and many future statisticians, learn R as their first programming language, instead of learning real programming languages such as Python, Lisp, or even C and FORTRAN. It seems realistic to worry at least somewhat about the future, and to anticipate the possibility that all of those thousands of flowers that are now blooming may wilt rather quickly.

Open Source and Reproducibility

One of the consequences of the computer and Internet revolution is that more and more scientists promote open source software and reproducible research. Science should be, per definition, both open and reproducible. In the context of statistics (Gentleman and Temple-Lang 2004) this means that the published article or report is not the complete scientific result. In order for the results to be reproducible, we should also have access to the data and to a copy of the computational environment in which the calculations were made.

Publishing is becoming more open, with e-journals, preprint servers, and open access. Electronic publishing makes both open source and reproducibility more easy to realize. The Journal of Statistical Software, at http://www.jstatsoft.org, the only journal that publishes and reviews statistical software, insists on complete code and completely reproducible examples. Literate Programming systems such as Sweave, at http://www.stat.uni-muenchen.de/~leisch/Sweave/, are becoming more popular ways to integrate text and computations in statistical publications.

We started this overview of statistical software by indicating that the computer revolution has driven much of the recent development of statistics, by increasing the size and availability of data. Replacement of mainframes by minis, and eventually by powerful personal computers, has determined the directions in the development of statistical software. In more recent times the Internet revolution has accelerated these trends, and is changing the way scientific knowledge, of which statistical software is just one example, is disseminated.

About the Author

Dr. Jan de Leeuw is Distinguished Professor and Chair, Department of Statistics, UCLA. He has a 1973 Ph.D. in

Social Sciences from the University of Leiden, Netherlands. He came to UCLA in 1987, after leading the Department of Data Theory at the University of Leiden for about 10 years. He is Elected Fellow, Royal Statistical Society (1984), Elected Member, International Statistical Institute (1986), Corresponding Member, Royal Netherlands Academy of Sciences (1987), Elected Fellow, Institute of Mathematical Statistics (2001) and American Statistical Association (2001). Dr. de Leeuw is Editor-in-Chief, and Founding Editor of *Journal of Statistical Software*, and Editor-in-Chief, *Journal of Multivariate Analysis* (1997–). He is a Former President of the *Psychometric Society* (1987). Professor de Leeuw has (co-)authored over 550 papers, book chapters and reviews, including *Introducing Multilevel Modeling* (with Ita Kreft, Sage, 1998), and *Handbook of Multilevel Analysis* (edited with Erik Meijer, Springer, New York, 2007).

Cross References

- ►Analysis of Variance
- ►Behrens–Fisher Problem
- ►Chi-Square Tests
- ►Computational Statistics
- ►Multiple Imputation
- ►R Language
- ►Selection of Appropriate Statistical Methods in Developing Countries
- ►Spreadsheets in Statistics
- ►Statistical Analysis of Longitudinal and Correlated Data
- ►Statistical Consulting
- ►Statistical Fallacies: Misconceptions, and Myths
- ►Statistics and Climate Change

References and Further Reading

Becker RA (1994) A brief history of S. Technical report, AT&T Bell Laboratories, Murray Hill, N.J. URL http://www2.research.att.com/areas/stat/doc/94.11.ps

Chambers JM (2008) Software for data analysis: programming with R. Statistics and computing. Springer, New York, NY

Chambers JM, Hastie TJ (eds) (1992) Statistical models in S. Wadsworth, California

Francis I (1979) A comparitive review of statistical software. International Association for Statistical Computing, Voorburg, The Netherlands

Gentleman R, Temple-Lang D (2004) Statistical analyses and reproducible research. Bioconductor Project Working Papers 2. URL http://www.bepress.com/cgi/viewcontent.cgi?article=1001&context=bioconductor

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Gr Stat 5:299–314

Ripley BD (2002) Statistical methods *need* software: a view of statistical computing. Presentation RSS Meeting, September. URL http://www.stats.ox.ac.uk/~ripley/RSS2002.pdf

Tierney L (1990) LISP-STAT. An object-oriented environment for statistical computing and dynamic graphics. Wiley, New York

Valero-Mora PM, Udina F (2004) Special issue: lisp-stat: Past, present and future. J Stat Software 13, URL http://www.jstatsoft.org/v13

Venables WN, Ripley BD (1994) Modern applied statistics with S, 1st edn. Springer, New York

Venables WN, Ripley BD (2000) S Programming. Statistics and Computing. Springer, New York, NY

Statistical View of Information Theory

Adnan M. Awad
Professor
University of Jordan, Amman, Jordan

Information Theory has origins and applications in several fields such as: thermodynamics, communication theory, computer science, economics, biology, mathematics, probability and statistics. Due to this diversity, there are numerous information measures in the literature. Kullback (1978), Sakamoto et al. (1986), and Pardo (2006) have applied several of these measures to almost all statistical inference problems.

According to The Likelihood Principle, all experimental information relevant to a parameter θ is mainly contained in the likelihood function $L(\theta)$ of the underlying distribution. Bartlett's information measure is given by $-\log(L(\theta))$. Entropy measures (see ►Entropy) are expectations of functions of the likelihood. Divergence measures are also expectations of functions of likelihood ratios. In addition, Fisher-like information measures are expectations of functions of derivatives of the log-likelihood. DasGupta (2008, Chap. 2) reported several relations among members of these information measures. In sequential analysis, Wald (1947, p. 53) showed earlier that the average sample number depends on a divergence measure of the form

$$E_\theta\left[\log\frac{f(X,\theta_1)}{f(X,\theta_0)}\right]$$

where θ_0 and θ_1 are the assumed values of the parameter θ of the density function f of the random variable X under the null and the alternative hypothesis, respectively.

It is worth noting that, and from the point of view of decision making, the expected change in utility can be

used as a quantitative measure of the worth of an experiment. In this regard Bayes' rule can be viewed as a mechanism that processes information contained in data to update the prior distribution into the posterior probability distribution.

Furthermore, according to Jaynes' Principle of Maximum Entropy (1957), information in a probabilistic model is the available moment constraints on this model. This principle is in fact a generalization of Laplace's Principle of Insufficient Reason.

From a statistical point of view, one should concentrate on the statistical interpretation of properties of entropy-information measures with regard to the extent of their agreement with statistical theorems and to their degree of success in statistical applications.

The following provides a discussion of preceding issues with particular concentration on Shannon's entropy. For more details, the reader can consult the list of references.

1. Consider a discrete random variable X taking a finite number of values $\vec{X} = (x_1, \ldots, x_n)$ with probability vector $P = (p_1, \ldots, p_n)$. Shannon's entropy (information) of P or of X (1948) is given by

$$H(X) = H(P) = -\sum_{i=1}^{n} p_i \log(p_i).$$

The most common bases of the logarithm are 2 and e. With base 2, H is measured in bits whereas, in base e, the units of H are nats. In coding theory the base is 2 whereas, in statistics the base is e.

2. It is quite clear that $H(P)$ is symmetric in the components of the vector P. This implies that components of P can be rearranged to get different density functions which are either: symmetric, negatively skewed, positively skewed, unimodal or bimodal. Such distributions carry different information even though they all have same value of $H(P)$. Therefore, $H(P)$ is unable to reflect the information implied by the shape of the underlying distribution.
3. ▶Entropy of a discrete distribution is always positive while the differential entropy $H(f) = -\int_{-\infty}^{\infty} f(x) \log(f(x))\, dx$ of a continuous variable X with pdf f may take any value on the extended real line. This is due to the fact that the density $f(x)$ need not be less than one as in the discrete case. Thus, Shannon's entropy lacks the ability to give a proper assessment of information when the random variable is continuous. To overcome this problem, Awad (1987) introduced sup-entropy as $-E[\log(f(X)/s)]$, where s is the supremum of $f(x)$.
4. Based on a random sample $O_n = (X_1, \ldots, X_n)$ of size n from a distribution and according to Fisher (1925), a sufficient statistic T carries all information in the sample while any other statistic carries less information than T. The question that arises here is that: "Does Shannon's entropy agree with Fisher's definition of a sufficient statistic?". Let us consider the following two examples.

   First, let $Y : N(\theta, \sigma^2)$ denote a normal random variable with mean θ and variance σ^2. It can be shown that $H(Y) = \log(2\pi e\sigma^2)/2$ which is free of θ. Let O_n be a random sample of size n from $X : N(\theta, 1)$ then by the additivity property of Shannon's entropy, $H(O_n) = nH(X) = n\log(2\pi e)/2$. On the other hand, Shannon's entropy of the sufficient statistic $\overline{X}_n$ is $H(\overline{X}_n) = \log(2\pi e/n)/2 = H(X) - \log(n)/2$. Since $H(X)$ is positive, $H(O_n) \geq H(\overline{X}_n)$ with equality if $n = 1$, i.e., Shannon's entropy of sufficient statistic is less than that of the sample.

   Second, consider a random sample O_n of size n from a continuous uniform distribution on the interval $[0, \theta]$. Let $X_{1:n}$ and $X_{n:n}$ denote the minimum and the maximum ▶order statistics in O_n. It can be shown that $H(X_{1:n}) = H(X_{n:n})$, i.e., Shannon's entropy of sufficient statistic $X_{n:n}$ equals Shannon's entropy of a non-sufficient statistic $X_{1:n}$. These examples illustrate that Shannon's entropy does not agree with Fisher's definition of a sufficient statistic.
5. If $Y = \alpha + \beta X$, $\beta \neq 0$, then $H(Y) = H(X)$ when X is a discrete random variable. However, if X is continuous, $H(Y) = H(X) + \log(|\beta|)$. So, this result implies that two sufficient statistics T_1 and $T_2 = \beta T_1$ will carry (according to Shannon's entropy) unequal amounts of information, which contradicts the sufficiency concept.
6. Referring to the first example in (4), it is clear that Shannon's information in the sample mean is a decreasing function of the sample size n. This is in direct conflict with the usual contention that the larger the sample size is the more information one has. It is also interesting to recall in this regard Basu's example (1975), where a sample of size 2 is more informative (about an unknown parameter) than a sample of size 25. In fact, a rewording of Basu's conclusion is that some observations in the sample are more influential than others.

Acknowledgment

The author wish to thank Prof. Tarald O. Kvalseth and Prof. Miodrag Lovric for their careful reading and valuable suggestions that improved the presentation of the article.

About the Author

Adnan Awad graduated from Yale University, USA, with a PhD. in Statistics, 1978. He chaired the Department of Statistics, Yarmouk University, (1982–1985). He was past chair of the Mathematics Department, University of Jordan. He served as past Vice Dean of Faculty of Graduate studies, Jordan University (1998), and past Vice Dean of the Faculty of Research, Al-albayet University, Jordan (1999). He has authored and co-authored about 80 research papers and more than 20 text books. He supervised seven Ph.D and 28 M.Sc. theses. Moreover, he was a member of the UNESCO team, (1992–1996), of improving teaching mathematics in the Arab World. Professor Awad has been awarded the medal of High Research Evaluation at the Faculty of Science, Yarmouk University (1984), and Abdul-Hammed Shooman Prize for Young Arab Scientists in Mathematics and Computer Science, Jordan (1987), for his contributions to both Prediction Analysis and Information Theory.

Cross References

►Diversity
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Information Theory and Statistics
►Measurement of Uncertainty
►Sufficient Statistical Information
►Sufficient Statistics

References and Further Reading

Awad AM (1987) A statistical information measure. Dirasat (Science) 14(12):7–20
Bartlett MS (1936) Statistical information and properties of sufficiency. Proc R Soc London A 154:124–137
Basu D (1975) Statistical information and likelihood. Sankhya A 37(1):1–71
DasGupta A (2008) Asymptotic theory of statistics and probability. Springer Science Media, LLC
Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Philos Soc 22:700–725
Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630; 180:171–197
Kullback S (1978) Information theory and statistics. Gloucester, Peter Smith, MA
Lindley DV (1956) On the measure of information provided by an experiment. Ann Stat 27:986–1005
Pardo L (2006) Statistical inference based on divergence measures. Chapman and Hall, New York
Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. KTK
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423 and 623–656
Wald A (1947) Sequential analysis. Dover, New York

Statistics and Climate Change

Implication of Statisticians and Statistics Education

Parinbanu Kurji
Head of Biometry, Faculty of Agriculture
University of Nairobi, Nairobi, Kenya

What Does Climate Change Hold for the Future?

There is general agreement among experts that we can expect a rise in temperatures and an increase in the number of extreme events, but for other climate variables such as rainfall there is no clear prediction. However there does not seem to be any doubt that communities coping with poverty will be particularly vulnerable – this means developing countries like Africa will be the hardest hit (Cooper et al. 2006; Washington et al. 2006; Climate Proofing Africa, DFID 2005; Burton and van Aaist 2004). The climate change dialogue brings with it an enormous need for more and better climate data and greater rigor in its analysis. To understand both risks and opportunities associated with the season-to-season variability that is characteristic of current climates as well as changes in the nature of that variability due to climate change, there is need for all stakeholders, including the statistical community, policy makers, and scientists, to work together to propose appropriate strategies to counteract one and enhance the other. Such strategies must be based on scientific studies of climate risk and trend analyses and not fashionable perceptions or anecdotal evidence. Statisticians have a vital role to play here.

What Is Needed?

One of the ways of approaching this issue of climate change as it affects the people in the developing countries is through a better understanding of the season-to-season variability in weather that is a defined characteristic of current climate (*Climate: The statistical description in terms of means and variability of key weather parameters for a given area over a period of time – usually at least 30 years*) and using this to address future change. Managing current climate-induced risk is already an issue for farmers who practice rain-fed agriculture. Helping them to cope better with this risk while preparing for future change seems to be the best way of supporting the needy both for the present and for the future. Agriculture is one field where

the vagaries of climate have an impact but other fields such as health, construction, and transport among others would benefit equally from this approach.

Why Do Statisticians Need to Be Involved?

Meteorology departments are the custodians of climate data and, especially in many developing countries, data quality and management, rather than analysis, have been priority issues and the institutions have limited themselves mainly to providing data to users. There is now a move to shift from providing basic data and services to meeting increasingly challenging user needs.

Effective use of climatic summaries and especially applications require an understanding of statistical concepts underlying these summaries as well as proficiency in using and interpreting the advanced statistical techniques and models that are being suggested to understand climate change.

Statistics is the glue that brings the different disciplines together and statisticians need to form an integral part of multidisciplinary teams to understand, extend, and share knowledge of existing and upcoming technologies and statistical methods for development purposes.

Where Should Changes Occur?

The three areas where statisticians can be proactive in addressing the climate change issue are:

1. Working actively with researchers in various disciplines in guiding research to develop and test adaptation strategies.

   For example, if, as is expected, temperatures are going to rise, and this affects crop growth, it is now that research agendas must be set if we are to meet the new challenges. There needs to be a clear understanding about the implications of such conditions.
2. Being aggressively involved in building capacities of data producers and data users. At present the capacity in many developing countries for modeling and interpreting data is highly inadequate

   For example, creating awareness of the need for quantity, quality, and timeliness of climate data required for use in modeling climate processes and for using and extending these models in collaboration with agriculture scientists and extension workers.
3. Promoting changes in statistics training at all levels to meet the expanding needs.

   For example, innovative statistics curriculum at universities & colleges that mainstream climate data analysis and that emphasize understanding and application of concepts using a data-based approach.

Some Available Resources

Given the availability and affordability of computers today, they should now form an integral part of good statistics training. Among the many resources available to enhance statistics training in general, and training in climatic statistics in particular are:

- *CAST for Africa* (www.cast.massey.ac.nz), an electronic statistics textbook that provides an interesting interactive way of understanding statistical concepts with a number of real-life data sets from different disciplines. Climate CAST, which is an offshoot of this, provides the slant for exploring climatic data. The textbook goes from the very basic to reasonably complex topics.
- *Instat* (www.reading.ac.uk/ssc), a simple software package with a special climate menu and a number of useful guides in the help section to facilitate training as well as self study.
- *GenStat* (www.vsni.co.uk), a major statistical package, is an all-embracing data analysis tool, offering ease of use through comprehensive menu system reinforced with the flexibility of a sophisticated programming language. It has many useful facilities including analysis of extremes. The discovery version is provided free for nonprofit organizations while the latest version is available at very reasonable rates to training and research institutions. Here again there is wealth of information for the user in terms of guides, including a guide to climatic analyses, and tutorials and examples from diverse fields.
- *DSSAT* (www.icasa.net/dssat) and *ApSim* (www.apsim.info/apsim), crop simulation models, driven by long-term daily climatic data, which can be used to simulate realistic long-term field experiments. These are probably more useful at postgraduate or faculty levels but have great potential for statisticians working with agriculture scientists to explore possible scenarios without actually undertaking long costly field experiments.

Some Working Initiatives

- *Statistics Curriculum, at Faculty of Agriculture, University of Nairobi, Kenya*

  An innovative data-based problem-solving approach to service teaching for the Agriculture Faculty uses building blocks approach – from descriptive to modeling to application – to broaden and deepen

the students' understanding of how statistics is used in practice. The curriculum includes computer proficiency and soft skills as an integral part of the curriculum and exposes students to all types of data, including climatic data, which is not only important in its own right but also an important example of monitoring data. Examples of how climatic analyses have been incorporated into the service teaching of statistics are given by Kurji and Stern (2005).

- *Masters in Climate Data Analysis, at Science Faculty, Maseno University, Kenya*

  Currently there are a number of students who are working on their postgraduate degree with specific climate-related projects, both advancing the science, encouraging statisticians to embrace the new challenges of development, and building capacity in the field of climate analysis.
- *Statistics for Applied Climatology (SIAC) at IMTR (Institute of Meteorological Training & Research), Kenya*

  This is a regional program run by the Institute for groups comprising officers from National Met services and Agriculture Research Scientists to develop statistical skills and build networks for further collaborative work. The course has two components, a 6-week e-learning course followed by a 4-week face-to-face course, which culminates in a project that can be continued after the participants return to their bases.

Cross References

►Agriculture, Statistics in
►Mathematical and Statistical Modeling of Global Warming
►Role of Statistics
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistics Education

References and Further Reading

Burton I, van Aaist M (2004) Look before you leap: a risk management approach for incorporating climate change adaptation into World Bank Operations. World Bank Monograph, Washington (DC), DEV/GEN/37 E. 10

Cooper PJM, Dimes J, Rao KPC, Shapiro B, Shiferaw B, Twomlow S (2006) Coping better with current climatic variability in the rain-fed farming systems of sub-Saharan Africa: a dress rehearsal for adapting to future climate change? Global theme on agro-ecosystems Report no. 27. International Crops Research Institute for the Semi-Arid Tropics, PO Box 29063-00623, Nairobi, Kenya, 24pp

DFID (2005) Climate proofing Africa: climate and Africa's development challenge. Department for International Development, London

Kurji P, Stern RD (2005) Teaching statistics using climatic data. http://www.ssc.rdg.ac.uk/bucs/MannafromHeaven.pdf

Washington R, Harrison M, Conway D, Black E, Challinor A, Grimes D, Jones R, Morse A, Kay G, Todd M (2006) African climate change: taking the shorter route. Bull Am Meteorol Soc 87: 1355–1366

Statistics and Gambling

Kyle Siegrist
Professor
University of Alabama in Huntsville, Huntsville, AL, USA

Introduction

Statistics can broadly be defined as the science of decision-making in the face of (random) uncertainty. Gambling has the same definition, except in the narrower domain of a gambler making decisions that affect his fortune in games of chance. It is hardly surprising, then, that the two subjects are closely related. Indeed, if the definitions of "game," "decision," and "fortune" in the context of gambling are sufficiently broadened, the two subjects become almost indistinguishable.

Let's review a bit of the history of the influence of gambling on the development of probability and statistics. First, of course, gambling is one of the oldest of human activities. The use of a certain type of animal heel bone (called the *astragalus*) as a crude die dates to about 3500 BCE (and possibly much earlier). The modern six-sided die dates to about 2000 BCE.

The early development of probability as a mathematical theory is intimately related to gambling. Indeed, the first probability problems to be analyzed mathematically were gambling problems:

1. *De Mere's problem* (1654), named for Chevalier De Mere and analyzed by Blaise Pascal and Pierre de Fermat, asks whether it is more likely to get at least one six with 4 throws of a fair die or at least one double six in 24 throws of two fair dice.
2. The *problem of points* (1654), also posed by De Mere and analyzed by Pascal and Fermat, asks for the fair division of stakes when a sequence of games between two players (Bernoulli trials in modern parlance) is interrupted before its conclusion.
3. *Pepys' Problem* (1693), named for Samuel Pepys and analyzed by Isaac Newton, asks whether it is more likely to get at least one six in six rolls of a fair die or at least two sixes in 12 rolls of the die.

4. *The matching problem* (1708), analyzed by Pierre-Redmond de Montmort, is to find the probability that in a sequence of card draws, the value of a card is the same as the draw number.
5. *St. Petersburg Paradox* (1713), analyzed by Daniel Bernoulli, deals with a gambler betting on a sequence of coin tosses who doubles his bet each time he loses (and leads to a random variable with infinite expected value).

Similarly, the first books on probability were written by mathematician-gamblers to analyze games of chance: *Liber de Ludo Aleae* written sometime in the 1500s by the colorful Girolamo Cardano and published posthumously in 1663, and *Essay d'Analyse sur les Jeux de Hazard* by Montmort, published in 1708. See David 1998 and Epstein 1977 for more on the influence of gambling on the early development of probability and statistics.

In more modern times, the interplay between statistics and game theory has been enormously fruitful. Hypothesis testing, developed by Ronald Fisher and Karl Pearson and formalized by Jerzy Neyman and Egon Pearson is one of the cornerstones of modern statistics, and has a game-theory flavor. The basic problem is choosing between a presumed null hypothesis and a conjectured alternative hypothesis, with the decision based on the data at hand and the probability of a type 1 error (rejecting the null hypothesis when it's true). Influenced by the seminal work of John von Neumann and Oscar Morgenstern on game theory and economics (von Neumann 1944), the Neyman-Pearson hypothesis-testing framework was extended by Abraham Wald in the 1940s to *statistical decision theory* (Wald 1950). In this completely game-theoretic framework, the statistician (much like the gambler) chooses among a set of possible decisions, based on the data at hand according to some sort of value function. Statistical decision theory remains one of the fundamental paradigms of statistical inference to this day.

Bold Play in Red and Black

Gambling continue to be a source of interesting and deep problems in probability and statistics. In this section, we briefly describe a particularly beautiful problem analyzed by Dubins and Savage (1976). A gambler bets, at even stakes, on a sequence of Bernoulli trials (independent, identically distributed trials) with success parameter $p \in (0,1)$. The gambler starts with an initial fortune and must continue playing until he is ruined or reaches a fixed target fortune. (The last two sentences form the mathematical definition of *red and black*.) On each trial, the gambler can bet any proportion of his current fortune, so it's convenient to normalize the target fortune to 1; thus the space of fortunes is the interval $[0,1]$.

The gambler's goal is to maximize the probability $F(x)$ of reaching the target fortune 1, starting with an initial fortune x (thus, F is the value function in the context of statistical decision theory). The gambler's strategy consists of decisions on how much to bet on each trial. Since the trials are independent, the only information of use to the gambler on a given trial is his current fortune. Thus, we need only consider *stationary, deterministic strategies*. Such a strategy is defined by a *betting function* $S(x)$ that gives the amount bet on a trial as a function of the current fortune x.

Dubins and Savage showed that in the sub-fair case $\left(p \le \frac{1}{2}\right)$, an optimal strategy is *bold play*, whereby the gambler, on each trial, bets his entire fortune or the amount needed to reach the target (whichever is smaller). That is, the betting function for bold play is

$$S(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2} \\ 1 - x, & \frac{1}{2} \le x \le 1 \end{cases}$$

Conditioning on the first trial shows that the value function F for bold play satisfies the functional equation

$$F(x) = \begin{cases} pF(2x), & x \in \left[0, \frac{1}{2}\right] \\ p + (1-p)F(2x-1), & x \in \left[\frac{1}{2}, 1\right] \end{cases} \tag{1}$$

with boundary conditions $F(0) = 0$, $F(1) = 1$. Moreover, F is the unique bounded solution of (1) satisfying the boundary conditions. This functional equation is one of the keys in the analysis of bold play. In particular, the proof of optimality involves showing that if the gambler starts with some other strategy on the first trial, and then plays boldly thereafter, the new value function is no better than the value function with bold play.

Interestingly, as Dubins and Savage also showed, bold play is not the unique optimal strategy. Consider the following strategy: Starting with fortune $x \in \left[0, \frac{1}{2}\right)$, the gambler plays boldly, but with the goal of reaching $\frac{1}{2}$. Starting with fortune $x \in \left(\frac{1}{2}, 1\right]$, the gambler plays boldly, but with the goal of not falling below $\frac{1}{2}$. In either case, if the gambler's fortune reaches $\frac{1}{2}$, he plays boldly and bets $\frac{1}{2}$. Thus, the betting function S_2 for this new strategy is related to the betting function S of bold play by

$$S_2(x) = \begin{cases} \frac{1}{2}S(2x), & 0 \le x < \frac{1}{2} \\ \frac{1}{2}S(2x-1), & \frac{1}{2} < x \le 1 \\ \frac{1}{2}, & x = \frac{1}{2} \end{cases}$$

By taking the three cases $x \in \left[0, \frac{1}{2}\right)$, $x = \frac{1}{2}$, and $x \in \left(\frac{1}{2}, 1\right]$, it's easy to see that that the value function F_2 for strategy S_2 satisfies the functional equation (1). Trivially the boundary conditions are also satisfied, so by uniqueness, $F_2 = F$ and thus S_2 is also optimal.

Once one sees that this new strategy is also optimal, it's easy to construct an entire sequence of optimal strategies. Specifically, let $S_1 = S$ denote the betting function for ordinary bold play and then define S_n recursively by

$$S_{n+1}(x) = \begin{cases} \frac{1}{2} S_n(2x), & 0 \le x < \frac{1}{2} \\ \frac{1}{2} S_n(2x-1), & \frac{1}{2} < x \le 1 \\ \frac{1}{2}, & x = \frac{1}{2} \end{cases}$$

Then S_n has the same value function F as bold play and so is optimal for each n. Moreover, if $x \in (0,1)$ is not a binary rational (that is, does not have the form $\frac{k}{2^n}$ for some k and n), then there exist optimal strategies that place arbitrarily small bets when the fortune is x. This is a surprising result that seems to run counter to a naive interpretation of the law of large numbers.

Bold play in red and black leads to some exotic functions of the type that are not usually associated with a simple, applied problem. The value function F can be interpreted as the distribution function of a random variable X (the variable whose binary digits are the complements of the trial outcomes). Thus F is continuous, but has derivative 0 almost everywhere if $p \ne \frac{1}{2}$ (singular continuous). If $p = \frac{1}{2}$, X is uniformly distributed on $[0,1]$ and $F(x) = x$. If $G(x)$ denotes the expected number of trials under bold play, starting with fortune x, then G is discontinuous at the binary rationals and continuous at the binary irrationals.

Finally, note that when the gambler plays boldly, his fortune process follows the deterministic map $x \mapsto 2x \bmod 1$, until the trial that ends the game (with fortune 0 or 1). Thus, bold play is intimately connected with a discrete dynamical system. This connection leads to other interesting avenues of research (see Pendergrass and Siegrist 2001).

About the Author

Kyle Siegrist is Professor of Mathematics at the University of Alabama in Huntsville, USA. He was Chair of the Department of Mathematical Sciences from 2001 to 2005 and was Editor of the *Journal of Online Mathematics at Its Applications* from 2005 to 2009. He is the author of over 30 journal articles and one book. He is the principle developer of Virtual Laboratories in Probability and Statistics, a web project that has twice received support from the US National Science Foundation.

Cross References

►Actuarial Methods
►Components of Statistics
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Martingales
►Monty Hall Problem: Solution
►Probability Theory: An Outline
►Probability, History of
►Significance Testing: An Overview
►St. Petersburg Paradox
►Uniform Random Number Generators

References and Further Reading

Blackwell D, Girshick MA (1979) Theory of games and statistical decisions. Dover, New York

David FN (1998) Games, gods and gambling, a history of probability and statistical ideas. Dover, New York

Dubins LE, Savage LJ (1976) Inequalities for stochastic processes (how to gamble if you must). Dover, New York

Epstein RA (1977) The theory of gambling and statistical logic. Academic, New York

Pendergrass M, Siegrist K (2001) Generalizations of bold play in red and black. Stoch Proc Appl 92

Savage LJ (1972) The foundations of statistics. Dover, New York

von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton

Wald A (1950) Statistical decision functions. Wiley, New York

Statistics and the Law

Mary W. Gray
Professor
American University, Washington DC, USA

The role of the statistician in litigation has much in common with that of a consultant in any field. To be an effective expert witness, we should be certain that we know what questions must be answered and what data will be required in order to answer them. Other guidelines include

- Promoting and preserving the confidence of the client and the public without exaggerating the accuracy or explanatory power of the data
- Avoiding unrealistic expectations and not promising more than you can deliver
- Being responsible and accountable, guarding your reputation

- Providing adequate information to permit methods, procedures, techniques, and findings to be assessed
- Addressing rather than minimizing uncertainty

However, the statistician must understand that litigation is an adversarial process; one must consider the strategy of the other side and be prepared for what is likely to be presented. The keys to effective statistical evidence are

- Early involvement by the statistician (as is the case in any situation)
- Adequate data
- Clarity of presentation
- Effective supplemental anecdotal evidence (not the task of the statistician, but an important complement to it)
- Understanding of the statistics by the litigator
- Recognizing that the statistician cannot reach legal conclusions nor can s/he be an advocate (for anything other than statistics!)

In the United States statistical evidence has been used in cases involving

- Race, sex, and age discrimination in employment and education
- Evidence-based medicine
- Environmental effects of business practices
- DNA, ear prints, bullet composition
- Death penalty
- Product liability
- Intellectual property and many other issues
- On the international scene, statistical evidence was used in the war crimes trial of Milosevic and in other human rights cases.

The techniques used span the range of statistical methodology from descriptive statistics to *t*-tests to regression (nearly ubiquitous), non-parametric tests, capture-recapture, urn models, change point analysis, multiple systems analysis, Mantel-Hanszel tests to Bayesian techniques (not generally popular with the courts) and a variety of other sophisticated methods. Courts have a great deal of difficulty with the concept of sampling, especially when the sample is very small in comparison with a population. They also often have difficulty in seeing the applicability of statistics to an individual case. For example, evidence that, all else being equal, the death penalty was far more likely to be imposed when the victim was white than when the victim was black, has not kept individuals whose victims were white from being sentenced to death.

An important observation to keep in mind is that an expert with a newly-developed technique may not fare well in court. The usual standard for admission of statistical or other scientific evidence is that

1. The testimony is based upon sufficient facts or data,
2. The testimony is the product of reliable principles and methods, and
3. The witness has applied the principles and methods reliably to the facts of the case

Peer-reviewed publication usually meets the second requirement.

The classic example of the ▶misuse of statistics is in *People v. Collins* (1968), where the following analysis sent Malcolm Collins to prison. Witnesses reported various characteristics, characteristics that Malcolm and Janet Collins had, and the prosecutor got the expert to agree to certain hypothetical probabilities as follows (expert witnesses can testify about their opinions based on hypotheses).

Characteristic	Probability
Partly yellow automobile	1/10
Man with mustache	1/4
Woman with ponytail	1/10
Blond woman	1/3
Black man with beard	1/10
Interracial couple in a car	1/1000

Then the prosecutor said: the probability of having all of these characteristics is 1/12,000,000, overriding the expert's objection about their lack of independence. He continued: since there are 12,000,000 people in metropolitan Los Angeles Malcolm and Janet Collins must be the only couple with these characteristics and thus the perpetrators of the mugging in question. In addition to the problem with independence, of course, the probability of "more than one given at least one" in a Poisson distribution turns out to be .43, hardly the "beyond a reasonable doubt" required for a criminal conviction. The unfortunate Malcolm spent some time in prison before his conviction was overturned on appeal, as did the Garrett Wilson of

Maryland v. Wilson (2002), where not only was the probability of two children dying of Sudden Infant Death Syndrome similarly miscalculated, but the prosecutor argued not only that there was a low probability that two deaths would occur in one family but that there was a low probability that the defendant was innocent (This is called the "prosecutor's fallacy."). Analogous bad statistics in the UK led to the physician who testified about statistics being stricken from the registry and 250 prior convictions being reviewed. Unfortunately one of the victims of the erroneous testimony, faced with a ruined career as a solicitor, committed suicide when eventually released from prison.

But there are better results: statistics in cases I have worked on helped convince the courts that similarly situated women and men should receive equal pensions and that women's sports teams should be supported in colleges and universities as well as are men's. In the former case a man who had the same accumulation of pension funds in a defined contribution plan as a woman, was getting 15% more in monthly benefits on the stated grounds that (statistically speaking!) women live longer than men. The U.S. law clearly stated that discrimination on the basis of sex in employment-related matters such as pensions was forbidden, but the pension fund administrators insisted that the discrimination was on the basis of longevity, admitting of course that no individual woman could be expected to live long than any individual man. We showed that of a cohort of 1000 men and women at age 65, 7% of the population would be women could be expected to live longer than men with whom they could be matched and 7% of the population would be men who would die young, unmatched by women's early deaths. Hence 86% of the population could be paired up as to age at death – i.e., 86% of the men and women "died at the same age" (for statistical purposes). Thus for 86% of the population, those "similarly situated with respect to longevity," men and women were being treated differently. This together with the fact that, at least at the time (more than 20 years ago) men indulged in more voluntary life-shortening behavior like smoking and drinking to excess and the – what seemed to many – clear statutory mandate of equal treatment, convinced the courts.

In the sports case it was simply that 51% of the undergraduate students at Brown University were women, while only 39% of the student athletes were. The probability of such a disparity were it due to chance was about 1 in a million. Thus the courts found that the distribution of athletes by sex was not "substantially proportionate" to the distribution of students by sex. Statistical significance isn't everything, but in this case it prevented the cancellation of university support for some of the women's teams.

My late husband used to say that mathematics and the law both have axiom systems – it is just that the law's is inconsistent. Sometimes we all feel that way, but statistics can sometimes help bring justice.

About the Author

Dr. Mary W. Gray is a Professor and Chair, Department of Mathematics and Statistics, American University, Washington DC. She is the founding President of the Association for Women in Mathematics and past President of the Caucus for Women in Statistics. She was the Chair of the Department of Mathematics and Statistics (1977–1981, 1983–1985, 2001–2003). In 1976 Dr Gray was elected the second female Vice President of the American Mathematical Society (70 years after Charlotte Scott became the first female Vice President). In 1993 she became Chair of the USA Board of Directors of Amnesty International. She is an Elected member of the International Statistical Institute and a Fellow of the American Statistical Association, the American Association for the Advancement of Science, and the Association for Women in Science. She has authored and co-authored more than 100 papers and 2 books. Professor Gray has received the (U.S.) Presidential Award for Excellence in Science, Technology, Engineering and Mathematics Mentoring, the Lifetime Mentoring Award of the American Association for the Advancement of Science, and three honorary doctorates. Professor Gray has mentored twenty-three students through successful dissertations in mathematics, including fourteen women and eight African-American students. She has lectured throughout the United States, Europe, Latin America and the Middle East. She is a member of the District of Columbia and U.S. Supreme Court Bars. Currently, she is an Associate editor for the *International Journal of Surgery*.

Cross References

►Forensic DNA: Statistics in
►Misuse of Statistics
►Presentation of Statistical Testimony
►Statistical Evidence
►Statistical Significance

References and Further Reading

Asher J, Banks D, Scheuren F (eds) (2008) Statistical methods for human rights. Springer-Verlag, New York

Ball P, Asher J (2002) Statistics and Slobodan: using data analysis and statistics in the war crimes trial of former president Milosevic. Chance 15:17–24

Fienberg S, Kadane JB (1983) The presentation of Bayesian statistical analyses in legal proceedings. The Statistician 32:88–108
Fienberg S (ed) (1989) The evolving role of statistical assessments in the courts. Springer, New York
Fienberg SE, Krislov SH, Straf ML (1995) Understanding and evaluating statistical evidence in litigation. Jurimetrics Journal 36:1–32
Finkelstein MO, Levin B (2001) Statistics for lawyers, 2nd edn. Springer-Verlag, New York
Gastwirth JL (ed) (2000) Statistical science in the courtroom. Springer-Verlag, New York
Gray MW (1993) Can statistics tell us what we do not want to hear? The case of complex salary structures. Stat Sci 8:144–179
Gray MW (1996) The concept of "substantial proportionality" in Title IX athletics cases. Duke J Gender Soc Policy 3:165–185

Statistics Education

Richard L. Scheaffer
Professor Emeritus
University of Florida, Gainesville, FL, USA

Overview

Statistics education at all levels, school, undergraduate, graduate, and in the workplace, has been the subject of much debate over most of the 20th century and into the 21st. Proposals to make statistics a part of everyone's basic education surfaced in the 1930s and 1940s, but gained little traction. World War II forced a renewed emphasis on scientific thinking and statistics gained attention as an essential component of applied science and industrial management. This led to the few existing graduate programs in statistics being expanded and new ones being developed at various universities around the world, a trend that went on for about the subsequent forty years. Some of these programs emphasized application and some theory, but as the need for statistics in many different fields (business, engineering, health sciences, social sciences, to name a few) became essential and the advent of electronic computing made it possible to meet those needs, graduate programs in statistics tended to merge toward a combination of application and theory, a very healthy trend indeed.

During that same period, introductory undergraduate courses were developed, but these courses stayed on the theory track perhaps too long and only since about 1980 have been giving more attention to applications emphasizing data analysis, again with the assistance of ubiquitous computing. Work beyond the introductory course has not kept pace with the need; even today most colleges and universities offer little in the way of undergraduate statistics beyond the basic course.

Although overtures to making statistics a part of the school curriculum were advanced prior to the 1940s, nothing in that arena really took root until the early 1980s as well. Today, there is great debate on the place of statistics in the school curriculum, but most educators agree that it should be included in the broader picture of mathematical sciences to which all school students should be exposed before moving on to college or the workplace.

An enlightened 21st century view of the role of statistics in society was presented quite clearly in a recent article by Hal Varian of Google:

> The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. (The McKinsey Quarterly, January 2009)

This view of the importance of statistics is becoming the predominant one among those affecting education in the mathematical sciences, and it appears that statistics education is on an upward swing as the information age continues.

University Education in Statistics

The American Statistical Association (http://www.amstat.org/) has links to lists that contain information on over 300 college and university programs in statistics around the world. This is a relatively small number, compared to, say, mathematics, and many of the programs are small or highly specialized (►biostatistics, for example). In the United States, the nearly one hundred graduate programs in statistics produced about 410 doctoral degrees and 408 master's degrees in the 2006–2007 academic year. A much smaller number of bachelors degree programs produced about 445 degrees in that same year. These numbers are underestimates, especially at the master's level, as they come from a survey of mathematical science departments conducted by the American Mathematical Society (http://www.ams.org/), but they do give a perspective on the relatively small numbers of degrees awarded in statistics at all levels. Yet, the number of job opportunities in statistics remains large even in times of economic downturn, especially for those with at least a master's degree

in the subject, and the number of degrees awarded lags behind demand.

Enrollments and other details on the undergraduate teaching of statistics in the United States can be found at in the CBMS 2005 Survey: *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States* (http://www.cbmsweb.org/). Details on current thinking in the teaching of statistics at the college level can be found in one of two journals, the *Journal of Statistics Education* (http://www.amstat.org/PUBLICATIONS/JSE/) and the *Statistics Education Research Journal* of the International Association for Statistics Education (IASE) (http://www.stat.auckland.ac.nz/~iase/). The former is directed toward experiences with teaching practices in the classroom, often including useful data sets, while the latter is directed toward research on effective teaching and learning of statistics. A good resource on all aspects of undergraduate statistics education can be found at the Consortium for Advancing Undergraduate Statistics Education (CAUSE) (http://www.causeweb.org/).

School Education in Statistics

The modern era of statistics education at the school level dates from the late 1970s, when the United Kingdom, Australia, New Zealand and Sweden led the way in developing educational programs and materials that were effective in enlisting the interest of school children (as well as their teachers) in data analysis. The journal *Teaching Statistics* (http://ts.rsscse.org.uk/), now a product of the Royal Statistical Society's Center for Statistics Education, was an outcome of those efforts in the UK and still remains a premier source of information on effective teaching of statistics in the schools. These efforts influenced work in the United States that led the National Council of Teachers of Mathematics (NCTM) (http://www.nctm.org/) to place an emphasis on data analysis in their *Principles and Standards for School Mathematics*, first published in 1989 and revised in 2000.

Over the years, national and international assessments of school mathematics have included increasingly larger emphases on data analysis, statistics and probability. In its 2006 framework, the OECD Program for International Student Assessment (PISA) (http://www.pisa.oecd.org/) lists Uncertainty as one of the four main areas of mathematics, along with Space and shape, Change and relationships, and Quantity. There description of this area is enlightening:

> As an overarching idea, *uncertainty* suggests two related topics: data and chance. These phenomena are respectively the subject of mathematical study in statistics and probability. Relatively recent recommendations concerning school curricula are unanimous in suggesting that statistics and probability should occupy a much more prominent place than has been the case in the past. Specific mathematical concepts and activities that are important in this area are collecting data, data analysis and display/visualization, probability and inference.

For the United States, the 2009 framework of the National Assessment of Educational Progress (NAEP) (http://www.nagb.org/publications/frameworks/math-framework09.pdf) gives data analysis, statistics and probability 25% of the weight of questions at the high school level, in connection with number properties (10%), measurement and geometry (30%) and algebra (35%).

As to content emphases, *the Guidelines for Assessment and Instruction in Statistics Education* (GAISE) (http://www.amstat.org/education/gaise/) report of the American Statistical Association has been instrumental in shaping the revision of mathematics standards for many states and some other countries. GAISE views statistics as a problem-solving process built around the steps of:

- Formulate questions
- Collect data
- Analyze data
- Interpret results

Its guiding principles for teaching statistics are:

- Conceptual understanding takes precedence over procedural skill.
- Active learning is key to the development of conceptual understanding.
- Real-world data must be used wherever possible in statistics education.
- Appropriate technology is essential in order to emphasize concepts over calculations.
- All four steps of the investigative process should be encountered at each grade level.
- The illustrative investigations should show situations in which the statistics is essential to the answering of a question, not just an add-on.
- Such investigations should be tied to the mathematics that they illustrate, motivate and emphasize.

Statistics in the Workplace

As Hal Varian expressed it in the article cited above, "I keep saying the sexy job in the next ten years will be statisticians." There seems to be no end of the demand for

statisticians, or those trained in statistics, so long as they can combine theoretical knowledge and problem-solving skills with the ability to do practical work with data and computers. Another manifestation of the huge need for statistical knowledge lies in the area of productivity and product improvement in industry, as reflected by the interest and excitement that surrounds the Six Sigma program. (See the American Society for Quality, Six Sigma program at http://www.asq.org/learn-about-quality/six-sigma/overview/overview.html.)

Statistics has a bright future, and statistics education must expand and adapt to meet the increasing needs of a world economy that runs on data.

About the Author

Dr. Richard Scheaffer is Professor Emeritus in statistics at Department of Statistics, Florida State University. He was Chairman of the Department for a period of 12 years. He has published numerous papers in the statistical literature and is co-author of five textbooks covering aspects of sampling, probability and mathematical statistics. In recent years, he focused on statistics education throughout the school and college curriculum. He was one of the developers of the Quantitative Literacy Project in the United States that formed the basis of the data analysis emphasis in the mathematics curriculum standards recommended by the National Council of Teachers of Mathematics. He continues to work on educational projects at the elementary, secondary and college levels, and served as the Chief Faculty Consultant for the Advanced Placement Statistics Program in the United States during its first two years (1997–1998). Dr. Scheaffer is a Fellow and Past President of the American Statistical Association (2001), from whom he has received a Founder's Award.

Cross References

►Business Statistics
►Careers in Statistics
►Data Analysis
►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Rise of Statistics in the Twenty First Century
►Role of Statistics in Advancing Quantitative Education
►Statistical Literacy, Reasoning, and Thinking
►Statistics and Climate Change
►Statistics: Nelder's view

References and Further Reading

American Mathematical Society: http://www.ams.org/
American Society for Quality, Six Sigma
American Statistical Association, Guidelines for Assessment and Instruction in Statistics Education (GAISE): http://www.amstat.org/education/gaise/
Consortium for advancing undergraduate statistics education (CAUSE): http://www.causeweb.org/
http://www.asq.org/learn-about-quality/six-sigma/overview/overview.html
International Association for Statistics Education (IASE), Statistics Education Research Journal: http://www.stat.auckland.ac.nz/~iase/
Journal of Statistics Education: http://www.amstat.org/PUBLICATIONS/JSE/
National Council of Teachers of Mathematics (NCTM): http://www.nctm.org/
Statistical abstract of undergraduate programs in the mathematical sciences in the United States: http://www.cbmsweb.org/
Teaching statistics: http://ts.rsscse.org.uk/
Conference Board of the Mathematical Sciences (CBMS) 2005 Survey
National Assessment of Educational Progress (NAEP), Mathematics framework for 2009: http://www.nagb.org/publications/frameworks/math-framework09.pdf
OECD Programme for International Student Assessment (PISA), A framework for PISA 2006: http://www.pisa.oecd.org/

Statistics of Extremes

Anthony C. Davison
Professor
Ecole Polytechnique Fédérale de Lausanne,
EPFL-FSB-IMA-STAT, Lausanne, Switzerland

Introduction

Statistics of extremes concerns the occurrence of rare events: catastrophic flooding due to very high tides or landslides following unusually heavy rain, structural failure of dams and bridges, massive earthquakes, stock market crashes, and so forth. It has applications in many domains of engineering, in meteorology, hydrology and other earth sciences, in telecommunications, in finance and insurance – indeed, in any domain in which major risks arise due to unusual events or combinations thereof. In applications the available data are often very limited in relation to the event of interest, so a key issue is the validity of extrapolation far into the tail of a distribution, based on data that are less extreme. This is usually formulated mathematically in terms of stability properties that reasonable models ought to possess, and these properties place strong restrictions on the families of distributions on

which extrapolation should be based. The relevance of such properties to an application must be carefully considered, and any relevant subject-matter knowledge incorporated, if wholly inappropriate extrapolation is to be avoided.

Maxima

Consider the maximum $M_k = \max(X_1, \ldots, X_k)$ of independent identically distributed continuous random variables $X_1, \ldots, X_k$ from a distribution F whose upper support point is $x_{\max} = \sup\{x : F(x) < 1\}$. In analogy with the central limit theorem (see ▶Central Limit Theorems), we seek a useful limiting distribution for M_k as $m \to \infty$. The distribution function of M_k is $F^k(x)$, but this converges to a degenerate distribution putting unit mass at $x_{\max}$, so instead we consider the sequence of linearly rescaled maxima $Y_k = (M_k - b_k)/a_k$ for $b_k \in \mathbb{R}$ and $a_k > 0$, and ask whether the sequences $\{a_k\}, \{b_k\}$ can be chosen so that a non-degenerate limiting distribution exists. Remarkably it can be shown that if such a limit exists, it must lie in the generalized extreme-value family

$$H(y) = \exp\left\{-\left[1 + \xi\left(\frac{y-\eta}{\tau}\right)\right]_+^{-1/\xi}\right\}, \quad -\infty < \eta, \xi < \infty, \tau > 0, \tag{1}$$

where $x_+ = \max(x, 0)$. This result, known as the extremal types theorem, provides strong motivation for the use of (1) when modeling maxima, in analogy with the use of the Gaussian distribution for averages. Note however the conditional nature of the theorem: there is no guarantee that such a limiting distribution will exist in practice. The connection with the stability properties mentioned above is that (1) is the entire class of so-called max-stable distributions, i.e., those satisfying the natural functional stability relation $H(y)^m = H(b_m + a_m y)$ for suitable sequences $\{a_m\}, \{b_m\}$ for all $m \in \mathbb{N}$.

The parameters η and τ in (1) are location and scale parameters. The shape parameter ξ plays a central role, as it controls the behavior of the upper tail of the distribution H. Taking $\xi > 0$ gives distributions with heavy upper tails and taking $\xi < 0$ gives distributions with a finite upper endpoint, while the Gumbel distribution function $\exp\{-\exp[-(y-\eta)/\tau]\}$ valid for $-\infty < y < \infty$ emerges as $\xi \to 0$. Fisher and Tippett (1928) derived these three classes of distributions, which are known as the Gumbel or Type I class when $\xi = 0$, the Fréchet or Type II class when $\xi > 0$, and the (negative or reversed) Weibull or Type III class when $\xi < 0$. The appearance of the ▶Weibull distribution signals that there is a close link with reliability and with survival analysis, though in those contexts the behavior of minima is typically the focus of interest. Since $\min(X_1, \ldots, X_k) = -\max(-X_1, \ldots, -X_k)$, results for maxima may readily be converted into results for minima; for example, the extremal types theorem implies that if a limiting distribution for linearly rescaled minima exists, it be of form $1 - H(-y)$. Below we describe the analysis of maxima, but the ideas apply equally to minima.

Application

A typical situation in environmental science is that n years of daily observations are available, and then it is usual to fit the generalized extreme-value distribution (1) to the n annual maxima, effectively taking $k = 365$ and ignoring any seasonality or dependence in the series. The fitting is typically performed by maximum likelihood estimation or by Bayesian techniques. The method of moments is generally quite inefficient relative to maximum likelihood because (1) has a finite rth moment only if $r\xi < 1$. Often in environmental applications it is found that $|\xi| < 1/2$, but in financial applications second and even first moments may not exist. Probability weighted moments fitting of (1) is quite widely performed by hydrologists, but unlike likelihood estimation, this method is too inflexible to deal easily more complex settings, for example trend in location or censored observations.

The parameters of (1) are rarely the final goal of the analysis, which usually focuses on quantities such as the $1/p$-year return level, i.e., the level exceeded once on average every $1/p$ years; here $0 < p < 1$. The quantity $1/p$ is known as the return period and is important in engineering design. The usual return level estimate is the $1-p$ quantile of (1),

$$y_{1-p} = \eta + \frac{\tau}{\xi}\left\{[-\ln(1-p)]^{-\xi} - 1\right\},$$

with parameters replaced by estimates. Analogous quantities, the value at risk and expected shortfall, play a central role in the regulation of modern financial markets. Two major concerns in practice are that inference is often required for a return period much longer than the amount of data available, i.e., $np \ll 1$, and that the fitted distribution is very sensitive to the values of the most extreme observations; these difficulties are inherent in the subject.

Threshold Exceedances

The use of annual maxima alone seems to be wasteful of data: much sample information is ignored. A potentially more efficient approach may be based on the following characterization. Let $X_1, \ldots, X_{nk}$ be a set of nk independent identically distributed random variables, and consider the planar point pattern with points at (x, y) coordinates $(j/(nk+1), a_k(X_j - b_k))$, $j = 1, \ldots, nk$.

Then provided a_k and b_k are chosen so that the limiting distribution for $(M_k - b_k)/a_k$ as $k \to \infty$ is given by expression (1), the empirical point pattern above a high threshold t will converge to a nonhomogeneous Poisson process (see ►Poisson Processes) with measure

$$\Lambda\{(x_1, x_2) \times (u, \infty)\} = \exp\left[-n(x_2 - x_1)\left(1 + \xi\frac{u-\eta}{\tau}\right)_+^{-1/\xi}\right], \quad 0 < x_1 < x_2 < 1, u > t. \tag{2}$$

A variety of results follow. For example, on noting that the rescaled maximum of k observations, M_k, is less than $y > t$ only if there are no points in the set $(0, 1/n) \times (y, \infty)$, (2) immediately gives (1). The model (2) shows that if N observations, $y_1, \ldots, y_N$, exceed a threshold $u > t$ over a period of n years, their joint probability density function is

$$\exp\left[-n\left(1 + \xi\frac{u-\eta}{\tau}\right)_+^{-1/\xi}\right]\prod_{j=1}^{N}\frac{1}{\tau}\left(1 + \xi\frac{y_j - \eta}{\tau}\right)_+^{-1/\xi-1},$$

which can be used as a likelihood for η, τ, and ξ. Maximum likelihood inference can be performed numerically for this point process model (see ►Point Processes) and regression models based on it. A popular and closely related approach is the fitting of the generalized Pareto distribution

$$\Pr(X \le t + y \mid X > t) = G(y) = 1 - (1 + \xi y/\tau)_+^{-1/\xi}, \quad y > 0; \tag{3}$$

to the exceedances over the threshold t. As $\xi \to 0$ expression (3) becomes the exponential distribution with mean τ, which here occupies the same central role as the Gumbel distribution for maxima. The distribution (3) has the stability property that if $X \sim G$, then conditional on $X > u$, $X - u$ also has distribution G, but with parameters ξ and $\tau_u = \tau + u\xi$. The conditioning in (3) appears to remove dependence on the location parameter η, but this is illusory because the probability of an exceedance of t must be modeled in this setting.

One important practical matter is the choice of threshold t. Too high a value for t will result in loss of information about the process of extremes, while too low a value will lead to bias because the point process model applies only asymptotically for high thresholds. The value of t is usually chosen empirically, by calculating parameter estimates and other quantities of interest for a number of thresholds and choosing the lowest above which the results appear to be stable. In practice the threshold exceedances are typically dependent owing to clustering of rare events, and this is usually dealt with by identifying clusters of exceedances, and fitting (3) to the cluster maxima, a procedure that may be justified using the asymptotic theory.

Dependence

The discussion above has assumed that the data are independent, but this is rare in practice. Fortunately there is a well-developed probabilistic theory of extremes for stationary dependent continuous time series. To summarize: under mild conditions on the dependence structure, the limiting distribution (1) again emerges as the limit for the maximum, but with a twist. Suppose that $X_1, \ldots, X_k$ are consecutive observations from such a series, that $X_1^*, \ldots, X_k^*$ are independent observations with the same marginal distribution, F, and that M_k and M_k^* are the corresponding maxima. Then it turns out that there exist sequences $\{a_k\}$ and $\{b_k\}$ such that $\left(M_k^* - b_k\right)/a_k$ has limiting distribution H if and only if $(M_k - b_k)/a_k$ has limiting distribution H^θ, where the parameter $\theta \in (0, 1]$ is known as the extremal index (Leadbetter et al. 1983). This quantity has various interpretations, the most direct being that θ^{-1} is the mean size of the clusters of extremes that appear in dependent data. The case $\theta = 1$ corresponds to independence but also covers many other situations: for example, Gaussian autoregressions of order p also have $\theta = 1$. This raises a general problem in the statistics of extremes, that of the relevance of asymptotic arguments to applications: this result indicates that extremely rare events will occur singly, but for levels of interest, there may be appreciable clustering that must be modeled.

Further Reading

The probabilistic basis of extremes is discussed from different points of view by Galambos (1987), Resnick (1987) and de Haan and Ferreira (2006), and Resnick (2006) discusses the closely related topic of heavy-tailed modeling. A historically important book on statistics of extremes is Gumbel (1958). Coles (2001) and Beirlant et al. (2004) give modern accounts, the former focusing exclusively on modeling using likelihood methods, and the latter taking a broader approach. Embrechts et al. (1997) give an discussion oriented towards finance, while Castillo (1988) is turned towards applications in engineering; as mentioned above there is a close connection to the extensive literature on survival analysis and reliability modeling. The essays in Finkenstädt and Rootzén (2004) provide useful overviews of various topics in extremes.

One important topic not discussed above is multivariate extremes, such as the simultaneous occurrence of rare events in many financial time series, or environmental events such as heatwaves or severe rainstorms. Much current research activity is devoted to this domain, which has

obvious implications for ►risk analysis and management. In addition to the treatments in the books cited above, Kotz and Nadarajah (2000) provide extensive references to the early literature on multivariate extremes. Balkema and Embrechts (2007) take a more geometric approach.

The journal *Extremes* (http://www.springer.com/statistics/journal/10687) provides an outlet for both theoretical and applied work on extremal statistics and related topics.

About the Author

Professor Davison is Editor of *Biometrika* (2008–). He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics, an elected member of the International Statistical Institute, and a Chartered Statistician. Professor Davison has published on a wide range of topics in statistical theory, methods and applications. He has also co-written highly-regarded books, including *Bootstrap Methods and their Application* (with D.V. Hinkley, Cambridge University Press, Cambridge, 1997) and *Statistical Models* (Cambridge University Press, Cambridge, 2003). In 2009, Professor Davison was awarded a *laurea honoris causa* in Statistical Science by the University of Padova, Italy.

Cross References

►Environmental Monitoring, Statistics Role in
►Extreme Value Distributions
►Fisher-Tippett Theorem
►Generalized Extreme Value Family of Probability Distributions
►Generalized Weibull Distributions
►Insurance, Statistics in
►Methods of Moments Estimation
►Point Processes
►Poisson Processes
►Quantitative Risk Management
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistical Modeling of Financial Markets
►Testing Exponentiality of Distribution
►Weibull Distribution

References and Further Reading

Balkema G, Embrechts P (2007) High risk scenarios and extremes. European Mathematical Society, Zürich
Beirlant J, Goegebeur Y, Teugels J, Segers J (2004) Statistics of extremes: theory and applications. Wiley, New York
Castillo E (1988) Extreme value theory in engineering. Academic, New York
Coles SG (2001) An introduction to statistical modeling of extreme values. Springer, New York
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
Finkenstädt B, Rootzén H (eds) (2004) Extreme values in finance, telecommunications, and the environment. Chapman and Hall/CRC, New York
Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proc Camb Philos Soc 24:180–190
Galambos J (1987) The asymptotic theory of extreme order statistics, 2nd edn. Krieger, Melbourne, FL
Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York
Kotz S, Nadarajah S (2000) Extreme value distributions: theory and applications. Imperial College Press, London
Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York
Resnick SI (1987) Extreme values, regular variation and point processes. Springer, New York
Resnick SI (2006) Heavy-tail phenomena: probabilistic and statistical modeling. Springer, New York

Statistics on Ranked Lists

Michael G. Schimek
Professor
Medical University of Graz, Graz, Austria

Introduction

In various fields of application, we are confronted with lists of distinct objects in rank order because we can always rank objects according to their position on a scale. When we have variate values (interval or ratio scale), we might replace them by corresponding ranks. In the latter case, there is a loss of accuracy but a gain in generality. The ordering might be due to a measure of strength of evidence or to an assessment based on expert knowledge or a technical device. Taking advantage of the generality of the rank scale, we are in the position of ranking objects which might otherwise not be comparable across lists, for instance, because of different assessment technologies or levels of measurement error. This is a direct result of the fact that rankings are invariant under the stretching of the scale.

In this article, we focus primarily on statistics for two ranked lists comprising all elements of a set of objects (i.e., no missing elements). Due to limited space, we will not discuss methods for m lists in detail but give an example at the end and some references. Let us assume two (but it could be

up to m) assessors, one of which ranks N distinct objects according to the extent to which a particular attribute is present. The ranking is from 1 to N, without ties. The other assessor also ranks the objects from 1 to N. Historically, the goal of rank order statistics was to have a handle that allows the avoidance of the difficulty of setting up an objective scale in certain applications such as in psychometrics. It all started about 100 years ago with seminal work of the psychologist and statistician Charles E. Spearman (1863–1945) aiming at a measure of association between ranked lists. Nowadays, there are four primary tasks when analyzing rank scale data: (1) measuring association between ranked lists, (2) measuring distance between ranked lists, (3) identification of significantly overlapping sublists (estimation of the point of degeneration of paired rankings into noise), and (4) aggregation of ranked full lists or sublist.

Association between Ranked Lists

Suppose we have $N = 10$ major cities ranked according to a measure of air pollution (e.g., particulate matter) and the prevalence of respiratory disease (Table 1).

We are interested in the degree of association between these two rankings representing air pollution and disease prevalence. Such a measure of association is the Kendall's τ coefficient (Kendall 1938, 1942). Let us consider any pair of objects (o_i, o_j). Is the pair in direct order, we score for this pair +1, is it in inverse order, we score for this pair –1. Then the scores obtained for the two lists for a fixed pair of objects are multiplied, giving a common score. This procedure is performed for all $\frac{1}{2}N(N-1)$ possible pairs (45 in this example). Finally, the total of the positive scores, say P, and of the negative scores, say Q, is calculated. The overall score $S = P + Q$ is divided by the maximum possible score (the value that S takes when all rankings are identical). This heuristic procedure defines the τ coefficient which in our example is $\tau = 0.644$. A zero value would indicate independence (no association). τ takes 1 for complete agreement and –1 for complete disagreement. In practice, there are more efficient ways to calculate τ. The coefficient can be interpreted as a measure of concordance between two sets of N rankings (P is the number of concordant pairs, Q of discordant pairs, and S is the excess of concordant over discordant pairs) as well as a coefficient of disarray (minimum moves necessary to transform the second list into the natural order of the first one by successively interchanging pairs of neighbors).

Statistics on Ranked Lists. Table 1 Example of two rankings for ten cities ordered according to pollution rank

City (object)	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
Pollution	1	2	3	4	5	6	7	8	9	10
Disease	3	1	2	5	8	6	4	9	7	10

Another famous measure of association is Spearman's ρ, also called rank correlation coefficient (Spearman 1904). Let d_i be the difference between the ranks in the two lists for object o_i (for the N objects these differences sum to zero). The coefficient is of the form

$$\rho = 1 - \frac{6\sum_i d_i^2}{N^3 - N}. \quad (1)$$

When two rankings are identical, it follows from (1) that $\rho = 1$, in the case of reverse order we have $\rho = -1$ (in our example $\rho = 0.818$). Q, the total of the negative scores for Kendall's τ coefficient, is equivalent to the number of pairs which occur in different orders in the two lists forming so-called inversions. Thus τ is a linear function of the number of inversions and ρ can be interpreted as a coefficient of inversion when each inversion is weighted. If a pair of ranks (i, j) is inverted $(i < j)$, we score $(j - i)$ for any inversion, then the sum of all such scores totals to V. One can show that

$$\rho = 1 - \frac{12V}{N^3 - N},$$

where V can also be expressed as $\frac{1}{2}\sum_i d_i^2$.

A detailed account of rank correlation methods summarizing the classical literature up to 1990 can be found in Kendall and Gibbons (1990). Around that time there was little interest in procedures for ranked data, some of them, like Spearman's L_1-based footrule (Spearman 1906), were almost unknown in the statistical community because of technical and computational shortcomings, as well as a lack of relevance for common applications. Most recently, there has been a dramatic shift in relevance because of emerging technologies producing huge amounts of ranked lists, such as Web search engines offering selected server-based information and high-throughput techniques in genomics providing insight into gene expression. These and others have given rise to new developments concerning the statistical handling of rank scale information. An essential aspect is the measurement of distance between ranked lists.

Distance between Ranked Lists

The most popular distance measure is Kendall's τ intrinsic to his already introduced measure of association. It is equal to the number of adjacent pairwise exchanges required to convert one ranking to another. Let us have two

permutations τ and τ' of a set O of objects. Then Kendall's τ distance is given by

$$K(\tau, \tau') = \sum_{\{i,j\} \in O} K_{i,j}(\tau, \tau'),$$

where $K_{i,j}(\tau, \tau')$ takes 0 if the orderings of the ranks of objects i and j agree in the two lists and otherwise 1. Its maximum is $\frac{1}{2}N(N-1)$ where N is the list length.

An alternative measure of distance is Spearman's footrule (related to the Manhattan distance for variate values). Let us again assume two permutations τ and τ' of a set O of objects. Spearman's footrule distance is the sum of the absolute differences between the ranks of the two lists over the N elements in O,

$$S(\tau, \tau') = \sum_{i=1}^{N} |R_\tau(o_i) - R_{\tau'}(o_i)|,$$

where $R_\tau(o_i)$ is the rank of object o_i in list τ, and $R_{\tau'}(o_i)$ in list τ', respectively. As can be seen from the above formulae, Spearman's footrule takes the actual rankings of the elements into consideration, whereas, in Kendall's τ only relative rankings matter. The maximum Spearman's distance is $\frac{1}{2}|N|^2$ for N even, and $\frac{1}{2}(|N| + 1)(|N| - 1)$ for N odd, which corresponds to the situation in which the two lists are exactly the reverse of each other.

For a mathematical theory of distance measures, we refer to Fagin et al. (2003). Recent developments as well as novel applications are discussed in Schimek et al. (2011).

Degeneration of Rankings into Noise

Typically, when the number N of objects is large or even huge, it is unlikely that consensus between two rankings of interest prevails. Only the top-ranked elements might be relevant. For the remainder objects their ordering is more or less at random. This is not only true for surveys of consumer preferences but also for many other applications of topical interest such as the ▶meta-analysis of gene expression data from several laboratories. In many instances, we observe a general decrease of the probability for consensus rankings with increasing distance from the top rank position. Typically, there is reasonable conformity in the rankings for the first, say k, elements of the lists, motivating the notion of *top-k ranked lists*.

The statistical challenge is to identify the length of the top list. So far, heuristics have been used in practice to specify k. Recently Hall and Schimek (2010) could derive a moderate deviation-based inference procedure for random degeneration in paired ranked lists. The result is an estimate $\hat{k}$ for the length of the so-called partial (top-k) list. Such an inference procedure is not straightforward since the degree of correspondence between ranked lists (full or partial) is not necessarily high, due to various irregularities of the assessments.

Let us define a sequence of indicators, where $I_j = 1$ if the ranking given by the second assessor to the object ranked j by the first assessor, is not distant more than δ index positions from j, and otherwise $I_j = 0$. Further, let us assume (1) independent Bernoulli random variables $I_1, \ldots, I_N$, with $p_j \geq \frac{1}{2}$ for each $j \leq j_0 - 2$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$ for $j \geq j_0$; (2) a general decrease of p_j for increasing j that does not need to be monotone. The index j_0 is the point of degeneration into noise and needs to be estimated ($\hat{j}_0 - 1 = \hat{k}$). Then for a pilot sample size v a constant $C > 0$ is chosen such that $z_v \equiv \left(C v^{-1} \log v\right)^{1/2}$ is a moderate-deviation bound for testing the null hypothesis H_0 that $p_k = \frac{1}{2}$ for v consecutive values of k, versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k. In particular, it is assumed that H_0 applies to the v consecutive values of k in the respective series defined by

$$\hat{p}_j^+ = \frac{1}{v} \sum_{\ell=j}^{j+v-1} I_\ell \quad \text{and} \quad \hat{p}_j^- = \frac{1}{v} \sum_{\ell=j-v+1}^{j} I_\ell,$$

where $\hat{p}_j^+$ and $\hat{p}_j^-$ are estimates of p_j computed from the v data pairs I_ℓ for which ℓ lies immediately to the right of j, or immediately to the left of j, respectively. We reject H_0 if and only if $\hat{p}_j^\pm - \frac{1}{2} > z_v$. Under H_0, the variance of $\hat{p}_j^\pm$ equals $(4v)^{-1}$ (this implies $C > \frac{1}{4}$). Taking advantage of this inference procedure, the complex decision problem is solved via an iterative algorithm, adjustable for irregularity in the rankings.

Aggregation of Ranked Lists

The task of rank aggregation is to provide consensus rankings (majority preferences) of objects across lists, thereby producing a conforming subset of objects O^*. The above described inference procedure facilitates rank aggregation because it helps to specify the partial list length k which means a substantial reduction in the associated computational burden. As a matter of fact, list aggregation by means of brute force is limited to the situation where N is unrealistically small. The approach proposed in Lin and Ding (2009) which we describe below, outperforms most of the aggregation techniques so far but for large sets O, the specification of k beforehand remains crucial. It is a stochastic search algorithm that provides an optimal solution, i.e., a consolidated list of objects, for a given distance measure such as Kendall's τ or Spearman's footrule, to be precise, for their penalized versions because of the partial nature of the input lists (for details see Schimek et al. 2011). Lin's and Ding's algorithm is preferable to those that do not aim to optimize any criterion, thus only providing approximate

solutions under unknown statistical properties (examples are Dwork et al. 2001, DeConde et al. 2006).

Let us assume a random matrix $(\mathbf{X})_{N\times k}$ with elements 0 and 1 with the constraints of its columns summing up to 1 and its rows summing up to, at most, 1. Under this setup, each realization of $\mathbf{X}$, x, uniquely determines an ordered list (permutation) of length k by the position of 1's in each column from left to right. Let $\mathbf{p} = (p_{jr})_{N\times k}$ denote the corresponding probability matrix (each column sums to 1). For each column variable, $\mathbf{X}_r = (X_{1r}, X_{2r}, \ldots, X_{Nr})$, a ►multinomial distribution with sample size 1 and probability vector $\mathbf{p}_r = (p_{1r}, p_{2r}, \ldots, p_{Nr})$ is assumed. Then the probability mass function is of the form

$$P_{\mathbf{v}}(x) \propto \prod_{j=1}^{N}\prod_{r=1}^{k}(p_{jr})^{x_{jr}} I\left(\sum_{r=1}^{k} x_{jr} \le 1, 1 \le j \le N; \sum_{j=1}^{N} x_{jr} = 1, 1 \le r \le k\right).$$

Any realization x of $\mathbf{X}$ uniquely determines the corresponding top-k candidate list without reference to the probability matrix $\mathbf{p}$. The idea is to construct a stochastic search algorithm to find an ordering x^* that corresponds to an optimal τ^* satisfying the minimization criterion. Lin and Ding (2009) use a cross-entropy Monte Carlo technique in combination with an Order Explicit algorithm (since the orders of the objects in the optimal list are explicitly given in the probability matrix $\mathbf{p}$). Cross-entropy Monte Carlo is iterating between two steps: a simulation step in which random samples from $P_{\mathbf{v}}(x)$ are drawn, and an update step producing improved samples increasingly concentrating around an x^* corresponding to an optimal τ^*.

Let us finally illustrate the application of the inference procedure together with rank aggregation as outlined in this paper. We simulated $m = 5$ ranked lists τ_j of gene expression data ($N = 60$ genes) from a known central ranking as outlined in DeConde et al. (2006). The length of the top-k list was set to 10. In Table 2, we display the input lists and the output top-k list for $\delta = 10$ and $\nu = 16$, applying the (penalized) Kendall's τ distance. We obtained an estimated $\hat{k} = 8$ instead of the true $k = 10$. Most objects ranked in input position 9 and 10 are displaced due to irregular (random) assignments. Therefore our procedure was short-cutting the top-ranked elements for the sake of clear separation. However, a longer partial list could have been obtained by parameter adaptations in the moderate deviation-based inference procedure. All calculations were carried out with the R package TopKLists of the author and collaborators.

Statistics on Ranked Lists. Table 2 Example of the aggregation of five rankings of $N = 60$ objects (genes) and the consensus top-ranking set of $\hat{k} = 8$ objects

	Input lists					Output list
Rank	τ_1	τ_2	τ_3	τ_4	τ_5	τ^*
1	o_8	o_1	o_{12}	o_4	o_3	o_2
2	o_{10}	o_5	o_{45}	o_{10}	o_7	o_5
3	o_4	o_{37}	o_2	o_6	o_5	o_4
4	o_7	o_4	o_5	o_2	o_{46}	o_6
5	o_{50}	o_6	o_9	o_9	o_2	o_8
6	o_6	o_{20}	o_8	o_3	o_8	o_{10}
7	o_{40}	o_2	o_6	o_7	o_{32}	o_3
8	o_{55}	o_{34}	o_3	o_1	o_{41}	o_7
9	o_{33}	o_{47}	o_{28}	o_{40}	o_{44}	–
10	o_{21}	o_{44}	o_{26}	o_{11}	o_1	–
11	o_{15}	o_{19}	o_{60}	o_{46}	o_9	–
12	o_{14}	o_{57}	o_{38}	o_{16}	o_{55}	–
13	o_{54}	o_{46}	o_1	o_{54}	o_{42}	–
14	o_{13}	o_8	o_{41}	o_{43}	o_{40}	–
15	o_{53}	o_{36}	o_{15}	o_{35}	o_{27}	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	o_{35}	o_{16}	o_{39}	o_{48}	o_{39}	–

About the Author

Professor Michael G. Schimek is Past Vice President of the International Association for Statistical Computing, Member of the International Statistical Institute, Fellow and Chartered Statistician of the Royal Statistical Society, as well as Adjunct Professor of Masaryk University in Brno (Czech Republic).

Cross Reference

►Distance Measures
►Kendall's Tau
►Measures of Dependence
►Moderate Deviations
►Ranks

References and Further Reading

DeConde RP et al (2006) Combined results of microarray experiments: a rank aggregation approach. Stat Appl Genet Mol Biol 5(1):Article 15

Dwork C et al (2001) Rank aggregation methods for the Web. http://www10.org/cdrom/papers/577/

Fagin R, Kumar R, Sivakumar D (2003) Comparing top-k lists. SIAM J Discrete Math 17:134–160

Hall P, Schimek MG (2010) Moderate deviation-based inference for random degeneration in paired rank lists. Submitted manuscript

Kendall M (1938) A new measure of rank correlation. Biometrika 30:91–93

Kendall M (1942) Note on the estimation of a ranking. J R Stat Soc A 105:119–121

Kendall M, Gibbons JD (1990) Rank correlation methods. Edward Arnold, London

Lin S, Ding J (2009) Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics 65:9–18

Schimek MG, Lin S, Wang N (2011) Statistical integration of genomic data. Springer, New York (forthcoming)

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Spearman C (1906) A footrule for measuring correlation. Brit J Psychol 2:89–108

Statistics Targeted Clinical Trials Stratified and Personalized Medicines

Aboubakar Maitournam
University Abdou Moumouni of Niamey, Niamey, Niger

The rapid breakthroughs in genomics-based technologies like DNA sequencing, microarrays for gene expression and mRNA transcript profiling, comparative genomic hybridization (CGH), and mass spectrometry for protein characterization and identification of metabolic and regulatory pathways and networks announce the advent of stratified medicine and its immediate corollary called personalized medicine. Both stratified and personalized medicine are in their infancy. But, they already raise statistical and stochastic modeling challenges partially handled by the growing multidisciplinary field of ▶bioinformatics.

Statistics, Targeted Clinical Trials, and Stratified Medicine

With the actual progress in the burgeoning field of genomic science, most of the common diseases like cancer can be stratified at the molecular level. The aim is to refine disease taxonomies and to allocate patients to molecularly targeted therapy subgroups based on prognostic and predictive biomarkers. This will improve the efficiency of the treatment by adapting it to the patient prognostic profile. However, molecularly targeted therapy benefits only a subset of patients (Betensky et al. 2002). The refinement of the disease classification is based on gene expression transcript profiling, and the prediction of which patients will be more responsive to the experimental treatment than to the control regimen may be based on a molecular assay measuring, for example, expression of targeted proteins.

For stratified medicine, both molecular signatures of patients and of the diseases can be used, firstly for stratification of patients into responder and nonresponder groups and, secondly, in the near future also for individualized therapy. Stratification of patients into responder or nonresponder groups based on theranostics (molecular diagnosis assays) is the basis of stratified medicine. This implies that the first steps toward stratified medicine are randomized clinical trials for the evaluation of molecularly targeted therapy called targeted clinical trials (Simon 2004). Targeted clinical trials have eligibility restricted to patients predicted to be responsive to the molecularly targeted drug.

In a modeling of phase III randomized clinical trials for the evaluation of molecularly targeted therapy, (Maitournam and Simon 2004 and Simon and Maitournam 2004) established that the targeted clinical trial design is more efficient than a conventional untargeted design with broad eligibility. They evaluated relative efficiencies, e_1 and e_2, of the two designs, respectively, with respect to the number of patients required for randomization $\left(e_1 = \frac{n}{n_T}\right)$ and relatively to the number required for screening

$$\left(e_2 = n \Big/ \left(n_T \Big/ \left((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma) \right) \right) \right),$$

where $2n$ is the total number of randomized patients for untargeted design, $2n_T$ is that of targeted design, λ_{spec} and λ_{sens} are the specificity and the sensitivity of the molecular diagnosis assay, and γ is the proportion of not responders in the referral population. Indeed, for untargeted design, n patients are allocated to control group and n other patients to treatment group. Consequently, the total number of randomized patients for untargeted design is $2n$. In the same way, for targeted design the total number of randomized patients is $2n_T$. Thus, the relative efficiencies are respectively

$$e_1 = \frac{2n}{2n_T} = \frac{n}{n_T}$$

and

$$e_2 = \frac{2n}{\left(2n_T/((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma))\right)} = \frac{n}{\left(n_T/((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma))\right)}$$

They derived explicit formulas for calculating the above relative efficiencies, in the case of continuous outcome based on normal mixture, and in the binary case by using the Ury and Fleiss formula. In the continuous case, outcomes are also compared by using a two-sample Wilcoxon test, and in that nonparametric setting relative efficiencies are evaluated by Monte Carlo simulation. Online efficiency calculation for binary case is available at (http://linus.nci.nih.gov/brb/samplesize/td.html).

However, some statistical challenges related to the design of targeted clinical trials remain. For example, analytical expressions of relative efficiencies of targeted versus untargeted clinical trial designs for continuous outcomes are not trivial in the nonparametric and Bayesian settings. Furthermore, the conventional statistical challenges raised by genomics and microarrays (see Simon et al. 2003 and Sebastini et al. 2003) like experimental design, data quality, normalization, choice of data analysis method, correction of multiple hypotheses testing, validation of cluster, and classifier (see Simon et al. 2003 for a comprehensive synthesis) slow the progress of theranostics and subsequently that of targeted clinical trials and stratified medicine. The latter announces the advent of Personalized Medicine.

Statistics and Personalized Medicine

Personalized medicine (Langreth and Waldholz 1999) is in a restrictive and ideal sense, the determination of the right dose at the right time for the right patient or the evaluation of his predisposition to disease by using genomics-based technologies and his genomic makeup. More precisely, personalized medicine relies on patient polymorphic markers like single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTR), short tandem repeats (STRs), and other mutations (Bentley 2004). Personalized medicine is sometimes mistaken as stratified medicine. In fact, stratified medicine is the precursor of personalized medicine.

Personalized medicine is opening huge opportunities for mathematical formalization sketched, for example, for molecular biology of DNA (Carbone and Gromov, 2001). Indeed, the upcoming era of personalized medicine coincides with the actual era of data (Donoho 2000) characterized by massive records of various individual data generated almost continuously. Individual i will thus be identified as a high-dimensional heterogeneous vector $(X_{i1}, \ldots, X_{im})$, where m is an integer, the $X_{ij}, j = 1, \ldots, m$; are deterministic or random qualitative and quantitative variables. The latter are for instance: biometric and genomic fingerprints, family records, age, gender, height, weight, diseases status, diet, medical images, personal medical history, family history, conventional prognostic profiles, and so on.

However, as personalized medicine will rely on huge technological infrastructures, it will generate a lot of data at the individual level. This will lead to enormous problems of:

- Correlation
- Multiple hypotheses testing
- Sensitivity and specificity of molecular diagnosis tools
- Choice of metrics for comparisons between individuals and between individuals and databases
- Integration of heterogeneous data and, subsequently, qualitative and quantitative standardization.

Acknowledgment

The author thanks Dr. Carmen Buchrieser, Senior Researcher at Pasteur Institute, for reviewing the manuscript.

About the Author

Dr. Aboubakar Maitournam held several positions at Pasteur Institute as postdoc and researcher at the interface of Genomics, Imaging, Bioinformatics, and Statistics. He contributed to the publication of *Listeria monocytogenes* genome and to the setting up of statistical methods for analysis of gene expression data at the Genopole of Pasteur Institute. His latest works at National Institute of Health, Biometric Research Branch (Bethesda, USA) focused on the statistical design of targeted clinical trials for the evaluation of molecularly targeted therapies. Currently Dr. Aboubakar Maitournam is a faculty member of department of Mathematics and Computer Sciences (University Abdou Moumouni of Niamey, Niger). Dr. Aboubakar Maitournam was also appointed as Director of Statistics (2008–2010) at the Ministry of Competitiveness and Struggle Against High Cost Life (Niger). Dr. Maitournam is a member of SPAS (Statistical Pan African Society). He contributes regularly to a Nigerien weekly newspaper called *Le Républicain* with papers for general public related to information era, statistical process control and competitiveness, genomics and statistics, mathematics, and society.

Cross References

►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest

►Medical Research, Statistics in
►Monte Carlo Methods in Statistics

References and Further Reading

Bentley DR (2004) Genomes for medicine. Nature 429:440–445
Betensky RA, Louis DN, Cairncross JG (2002) Influence of unrecognized molecular heterogeneity on randomized trials. J Clin Oncol 20(10):2495–2499
Carbone A, Gromov M (2001) Mathematical slices of molecular biology. La Gazette des Mathématiciens, Société Mathématique de France, special edition, 11–80
Donoho D (2000) High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Mémoire, Stanford University
Langreth R, Waldholz M (1999) New era of personalized medicine-targeting drugs for each unique genetic profile. Oncologist 4:426–427
Maitournam A, Simon R (2004) On the efficiency of targeted clinical trials. Stat Med 24:329–339
Sebastini P, Gussoni E, Kohane IS, Ramoni MF (2003) Statistical challenges in functional genomics. Stat Sci 18(1):33–70
Simon R (2004) An agenda for clinical trials: clinical trials in the genomic era. Clin Trials 1:468–470
Simon R, Maitournam A (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10: 6759–6763
Simon RM, Korn EI, McShane LM, Radmacher MD, Wright GW, Zhao Y (2003) Design and analysis of DNA microarray investigations. Springer, New York

Statistics, History of

Oscar Sheynin (assisted by Miodrag Lovric)
Berlin, Germany
Faculty of Economics, University of Kragujevac, Kragujevac, Serbia

Statistics: Origin of that Term

Many authors discussed this, notably Karl Pearson (1978). It is widely believed that the term statistics originated from the Latin *Status* (situation, condition) of population and economics; in late Latin, the same term meant State. Another root of the term comes from the Italian word *stato* (state), and a *statista* (a person who deals with affairs of state). According to Kendall (1960:447) the first use of the word statistics "occurs in a work by an Italian historian Girolamo Ghilini, who in 1589 refers to an account of *civile, politica, statistica e militare scienza*." In 1587 Giovanni Botero described the political structure of several states in his *Della ragione di stato* (English translation 1956) latinized as *De Disciplina status*. Humboldt (1815) wrote "political arithmetic (see Staatswissenschaft and Political Arithmetic) or, in latino-barbare (late Latin), statistics."

None of the above belonged to statistics or statisticians in the modern sense and the same is true for later sources: Shakespeare's *Hamlet* (1601), Helenus Politanus' (1672) *Microscopium statisticum*, and for Hermann Conring's lectures (from 1660, published 1673).

In English, the word *statist* appeared in Shakespeare's *Hamlet*, Act V, Scene 2 (c. 1601), and *Cymbeline*, Act II, Scene 4 (c. 1610), and the word *statistics* was first introduced into English in 1770 by W. Hooper in his translation of J. F. Von Bielfeld's *The elements of universal erudition, Containing an analytical argument of the sciences, polite arts, and belles letters* (3 vols): "The science, that is called *statistics, teaches us what is the political arrangement of all the modern states of the known world*." (vol 3, p 269). The word statistics was used again in this old sense in 1787 by E. A. W. Zimmermann in his book *A Political Survey of the Present State of Europe*. According to Karl Pearson (1978:2), John Sinclair was the first who had attached modern meaning to the word statistics in *The statistical account of Scotland drawn up from the communications of the ministers of the different parishes* (21 vols, 1791–1799).

Staatswissenschaft and Political Arithmetic

The Staatswissenschaft or University statistics was born in Germany in the mid-seventeenth century and a century later Achenwall established its Göttingen school which described various aspects of a given state, mostly without use of numbers. His successor Schlözer (1804:86) coined a pithy saying: *History is statistics flowing, and statistics is history standing still*. His followers adopted it as the definition of statistics (which did not involve studies of causes and effects).

Also during that time political arithmetic had appeared (Graunt, Petty). It widely used numbers and elementary stochastic considerations and discussed causes and relations, thus heralding the birth of statistics. Graunt (1662/1899) stated that it was necessary to know "how many people there be" of each sex, age, religion, trade, etc. (p. 396), provided appropriate estimates (sometimes quite wrongly), especially concerning ►medical statistics. He was able to use sketchy and unreliable statistical data for estimating the population of London and England as well as the influence of various diseases on mortality and attempted to discover regularities in the movement of population. Contradicting the prevailing opinion, he established that both sexes were approximately equally numerous and derived a rough estimate of the sex ratio

at birth (p. 389). Graunt also reasonably noted that mortality from syphilis was underestimated because of moral considerations (p. 356). Graunt doubted, however, that statistical investigations were needed for anyone except the King and his main ministers (p. 397).

He also compiled the first ever mortality table (p. 387); although rather faulty but of great methodological importance, it was applied by Jakob Bernoulli and Huygens.

One of the main subjects of political arithmetic was indeed population statistics, and it certainly confirmed that "In a multitude of people is the glory of a king, but without people a prince is ruined" (Proverbs 14:28). And here is another link between the Old Testament and that new discipline: Moses sent spies to the land of Canaan to find out "whether the people [there] are strong or weak, whether they are few or many, […] whether the land is rich or poor […]" (Numbers 13: 17–20).

Tabular statistics which appeared in the mid-eighteenth century could have served as a link between the two new disciplines, but its representatives were being scorned as "slaves of tables" (Knies 1850:23). However, in the 1680s Leibniz recommended to compile "statistical tables" with or without numbers and wrote several papers belonging to both those disciplines. They were first published in the nineteenth century, then reprinted (Leibniz 1986).

Numerical description of phenomena without studying causes and effects also came into being. The London Statistical Society established in 1834 declared that all conclusions "shall admit of mathematical demonstrations" (which was too difficult to achieve), and stipulated that statistics did not discuss causes and effects (which was impossible to enforce) (see Anonymous 1839). Louis (1825) described the *numerical method* which was actually applied previously. Its partisans (including D'Alembert) advocated compilation of numerical data on diseases, scarcely applied probability, and believed that theory was hardly needed.

A similar attitude had appeared in other natural sciences; the astronomer Proctor (1872) plotted 324 thousand stars on his charts wrongly stating that no underlying theory was necessary. Compilation of statistical yearbooks, star catalogues, etc., can be mentioned as positive examples of applying the same method, but they certainly demand preliminary discussion of data. Empiricism underlying the numerical method was also evident in the Biometric school (The Two Streams of Statistical Thought).

The Staatswissenschaft continued to exist, although in a narrower sense; climate, for example, fell away. At least in Germany it is still taught at universities, certainly includes numerical data, and studies causes and effects. It thus is partly the application of the statistical method to various disciplines and a given state. Chuprov's opinion (1909/1959:50, 1922:339) that the Staatswissenschaft will revive, although with an emphasis on numbers, and determine the essence of statistics was partly wrong: that science did not at all die, neither does it determine statistics.

Statistics and the Statistical Method: The Theory of Errors

Kolmogorov and Prokhorov 1982 defined mathematical statistics as a branch of mathematics devoted to systematizing, processing, and utilizing statistical data, i.e., the number of objects in some totality. Understandably, they excluded the collection of data and their exploratory analysis. The latter is an important stage of theoretical statistics which properly came into being in the mid-twentieth century. Debates about mathematical versus theoretical statistics can be resolved by stating that both data analysis and collection of data only belong to the latter and determine the difference between it and the former.

The first definition of the theory of statistics (which seems to be almost the same as theoretical statistics) worth citing is due to Butte (1808:XI): It is a science of understanding and estimating statistical data, their collection, and systematization. It is unclear whether Butte implied applications of statistics as well. Innumerable definitions of statistics (without any adjectives) had been offered beginning with Schlözer (Staatswissenschaft and Political Arithmetic), but the above suffices, and I only adduce the definition of its aims due to Gatterer (1775:15) which seems partly to describe both political arithmetic and the new Staatswissenschaft (Staatswissenschaft and Political Arithmetic): To understand the state of a nation by studying its previous states.

The statistical method is reasoning based on mathematical treatment of numerical data and the term is mostly applied to data of natural sciences. The method underwent two previous stages. During the first one, statements based on unrecorded general notions were made, witness an aphorism (Hippocrates 1952): Fat men are apt (!) to die earlier than others. Such statements express qualitative correlation quite conforming to the qualitative nature of ancient science.

The second stage was distinguished by the availability of statistical data (Graunt). The present, third stage began by the mid-nineteenth century when the first stochastic criteria for checking statistical inferences had appeared (Poisson, see Sheynin 1978, Sect. 5.2). True, those stages are not really separated one from another: even ancient astronomers had collected numerical observations.

Most important discoveries were made even without such criteria. Mortality from cholera experienced by those whose drinking water was purified was eight times lower than usual (Snow 1855:74–86) which explained the spread

of cholera. Likewise, smallpox vaccination (Jenner 1798) proved absolutely successful.

The theory of errors belongs to the statistical method. Its peculiar feature is the use of the "true value" of the constants sought. Fourier (1826/1890:533–534) defined it as the limit of the arithmetic mean of observations which is heuristically similar to the frequentist definition of probability and which means that residual systematic errors are included in that value.

From its birth in the second half of the eighteenth century (Simpson, Lambert who also coined that term (1765, Sect. 321)) to the 1920s it constituted the main field of application for the probability theory, and mathematical statistics borrowed its principles of maximal likelihood (Lambert 1760, Sect. 303) and least variance (Gauss 1823, Sect. 17) from it (from the theory of errors).

Gauss' first justification of the method of ►least squares (1809) for adjusting "indirect observations" (of magnitudes serving as free terms in a system of redundant linear algebraic equations with unknowns sought and coefficients provided by the appropriate theory) was based on the (independently introduced) principle of maximum likelihood and on the assumption that the arithmetic mean of the "direct observations" was the best estimator of observations. He abandoned that approach and offered a second substantiation (1823), extremely difficult to examine, which rested on the choice of least variance. Kolmogorov (1946) noted in passing that it was possible to assume as the starting point minimal sample variance (whose formula Gauss had derived) – with the method of least squares following at once!

Gauss (1823, Sect. 2) stated that he only considered random errors. Quite a few authors had been favoring this second substantiation; best known is Markov (1899/1951:247) who (p. 246) nevertheless declared that the method of least squares was not optimal in any sense. On the contrary, in case of normally distributed errors it provides jointly efficient estimators (Petrov 1954).

One of the previous main methods for treating indirect observations was due to Boscovich (Cubranic 1961, 1962; Sheynin 1971) who participated in the measurement of a meridian arc. In a sense it led to the median. Already Kepler (Sheynin 2009, Sect. 1.2.4) indirectly considered the arithmetic mean "the letter of the law." When adjusting indirect observations, he likely applied elements of the minimax method (choosing a "solution" of a redundant system of equations that corresponded to the least maximal absolute residual free term) and of statistical simulation: He corrupted observations by small arbitrary "corrections" so that they conform to each other. Ancient astronomers regarded observations as their private property, did not report rejected results, and chose any reasonable estimate. Errors of observation were large, and it is now known that with "bad" distributions the arithmetic mean is not better (possibly worse) than a separate observation.

Al-Biruni, the Arab scholar (10*th*–11*th* cc.) who surpassed Ptolemy, did not yet keep to the arithmetic mean but chose various estimators as he saw fit (Sheynin 1992).

There also exists a determinate theory of errors which examines the entire process of measurement without applying stochastic reasoning and which is related to the exploratory data analysis and experimental design. Ancient astronomers selected optimal conditions for observation, when errors least influenced the end result (Aaboe and De Solla Price 1964). Bessel (1839) found out where should the two supports of a measuring bar be situated to ensure the least possible change of its length due to its weight. At least in the seventeenth century, natural scientists including Newton gave much thought to suchlike considerations. Daniel Bernoulli (1780) expressly distinguished random and systematic errors. Gauss and Bessel originated a new stage in experimental science by assuming that each instrument was faulty unless and until examined and adjusted.

Another example: the choice of the initial data. Some natural scientists of old mistakenly thought that heterogeneous material could be safely used. Thus, the English surgeon Simpson (1847–1848/1871:102) vainly studied mortality from amputations performed in many hospitals during 45 years. On the other hand, conclusions were sometimes formulated without any empirical support. William Herschel (1817/1912:579) indicated that the size of a star randomly chosen from many thousands of them will hardly differ much from their mean size. He did not know that stars enormously differed in size so that their mean size did not really exist and in any case nothing follows from ignorance: *Ex nihilo nihil*!

Jakob Bernoulli, De Moivre, Bayes: Chance and Design

The theory of probability emerged in the mid-seventeenth century (Pascal, Fermat) with an effective introduction of expectation of a random event. At first, it studied games of chance, then (Halley 1694) tables of mortality and insurance, and (Huygens 1699) problems in mortality. Halley's research, although classical, contained a dubious statement. Breslau, the city whose population he studied, had a yearly rate of mortality equal to 1/30, the same as in London, and yet he considered it as a statistical standard. If such a concept is at all appropriate, there should be standards of several levels.

Equally possible cases necessary for calculating chances (not yet probabilities) were lacking in those applications, and Jakob Bernoulli (1713, posthumously) proved

that posterior statistical chances of the occurrence of an event stochastically tended to the unknown prior chances. In addition, his law of large numbers (the term was due to Poisson) determined the rapidity of that process; Markov (1900/1924:44–52) improved Bernoulli's crude intermediate calculations and strengthened his estimate. Pearson (1925) achieved even better results, but only by applying the Stirling formula unknown to Bernoulli (as did Markov providing a parallel alternative improvement on pp 102–115). Pearson also unreasonably compared Bernoulli's estimate with the wrong Ptolemaic system of the world. He obviously did not appreciate theorems of existence (of the limiting property of statistical chances).

Statisticians never took notice of that rapidity, neither did they cite Bernoulli's law if not sure that the prior probability really existed and they barely recognized the benefits of the theory of probability (and hardly mentioned the more powerful forms of that law due to Poisson and Chebyshev). They did not know or forgot that mathematics as a science did not depend on the existence of its objects of study. The actual problem was to investigate whether the assumptions of the *Bernoulli trials* (their mutual independence and constancy of the probability of the studied event) were obeyed, and it was Lexis (The Two Streams of Statistical Thought) who formulated it. The previous statement of Cournot (1843; Sect. 86), whose outstanding book was not duly appreciated, that prior probability can be replaced by statistics in accord with *the Bernoulli's principle* was unnoticed.

The classical definition of probability, due to De Moivre (1738, Introduction) rather than to Laplace, with its equally possible cases is still with us. The axiomatic approach does not help statisticians and, moreover, practitioners have to issue from data, hence from the Mises frequentist theory developed in the 1930s which is not, however, recognized as a rigorous mathematical discovery.

Arbuthnot (1712) applied quite simple probability to prove that only Divine Providence explained why during 82 years more boys were invariably born in London than girls since the chances of a random occurrence of that fact were quite negligible. Cf. however the D'Alembert–Laplace problem: a long word is composed of printer's letters; was the composition random? Unlike D'Alembert, Laplace (1814/1995:9) decided that, although all the arrangements of the letters were equally unlikely, the word had a definite meaning, and therefore composed with an aim. His was a practical solution of a general and yet unsolved problem: to distinguish between a random and a determinate finite sequence of unities and zeros.

Arbuthnot could have noticed that Design was expressed by the binomial law, but it was still unknown. Even its introduction by Jakob Bernoulli and later scientists failed to become generally accepted: philosophers of the eighteenth century almost always only understood randomness in the "uniform" sense.

While extending Arbuthnot's study of the sex ratio at birth, De Moivre (1733) essentially strengthened the law of large numbers by proving the first version of the central limit theorem (see ►Central Limit Theorems) thus introducing the normal distribution, as it became called in the end of the nineteenth century. Laplace offered a somewhat better result, and Markov (1914/1951:511) called their proposition the *De Moivre–Laplace theorem*.

De Moivre devoted the first edition of his *Doctrine of Chances* (1718) to Newton, and there, in the Dedication, reprinted in 1756 (p. 329), we find his understanding of the aims of the new theory: separation of chance from Divine design, not yet the study of various and still unknown distributions, etc.

Such separations were being made in everyday life even in ancient India in cases of testimonies (Bühler 1886/1967:267). A misfortune encountered by a witness during a week after testifying was attributed to Divine punishment for perjury and to chance otherwise.

Newton himself (manuscript 1664–1666/1967:58–61) considered geometric probability and statistical estimation of the probability of various throws of an irregular die.

Bayes (1763), a memoir with a supplement published next year (Price and Bayes 1764), influenced statistics not less than Laplace. The so-called ►Bayes' theorem actually introduced by Laplace (1814/1995:10) was lacking there, but here is in essence his pertinent problem: a_i urns $(i = 1, 2)$ contain white and black balls in the ratio of α_i/β_i. A ball is extracted from a randomly chosen urn, determine the probability of its being white. The difficulty here is of a logical nature: may we assign a probability to an isolated event? This, however, is done, for example, when considering a throw of a coin. True, prior probabilities such as $\alpha_i/(\alpha_i + \beta_i)$ are rarely known, but we may keep to Laplace's principle (1803:xi): adopt a hypothesis and repeatedly correct it by new observations – if available!

Owing to these difficulties English and American statisticians for about 30 years had been abandoning the Bayes approach, but then (Cornfield 1967) the Bayes theorem *had returned from the cemetery*.

The main part of the Bayes memoir was his stochastic estimation of the unknown prior probability of the studied event as the number of *Bernoulli trials* increased. This is the inverse problem as compared with the investigations of Bernoulli and De Moivre, and H. E. Timerding, the Editor of the German translation of Bayes (1908), presented his result as a limit theorem. Bayes himself had not done it

for reasons concerned with rigor: unlike other mathematicians of his time (including De Moivre), he avoided the use of divergent series. Bayes' great discovery also needed by statisticians was never mentioned by them. Great, because it did not at all follow from previous findings and concluded the creation of the initial version of the theory of probability.

Both Bernoulli and De Moivre estimated the statistical probability given its theoretical counterpart and declared that they had at the same time solved the inverse problem (which Bayes expressly considered). Actually, the matter concerned the study of two different random variables with differing variances (a notion introduced by Gauss 1823), and only Bayes understood that the De Moivre formula did not ensure a good enough solution of the inverse problem.

Statistics in the Eighteenth Century

Later statisticians took up De Moivre's aim (Jakob Bernoulli, De Moirre, Bayes: Chance and Design) who actually extended Newton's idea of discovering the Divinely provided laws of nature. They, and especially Süssmilch, made the next logical step by attempting to discover the laws of the movement of population, hence to discern the pertinent Divine design. Euler essentially participated in compiling the most important chapter of the second edition, 1761–1762, of Süssmilch (1741), and Malthus (1798) picked up one of its conclusions, viz., that population increases in a geometric progression.

Süssmilch also initiated moral statistics by studying the number of marriages, of children born out of wedlock, etc. Its proper appearance was connected with A. M. Guerry and A. Quetelet (1830s and later).

Euler published a few elegant and methodically important memoirs on population statistics and introduced such concepts as increase in population and period of its doubling (see Euler 1923). Also methodically interesting were Lambert's studies of the same subject. When examining the number of children in families he (1772, Sect. 108) arbitrarily increased by a half their total number as given in his data likely allowing for stillbirths and mortality.

Most noteworthy were Daniel Bernoulli's investigations of several statistical subjects. His first memoir was devoted to inoculation (1766), to not a quite safe communication of a mild form of the deadly smallpox from one person to another (Jenner introduced vaccination of smallpox at the turn of that century) and proved that it lengthened mean life by two years plus and was thus highly beneficial (in the first place, for the nation). Then, he investigated the duration of marriages (1768), which was necessary for insurance depending on two lives. Finally, he (1770–1771) turned to the sex ratio at birth. He evidently wished to discover the *true value* of the ratio of male/female births (which does not really exist) but reasonably hesitated to make a final choice. However, he also derived the normal distribution although without mentioning De Moivre whose statistical work only became known on the Continent by the end of the nineteenth century.

Laplace (1812, Chapter 6) estimated the population of France by sampling (New Times: Great Progress and the Soviet cul-de-sac) and studied the sex ratio at birth. In this latter case he introduced *functions of very large numbers* (of births a and b) $x^a(1-x)^b$ and managed to integrate them. As usual, he had not given thought to thoroughly presenting his memoirs. While calculating the probability that male births will remain prevalent for the next 100 years, he did not add *under the same conditions of life*; and the final estimate of France's population was stated carelessly: Poisson, who published a review of that classic, mistakenly quoted another figure. Laplace's *Essai philosophique* (1814) turned general attention to probability and statistics.

The Theory of Probability and Statistics: Quetelet

Both Cournot (1843) and Poisson (1837) thought that mathematics should be the base of statistics. Poisson with coauthors (1835) were the first to state publicly that statistics was "the functioning mechanism of the calculus of probability" and had to do with mass observations. The most influential scholars of the time shared the first statement and likely the second as well. Fourier, in a letter to Quetelet (1869, t. 1, p 103) written around 1820, declared that statistics must be based on *mathematical theories*, and Cauchy (1845/1896:242) maintained that statistics provided means for judging doctrines and institutions and should be applied "avec tout la rigueur."

However, Poisson and Gavarret, his former student who became a physician and the author of the first book on medical statistics (1840), only thought about large numbers (e.g., when comparing two empirical frequencies) and a German physician Liebermeister (ca. 1877) complained that the alternative, i.e., the mathematical statistical approach was needed.

The relations between statistics and mathematics remained undecided. The German statistician Knapp (1872:116–117) declared that placing colored balls in Laplacean urns was not enough for shaking scientific statistics out of them. Much later mathematicians had apparently been attempting to achieve something of the

sort since Chuprov (1922:143) remarked that "Mathematicians playing statistics can only be overcome by mathematically armed statisticians." In the nineteenth, and the beginning of the twentieth century statisticians had still been lacking such armament.

Quetelet, who dominated statistics for several decades around the mid-nineteenth century, popularized the theory of probability. He tirelessly treated statistical data, attempted to standardize population statistics on an international scale, initiated anthropometry, declared that statistics ought to help foresee how various innovations will influence society, and collected and systematized meteorological data. Being a religious person, he (1846:259) denied any evolution of organisms which to some extent explains why Continental statisticians were far behind their English colleagues in studying biological problems. And Quetelet was careless in his writings so that Knapp (1872:124) stated that his spirit was rich in ideas but unmethodical and therefore un-philosophical. Thus, Quetelet (1836, t. 1, p 10) stated without due justification that the crime rate was constant although he reasonably but not quite expressly added: under invariable social conditions.

Quetelet paid attention to preliminary treatment of data and thus initiated elements of the exploratory data analysis (Statistics and the Statistical Method: The Theory of Errors); for example, he (1846:278) maintained that a too detailed subdivision of the material was a *charlatanisme scientifique*. He (1848:38) introduced the concept of Average man both in the impossible physical sense (e.g., mean stature and mean weight cannot coexist) and in the moral sphere, attributed to him mean inclinations to crime (1836, t. 2, p 171) and marriage (1848, p 77) and declared that that fictitious being was a specimen of mankind (1832, p 1).

Only in passing did he mention the Poisson law of large numbers, so that even his moral mean was hardly substantiated. Worse, he had not emphasized that the inclinations should not be attributed to individuals, and after his death German statisticians, without understanding the essence of the matter, ridiculed his innovations (and the theory of probability in general!) which brought about the downfall of *Queteletism*.

Fréchet (1949) replaced the Average man by *homme typique*, by an individual closest to the average. In any case, an average man (although not quite in Quetelet's sense) is meant when discussing per capita economic indications.

New Times: Great Progress and the Soviet cul-de-sac

In the main states of Europe and America statistical institutions and/or national statistical societies, which studied and developed population statistics, came into being during the first five decades of the nineteenth century. International statistical congresses aiming at unifying official statistical data had been held from 1851 onward, and in 1885 the still active International Statistical Institute was established instead.

A century earlier Condorcet initiated and later Laplace and Poisson developed the application of probability for studying the administration of justice. The French mathematician and mechanician Poinsot (1836) declared that calculus should not be applied to subjects permeated by imperfect knowledge, ignorance, and passions, and severe criticism was leveled at applications to jurisprudence for tacitly assuming independence of judges or jurors: "In law courts people behave like the*moutons de Panurge*" (Poincaré 1912:20). Better known is Mill's declaration (1843/1886:353): Such applications disgrace mathematics. Laplace (1812, Supplement of 1816/1886:523) only once and in passing mentioned that assumption.

However, stochastic reasoning can provide a "guideline" for determining the number of witnesses and jurors (Gauss, before 1841/1929:201–204) and the worth of majority verdicts. Poisson (1837:4) introduced the mean prior (statistically justified) probability of the defendant's guilt, not to be assigned to any individual and akin to Quetelet's inclination to crime. Statistical data was also certainly needed here. Quetelet (1836, t. 2, p 313) studied the rate of conviction as a function of the defendant's personality, noted that in Belgium the rate of conviction was considerably higher than in France (1833:18) and correctly explained this by the absence, in the former, of the institution of jurors (1846:334).

Statistical theory was also invariably involved in jurisprudence in connection with errors of the first and second kind. Thus (Sheynin 2009:17), the Talmud stipulated that a state of emergency (leading to losses) had to be declared in a town if a certain number of its inhabitants died during three consecutive days. Another example pertaining to ancient India is in Jakob Bernoulli, De Moirre, Bayes: Chance and Design.

A number of new disciplines belonging to natural science and essentially depending on statistics had appeared in the nineteenth century. *Stellar statistics* was initiated earlier by William Herschel (1784:162) who attempted to catalogue all the visible stars and thus to discover the form of our (finite, as he thought at the time) universe. In one section of the Milky Way he replaced counting by sample estimation (p. 158). He (1783) also estimated the parameters of the Sun's motion by attributing to it the common component of the proper motion of a number of stars. Galileo (1613) applied the same principle for estimating the

period of rotation of the Sun about its axis: he equated it with the (largely) common period of rotation of sunspots.

Most various statistical studies of the solar system (Cournot 1843) and the starry heaven (F. G. W. Struve, O. Struve, Newcomb) followed in the mid-nineteenth century and later (Kapteyn). Newcomb (Sheynin 2002) processed more than 62 thousand observations of the Sun and the planets and revised astronomical constants. His methods of treating observations were sometimes quite unusual. Hill and Elkin (1884:191) concluded that the "great Cosmical questions" concerned not particular stars, but rather their average parallaxes and the general relations between star parameters.

Daniel Bernoulli was meritorious as the pioneer of *epidemiology* (Statistics in the Eighteenth Century). It came into being in the nineteenth century mostly while studying cholera epidemics. The other new disciplines were *public hygiene* (the forerunner of ecology), *geography of plants, zoogeography, biometry, and climatology*.

Thus, in 1701 Halley published a chart of North Atlantic showing (contour) lines of equal magnetic declination, and Humboldt (1817) followed suit by inventing lines of equal mean yearly temperatures (isotherms) replacing thousands of observations and thus separating climatology from meteorology. These were splendid examples of exploratory data analysis (Statistics and the Statistical Method: The Theory of Errors). Also in meteorology, a shift occurred from studying mean values (Humboldt) to examining deviations from them, hence to temporal and spatial distributions of meteorological elements.

Statistics ensured the importance of public hygiene. Having this circumstance in mind, Farr (1885:148) declared that "Any deaths in a people exceeding 17 in 1,000 annually are unnatural deaths." Data pertaining to populations in hospitals (*hospitalism*, mortality due to bad hygienic conditions), barracks, and prisons were collected and studied, causes of excessive mortality indicated and measures for preventing it made obvious.

At least medicine had not submitted to statistics without opposition since many respected physicians did not understand its essence or role. A staunch supporter of "rational" statistics was Pirogov, a cofounder of modern surgery and founder of military surgery. He stressed the difficulty of collecting data under war conditions and reasonably interpreted them.

Around the mid-nineteenth century, statistics essentially fostered the introduction of anesthesia since that new procedure sometimes led to serious complications. Another important subject statistically studied was the notorious hospitalism, see above.

Biometry indirectly owed its origin to Darwin, witness the Editorial in the first issue of *Biometrika* in 1902: "The problem of evolution is a problem of statistics. […] Every idea of Darwin […] seems at once to fit itself to mathematical definition and to demand statistical analysis."

Extremely important was the recognition of the statistical laws of nature (theory of evolution, in spite of Darwin himself), kinetic theory of gases (Maxwell), and stellar astronomy (Kapteyn). And the discovery of the laws of heredity (Mendel 1866) would have been impossible without statistics. Methodologically these laws were based on the understanding that randomness in individual cases becomes regularity in mass (Kant, Laplace, and actually all the stochastic laws).

Laplace (1814; English translation 1995:2) declared that randomness was only occasioned by our failure to comprehend all the natural forces and by the imperfection of analysis, and he was time and time again thought only to recognize determinism. However, the causes he mentioned were sufficiently serious; he expressly formulated *statistical determinism* (e.g., stability of the relative number of dead letters, an example of transition from randomness to regularity); and his work in astronomy and theory of errors was based on the understanding of the action of random errors. It is also opportune to note here that randomness occurs in connection with unstable movement (Poincaré) and that a new phenomenon, chaotic behavior (an especially unpleasant version of instability of motion), was discovered several decades ago. Finally, Laplace was not original: Maupertuis (1756:300) and Boscovich (1758, Sect. 385) preceded him.

In the nineteenth century, but mostly perhaps in the twentieth, the statistical method penetrated many other sciences and disciplines beyond natural sciences so that it is now difficult to say whether any branch of knowledge can manage without it.

There are other points worth mentioning. *Correlation theory* continued to be denied even in 1916 (Markov), actually because it was not yet sufficiently developed. Its appearance (Galton, Pearson) was not achieved at once. In 1865–1866 the German astronomer and mathematician Seidel quantitatively estimated the dependence of the number of cases of typhoid fever on the level of subsoil water and precipitation but made no attempt to generalize his study. And in the 1870s several scientists connected some terrestrial phenomena with solar activity but without providing any such estimates.

According to Gauss (1823:18), for series of observations to be independent, it was necessary for them not to contain common measurements, and geodesists without referring to him have been intuitively keeping to his viewpoint.

For two series of about m observations each, n of them common to both, the measure of their interdependence was thought to be n/m. Kapteyn (1912) made the same proposal without mentioning anyone.

Estimation of precision was considered superfluous (Bortkiewicz 1894–1896, Bd 10, pp 353–354): it is a *luxury* as opposed to the statistical feeling. *Sampling* met with protracted opposition although even in 1812 the German statistician Lueder (Lueder 1812:9) complained about the appearance of "legions" of numbers. In a crude form, it existed long ago, witness the title of Stigler (1977). In the seventeenth century in large Russian estates it was applied for estimating the quantity of the harvested grain, and, early in the next century Marshal Vauban, the *French Petty*, made similar estimations for France as a whole.

No wonder that Laplace, in 1786, had estimated the population of France by sampling, and, much more important, calculated the ensuing error. True, Pearson (1928) discovered a logical inconsistency in his model. As a worthy method, sampling penetrated statistics at the turn of the nineteenth century (the Norwegian statistician Kiaer) and Kapteyn (1906) initiated the study of the starry heaven by stratified sampling, but opposition continued (Bortkiewicz 1901).

The *study of public opinion* and *statistical control of quality of industrial production*, also based on sampling, had to wait until the 1920s (true, Ostrogradsky (1848) proposed to check samples of goods supplied in batches), and *econometrics* was born even later, in the 1930s.

A curious side issue of statistics, *sociography*, emerged in the beginning of the twentieth century. It studies ethnic, religious, etc., subgroups of society, does not anymore belong solely to statistics, and seems not yet to be really scientific. And in sociology it became gradually understood that serious changes in the life of a society or a large commercial enterprise should be based on preliminary statistical studies.

Soviet statistics became a dangerous pseudoscience alienated from the world (Sheynin 1998). Its main goal was to preserve appearances by protecting Marxist dogmas from the pernicious influence of contemporary science and it frustrated any quantitative studies of economics and banished mathematics from statistics. In 1909, Lenin called Pearson a Machian and an enemy of materialism which was more than enough for Soviet statisticians to deny the work of the Biometric school lock, stock, and barrel.

Culmination of the success in that direction occurred in 1954, during a high-ranking conference in Moscow. Its participants even declared that statistics did not study mass random phenomena which, moreover, did not possess any special features. Kolmogorov, who was present at least for his own report, criticized Western statisticians for adopting unwarranted hypotheses…

Soviet statisticians invariably demanded that quantitative investigations be inseparably linked with the qualitative content of social life (read: subordinated to Marxism), but they never repeated such restrictions when discussing the statistical method as applied to natural sciences.

The Two Streams of Statistical Thought

Lexis (1879) proposed a distribution-free test for the equality of probabilities of the studied event in a series of observations, the ratio Q of the standard deviation of the frequency of the occurrence of the studied event, as calculated by the Gauss formula, to that peculiar to the ►binomial distribution. That ratio would have exceeded unity had the probability changed; been equal to unity otherwise, all this taking place if the trials were independent; and been less than unity for interdependent trials. Lexis (1879, Sect. 1) also qualitatively isolated several types of statistical series and attempted to define stationarity and trend.

Bortkiewicz initiated the study of the expectation of Q and in 1898 introduced his celebrated law of small numbers which actually only essentially popularized the barely remembered Poisson distribution. In general, his works remain insufficiently known because of his pedestrian manner, excessive attention to detail, and bad composition which he refused to improve. Winkler (1931:1030) quoted his letter (date not given) stating that he expected to have five readers (!) of his (unnamed) contribution.

Markov and mostly Chuprov (1918–1919) refuted the applicability of Q but anyway Lexis put into motion the Continental direction of statistics by attempting to base statistical investigations on a stochastic basis. Lexis was not, however, consistent: even in 1913 he held that the law of large numbers ought to be justified by empirical data. Poisson can be considered the godfather of the new direction.

On the other hand, the Biometric school with its leader Pearson was notorious for disregarding stochastic theory and thus for remaining empirical. Yet he developed the principles of correlation theory and contingency, introduced *Pearsonian* curves for describing asymmetrical distributions, devised the most important chi-square test (see ►Chi-Square Tests), and published many useful statistical tables. To a large extent his work ensured the birth of mathematical statistics.

Pearson successfully advocated the application of the new statistics in various branches of science and studied his own discipline in the context of general history (1978, posthumous). There (p 1) we find: "I do feel how wrongful

it was to work for so many years at statistics and neglect its history." He acquired many partisans and enemies (including Fisher). Here is Newcomb in a letter to Pearson of 1903 (Sheynin 2009, Sect. 10.9.4) and Hald (1998:651): "You are the one living author whose production I nearly always read when I have time […] and with whom I hold imaginary interviews […]"; "Between 1892 and 1911 [he] created his own kingdom of mathematical statistics and biometry in which he reigned supremely, defending its ever expanding frontiers against attacks."

Nevertheless, the work of his school was scorned by Continental scientists, especially Markov, the apostle of rigor. Chuprov, however, tirelessly, although without much success, strove to unite the two streams of statistical thought. Slutsky also perceived the importance of the Biometric school. He (1912) expounded its results and, although only in a letter to Markov of 1912, when he was not yet sufficiently known, remarked that Pearson's shortcomings will be overcome just as it happened with the nonrigorous mathematics of the seventeenth and eighteenth centuries.

Chuprov also achieved important results, discovering for example finite exchangeability (Seneta 1987). He mainly considered problems of the most general nature, hence inevitably derived unwieldy and too complicated formulas, and his contributions were barely studied. In addition, his system of notations was horrible. In one case he (1923:472) applied two-storey superscripts and, again, two-storey subscripts in the same formula!

Markov, the great mathematician, was to some extent a victim of his own rigidity. Even allowing for the horrible conditions in Russia from 1917 to his death in 1922, it seems strange that he failed, or did not wish to notice the new tide of opinion in statistics (and even in probability theory).

Mathematical Statistics

In what sense is mathematical statistics different from biometry? New subjects have been examined such as sequential analysis, the treatment of previously studied problems (sampling, time series, hypothesis testing) essentially developed, links with probability theory greatly strengthened (Pearson's empirical approach is not tolerated anymore). New concepts have also appeared and this seems to be a most important innovation. Fisher (1922) introduced statistical estimators with such properties as consistency, efficiency, etc., some of which go back to Gauss who had used and advocated the principle of unbiased minimum variance.

It is known that the development of mathematics has been invariably connected with its moving ever away from Nature (e.g., to imaginaries) and that the more abstract it was becoming, the more it benefited natural sciences. The transition from true values to estimating parameters was therefore a step in the right direction. Nevertheless, the former, being necessary for the theory of errors, are still being used in statistics, and even for objects not existing in Nature, see Wilks (1962, Sect. 10.1), also preceded by Gauss (1816, Sects. 3 and 4) in the theory of errors.

Rao (*Math. Rev.* 2005k:62007) noted that modern statistics has problems with choosing models, measuring uncertainty, testing hypotheses, and treating massive sets of data, and, in addition, that statisticians are not acquiring sufficient knowledge in any branch of natural science.

About the Author

Oscar Sheynin was born in Moscow, 1925. He graduated from the Moscow Geodetic Institute and Mathematical-Mechanical Faculty of Moscow State University, and he is Candidate of Sciences, Physics and Mathematics. He was working as a geodesist in the field, then taught mathematics, notably at the Plekhanov Institute for National Economy (Moscow) as Dozent. From 1962 to this day, he independently studies history of probability and statistics and since 1991 he has been living in Germany. He is a Member of International Statistical Institute (1975), Full Member, International Academy of History of Science (1995), and of the Royal Statistical Society. He has published more than 130 papers including 25 in the *Archive for History of Exact Sciences* and a joint paper on probability in the nineteenth century with Boris Gnedenko. Much more can be found at www.sheynin.de.

Cross References

►Astrostatistics
►Bayes' Theorem
►Foundations of Probability
►Laws of Large Numbers
►Least Squares
►Medical Statistics
►Normal Distribution, Univariate
►Poisson Distribution and Its Application in Statistics
►Probability, History of
►Sex Ratio at Birth
►Statistical Publications, History of

References and Further Reading

Aaboe A, De Solla Price DJ (1964) Qualitative measurements in antiquity. In: Mélanges A. Koyré, t. 1: L'aventure de la science. Hermann, Paris, pp 1–20

Anchersen JP (1741) Descriptio statuum cultiorum in tabulis. Otto Christoffer Wenzell, Copenhagen/Leipzig

Anonymous (1839) Introduction. J Stat Soc Lond 1:1–5
Arbuthnot J (1710/1712) An argument for Divine Providence taken from the constant regularity observed in the birth of both sexes. Philos Trans R Soc Lond (repr Kendall MG, Plackett RL (eds) (1997) Studies in the history of statistics and probability, vol 2. Griffin, High Wycombe, pp 30–34)
Bayes T (1763, published 1764) An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. Philos Trans R Soc Lond 53:370–418
Bayes T (1908) Versuch zur ösung eines problems der wahrscheinichkeitsrechnung. Herausgeber, Timeding HE. Ostwald Klassiker No. 169. Leipzig:Engelmann
Bernoulli D (1766) Essai d'une nouvelle analyse de la mortalité cause par la petite vérole etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 235–267
Bernoulli D (1768) De duratione media matrimoniorum etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 290–303; Sheynin O (2004) Probability and statistics. Russian Papers. Berlin, pp 17–31 (translated from Russian)
Bernoulli D (1770–1771) Mensura sortis ad fortuitam successionem rerum naturaliter contingentium applicata. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 326–360
Bernoulli D (1780) Specimen philosophicum de compensationibus horologicis etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 376–390
Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2. Basel
Bernoulli J (1713) Ars Conjectandi. Werke, Bd 3 (1975, Birkhäuser, Basel, pp 107–259); German trans: (1899) Wahrscheinlichkeitsrechnung (1999, Thun/Frankfurt am Main); English trans of pt 4: Bernoulli J (2005) On the law of large numbers. Berlin. Available at http://www.sheynin.de
Bessel FW (1839) Einfluß der Schwere auf die Figur eines . . . Stabes. In: Bessel FW (1876) Abhandlungen, Bd 3. Wilhelm Engelmann, Leipzig, pp 275–282
Bortkiewicz L (1894–1896) Kritische Betrachtungen zur theoretischen Statistik. Jahrbücher f. Nationalökonomie u. Statistik, 3. Folge, 8:641–680, 10:321–360, 11:701–705
Bortkiewicz L (1898) Das Gesetz der kleinen Zahlen. Leipzig
Bortkiewicz L (1904) Anwendung der Wahrscheinlichkeitsrechnung auf Statistik. Enc Math Wiss 1:821–851
Boscovich R (1758, in Latin/1966) Theory of natural philosophy. MIT Press, Cambridge. Translated from edition of 1763
Bühler G (ed) (1886) Laws of Manu. Clarendon Press, Oxford (repr 1967)
Butte W (1808) Die Statistik als Wissenschaft. Landshut
Cauchy AL (1845) Sur le secours que les sciences du calcul peuvent fournir aux sciences physiques ou même aux sciences morales. Oeuvr Compl 1 (1896), t. 9. Paris, pp 240–252
Chuprov (Tschuprow) AA (1909, in Russian) Essays on the theory of statistics. Sabashnikov, Saint Petersburg (repr State Publishing House, Moscow, 1959)
Chuprov (Tschuprow) AA (1918–1919) Zur Theorie der Stabilität statistischer Reihen. Skand Aktuarietidskrift 1:199–256; 2: 80–133
Chuprov (Tschuprow) AA (1922) Review of books. Nordisk Statistisk Tidskrift 1:139–160, 329–340
Chuprov (Tschuprow) AA (1923) On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, Metron 2:461–493, 646–683
Cornfield J (1967) The Bayes theorem. Rev Inter Stat Inst 35:34–49
Cournot AA (1843) Exposition de la théorie des chances et des probabilités. Hachette, Paris (repr 1984)
Cubranic N (1961) Geodetski rad R. Boscovica. Zagreb
Cubranic N (1962) Geodätisches Werk R. Boscovic's. In: Actes Symp. Intern. Boscovic. Beograd, 1962, pp 169–174
De Moivre A (1718) Doctrine of chances. W. Pearson, London (2nd edn: 1738, 3rd edn: 1756; last edn repr Chelsea, New York, 1967)
De Moivre A (1733, in Latin) A method of approximating the sum of the terms of the binomial $(a + b)^n$ expanded into a series from whence are deduced some practical rules to estimate the degree of assent which is to be given to experiments. Translated by De Moivre A and inserted in his book (1738, 1756), pp 243–254 in 1756
Euler L (1923) Opera omnia 1, t. 7. Leipzig
Farr W (1885) Vital Statistics: A memorial volume of selections from the reports and writings of William Farr MD, DCL, CB, F.R.S.N. (N A Humphreys, ed.). Sanitary Institute of London, London. (Reprinted 1975 by Scarecrow Press, Metuchen, NJ.)
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A 222:309–368
Fourier JBJ (1826) Sur les resultats moyen déduits d'un grand nombre d'observations. Oeuvr (1890), t. 2. Paris, pp 525–545
Fréchet M (1949) Réhabilitation de la notion statistique de l'homme moyen. In: Fréchet M (1955) Les mathématiques et les concret. Presses Universitaires de France, Paris, pp 317–341
Galilei G (1613, in Italian) History and demonstrations concerning sunspots etc. In: Galilei G (1957) Discoveries and opinions of Galilei. Garden City, pp 88–144
Gatterer JC (1775) Ideal einer allgemeinen Weltstatistik. Göttingen
Gauss CF (1809) Theoria Motus Corporum Coelestium in Sectionibus Conicis solem Ambientum. Perthes und Besser, Hamburg (English trans: Davis CH (1857) Theory of the motion of the heavenly bodies moving about the sun in conic sections. Little, Brown, Boston (repr Mineola, Dover, (2004))
Gauss CF (1816) Bestimmung der Genauigkeit der Beobachtungen (repr (1880) Carl Friedrich Gauss Werke 4, 109–117. Königliche Gesellschaft der Wissenschaften, Göttingen; English trans: David HA, Edwards AWF (2001) The determination of the accuracy of observations. In: Annotated readings in the history of statistics. Springer, New York, pp 41–50)
Gauss CF (1823) Theoria combinationis observationum erroribus minimis obnoxiae (repr (1880) Carl Friedrich Gauss Werke 4, 1–53. Königliche Gesellschaft der Wissenschaften, Göttingen; English trans: Stewart GW (1995) Theory of the combination of observations least subject to errors. SIAM, Philadelphia)
Gauss CF (1887) Abhandlungen zur Methode der kleinsten Quadrate (repr Vaduz, 1998), Berlin
Gauss CF (1929) Werke, Bd 12. Göttingen/Berlin
Gavarret J (1840) Principes généraux de statistique médicale. Paris
Graunt J (1662) Natural and political observations made upon the bills of mortality. In: Petty W (1899) Economic writings, vol 2, pp 317–435 with Graunt's additions of 1665 and 1676. The Writings were reprinted: Fairfield, 1986; London, 1997. Many other editions of Graunt, e.g., Baltimore, 1939
Hald A (1998) History of mathematical statistics from 1750 to 1930. New York
Halley E (1694) An Estimate of the degrees of mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives Philosophical Transactions of the Royal Society of

London (17): 596–610 and 654–656. (Reprinted, edited with an introduction by Reid LJ, Baltimore, MD: The Johns Hopkins Press 1942)
Herschel W (1783) On the proper motion of the Sun. In: Herschel W (1912) Scientific Papers, vol 1. London, pp 108–130
Herschel W (1784) Account of some observations. In: Herschel W (1912) Scientific Papers, vol 1. London, pp 157–166
Herschel W (1817) Astronomical observations and experiments etc. In: Herschel W (1912) Scientific Papers, vol 2. London, pp 575–591
Hill D, Elkin WL (1884) Heliometer-determination of stellar parallax. Mem R Astron Soc 48, the whole pt 1
Hippocrates (1952) Aphorisms. In: Great books of the western world, vol 10. Encyclopaedia Britannica, Chicago, pp 131–144
Humboldt A (1815) Prolegomena. In: Bonpland A, Humboldt A, Kunth KS. Nova genera et species plantarum etc., vol 1. Russian trans: 1936, Paris
Humboldt A (1817) Des lignes isothermes. Mém Phys Chim Soc Arcueil 3:462–602
Huygens C (1699) Correspondence. Oeuvr Compl (1895), t. 14. La Haye
Jenner E (1798) An inquiry into the causes and effects of the variolae vaccinae, a disease discovered in some of the Western counties of England, particularly Gloucestershire, and known by the name of the cow pox. Sampson Low, London, for the author. In: The three original publications on vaccination against smallpox, vol XXXVIII, pt 4: the Harvard Classics (1909–1914). P.F. Collier, New York
Kapteyn JC (1906) Plan of selected areas. Groningen
Kapteyn JC (1912) Definition of the correlation-coefficient. Monthly Notices R Astron Soc 72:518–525
Kendall MG (1960) Studies in the history of probability and statistics. X. Where shall the history of statistics begin? Biometrika 47(3–4):447–449
Knapp GF (1872) Quetelet als Statistiker. Jahrbücher f. Nationalökonomie u. Statistik 18:89–124
Knies CGA (1850) Die Statistik als selbstständige Wissenschaft. Kassel
Kolmogorov AN (1946, in Russian) Justification of the method of least squares. Selected works (1992), vol 2. Kluwer, Dordrecht, pp 285–302
Kolmogorov AN, Prokhorov (1982 in Russian) Mathematical Statistics In: Vinogradov IM (ed) Soviet Matematicheskaya entsiklopediya (Encyclopaedia of Mathematics), vol 3, Moscow, 576–581
Lambert JH (1760) Photometria (in Latin). Augsburg. Cited statement omitted from German translation
Lambert JH (1765) Anmerkungen und Zusätze zur practischen Geometrie. In: Lambert JH. Beyträge, Tl. 1, pp 1–313
Lambert JH (1765–1772) Beyträge zum Gebrauche der Mathematik und deren Anwendung, Tl. 1–3. Berlin
Lambert JH (1772) Anmerkungen über die Sterblichkeit, Todtenlisten, Geburthen and Ehen. In: Lambert JH. Beyträge, Tl. 3, pp 476–569
Laplace PS (ca. 1803) Traité de Mécanique Céleste, t. 3. Oeuvr Compl (1878), t. 3. Paris. Translation by Bowditch N (1832) Celestial mechanics. New York, 1966
Laplace PS (1812) Théorie analytique des probabilités. Oeuvr Compl (1886), t. 7. Paris
Laplace PS (1814) Essai philosophique sur les probabilités. Oeuvr Compl (1886), t. 7, No. 1, separate paging (English trans: New York, 1995)
Leibniz GW (1986) Sämmtl. Schriften und Briefe, 4, Bd 3. Berlin
Lexis W (1879) Über die Theorie der Stabilität statistischer Reihen. Jahrbücher f. Nationalökonomie u. Statistik 32:60–98 (repr Lexis W (1903) Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik. Jena, pp 170–212)
Lexis W (1913) Review of book. Schmollers Jahrbuch f. Gesetzgebung, Verwaltung u. Volkswirtschaft im Deutschen Reich 37:2089–2092
Liebermeister C (ca. 1877) Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. Sammlung klinischer Vorträge No. 110 (Innere Med. No. 39), Leipzig, pp 935–962
Louis PCA (1825) Recherches anatomico-pathologiques sur la phtisie. Paris
Lueder AF (1812) Kritik der Statistik und Politik. Göttingen
Malthus TR (1798) Essay on the principle of population. Works (1986), vol 1. Pickering, London
Markov AA (1899, in Russian) On the law of large numbers and the method of least squares. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 231–251
Markov AA (1900, in Russian) Calculus of probability. Later editions: 1908, 1913 and posthumous, Moscow, 1924 (German trans: Leipzig/Berlin, 1912) Academy of Sciences, St. Peterburg
Markov AA (1914, in Russian) On Jakob Bernoulli's problem. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 511–521
Markov AA (1916, in Russian) On the coefficient of dispersion. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 523–535
Markov AA (1951) Izbrannye Trudy (Selected works). Academy of Sciences, USSR
Maupertuis PLM (1756) Sur la divination. Oeuvres, t. 2. Lyon, pp 298–306
Mendel JG (1866, in German) Experiments in plant hybridization. In: Bateson W (1909) Mendel's Principles of Heredity. Cambridge University Press, Cambridge, pp 317–361 (repr 1913)
Mill JS (1843) System of logic. London (repr 1886)
Newton I (1967) Mathematical papers, vol 1. Cambridge University Press, Cambridge
Ostrogradsky MV (1848) Sur une question des probabilités. Bull Cl Phys-Math Acad Imp Sci St Pétersb 6(21–22):321–346
Pearson K (1925) James Bernoulli's theorem. Biometrika 17:201–210
Pearson K (1928) On a method of ascertaining limits to the actual number of individuals etc. Biometrika 20A:149–174
Pearson K (1978) History of statistics in the 17th and 18th centuries against the changing background of intellectual, scientific, and religious thought. Pearson ES (ed) Lectures 1921–1933. Griffin, London
Petrov VV (1954, in Russian) On the method of least squares and its extreme properties. Uspekhi Matematich Nauk 1:41–62
Poinsot L (1836) A remark, in Poisson (1836) p 380 http://www.archive.org/stream/comptesrendusheb02acad#page/380/mode/2up
Poisson S-D (1836) Note sur la loi des grands nombres. C r Acad Sci 2:377–382 http://www.archive.org/stream/comptesrendusheb02acad#page/380/mode/2up
Poisson S-D (1837) Recherches sur la probabilité des jugements etc. Paris (repr Paris, 2003)

Poisson S-D, Dulong PL, Double (1835) Rapports: Recherches de Statistique sur l'affection calculeuse, par M. Le docteur Civiale. C r Acad Sci Paris 1:167–177 (Statistical research on conditions caused by calculi by Doctor Civiale translated for the Int J Epidemiol by Swaine Verdier A, 2001, 30:1246–1249)
Politanus H (1672) Microscopium statisticum quo status imperii Romano-Germanici cum primis extraordinarius, ad vivum repraesentatur
Price RA, Bayes T (1764, published 1765) A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances. Published in the Philosophical Transactions, Vol. LIII. Communicated by the Rev. Mr. Richard Price, in a letter to Mr. John Canton, M. A. F. R. S. Philos Trans 54:296–325
Proctor RA (1872) On star-grouping. Proc R Instn Gr Brit 6:143–152
Quetelet A (1832) Recherches sur la loi de la croissance de l'homme. Mém Acad R Sci Lettre Beaux-Arts Belg 7, pp 32
Quetelet A (1833) Statistique des tribunaux de la Belgique. Bruxelles (Coauthor, Smits E)
Quetelet A (1836) Sur l'homme, tt. 1–2. Bruxelles
Quetelet A (1846) Lettres sur la théorie des probabilités. Bruxelles
Quetelet A (1848) Du système social. Paris
Quetelet A (1869) Physique sociale etc., tt. 1–2. Bruxelles (Bruxelles, 1997)
Schlözer AL (1804) Theorie der Statistik. Göttingen
Seidel L (1865) Über den Zusammenhang zwischen den Häufigkeit der Typhus-Erkrankungen und dem Stande des Grundwassers. Z Biol 1:221–236
Seidel L (1866) Vergleichung der Schwankung der Regenmengen mit den Schwankungen in der Häufigkeit des Typhus. Z Biol 2: 145–177
Seneta E (1987) Chuprov on finite exchangeability, expectation of ratios and measures of association. Hist Math 14:243–257
Sheynin O (1971) O dva neobjavljena spisa R. Boskovica iz teorije verovatnoce. Dijalektika 2(Godina 6):85–93
Sheynin O (1978) Poisson's work in probability. Arch Hist Ex Sci 18:245–300
Sheynin O (1992) Al-Biruni and the mathematical treatment of observations. Arabic Sci Philos 2:299–306
Sheynin O (1998) Statistics in the Soviet epoch. Jahrbücher f. Nationalökonomie u. Statistik 217:529–549
Sheynin O (1999) Statistics, definitions of. In: Kotz S (ed) Encyclopedia of statistical sciences, update vol 3. New York, pp 704–711 (repr 2nd edn (2006) of that encyclopedia, vol 12. Hoboken, pp 8128–8135)
Sheynin O (2009) Theory of probability. Historical essay. Berlin. Available at: http://www.sheynin.de and Google
Simpson JY (1847–1848) Anaesthesia. In: Simpson JY (1871) Works, vol 2. Adam and Charles Black, Edinburgh, pp 1–288
Slutsky EE (1912, in Russian) Theory of correlation etc. Kiev
Snow J (1855) On the mode of communication of cholera. Churchill, London (repr (1965) Snow on cholera. Hafner, New York, pp 1–139)
Stigler SM (1977) Eight centuries of sampling inspection: the trial of the pyx. J Am Stat Assoc 72:493–500
Süssmilch JP (1741) Göttliche Ordnung. Berlin. Many later editions
Wilks SS (1962) Mathematical statistics. Wiley, New York
Winckler W (1931) Ladislaus von Bortkiewicz. Schmollers Jahrbuch f. Gesetzgebung, Verwaltung u. Volkswirtschaft im Deutschen Reich 55:1025–1033
Yule GU (1905) The introduction of the words "statistics", "statistical" into the English language. J R Stat Soc 68:391–396

Statistics: An Overview

David Hand
Professor, President of the Royal Statistical Society (2008–2009, 2010)
Imperial College, London, UK

One can define statistics in various ways. My favorite definition is bipartite:

- *Statistics is both the science of uncertainty and the technology of extracting information from data.*

This definition captures the two aspects of the discipline: that it is about understanding (and indeed manipulating) chance, and also about collecting and analyzing data to enable us to understand the world around us. More specifically, of course, statistics can have different aims, including prediction and forecasting, classification, estimation, description, summarization, decision-making, and others.

Statistics has several roots, which merged to form the modern discipline. These include (1) the theory of probability, initially formalized around the middle of the seventeenth century in attempts to understand games of chance, and then put on a sound mathematical footing with Kolmogorov's axioms around 1930; (2) surveys of people for governmental administrative and economic purposes, as well as work aimed at constructing life tables (see ►Life Table) for insurance purposes (see ►Insurance, Statistics in); and (3) the development of arithmetic methods for coping with measurement errors in areas like astronomy and mechanics, by people such as Gauss, in the eighteenth and nineteenth centuries.

This diversity of the roots of statistics has been matched by the changing nature of discipline. This is illustrated by, for example, the papers which have appeared in the journal of the Royal Statistical Society (the journal was launched in 1838). In the earlier decades, there was a marked emphasis on social matters, which gradually gave way around the turn of the century, to more mathematical material. The first half of the twentieth century then saw the dramatic development of deep and powerful ideas of statistical inference, which continue to be refined to the present day. In more recent decades, however, the computer has had an equally profound impact on the discipline. Not only has this led to the development of entirely new classes of methods, it has also put powerful tools into the hands of statistically unsophisticated users – users who do not understand the deep mathematics underlying the tools. As might be expected, this can be a mixed blessing: powerful tools in hands which understand and know how to use them properly can be a tremendous asset, but those

same tools in hands which can misapply them may lead to misunderstandings.

Although the majority of statisticians are still initially trained in university mathematics departments (with statistics courses typically being part of a mathematics degree), statistics should not be regarded as a branch of mathematics – just as physics, engineering, surveying, and so on have a mathematical base but are not considered as branches of mathematics. Statistics also has a mathematical base, but modern statistics involves many other intrinsically non-mathematical ideas.

An illustration of this difference is given by the contrast between probability (properly considered as a branch of mathematics – based on an axiom system) and statistics (which is not axiomatic). Given a system or process which is producing data, probability theory tells us what the data will be like. If we repeatedly toss a fair coin, for example, probability theory tells us about the properties of the sequences of heads and tails we will observe. In contrast, given a set of data, statistics seeks to tell us about the properties of the system which generated the data. Since, of course, many different systems could typically have generated any given data set, statistics is fundamentally *inductive*, whereas probability is fundamentally *deductive*.

At its simplest level, statistics is used to describe or summarize data. A set of 1,000 numerical values can be summarized by their mean and dispersion – though whether this simple two-value summary will be adequate will depend on the purpose for which the summary is being made. At a much more sophisticated level, official statistics are used to describe the properties of the entire population and economy of a country: the distribution of ages, how many are unemployed, the Gross National Product, and so on. The effective governance of a country, management of a business, operation of an education system, running of a health service, and so on, all depend on accurate descriptive statistics, as well as on statistical extrapolations of how things are likely to change in the future.

Often, however, mere descriptions are not enough. Often the observed data are not the entire population, but are simply a sample from this population, and the aim is to infer something about the entire population. Indeed, often the "entire population" may not be well-defined; what, for example, would be the entire population of possible measurements of the speed of light in repeated experiments? In such cases, the aim is to use the observed sample of values as the basis for an estimate of the "true underlying" value (of the speed of light in this example).

A single "point" estimate is all very well, but we must recognize that if we had chosen a different sample of values we would probably have obtained a different estimate – there is uncertainty associated with our estimate. A point estimate can be complemented by indicating the range of this uncertainty: indicating how confident we can be that the true unobserved value lies in a specified interval of values. Basic rules of probability tell us that increasing the sample size allows us to narrow down this range of uncertainty (provided the sample is collected in a certain way), so that we can be as confident as we wish (or as we can afford) about the unknown true value.

Estimation is one aspect of statistics, but often one has more pointed questions. For example, one might be evaluating a new medicine, and want to test whether it is more effective than the current drug of choice. Or one might want to see how well the data support a particular theory – that the speed of light takes a certain specified value, for example. Since, in the first example, people respond differently, and, in the second, measurement error means that repeated observations will differ, the data will typically consist of several observations – a sample, as noted above – rather than just one. Statistical *tests*, based on the sample, are then used to evaluate the various theories. *Hypothesis testing* methods (Neyman-Pearson hypothesis tests) are used for comparing competing explanations for the data (that the proposed new medicine is more effective than or is as effective as the old one, for example). Such tests use probability theory to calculate the chance that some summary statistic of the data will take values in given ranges. If the observed value of the summary statistic is very unlikely under one hypothesis, but much more likely under the other, one feels justified in rejecting the former and accepting the latter. *Significance testing* methods (Fisherian tests) are used to see how well the observed data match a particular given hypothesis. If probability calculations show that one is very unlikely to obtain a value at least as extreme as the observed value of the summary statistic then this is taken as evidence against the hypothesis.

Such testing approaches are not uncontroversial. Intrinsic to them is the calculation of how often one would expect to obtain such results in repeated experiments, assuming that the data arose from a distribution specified by a given hypothesis. They are thus based on a particular interpretation of probability – the *frequentist* view. However, one might argue that hypothetical repeated experiments are all very well, but in reality we have just the one observed set of data, and we want to draw a conclusion using that one set. This leads to ►Bayesian statistics. Bayesian statistics is based on a different interpretation of probability – the *subjective* view. In this view, probability is regarded as having no external reality, but rather as a degree of belief. In particular, in the testing context, the different values of the parameters of the distribution producing the data are themselves assumed to take some

distribution. In this approach to inference, one then uses the data to refine one's beliefs about the likely form of the distribution of the parameters, and hence of the distribution from which the data were generated.

The *likelihood function* plays an important role in all schools of inference; it is defined as the probability of obtaining the observed data, viewed as a function of the parameters of the hypothesized distribution. The likelihood function is used in Bayesian inference to update one's initial beliefs about the distribution of the parameters. A further school of statistics, the *likelihood* school, focuses attention on the likelihood function, on the grounds that it is this which contains all the relevant information in the data. Comparative discussions of the various schools of inference, along with the various profound concepts involved, are given by Barnett (1999) and Cox (2006).

The choice of the term "Bayesian" to describe a particular school of inference is perhaps unfortunate: ►Bayes' theorem is accepted and used by all schools. The key distinguishing feature of Bayesian statistics is the subjective interpretation of probability and the interpretation of the parameters of the distributions as random variables themselves.

The differences between the various schools of inference have stimulated profound, and sometimes fierce debates. Increasingly, however, things seem to be moving towards a recognition that different approaches are suited to different questions. For example, one might distinguish between what information the data contain, what we should believe after having observed the data, and what action we should take after having observed the data.

Thus far I have been talking about data without mentioning how it was collected. But data collection is a key part of statistical science. Properly designed data collection strategies lead to faster, cheaper collection, and to more accurate results. Indeed, poorly designed data collection strategies can completely invalidate the conclusions. For example, an experiment to compare two medicines in which one purposively gave one treatment to the sicker patients is unlikely to allow one to decide which is the more effective treatment. Sub-disciplines of statistics such as *experimental design* and *survey sampling* are concerned with effective data collection strategies. Experimental design studies situations in which it is possible to manipulate the subject matter: one can choose which patient will get which treatment, one can control the temperature of a reaction vessel, etc. Survey design is concerned with situations involving observational data, in which one studies the population as it is, without being able to intervene: in a salary survey, for example, one simply records the salaries. Observational data are weaker in the sense that causality cannot be unambiguously established: with such data there is always the possibility that other factors have caused an observed correlation. With experimental data, on the other hand, one can ensure that the only difference between two groups is a controlled difference, so that this must be the cause of any observed outcome difference. Key notions in experimental design are control groups, so that like is being compared with like, and random assignment of subjects to different treatments. A key notion in survey sampling is the random selection of the sample to be analyzed. In both cases, ►randomization serves the dual roles of reducing the chance of biases which could arise (even subconsciously) if purposive selection were to be used (as in the example of giving one treatment to sicker patients), and permitting valid statistical inference.

Once the data set has been collected, one has to analyze it. There exist a huge number of statistical data analysis tools. A popular misconception is that one can think of these tools as constituting a toolbox, from which one chooses that tool which best matches the question one wishes to answer. This notion has probably been promoted by the advent of powerful and extensive software packages, such as SAS and SPSS, which have modules structured around particular analytic techniques. However, the notion is a misleading one: in fact, statistical techniques constitute a complex web of related ideas, with, for example, some being special cases of others, and others being variants applied to different kinds of data. Rather than a toolbox, it is better to think of statistics as a language, which enables one to construct a way to answer any particular scientific question. This perspective is illustrated by statistical languages such as Splus and R. Statistical tools are underwritten by complex and powerful theory, which ties them together in various ways. For example:

- We can compare two groups using a t-test.
- If we are uneasy about the t-test assumptions, we might use a nonparametric alternative, or perhaps a ►randomization test.
- The t-test can be generalized to deal with more than two groups, as in ►analysis of variance.
- And it can be generalized to deal with a continuous "independent" variable in regression.
- Analysis of variance and regression are each special cases of ►analysis of covariance.
- And all these are examples of linear models.
- Linear models can be extended by generalizing the assumed distributional forms, in ►generalized linear models.
- Analysis of variance itself can be generalized to the multivariate situation in multivariate analysis

of variance (see ►Multivariate Analysis of Variance (MANOVA)) and the general linear model (see ►General Linear Models).
- And linear discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) can be regarded as a special case of multivariate analysis of variance.
- Linear discriminant analysis is a special case of supervised classification, with other such tools being ►logistic regression, ►neural networks, support vector machines, recursive partitioning classifiers, and so on.
- And on and on.

There are some very important subdomains of statistics which have been the focus of vast amounts of work, because of the importance of the problems with which they deal. These include (but are certainly not limited to) areas such as time series analysis, supervised classification, nonparametric methods, latent variable models, neural networks, belief networks, and so on.

Certain important theoretical ideas pervade statistical thinking. I have already referred to the likelihood function as a central concept in inference. Another example is the concept of overfitting. When one seeks to model a sample of observations with a view to understanding the mechanism which gave rise to it, it is important to recognize that the sample is just that, a sample. A different sample would probably be rather different from the observed sample. What one is really seeking to do is capture the common underlying characteristics of the various possible samples, not the peculiar characteristics of the sample one happens to have drawn. Too close a fit of a model to the observed data risks capturing the idiosyncrasies of these data. There are various strategies for avoiding this, including smoothing a model, using a weaker model, averaging multiple models based on subsets of the data or random perturbations of it, adding a penalization term to the measure of goodness of fit of the model to the data so that overfitting is avoided, and others.

I have already noted how the discipline of statistics has evolved over the past two centuries. This evolution is continuing, driven by the advent of new application areas (e.g., ►bioinformatics, retail banking, etc.) and, perhaps especially, the computer. The impact of the computer is being felt in many ways. A significant one is the appearance of very large data sets – in all domains, from telecommunications, through banking and supermarket sales, to astronomy, genomics, and others. Such large data sets pose new challenges. These are not merely housekeeping ones of keeping track of the data, and of the time required to analyze them, but also new theoretical challenges. Closely related to the appearance of these very large data sets is the growth of interest in *streaming* data: data which simply keep on coming, like water from a hose. Again, such data sets are ubiquitous, and typically require real-time analysis.

The computer has also enabled significant advances through computer intensive methods, such as ►bootstrap methods and ►Markov chain Monte Carlo. Bootstrap methods approximate the relationship between a sample and a population in terms of the observed relationship between a subsample and the sample. They are a powerful idea, which can be used to explore properties of even very complex estimators and procedures. Markov chain Monte Carlo methods (see ►Markov Chain Monte Carlo) are simulation methods which have enabled the practical implementation of Bayesian approaches, which were otherwise stymied to a large extent by impractical mathematics.

Graphical displays have long been a familiar staple of statistics – on the principle that a picture is worth a thousand words, provided it is well-constructed. Computers have opened up the possibility of interactive dynamic graphics for exploring and displaying data. However, while some exciting illustrations exist, the promise has not yet been properly fulfilled – though this appears to be simply a matter of time.

Another important change driven by the computer has been the advent of other data analytic disciplines, such as machine learning, ►data mining, image processing, and pattern recognition (see ►Pattern Recognition, ►Aspects of and Statistical Pattern Recognition Principles). All of these have very considerable overlaps with statistics – to the extent that one might regard them as part of "greater statistics," to use John Chambers's phrase (Chambers 1993). Such disciplines have their own emphasis and flavor (e.g., data mining being concerned with large data sets, machine learning with an emphasis on algorithms rather than models, etc.) but it is futile to try to draw sharp distinctions between them and statistics.

From an external perspective, perhaps the single most striking thing about statistics is how pervasive it is. One cannot run a country effectively without measures of its social and economic characteristics, without knowing its needs and resources. One cannot run a corporation successfully without understanding its customer base, its manufacturing and service operations, and its workforce. One cannot develop new medicines without rigorous clinical trials. One cannot control epidemics without forecasting and extrapolation models. One cannot extract information from physics or chemistry experiments without proper statistical techniques for analyzing the resulting data. And so on and on. All of these require measurements, projections, and understanding based on statistical analysis. The fact is that the modern world is a very complex place. Statistical methods are vital tools for understanding

its complexity, grasping its subtleties, and coping with its ambiguities and uncertainties.

An excellent overview of statistics is given by Wasserman (2004), and a short introduction describing the power and fascination of the modern discipline is given by Hand (2008). Aspects of the modern discipline are set in context in Hand (2009).

About the Author

David Hand is Professor of Statistics at Imperial College, London. He previously held the Chair of Statistics at the Open University. Professor Hand is a Fellow of the Royal Statistical Society and of the British Academy, an Honorary Fellow of the Institute of Actuaries, and a Chartered Statistician. He is a past-president of the International Federation of Classification Societies, and was president of the Royal Statistical Society for the 2008–2009 term, and again in 2010. He is the second person to serve twice since Lord George Hamilton, in 1915. He was Joint Editor of the Journal of the Royal Statistical Society Series C, Applied Statistics (1989–1992). He is founding editor of *Statistics and Computing* (1991–2001). David Hand has received various awards and prizes for his research including, the Thomas L. Saaty Prize for Applied Advances in the Mathematical and Management Sciences (2001), the Royal Statistical Society's Guy Medal in Silver (2002), the IEEE International Conference on Data Mining award for Outstanding Contributions (2004) and a Royal Society Wolfson Research Merit Award (2006–2010). Professor Hand has (co-)authored over 300 papers and 26 books.

Cross References

►Agriculture, Statistics in
►Astrostatistics
►Banking, Statistics in
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bioinformatics
►Biopharmaceutical Research, Statistics in
►Biostatistics
►Business Statistics
►Careers in Statistics
►Chemometrics
►Components of Statistics
►Computational Statistics
►Confidence Interval
►Decision Theory: An Overview
►Demography
►Econometrics
►Economic Statistics
►Environmental Monitoring, Statistics Role in
►Estimation: An Overview
►Federal Statistics in the United States, Some Challenges
►Fraud in Statistics
►Industrial Statistics
►Information Theory and Statistics
►Insurance, Statistics in
►Marine Research, Statistics in
►Medical Statistics
►Misuse and Misunderstandings of Statistics
►National Account Statistics
►Philosophical Foundations of Statistics
►Prior Bayes: Rubin's View of Statistics
►Promoting, Fostering and Development of Statistics in Developing Countries
►Psychiatry, Statistics in
►Psychology, Statistics in
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Significance Testing: An Overview
►Social Statistics
►Sociology, Statistics in
►Sport, Statistics in
►Statistical Distributions: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Genetics
►Statistical Inference
►Statistical Inference in Ecology
►Statistical Inference: An Overview
►Statistical Methods in Epidemiology
►Statistical Modeling of Financial Markets
►Statistical Modelling in Market Research
►Statistical Quality Control
►Statistical Software: An Overview
►Statistics and Climate Change
►Statistics and Gambling
►Statistics and the Law
►Statistics Education
►Statistics, History of
►Statistics: Controversies in Practice
►Statistics: Nelder's view
►Tourism Statistics

References and Further Reading

Chambers JM (1993) Greater or lesser statistics: a choice for future research. Stat Comput 3:182–184

Hand DJ (2008) Statistics: a very short introduction. Oxford University Press, Oxford

Hand DJ (2009) Modern statistics: the myth and the magic (RSS Presidential Address). J R Stat Soc A 172:287–306

Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge

Barnett V (1999) Comparative statistical inference, 3rd edn. Wiley, Chichester
Wasserman L (2004) All of statistics: a concise course in statistical inference. Springer, New York

Statistics: Controversies in Practice

William Notz
Professor
The Ohio State University, Columbus, OH, USA

Controversies may arise when statistical methods are applied to real problems. The reasons vary, but some possible sources are (1) the user fails to appreciate the limitations of the methods and makes claims that are not justified, (2) the use of statistical methods is affected by non-statistical considerations, and (3) researchers disagree on the appropriate statistical methods to use. In what follows, we provide examples of controversies involving all these sources. The references allow readers to explore these examples in more detail. We hope that this article will help readers identify and assess controversies that they encounter in practice.

Example 1: Web Surveys

Using the Internet to conduct "Web surveys" is becoming increasingly popular. Web surveys allow one to collect large amounts of survey data at lower costs than traditional methods. Anyone can put survey questions on dedicated sites offering free services, thus large-scale data collection is available to almost every person with access to the Internet. Some argue that eventually Web surveys will replace traditional survey methods.

Web surveys are not easy to do well. Problems faced by those who conduct them include (1) participants may be self-selected, (2) certain members of the target population may be systematically underrepresented and (3) nonresponse. These problems are not unique to Web surveys, but how to overcome them in Web surveys is not always clear. For a more complete discussion, see Couper (2000).

Controversy arises because those who do Web surveys may make claims about their results that are not justified. The controversy can be seen in the Harris Poll Online. The Harris Poll Online has created an online research panel of over 6 million volunteers, consisting "of a diverse cross-section of people residing in the United States, as well as in over 200 countries around the world" (see www.harrispollonline.com/question.asp). When the Harris Poll Online conducts a survey, a probability sample is selected from the panel and statistical methods are used to weight the responses and provide assurance of accuracy and representativeness. As a result, the Harris Poll Online believes their results generalize to some well-defined larger population. But the panel members (and hence participants) are self-selected, and no weighting scheme can account for all the ways in which the panel is different from the target population.

Example 2: Accessibility of Data

Research in many disciplines involves the collection and analysis of data. In order to assess the validity of the research, it may be important for others to verify the quality of the data and its analysis. Scientific journals, as a rule, require that published experimental findings include enough information to allow other researchers to reproduce the results. But how much information is enough? Some argue that all data that form the basis for the conclusions in a research paper should be publicly available, or at least available to those who review the research for possible publication.

Controversy arises because of non-statistical considerations. Data collection can be time consuming and expensive. Researchers expect to use the data they collect as the basis for several research papers. They are reluctant to make it available to others until they have a chance to fully exploit the data themselves.

An example of this controversy occurred when mass spectrometry data from a sample of a fossilized femur of a *Tyrannosaurus rex* indicated that fragments of protein closely matched sequences of collagen, the most common protein found in bones, from birds (see Asara et al. 2007 and Schweitzer et al. 2007). This was the first molecular confirmation of the long-theorized relationship between dinosaurs and birds. Many researchers were skeptical of the results (see, for example, Pevzner et al. 2008). They questioned the quality of the data, the statistical analyses, and doubted that collagen could survive so long, even partially intact. Critics demanded that all the data be made publicly available. Eventually researchers posted all the spectra in an online database. Although there was evidence that some of the data may have been contaminated, a reanalysis (see Bern et al. 2009) supported the original findings.

Example 3: Placeboes in Surgery

Randomized, double-blind, placebo-controlled trials are the gold standard for evaluating new medical interventions and are routinely used to assess new medical therapies. However, only a small percentage of studies of surgery use randomized comparisons. Surgeons think their operations succeed, but even if the patients are helped, the placebo

effect may be responsible. To find out, one should conduct a proper experiment that includes a "sham surgery" to serve as a placebo. See Freeman et al. (1999) and Macklin (1999) for discussion of the use of placeboes in surgery trials.

The use of placeboes in surgery trials is controversial. Arguments against the use of placeboes include nonstatistical considerations. Placebo surgery always carries some risk, such as postoperative infection. A fundamental principle is that "the interests of the subject must always prevail." Even great future benefits cannot justify risks to subjects today unless those subjects receive some benefit. No doctor would do a sham surgery as ordinary therapy, because there is some risk. If we would not use it in medical practice, it is not ethical to use it in a clinical trial. Do these arguments outweigh the acknowledged benefits of a proper experiment?

Example 4: Hypothesis Testing in Psychology Research

Research studies in many fields rely on tests of significance. Custom may dictate that results should be significant at the 5% level in order to be published. Overreliance on statistical testing can lead to bad habits. One simply formulates a hypothesis, decides on a statistical test, and does the test. One may never look carefully at the data. The limitations of tests are so severe, the risks of misinterpretation so high, and bad habits so ingrained, that some critics in psychology have suggested significance tests be banned from professional journals in psychology.

Here the controversy involves the appropriate statistical method. To help resolve the controversy, the American Psychological Association appointed a Task Force on Statistical Inference. The Task Force did not want to ban tests. Its report (see Wilkinson 1999) discusses good statistical practice in general. Regarding hypothesis testing, the report states "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p-value or, better still, a confidence interval. … Always provide some effect-size estimate when reporting a p-value." Although banning tests might eliminate some abuses, the committee thought there were enough counterexamples to justify forbearance.

About the Author

William Notz has served as Editor of *Technometrics* and Editor of the *Journal of Statistics Education*. He is a Fellow of the American Statistical Association. He is co-author (with David Moore) of the book, *Statistics Concepts and Controversies* (W.H. Freeman and Company 7th edition). He has served as Acting Chair of the Department of Statistics and Associate Dean of the College of Mathematical and Physical Sciences at the Ohio State University.

Cross References

►Clinical Trials: An Overview
►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Internet Survey Methodology: Recent Trends and Developments
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Psychology, Statistics in
►P-Values

References and Further Reading

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. Science 316(5822):280–285

Bern M, Phinney BS, Goldberg D (2009) Reanalysis of Tyrannosaurus rex mass spectra. J Proteome Res, Article ASAP DOI: 10.1021/pr900349r, Publication Date (Web): July 15, 2009

Couper MP (2000) Web surveys: a review of issues and approaches. Public Opin Quart 64:464–494

Freeman TB, Vawter DE, Leaverton PE, Godbold JH, Hauser RA, Goetz CG, Olanow CW (1999) Use of placebo surgery in controlled trials of a cellular-based therapy for Parkinson's disease. New Engl J Med 341:988–992

Macklin R (1999) The ethical problems with sham surgery in clinical research. New Engl J Med 341:992–996

Pevzner PA, Kim S, Ng J (2008) Comment on Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. Science 321(5892):1040

Schweitzer MH, Suo Z, Avci R, Asara JM, Allen MA, Arce FT, Horner JR (2007) Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. Science 316(5822):277–280

Wilkinson L, Task Force on Statistical Inference, American Psychological Association, Science Directorate, Washington, DC, US (1999). Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54:594–604

Statistics: Nelder's View

John Nelder†
Formerly Visiting Professor
Imperial College, London, UK

Statistical Science is a Wonderful Subject

Many scientists in their training take a basic course in statistics, and from it most of them learn almost nothing

that will be useful to them in the practice of their science. In the wider world statistics has a bad name:

- There are lies, dams lies, and statistics
  You can prove anything with statistics
  and so on.

I give here a personal view of my subject, what its components are, and what can be done with it. It should have not have a bad name; rather it should be regarded as a wonderful subject in which there are many new discoveries to be made.

Statistical Science

"Statistics" is an unfortunate term, because it can refer both to data and methods used to analyze those data. I, therefore, propose to use the term "Statistical science." It embraces all the techniques that can be used to make sense of figures. In principle it can be useful in the analysis of data from any scientific experiment or survey. It is above all a scientifically useful activity. A good statistical analysis will reveal; it will not obscure. I shall use the term "statistician" as a short form of "statistical scientist."

The Components of Statistical Science

Mathematics

The statistician must know some mathematics. Certain components are vital; for example, matrix algebra and methods for describing the structures of data. Remember always that mathematics, or parts of it, are tools for the statistician in his work. One part of mathematics is special and will be described separately.

Probability Theory

The statistician's use of probability theory is primarily for the construction of statistical models. These involve the use of probability distributions of one or more random variables to describe the assumed random components in the data, that is those aspects of the data that can only be described by their mass behavior. In addition models include what are described as fixed effects, that is effects that are assumed to stay constant across different data sets. In their statistics course scientists are usually introduced to the idea of statistical significance. Many come to believe that the sole purpose of a statistical analysis is to show that a difference between the effects of two treatments applied in an experiment is significant. The statistician knows that the size of a significant difference depends both on the size of the effect, the size of the sample and the underlying variation in the measurements. It is of course important that an experiment should be big enough to show clearly the differences it is sought to measure. Why is there this mistaken stress on the idea of statistical significance? I believe that it is because it gives the lecturer an opportunity to prove some mathematical theorems from probability theory.

Very often the lecturer is only interested in probability theory, whereas the statistician's interests are much wider. It is very important to stress that statistical science is not the same as probability theory.

Statistical Inference

Here we reach what I believe to be the heart of statistical science, namely what inferences may be legitimately made from the data we are analyzing. The components are the data we have, past data on a similar topic, and a statistical model for describing the data (we hope). When we have data from a number of experiments we shall be looking for effects that are constant across these experiments, in other word looking for statistical sameness rather than statistical differences. If we can find such effects we have extended the scope of our inferences about the effects in question.

How do we come by the statistical model that drives our inferences?

Sometimes there is a standard model from past work that has stood the test of time, but quite often the statistician has to draw on his own experience to formulate a suitable model.

The inference problem then becomes "given this model, defined by a set of unknown parameters, which values of those parameters do the data point to?"

The basic idea here is that of ▸likelihood, first introduced by Fisher in the 1920s. A likelihood is not a probability, and so requires new methods for its manipulation, not covered by probability theory. Unlike random variables, which can be integrated over any part of their distribution, likelihoods can be compared only at different ordinates. Fisher introduced the idea of maximum likelihood for defining the most likely values of the parameters given the data. However it may be that the model is unsuitable for describing the data; then the inference will be false. The statistician can test this by defining a goodness-of-fit statistic and testing the statistic against its null distribution. Models can be extended by adding terms, deleting terms, or exchanging terms, or by replacing a linear term by a smooth curve driven by the data, etc.

The Experimental Cycle

Both experiments and surveys need to be designed, and the statistician can be helpful at this stage if he is knowledgeable. The need for design is still not known as well as it should be: remember that double-blind trials, now

widely used in medicine, took 30 years to become accepted! The next stage is the execution of an experiment or survey. Many things can go wrong at this stage, biases introduced and so on, and a good statistician will be aware of such things. It is a good thing if all statisticians in their training actually do at least one experiment themselves, so that they get first-hand experience of the difficulties an experimenter may encounter. After execution comes analysis, which is often of major concern to the statistician. Output from analysis will include estimates of the effects of interest to the experimenter, together with estimates of their uncertainty.

Experiments rarely stand on their own, so finally a stage of consolidation is required, where results from the current experiment are compared with previous experiments on the same topic. This is often called ►meta-analysis, though I prefer the older combination of information. This completes the experimental cycle except for writing-up of the results; then the whole cycle can start again.

The Status of Bayesian Statistics

Many statisticians espouse methods based on Bayes's theorem for the analysis of experiments. In this framework there are no fixed effects and every parameter in the model is assigned a prior probability distribution. Much has been written about making these prior distributions uninformative etc., and some Bayesians regard these as purely subjective assessments. Given data, there is still no way of checking these prior assumptions. Various theorems can be proved from the Bayesian specification, but in my view these have nothing to do with the problems of scientific inference. Indeed I regard the problem given by Bayes in his original paper as much better described by a two-stage likelihood, than by a prior probability.

The Statistician and His Clients

A statistician will usually be working with other scientists who have statistical problems in the analysis of their data. The statistician must establish a close working relation with those he is helping, and to do this it is essential to learn some of the scientist's jargon. In my first job I had to learn at least six different jargons. The statistician should encourage his clients to learn something of his own jargon, so that his methods are not thought of as being some kind of magic!

Conclusion

Statistical science has a wider scope than any other science, because the idea of inference is not subject-dependent.

Its scope is therefore huge and its processes are continually both challenging and interesting. Remember only that statistical science is not the same as probability theory; it is much wider and (I think) much more interesting.

About the Author

John Ashworth Nelder was born on 8 October 1924 in Dulverton, Somerset, England. In 1950, he was appointed Head of the statistics section at the National Vegetable Research Station at Wellesbourne. In 1968, John succeeded Frank Yates as Head of statistics at Rothamsted Experimental Station, Harpenden. At Rothamsted, he started a collaboration with Robert Wedderburn that resulted in a seminal paper on Generalized Linear Models that has enormous impact on statistical analysis. During his time at Rothamsted, he was appointed as a visiting professor at Imperial College London (1972), which led to his collaboration with Peter McCullagh in writing a book, *Generalized Linear Models* (Chapman and Hall, 2nd edition 1989). Since his retirement in 1984, he had continued as a visiting professor in the Department of Mathematics at Imperial College, London. John Nelder had received many honors for his statistical work. He was awarded the Guy Medal in Silver of the Royal Statistical Society in 1977, and elected a Fellow of the Royal Society in 1981. He had served as President both the International Biometrics Society (1978–1979) and the Royal Statistical Society (1985–1986). In 1981, the Universitè Paul Sabatier, Toulouse, granted him an Honorary Doctorate. He had published over a 100 papers. Professor Nelder is also known for his contribution to statistical computing through designing and directing the development of the statistical software packages Genstat and GLIM. Professor Nelder received the Royal Statistical Society's Guy Medal in Gold in 2005.

John Nelder died on 7th August 2010 in Luton & Dunstable Hospital, Luton, UK, where he was recovering from a fall. He had sent his contributed entry on September 17 2009, adding the following text to his email: "Here is a first draft. I hope it may be useful."

"John Nelder was one of the most influential statisticians of his generation, having an impact across the entire range of statistics, from data collection, through deep theory, to practical implementation of software." (David Hand "Professor John Nelder FRS – Obituary", Reporter, Imperial College, London, 27 August 2010.)

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning

►Clinical Trials: An Overview
►Components of Statistics
►Effect Size
►Likelihood
►Medical Research, Statistics in
►Meta-Analysis
►Model Selection
►Prior Bayes: Rubin's View of Statistics
►Probability Theory: An Outline
►Statistical Consulting
►Statistical Design of Experiments (DOE)
►Statistical Inference
►Statistics: An Overview

Stem-and-Leaf Plot

Vesna Bucevska
Associate Professor, Faculty of Economics
Ss. Cyril and Methodious University, Skopje, Macedonia

A stem-and-leaf plot (or simply stemplot), was invented by John Tukey (The idea behind the stemplot can be traced back to the work of Arthur Bowley in the early 1900s.) in his paper "Some Graphic and Semigraphic Displays" in 1972. It is a valuable tool in exploratory data analysis, since it displays the relative density and shape of data. Therefore, it is used as an alternative to the histogram. In order to construct a stem-and-leaf plot the following steps have to be taken:

1. The data have to be sorted in ascending order.
2. The stem-and-leaf units have to be determined. This means that we have to define what will be the stems and what will be the leaves for observations of interest. Each stem can consist of any number of digits, but each leaf can have only a single digit.

Data are grouped according to their leading digits, called stems, which are placed on the left side of the vertical line, while on the right hand side of the vertical line in ascending order follow the final digits of each observation called leaves. We can illustrate the way to construct a stem-and-leaf plot using the following data set for number of customers per day in a shop:

5 12 8 20 14 16 17 23 27 22 22 25 31 34 42 39 44 53 44 50 62

First we have to sort data in ascending order:

5 8 12 14 16 17 20 22 22 23 25 27 31 34 39 42 44 44 50 53 62.

Let us decide that the stem unit is 10, and the leaf unit is 1. Thus, the stem-and-leaf has the following appearance:

Stem	Leaf
0 |	5 8
1 |	2 4 6 7
2 |	0 2 2 3 5 7
3 |	1 4 9
4 |	2 4 4
5 |	0 3
6 |	2

If a stem-and leaf is turned on its side, it looks like a histogram constructed from the digits of the data. It is important to list each stem even they do not have associated leaves. If a larger number of bins is desired then there may be two stems for each digit.

If some of the observations are not integers then these numbers have to be rounded. If there are some negative numbers in data set then a minus sign has to be put in front of the stem unit.

Typically in statistical software packages (like Minitab or Statgraphics) stem-and-leaf display is preceded by another column of numbers to the left of the plot. It represents depths, which give cumulative counts from the top and bottom of the table, stopping at the row that contains the median, and the number for this row is given in parentheses. Recalling the example given above, we obtain

	Stem	Leaf
2	0 |	5 8
4	1 |	2 4 6 7
(6)	2 |	0 2 2 3 5 7
9	3 |	1 4 9
6	4 |	2 4 4
3	5 |	0 3
1	6 |	2

Although the stem and leaf plot is very similar to histogram it has some advantages over it. First, it keeps data in

their original form and the values of each individual data can be recovered from the plot. Second, it can be easily constructed without using computer, especially when the data set we are dealing with is not very large (in a range from 15 to 150 data points). For very large data set the histogram is prefered.

Cross References

►Exploratory Data Analysis
►Sturges' and Scott's Rules

References and Further Reading

Becker WE, Harnett DK (1987) Business and economics statistics with computer applications. Addison-Wesley, Reading

Montgomery D (2005) Introduction to statistical quality control, 5th edn. Wiley, New York

Newbold P, Carlson WL, Thorne B (2007) Statistics for business and economics, 6th edn. Pearson, New Jersey

Tukey JW (1972) Some graphic and semigraphic displays. In: Bancroft TA (ed) Statistical papers in honor of George W. Snedecor. Iowa State University Press, Ames, pp 293–316

Step-Stress Accelerated Life Tests

Mohamed T. Madi
Professor of Statistics, Associate Dean
UAE University, Al-Ain, United Arab Emirates

Introduction

To ascertain the service life and reliability of a product, or to compare alternative manufacturing designs, life testing at normal conditions is clearly the most reliable method. Due to continual advances in engineering science and improvement in manufacturing designs, one often deals with products that are highly reliable with a substantially long life span. Electronic products and devices (e.g., toasters, washers, and electronic chips), for example, are expected to last over a period of time much longer than what laboratory testing would allow. In these situations, the standard life testing methods may require long and prohibitively expensive testing time in order to get enough failure data necessary to make inferences about its relationship with external stress variables.

In order to shorten the testing period, test units are subjected to conditions more severe than normal. Such accelerated life testing (ALT) results in shorter lives than would be observed under normal conditions. Commonly, each test unit is run to failure at a constant stress, then a model for the relationship between the life of the unit and the constant stress is fitted to the data. This relationship is then extrapolated to estimate the life distribution of the product and get the desired information on its performance under normal use. Stress factors can include humidity, temperature, vibration, voltage, load, or any other factor affecting the life of the units. For a recent account of work on accelerated testing and test plans, we refer the reader to Nelson (2005a, b).

When constant-stress testing is considered too lengthy, step-stress testing may be used to reduce the times to failure still further. Such testing involves starting a test unit at a specified low stress. If the unit does not fail in a specified time, then the stress on it is raised to a higher value and held for another specified time. The stress is repeatedly increased and held this way until failure occurs. The time in the step-stress pattern when a test unit fails is recorded as the data on that unit. Applications of this type of testing include metal fatigue under varying load in service, cryogenic cable insulation, and electronics applications to reveal failure modes (elephant testing), so they can be designed out of the product.

When more constraints on the length of a life test are present, some form of censoring is commonly adopted. If for example, removing unfailed items from the life test at prespecified times is adopted, we have type I censoring. Instead, if we terminate the life test at the time of a failed item and remove all remaining unfailed items from the test, we have type II censoring.

One advantage of step-stress accelerated life testing (SSALT) is that the experimenters need not start with a high stress that could be harsh for the product, hence avoiding excessive extrapolation of test results. The obvious drawback is that it requires stronger assumptions and more complex analysis, compared to constant-stress ALT.

The simplest form of SSALT is the partial ALT introduced by DeGroot and Goel (1979) and in which the products are first tested under use conditions for a period of time before the stress is increased and maintained at the higher level throughout the test. They modeled the effect of switching the stress from normal conditions stress to the single accelerated stress by multiplying the remaining lifetime of the item by some unknown factor $\alpha > 0$. They studied the issues of estimation and optimal design in the framework of Bayesian decision theory.

Another formulation of this type of ALT, called the cumulative exposure (CE) model, was proposed by Nelson (1980). It assumes that the remaining life of test units depends on the current cumulative fraction failed and current stress. Survivors will fail according to the cdf for that stress but starting at the previously accumulated fraction failed. Nelson (1980) and Miller and Nelson (1983) studied

maximum likelihood estimation (MLE) under this type of parametric model when the underlying distribution is taken to be the Weibull and exponential, respectively.

Bhattacharyya and Soejoeti (1988) proposed the tampered failure rate (TFR) model for SSALT. Their model assumes that a change in the stress has a multiplicative effect on the failure rate function over the remaining life. In the special setting of a two-step partially accelerated life test, and assuming that the initial distribution belongs to a two-parameter Weibull family, they studied MLE and derived the Fisher information matrix.

There are mainly two types of SSALTs: a simple SSALT where there is a single change of stress during the test and mutliple-step SSALT where change of the stress occurs more than once. Madi (1993) generalized the TFR model from the simple step-stress model to the multiple step-stress model.

Acceleration Models and Lifetime Distributions

Stress Functions

Unless a nonparametric approach is used (see Shaked and Singpurwalla (1983), McNichols and Padgett (1988), and Tyoskin and Krivolapov (1996)), an SSALT model (ALT model in general) consists of a theoretical life distribution whose parameters are functions of accelerating stress and unknown coefficients to be estimated from the test data. These simple relationships, called stress functions, are widely used in practice, and special cases include the Arrhenius, inverse power, and Eyring laws (see Nelson (1990)). For example, Nelson (1980) used the Weibull with parameters (α, β), as the lifetime distribution, where the scale parameter α depends on stress according to an inverse power law $\alpha(V) = (V_0/V)^p$.

Lifetime Distribution Under Step-stress Pattern

The Cumulative Exposure Model

The basic idea for this model, introduced by Nelson (1980), is to assume that the remaining life of specimens depends only on the current cumulative fraction failed and current stress, regardless of how the fraction accumulated. Specifically, if we let F_i denote the cumulative distribution function (cdf) of the time to failure under stress s_i, the cdf of the time to failure under a step-stress pattern, F_0, is obtained by considering that the lifetime t_{i-1}under s_{i-1} has an equivalent time u_i under s_i such that $F_{i-1}(t_{i-1}) = F_i(u_i)$. Then the model is built as follows:

We assume that the population cumulative fraction of specimens failing under stress s_1, in Step 1, is

$$F_0(t) = F_1(t),\ 0 \le t \le t_1$$

In Step 2, we write $F_2(u_1) = F_1(t_1)$ to obtain u_1 that is the time to failure that would have produced the population cumulative fraction failing under s_2. The population cumulative fraction of specimens failing in Step 2 by total time t is

$$F_0(t) = F_2(t - t_1 + u_1),\ t_1 \le t \le t_2$$

Similarly, in Step 3, the unit has survived Step 2 and we consider an equivalent time u_2 under s_3 such that

$$F_3(u_2) = F_2(t_2 - t_1 + u_1)$$

where $t_2 - t_1 + u_1$ is an equivalent time under s_2. Then we have

$$F_0(t) = F_3(t - t_2 + u_2),\ t_2 \le t \le t_3$$

In general, Step i has the equivalent start time u_{i-1} that is the solution of

$$F_i(u_{i-1}) = F_{i-1}(t_{i-1} - t_{i-2} + u_{i-2})$$

and

$$F_0(t) = F_i(t - t_{i-1} + u_{i-1}),\ t_{i-1} \le t \le t_i$$

Finally, the CE model can then be written as

$$F_0(t) = \begin{cases} F_1(t), & 0 \le t \le t_1 \\ F_2(t - t_1 + u_1), & t_1 \le t \le t_2 \\ F_3(t - t_2 + u_2), & t_2 \le t \le t_3 \\ \cdots & \cdots \\ \cdots & \cdots \\ F_i(t - t_{i-1} + u_{i-1}), & t_{i-1} \le t \le t_i \end{cases}$$

$u_0 = t_0 = 0$ and u_i is the solution of $F_{i+1}(u_i) = F_i(t_i - t_{i-1} + u_{i-1})$, for $i = 1, \ldots, m-1$.

If the stress function is taken to be the inverse power law and F_i is a ►Weibull distribution, then the cdf for the fraction of specimens failing by time t for the constant stress V_i is

$$F_i(t) = 1 - \exp[-\{t(V_i/V_0)^p\}^\beta],$$

and for $t_{i-1} \le t \le t_i$,

$$F_0(t) = 1 - \exp[-\{(t - t_{i-1} + u_{i-1})(V_i/V_0)^p\}^\beta].$$

The Tampered Failure Rate Model

Consider the experiment in which n units are simultaneously put on test at time $t_0 = 0$ to a stress setting x_1. Starting at time $t_2 > 0$, the surviving units are subjected to a higher stress level x_2 while in the time interval $[t_1, t_2)$. At time t_2, the stress is increased on the surviving units to x_3 over $[t_2, t_3)$ and so on until the kth and last time interval $[t_{k-1}, \infty)$, where the remaining units are subjected to x_k until they all fail. The TFR model assumes that the effect of changing the stress from x_{i-1} to x_i is to multiply the failure rate function by α_{i-1}. The resulting step-stress failure rate function is given by

$$\lambda^*(t) = \left(\prod_{i=0}^{j-1} \alpha_i\right) \lambda_1(t), \; t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

where $t_0 = 0, t_k = \infty$ and $\alpha_{-1} = \alpha_0 = 1$. The corresponding survival function is

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \overline{F}(t_i)^{(1-\alpha_i)\prod_{l=-1}^{i-1}\alpha_l}\right) \overline{F}(t)^{\prod_{i=0}^{j-1}\alpha_i},$$
$$t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

Substituting the Weibull survival function with scale parameter θ and shape parameter β, $\overline{F}(t) = \exp[-(y/\theta)^\beta]$, $\overline{F}^*(t)$ becomes

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \exp\left\{\left(\prod_{l=-1}^{i} \alpha_l\right)\left(\frac{t_i}{\theta}\right)^\beta - \left(\prod_{l=-1}^{i-1} \alpha_l\right)\left(\frac{t_i}{\theta}\right)^\beta\right\}\right)$$
$$\times \exp\left\{-\left(\prod_{i=0}^{j-1} \alpha_i\right)\left(\frac{t}{\theta}\right)^\beta\right\}$$

Putting $\delta_j = \theta\left(\prod_{i=0}^{j-1} \alpha_i\right)^{-\beta^{-1}}$, we have

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \exp\left\{(t_i/\delta_{i+1})^\beta - (t_i/\delta_i)^\beta\right\}\right)$$
$$\times \exp\left\{-(t/\delta_j)^\beta\right\},$$

which can be rewritten as

$$\overline{F}^*(t) = \exp\left\{\sum_{i=0}^{j-1}\left((t_i/\delta_{i+1})^\beta - (t_i/\delta_i)^\beta\right)\right\}$$
$$\times \exp\left\{-(t/\delta_j)^\beta\right\}, \quad t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

Inference

Different fitting methods can be used in the context of SSALT. They include maximum likelihood estimation, ►least squares, best linear unbiased, and graphical methods. MLE is used frequently because it is straightforward and yields approximate variances and confidence limits for the parameters and percentiles. The major drawback is the computational complexity. The estimators are rarely obtained in closed form and extensive iterative methods must be used to determine the MLE.

Recent inferential work based on maximum likelihood for the CE model under different censoring schemes include Gouno et al. (2004), Zhao and Elsayed (2005), Wu et al. (2006), Balakrishnan and Xie (2007a, b), and Balakrishnan and Han (2008). Madi (1993) considered the MLE for the multiple step-stress TFR model when the life distribution under constant stress is Weibull.

Optimal Designs

Different optimization criteria have been used to design SSALT plans. Most are based on the variance of the MLE of the parameter of interest (variance optimality) or the determinant of the Fisher information matrix (D-optimality). One question arising is is on the duration that items need to be exposed to each stress level.

For example, Miller and Nelson (1983) presented optimal design for simple SSALT under the assumption of an exponential distribution. Their optimization criterion is to minimize the asymptotic variance of the MLE of the mean at a specified design stress. This criterion leads to optimizing the levels of the first and the second test stresses and the time of stress change. Bai et al. (1989) extended their work to the case in which a prescribed censoring time is involved. Gouno et al. (2004) considered the multiple SSALT with equal duration steps τ and progressive type I censoring and addressed the problem of optimizing τ using variance optimality as well as D-optimality.

About the Author

Dr. Mohamed Madi is a Professor, Department of Statistics, and Associate Dean, College of Business and Economics, UAE University, United Arab Emirates. He was the Assistant Dean for Research and Director of the UAEU Research Affairs Unit for Internally Funded Projects (2005–2008). He has authored and coauthored more than 30 papers and one book. Professor Madi has received the College of Business and Economics 2008 Outstanding Senior Research Award. He is Associate editor for the *Journal of Statistical Theory & Applications*, USA, and the *Jordan Journal of Mathematics and Statistics*, Jordan.

Cross References

►Accelerated Lifetime Testing
►Censoring Methodology

►Degradation Models in Reliability and Survival Analysis
►Generalized Weibull Distributions
►Industrial Statistics
►Modeling Survival Data
►Ordered Statistical Data: Recent Developments
►Parametric and Nonparametric Reliability Analysis
►Significance Testing: An Overview
►Survival Data

References and Further Reading

Bai DS, Kim MS, Lee SH (1989) Optimum simple step-stress accelerated life tests with censoring. IEEE Trans Reliab 38:528–532
Balakrishnan N, Han D (2008) Exact inference for a simple step-stress model with competing risks for failure from exponential distribution under Type-II censoring. J Stat Plan Infer 138(12):4172–4186
Balakrishnan N, Xie Q (2007a) Exact inference for a simple step-stress model with Type-II hybrid censored data from the exponential distribution. J Stat Plan Infer 137(8):2543–2563
Balakrishnan N, Xie Q (2007b) Exact inference for a simple step-stress model with Type-I hybrid censored data from the exponential distribution. J Stat Plan Infer 137(11):3268–3290
Bhattacharyya GK, Soejoeti Z (1988) A tampered failure rate model for step-stress accelerated test. Commun Stat Theory Meth 18(5):1627–1643
DeGroot MH, Goel PK (1979) Bayesian estimation and optimal design in partially accelerated life testing. Nav Res Logist Q 26:223–235
Gouno E, Sen A, Balakrishnan N (2004) Optimal step-stress test under progressive Type-I censoring. IEEE Trans Reliab 53:383–393
Madi MT (1993) Multiple step-stress accelerated life test; the tampered failure rate model. Commun Stat Theory Meth 22(9):2631–2639
McNichols DT, Padgett WJ (1988) Inference for step-stress accelerated life tests under arbitrary right-censorship. J Stat Plan Infer 20(2):169–179
Miller R, Nelson W (1983) Optimum simple step-stress plans for accelerated life testing. IEEE Trans Reliab 32:59–65
Nelson W (1980) Accelerated life testing: step-stress models and data analysis. IEEE Trans Reliab 29:103–108
Nelson W (1990) Accelerated testing: statistical models, test, plans and data analyses. Wiley, New York
Nelson WB (2005a) A bibliography of accelerated test plans. IEEE Trans Reliab 54:194–197
Nelson WB (2005b) A bibliography of accelerated test plans. Part II. IEEE Trans Reliab 54:370–373
Shaked M, Singpurwalla ND (1983) Inference for step-stress accelerated life tests. J Stat Plan Infer 7(4):295–306
Tyoskin OI, Krivolapov SY (1996) Nonparametric model for step-stress accelerated life testing. IEEE Trans Reliab 45:346–350
Wu SJ, Lin YP, Chen YJ (2006) Planning step-stress life test with progressively type I group-censored exponential data. Stat Neerl 60:46–56
Zhao W, Elsayed EA (2005) A general accelerated life model for step-stress testing. IIE Trans 37:1059–1069

Stochastic Difference Equations and Applications

Alexandra Rodkina[2], Cónall Kelly[1,2]
[1]Professor and Head
University of the West Indies,
Mona Campus, Kingston, Jamaica
[2]University of the West Indies, Mona Campus, Kingston, Jamaica

A first-order difference equation of the form

$$x_{n+1} = F(n, x_n), \quad n \in \mathbb{N}, \tag{1}$$

may be used to describe phenomena that evolve in discrete time, where the size of the each generation is a function of that preceding. But the real world often refuses to conform to such a neat mathematical representation. Unpredictable effects can be included in the form of a sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$, and the result is a *stochastic difference equation*:

$$X_{n+1} = F(n, X_n) + G(n, X_n)\xi_{n+1}, \quad n \in \mathbb{N}. \tag{2}$$

The solution of (2) is a discrete time stochastic process adapted to the natural filtration of $\{\xi_n\}_{n\in\mathbb{N}}$. Stochastic difference equations also arise as discretizations of ►stochastic differential equations, though their asymptotic properties can be harder to analyze. Although a thorough introduction to the theory of deterministic difference equations can be found in Elaydi (2005) (for example), no comparable text exists for their stochastic counterparts. Nonetheless the recent development of powerful analytic tools is driving research efforts forward, and our understanding of discrete stochastic dynamics is growing. This has implications both for the modeling of real-world phenomena that evolve in discrete time, and the analysis of numerical methods for stochastic differential equations. Both are discussed in this article.

Mathematical biology is a good place to look for real-world phenomena that evolve in discrete time (see Murray 2002). Certain species, for example periodic cicadas and fruit flies, reproduce in non-overlapping generations, and the change in biomass from one generation to the next may be represented as a stochastic difference equation of the form

$$X_{n+1} = X_n\left[N(X_n) + Q(X_n)\xi_{n+1}\right], \quad n \in \mathbb{N}. \tag{3}$$

Notice that the form of (3) guarantees the existence of an equilibrium solution at $X \equiv 0$, corresponding to absence of the species. The sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$

captures random influences like disease and natural variability in fecundity between generations. In order to model predator-prey interaction, competition or mutualism, it is essential to have a good understanding of the role of the coefficient functions N and Q in the dynamics of systems of such equations. For example, an equilibrium solution displaying almost sure asymptotic stability indicates that a species is not viable in the long run, as its biomass will decay to an unsustainable level over time. In the stochastic context, *almost sure* means *with probability one* and is usually written *a.s.*

Theoretical tools for investigating the a.s. asymptotic stability of the equilibrium of the similar equation

$$X_{n+1} = X_n\left[1 + R(X_n) + Q(X_n)\xi_{n+1}\right], \quad n \in \mathbb{N}, \tag{4}$$

were developed in Appleby et al. (2009a), in the form of a semi-martingale convergence theorem and a discrete form of the Itô formula. It turns out that the relative speed of decay of R and Q close to equilibrium determines the a.s. asymptotic stability of the equilibrium. One consequence of this is that an unstable equilibrium in a deterministic system may be stabilized by an appropriate perturbation coefficient Q. In the special case where R and Q are polynomials, a more detailed description is possible. If the a.s. stability is a result of a dominant R then solutions decay at an exact power law rate, however if the system has been stabilized by a dominant Q no such rate is possible. Moreover, solutions can be shown to change sign a random (though finite) number of times, indicating that discrete equations with stabilizing noise may be inappropriate in the context of a population model: biomass is inherently nonnegative. The closely related question of the role played by R and Q in the oscillatory behavior of solutions of (4) was investigated in Appleby et al. (2010).

The influence of random perturbations can be hidden from any observer of a single trajectory. In Rodkina (2009) it was shown that when R and Q are polynomial, there exist solutions of (4) that, with arbitrarily high probability, converge to zero monotonically and inside a well-defined deterministic envelope. The fluctuations that ordinarily characterize the presence of random noise are absent. This phenomenon is impossible in continuous time, since solutions of stochastic differential equations have trajectories that are non-differentiable almost everywhere.

Stochastic difference equations also find applications in economic modeling. Consider a self regulating island economy in the tropics, and suppose one wishes to model the effects of the annual hurricane season on economic activity. The essential mechanism underlying dynamic equilibrium in an idealized model of such an economy can be represented by the equation

$$x_{n+1} = x_n + f(x_n), \quad n \in \mathbb{N}, \tag{5}$$

under appropriate conditions on f (see Appleby et al. (2008) for details).

The degree to which activity during a hurricane season influences such a model varies randomly from year to year, depending on the number and intensity of storm systems, and how close the centre of each storm passes to the island. These effects may be incorporated by adding the term $\sigma_n\xi_{n+1}$ at each iteration, where again $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of independent random variables, and each σ_n represents intensity of seasonal activity. Notice that including a state-independent perturbation in the model destroys the equilibrium.

In Appleby et al. (2008) it was shown that, if (5) is globally asymptotically stable, the perturbed model will eventually return to the vicinity of the former equilibrium, provided the intensity of seasonal activity converges to zero sufficiently quickly. However, no matter how effective the self-regulatory property of the system, if the seasonal activity fades out more slowly than a critical rate, which depends on the "heaviness" of the tails of the distribution of each ξ_n, then the system will not return to the former equilibrium. Hence (in this model), even if seasonal activity lessens each year, the economy may be prevented from settling back to near-equilibrium if the storms that do occur tend to be extremely violent. For models which are only locally stable in the absence of perturbations, the potential exists for an external shock to push a fundamentally stable economic situation over into instability.

Stochastic difference equations arise in numerical analysis, since they are the end product of the discretization of a stochastic differential equation. Consider

$$dX(t) = f(X(t))dt + g(X(t))dB(t), \quad t \geq 0, \tag{6}$$

where B is a standard Brownian motion. In general, solutions of (6) cannot be written in closed form; to explore their properties we can try to simulate them on a computer. Since computers are finite-state machines we must discretize the time set of (6) with, for example, a one-step Euler-Maruyama numerical scheme on a uniform mesh. This yields the stochastic difference equation

$$X_{n+1} = X_n + hf(X_n) + \sqrt{h}g(X_n)\xi_{n+1}, \quad n \in \mathbb{N}, \tag{7}$$

where $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. standard normal random variables, and h is the mesh size. A good discussion of numerical methods for stochastic differential equations may be found in Kloeden and Platen (1992).

But discretization can alter the very properties of (6) that we are trying to examine. For example, a geometric Brownian motion (see ►Brownian Motion and Diffusions) with positive initial value remains positive with probability one. However, the Euler-Maruyama discretization does not: discrete processes can jump across equilibrium given a sufficiently large input from the stochastic component. This is a concern as geometric Brownian motion is often used to model asset prices in financial markets, which (like biomass in the population model) are inherently nonnegative. However, the probability of positivity can be increased over a finite simulation interval by increasing the density of mesh-points.

Nonetheless, any practical simulation must be carried out with a fixed non-zero stepsize h, so it is also necessary to study the effect of discretization with fixed h on carefully chosen test equations with known dynamics. A linear stability analysis seeks to discover when the asymptotic stability of an equilibrium solution of the test equation is preserved after discretization. Direct analysis of solutions of the stochastic difference equation arising from the discretization is necessary. Since these solutions are stochastic processes, asymptotic stability may be defined in several ways, each of which speaks to a difference aspect of the process. For example, a.s. asymptotic stability is a property of almost all trajectories, whereas mean-square asymptotic stability is a property of the distribution.

The literature surrounding mean-square stability analysis of stochastic numerical methods is extensive. For example an analysis of the stochastic θ-method using a scalar geometric Brownian motion as test equation may be found in Higham (2000), with an extension to systems of two equations in Saito and Mitsui (2002), using a technique outlined in Kloeden and Platen (1992). By contrast, developments in a.s. asymptotic stability analysis are more recent: Rodkina and Schurz (2005) have investigated a.s. asymptotic stability for the θ-method applied to a scalar stochastic differential equation, and Higham et al. (2007) have shown that a.s. exponential asymptotic stability in systems of equations with linearly bounded coefficients can be recovered in a θ-discretisation for sufficiently small h. We anticipate an expansion of the literature in the coming years.

Finally, we comment that it is often possible to reproduce a specific continuous-time dynamic in a discrete stochastic process by through careful manipulation of the mesh, presenting two examples from the literature. First, a.s. oscillatory behavior in linear stochastic differential equations with a fading point delay has been reproduced in Appleby and Kelly (2006) using a pre-transformation of the differential equation and a mesh that contracts at the same rate as the delay function. Second, state-dependent meshes have been used to reproduce finite-time explosions in a discretization of (6) (see for example Dávila et al., 2005).

About the Authors

Professor Alexandra Rodkina is Head of the Department of Mathematics, University of the West Indies, Jamaica. She has authored and co-authored more than 200 papers and three books, and is a member of the Editorial Board of the *International Journal of Difference Equations*.

Dr. Cónall Kelly is a lecturer at the same department, and is the author of 13 papers. Together, Professor Rodkina and Dr. Kelly have published four papers and organized special sessions at three conferences.

Cross References

►Brownian Motion and Diffusions
►Statistical Aspects of Hurricane Modeling and Forecasting
►Stochastic Differential Equations

References and Further Reading

Appleby JAD, Kelly C (2006) Oscillation of solutions of a nonuniform discretisation of linear stochastic differential equations with vanishing delay. Dy Contin Discret Impuls Syst A 13B(suppl):535–550

Appleby JAD, Berkolaiko G, Rodkina A (2008) On local stability for a nonlinear difference equation with a non-hyperbolic equilibrium and fading stochastic perturbations. J Differ Equ Appl 14(9):923–951

Appleby JAD, Berkolaiko G, Rodkina A (2009a) Non-exponential stability and decay rates in nonlinear stochastic difference equations with unbounded noise. Stochastics 81(2):99–127

Appleby JAD, Kelly C, Mao X, Rodkina A (2010) On the local dynamics of polynomial difference equations with fading stochastic perturbations. Dy Contin Discret Impuls Syst A 17(3):401–430

Appleby JAD, Rodkina A, Schurz H (2010) Non-positivity and oscillations of solutions of nonlinear stochastic difference equations with state-dependent noise. J Differ Equ Appl 6(7):807–830

Dávila J, Bonder JF, Rossi JD, Groisman P, Sued M (2005) Numercial analysis of stochastic differential equations with explosions. Stoch Anal Appl 23(4):809–825

Elaydi S (2005) An introduction to difference equations, 3rd edn. Undergraduate Texts in Mathematics. Springer, New York

Hasminski RZ (1981) Stochastic stability of differential equations. Sijthoff and Noordhoff, Alpen aan den Rijn – Germantown, Md

Higham DJ (2000) Mean-square and asymptotic stability of the stochastic theta method. SIAM J Numer Anal 38:753–769

Higham DJ, Mao X, Yuan C (2007) Almost sure and moment exponential stability in the numerical simulation of stochastic differential equations. SIAM J Numer Anal 45:592–609

Kelly C, Rodkina A (2009) Constrained stability and instability of polynomial difference equations with state-dependent noise. Discret Contin Dyn Syst B 11(4):913–933

Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Springer, Berlin
Murray JD (2002) Mathematical biology 1: an introduction. Interdisciplinary applied mathematics, vol 17. Springer, New York
Rodkina A, Schurz H (2005) Almost sure asymptotic stability of drift implicit θ-methods for bilinear ordinary stochastic differential equation in R1. J Comput Appl Math 180:13–31
Saito Y, Mitsui T (2002) Mean-square stability of numerical schemes for stochastic differential systems. Vietnam J Math 30:551–560

Stochastic Differential Equations

Peter E. Kloeden
Professor
Institut fur Mathematik, Frankfurt, Germany

A scalar stochastic differential equation (SDE)

$$dX_t = f(t, X_t)\, dt + g(t, X_t)\, dW_t \tag{1}$$

involves a the Wiener process W_t, $t \geq 0$, which is one of the most fundamental ▶stochastic processes and is often called a Brownian motion (see ▶Brownian Motion and Diffusions). A Wiener process is a Gaussian process with $W_0 = 0$ with probability 1 and $\mathcal{N}(0, t-s)$-distributed increments $W_t - W_s$ for $0 \leq s < t$ where the increments $W_{t_2} - W_{t_1}$ and $W_{t_4} - W_{t_3}$ on non-overlapping intervals, (i.e., with $0 \leq t_1 < t_2 \leq t_3 < t_4$) are independent random variables. It follows from the Kolmogorov criterion that the sample paths of a Wiener process are continuous. However, they are nowhere differentiable.

Consequently, an SDE is not a differential equation at all, but only a symbolic representation for the stochastic integral equation

$$X_t = X_{t_0} + \int_{t_0}^{t} f(s, X_s)\, ds + \int_{t_0}^{t} g(s, X_s)\, dW_s,$$

where the first integral is a deterministic Riemann integral for each sample path. The second integral cannot be defined pathwise as a Riemann-Stieltjes integral because the sample paths of the Wiener process do not have even bounded variation on any bounded time interval, but requires a new type of stochastic integral. An Itô stochastic integral $\int_{t_0}^{T} f(t) dW_t$ is defined as the mean-square limit of sums of products of an integrand f evaluated at the left end point of each partition subinterval times $[t_n, t_{n+1}]$ the increment of the Wiener process, i.e.,

$$\int_{t_0}^{T} f(t) dW_t := \text{m.s.} - \lim_{N_\Delta \to \infty} \sum_{j=0}^{N_\Delta - 1} f(t_n)\ (W_{t_{n+1}} - W_{t_n}),$$

where $t_{n+1} - t_n = \Delta / N_\Delta$ for $n = 0, 1, \ldots, N_\Delta - 1$. The integrand function f may be random or even depend on the path of the Wiener process, but $f(t)$ should be independent of future increments of the Wiener process, i.e., $W_{t+h} - W_t$ for $h > 0$.

The Itô stochastic integral has the important properties (the second is called the Itô isometry) that

$$\mathbb{E}\left[\int_{t_0}^{T} f(t) dW_t\right] = 0,$$
$$\mathbb{E}\left[\left(\int_{t_0}^{T} f(t) dW_t\right)^2\right] = \int_{t_0}^{T} \mathbb{E}\left[f(t)^2\right] dt.$$

However, the solutions of Itô SDE satisfy a different chain rule to that in deterministic calculus, called the Itô formula, i.e.,

$$U(t, X_t) = U(t_0, X_{t_0}) + \int_{t_0}^{t} L^0 U(s, X_s)\, ds + \int_{t_0}^{t} L^1(s, X_s)\, dW_s,$$

where

$$L^0 U = \frac{\partial U}{\partial t} + f \frac{\partial U}{\partial x} + \frac{1}{2} g^2 \frac{\partial^2 U}{\partial x^2}, \qquad L^1 U = g \frac{\partial U}{\partial x}.$$

An immediate consequence is that the integration rules and tricks from deterministic calculus do not hold and different expressions result, e.g.,

$$\int_0^T W_s\, dW_s = \frac{1}{2} W_T^2 - \frac{1}{2} T.$$

There is another stochastic integral called the Stratonovich integral, for which the integrand function is evaluated at the mid-point of each partition subinterval rather than at the left end point. It is written with $\circ dW_t$ to distinguish it from the ▶Itô integral. A Stratonovich SDE is thus written

$$dX_t = f(t, X_t)\, dt + g(t, X_t) \circ dW_t.$$

Note that the Itô and Stratonovich versions of an SDE may have different solutions, e.g.,

$$dX_t = X_t\, dW_t \quad \Rightarrow X_t = X_0 e^{W_t - \frac{1}{2} t} \qquad \text{Itô}$$
$$dX_t = X_t \circ dW_t \Rightarrow X_t = X_0 e^{W_t} \qquad \text{Stratonovich}$$

However, the Itô SDE (1) has the same solutions as the Stratonovich SDE with the modified drift coefficient, i.e.,

$$dX_t = \underline{f}(t, X_t)\, dt + g(t, X_t) \circ dW_t, \qquad \underline{f} := f - \frac{1}{2} g \frac{\partial g}{\partial x}.$$

In particular, the Itô and Stratonovich versions of an SDE with additive noise, i.e., with g independent of x, are the same.

Stratonovich stochastic calculus has the same chain rule as deterministic calculus, which means that Stratonovich SDE can be solved with the same integration tricks as for ordinary differential equations. However, Stratonovich stochastic integrals do not satisfy the nice properties above for Itô stochastic integrals, nor does the Stratonovich SDE have the same direct connection with diffusion process theory as the Itô SDE, e.g., the coefficient of the Fokker-Planck equation correspond to those of the Itô SDE (1), i.e.,

$$\frac{\partial p}{\partial t} + f\frac{\partial}{\partial x} + \frac{1}{2}g^2\frac{\partial^2 p}{\partial x^2} = 0.$$

The Itô and Stratonovich stochastic calculi are both mathematically correct. Which one should be used is really a modeling issue, but once one has been chosen, the advantages of the other can be used through the above drift modification.

The situation for vector valued SDE and vector valued Wiener processes is similar. Details can be found in the given references.

About the Author

Peter Kloeden graduated with a B.A. (with First Class Honors) in Mathematics from Macquarie University in Sydney, Australia. He received his Ph.D. in Mathematics from the University of Queensland in 1975 under the supervision of Rudolf Vyborny. In 1995, he also received a Doctor of Science in Mathematics from the University of Queensland. After 20 years of teaching at various universities in Australia, he was appointed in 1997 to the Chair in Applied and Instrumental Mathematics at the Johann Wolfgang Goethe University in Frankfurt. Professor Kloeden received the 2006 W.T. and Idalia Reid Prize by the Society of Industrial and Applied Mathematics, USA, for his fundamental contributions to the theoretical and computational analysis of differential equations. In 2009 he was elected a Fellow of the Society of Industrial and Applied Mathematics. He is Associate editor of a number of journals including: *Journal of Nonlinear Analysis: Theory, Methods and Applications*, *Journal of Stochastic Analysis*, *SINUM*, *Discrete and Continuous Dynamical Systems – Series B*, *Nonlinear Dynamics and Systems Theory*, *Advances in Dynamical Systems and Applications (ADSA)*, *Stochastics and Dynamics*, *Journal of Stochastic Analysis and Applications*, *Advanced Nonlinear Studies* and *International Journal of Dynamical Systems and Differential Equations*.

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Itô Integral
►Numerical Methods for Stochastic Differential Equations
►Optimal Statistical Inference in Financial Engineering
►Sampling Problems for Stochastic Processes
►Stochastic Difference Equations and Applications
►Stochastic Modeling Analysis and Applications
►Stochastic Modeling, Recent Advances in
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Kloeden PE, Platen E (1992) The numerical solution of stochastic differential equations. Springer, Berlin (3rd revised edition, 1999)

Øksendal B (2003) Stochastic differential equations. an introduction with applications. Springer, Berlin (6th edition, Corr. 4th printing, 2007)

Stochastic Global Optimization

Anatoly Zhigljavsky
Professor, Chair in Statistics
School of Mathematics, Cardiff University, Cardiff, UK

Stochastic global optimization methods are methods for solving a global optimization problem incorporating probabilistic (stochastic) elements, either in the problem data (the objective function, the constraints, etc.), or in the algorithm itself, or in both.

Global optimization is a very important part of applied mathematics and computer science. The importance of global optimization is primarily related to the applied areas such as engineering, computational chemistry, finance and medicine amongst many other fields. For the state of the art in the theory and methodology of global optimization we refer to the "Journal of Global Optimization" and two volumes of the "Handbook of Global Optimization" (Horst and Pardalos 1995; Pardalos and Romeijn 2002). If the objective function is given as a "black box" computer code, the optimization problem is especially difficult. Stochastic approaches can often deal with problems of this kind much easier and more efficiently than the deterministic algorithms.

The problem of global minimization. Consider a general minimization problem $f(x) \to \min_{x \in X}$ with objective

function $f(\cdot)$ and feasible region X. Let x^* be a global minimizer of $f(\cdot)$; that is, x^* is a point in X such that $f(x^*) = f_*$ where $f_* = \min_{x \in X} f(x)$. Global optimization problems are usually formulated so that the structure of the feasible region X is relatively simple; this can be done on the expense of increased complexity of the objective function.

A global minimization algorithm is a rule for constructing a sequence of points $x_1, x_2, \ldots$ in X such that the sequence of record values $y_{on} = \min_{i=1\ldots n} f(x_i)$ approaches the minimum f_* as n increases. In addition to approximating the minimal value f_*, one often needs to approximate at least one of the minimizers x_*.

Heuristics. Many stochastic optimization algorithms where randomness is involved have been proposed heuristically. Some of these algorithms are based on analogies with natural processes; the well-known examples are evolutionary algorithms (Glover and Kochenberger 2003) and simulated annealing (Van Laarhoven and Aarts 1987). Heuristic global optimization algorithms are very popular in applications, especially in discrete optimization problems. Unfortunately, there is a large gap between practical efficiency of stochastic global optimization algorithms and their theoretical rigor.

Stochastic assumptions about the objective function. In deterministic global optimization, Lipschitz-type conditions on the objective function are heavily exploited. Much research have been done in stochastic global optimization where stochastic assumptions about the objective function are used in a manner similar to how the Lipschitz condition is used in deterministic algorithms. A typical example of a stochastic assumption of this kind is the postulation that $f(\cdot)$ is a realization of a certain stochastic process. This part of stochastic optimization is well described in Zhigljavsky and Zilinskas (2008), Chap. 4 and will not be pursued in this article.

Global random search (GRS). The main research in stochastic global optimization deals with the so-called global random search (GRS) algorithms which involve random decisions in the process of choosing the observation points. A general GRS algorithm assumes that a sequence of random points $x_1, x_2, \ldots, x_n$ is generated where for each $j \geq 1$ the point x_j has some probability distribution P_j. For each $j \geq 2$, the distribution P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on the results of the objective function evaluations at these points (the function evaluations may not be noise-free). The number of points $n, 1 \leq n \leq \infty$ (the stopping rule) can be either deterministic or random and may depend on the results of function evaluation at the points $x_1, \ldots, x_n$.

Three important classes of GRS algorithms. In the algorithm which is often called 'pure random search' (PRS) all the distributions P_j are the same (that is, $P_j = P$ for all j) and the points x_j are independent. In Markovian algorithms the distribution P_j depends only on the previous point x_{j-1} and $f(x_{j-1})$, the objective function value at x_{j-1}. In the so-called population-based algorithms the distributions P_j are updated only after a certain number of points with previous distribution have been generated.

Attractive features of GRS. GRS algorithms are very popular in both theory and practice. Their popularity is owed to several attractive features that many global random search algorithms share: (a) the structure of GRS algorithms is usually simple; (b) these algorithms are often rather insensitive to the irregularity of the objective function behavior, to the shape of the feasible region, to the presence of noise in the objective function evaluations, and even to the growth of dimensionality; (c) it is very easy to construct GRS algorithms guaranteeing theoretical convergence.

Drawbacks of GRS. Firstly, the practical efficiency of the algorithms often depends on a number of parameters, but the problem of the choice of these parameters frequently has little relevance to the theoretical results concerning the convergence of the algorithms. Secondly, for many global random search algorithms an analysis on good parameter values is lacking or just impossible. Thirdly, the convergence rate can be painfully slow, see discussion below. Improving the convergence rate (or efficiency of the algorithms) is a problem that much research in the theory of global random search is devoted to.

Main principles of GRS. A very large number of specific global random search algorithms exist, but only a few main principles form their basis. These principles can be summarized as follows:

(1) Random sampling of points at which $f(\cdot)$ is evaluated,
(2) Random covering of the space,
(3) Combination with local optimization techniques,
(4) The use of different heuristics including cluster-analysis techniques to avoid clumping of points around a particular local minima,
(5) Markovian construction of algorithms,
(6) More frequent selection of new trial points in the vicinity of "good" previous points,
(7) Use of statistical inference, and
(8) Decrease of randomness in the selection rules for the trial points.

In constructing a particular global random search method, one usually incorporates several of these principles, see Zhigljavsky and Zilinskas 2008 where all these principles are carefully considered.

Convergence of GRS. To establish the convergence of a particular GRS algorithm, the classical Borel-Cantelli theorem (see ►Borel–Cantelli Lemma and Its Generalizations) is usually used. The corresponding result can be formulated as follows, see Zhigljavsky and Zilinskas 2008, Theorem 2.1. Assume that $X \subseteq \mathbb{R}^d$ with $0 < \mathrm{vol}(X) < \infty$ and $\sum_{j=1}^{\infty} \inf P_j(B(x,\varepsilon)) = \infty$ for all $x \in X$ and $\varepsilon > 0$, where $B(x,\varepsilon) = \{y \in X : \|y-x\|_2 \le \varepsilon\}$ and the infimum is taken over all possible locations of previous points $x_1, \ldots, x_{j-1}$ and the results of the objective function evaluations at these points. Then with probability one, the sequence of points $x_1, x_2, \ldots$ falls infinitely often into any fixed neighborhood of any global minimizer.

In practice, a very popular rule for selecting the sequence of probability measures P_j is $P_j = \alpha_j P_0 + (1-\alpha_j) Q_j$, where $0 \le \alpha_j \le 1$, P_0 is the uniform distribution on X and Q_j is an arbitrary probability measure on X. In this case, the corresponding GRS algorithm converges if $\sum_{j=1}^{\infty} \alpha_j = \infty$.

Rate of convergence of PRS. Assume $X \subseteq \mathbb{R}^d$ with $\mathrm{vol}(X) = 1$ and the points $x_1, x_2, \ldots, x_n$ are independent and have uniform distribution on X (that is, GRS algorithm is PRS). The rate of convergence of PRS to the minimizer x_* is the fastest possible (for the worst continuous objective function) among all GRS algorithms. To guarantee that PRS reaches the ε-neighborhood $B(x_*,\varepsilon)$ of a point x_* with probability at least $1-\gamma$, we need to perform at least $n_* = \left\lceil -\log(\gamma) \cdot \Gamma\left(\frac{d}{2}+1\right) / \left(\pi^{\frac{d}{2}} \varepsilon^d\right) \right\rceil$ iterations, where $\Gamma(\cdot)$ is the Gamma-function. This may be a very large number even for reasonable values of d, ε and γ. For example, if $d = 10$ and $\varepsilon = \gamma = 0.1$ then $n_* \simeq 0.9 \cdot 10^{10}$. See Sect. 2.2.2 in Zhigljavsky and Zilinskas (2008) for an extensive discussion on convergence and convergence rates of PRS and other GRS algorithms.

Markovian GRS algorithms. In a Markovian GRS algorithm, the distribution P_j depends only on the previous point x_{j-1} and its function value $f(x_{j-1})$; that is, the sequence of points $x_1, x_2, \ldots$ constitutes a Markov chain (see ►Markov Chains). The most known Markovian GRS algorithms are the simulated annealing methods (Van Laarhoven and Aarts 1987). If a particular simulated annealing method creates a time-homogeneous Markov chain then the corresponding stationary distribution of this Markov chain is called Gibbs distribution. Parameters of the simulated annealing can be chosen so that the related Gibbs distribution is concentrated in a narrow neighborhood of the global minimizer x_*. The convergence to the Gibbs distribution can be very slow resulting in a slow convergence of the corresponding simulated annealing algorithm. The convergence of all Markovian GRS algorithms is generally slow as the information about the objective function obtained during the search process is used ineffectively.

Population-based methods. Population-based methods are very popular in practice (Glover and Kochenberger 2003). These methods generalize the Markovian GRS algorithms in the following way: rather than to allow the distribution P_j of the next point x_j to depend on the previous point x_{j-1}, it is now the distribution of a population of points (descendants, or children) depends on the previous population of points (parents) and the objective function values at these points. There are many heuristic arguments associated with these methods (Glover and Kochenberger 2003). There are also various probabilistic models of the population-based algorithms (Zhigljavsky 1991).

Statistical inference in GRS. The use of statistical procedures can significantly accelerate the convergence of GRS algorithms. Statistical procedures can be especially useful for defining the stopping rules and the population sizes in the population-based algorithms. These statistical procedures are based on the use of the asymptotic theory of extreme order statistics and the related theory of record moments. As an example, consider PRS and the corresponding sample $S = \{f(x_j), j = 1, \ldots, n\}$. This is an independent sample of values from the distribution with c.d.f. $F(t) = \int_{f(x) \le t} P(dx)$ and the support $[f_*, f^*]$, where $f^* = \sup_{x \in X} f(x)$. It can be shown that under mild conditions on f and P, this distribution belongs to the domain of attraction of the ►Weibull distribution, one of the ►extreme value distributions. Based on this fact, one can construct efficient statistical procedures for f_* using several minimal order statistics from the sample S.

For the theory, methodology and the use of probabilistic models and statistical inference in GRS, we refer to Zhigljavsky and Zilinskas (2008) and Zhigljavsky (1991).

Cross References

►Borel–Cantelli Lemma and Its Generalizations
►Markov Chains
►Weibull Distribution

About the Author

Professor Zhigljavsky is Director of the Center for Optimization and Its Applications at Cardiff University.

References and Further Reading

Glover F, Kochenberger GA (2003) Handbook on metaheuristics. Kluwer Academic, Dordrecht
Horst R, Pardalos P (eds) (1995) Handbook of global optimization. Kluwer Academic, Dordrecht
Pardalos P, Romeijn E (eds) (2002) Handbook of global optimization, vol 2. Kluwer Academic, Dordrecht

Van Laarhoven PJM, Aarts EHL (1987) Simulated annealing: theory and applications. D. Reidel, Dordrecht
Zhigljavsky A, Zilinskas A (2008) Stochastic global optimization. Springer, New York
Zhigljavsky A (1991) Theory of global random search. Kluwer Academic, Dordrecht

Stochastic Modeling, Recent Advances in

Christos H. Skiadas
Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece

The term Stochastic Modeling is related to the theory and applications of probability in the modeling of phenomena in real life applications. Stochastic is a term coming from the ancient Greek period and is related to "*stochastes*" (people who are philosophers or intellectuals, scientists in recent notation) and "*stochazomai*" (I am involved in highly theoretical and intellectual issues as are philosophy and science).

The term model accounts for the representation of the reality (a real situation) by a verbal, logical or mathematical form. It is clear that the model includes a part of the main characteristics of the real situation. As far as the real situation is better explained the model will be termed as successful or not.

The science or even the art to construct and apply a model to real situations is termed as modeling. It includes model building and model adaptation, application to specific data and even simulation; that is producing a realization of a real situation.

It is clear that it is essential to organise and apply a good method or even process of collecting, restoring, classifying, organising and fitting data related to the specific case; that is to develop the "data analysis" scientific field.

Modeling is related to the use of past data to express the future developments of a real system. To this end modeling accounts for two major intellectual and scientific schools; the school of determinism and the school of probabilistic or stochastic modeling.

Deterministic modeling is related to determinism; that is the expression of the reality with a modeling approach that uses the data from the past and could lead to a good and even precise determination of the future paths of a natural system. Determinism was a school of thought that was the basis of very many developments in various scientific fields last centuries. Deterministic models of innovation diffusion appear in (Skiadas 1985, 1986, 1987).

From the other part, it was clear from the very beginning even from the rising of philosophy and science from the ancient Greek period that the future was unpredictable (probabilistic) or even chaotic. However, the successful solutions of several problems last centuries, especially in physics, straighten determinism as a school of thought. Probabilistic methods came more recently with many applications. Of course the basic elements were developed during the last centuries but with only few applications. Some of the famous contributors are P.-S. Laplace and J.C.F. Gauss. A main development was done by studying and modeling the heat transfer by proposing and solving a partial differential equation for the space and time propagation of heat (see Fourier (1822, 1878) and Fick (1855)). However radical progress came by modeling the Brownian motion, Brown (1828), (see the seminal paper by Einstein (1905) followed by Smoluchowski (1906)). (See also ▶Brownian Motion and Diffusions)

Modeling by Stochastic Differential Equations

Time was needed to understand and introduce probabilistic ideas into differential equations; thus called ▶stochastic differential equations. This was achieved only during the twentieth century. Even more some very important details were missing. One important point had to do with calculus and how to apply calculus in stochastic differential equations. The solution came with Itô and his postulate that the infinitesimal second order terms of a stochastic process do not vanish thus accepting to apply rules of what is now called as the Itô calculus or stochastic calculus. Stochastic calculus is also proposed by others differentiating their work from Itô's calculus on the summation process applied in defining the stochastic integral (R.L. Stratonovich, P. Malliavin). Itô's proposition can be given in the following form useful to apply in stochastic differential equations, Oksendal (1989), Gardiner (1990):

$$df(x_t,t) = \frac{\partial f(x_t,t)}{\partial x_t}dx_t + \frac{1}{2}\frac{\partial^2 f(x_t,t)}{\partial {x_t}^2}(dx_t)^2$$

where x_t is a stochastic process over time t and $f(x_t,t)$ is a stochastic function of the specific process.

The above form for the function $f(x_t,t)$ usually is used as a transformation function to reduce a nonlinear stochastic differential equation to a linear one and thus finding a closed form solution.

Although the first proposal of a probabilistic differential equation is merely due to P. Langevin, in recent years it

was generally accepted the following stochastic differential equations form:

$$dx_t = \mu(x_t, t)dt + \sigma(x_t, t)dw_t,$$

where w_t is the so-called Wiener process. This is a stochastic process with mean value zero and variance 1 and is usually termed as the standard Wiener process with $N(0,1)$ property, the process is characterized by independent increments normally distributed.

By using the above Itô's rule and the appropriate transformation function the exact solutions of several nonlinear stochastic differential equations arise. Except of the usefulness of the exact solutions of stochastic differential equations when dealing with specific cases and applications their use is very important in order to check how precise the approximate methods of solution of stochastic differential equations are. A general method of solution was proposed by Kloeden et al. (1992, 1999, 2003). Related theoretical solutions with applications can be found in Skiadas et al. (1993, 1994), Giovanis and Skiadas (1995), Skiadas and Giovanis (1997), Skiadas (2010).

The main stochastic differential equations solved can be summarized into two categories: The stochastic differential equations with a multiplicative error term of the form: $dx_t = \mu(x_t, t)dt + \sigma(t)x_t dw_t$, frequently used in market applications, and the stochastic differential equations with non-multiplicative or additive error term of the form: $dx_t = \mu(x_t, t)dt + \sigma(t)dw_t$. In the later case there appear applications with a constant σ.

The most known model with a multiplicative error term is the so-called Black and Scholes (1973) model in finance: $dx_t = \mu x_t dt + \sigma x_t dw_t$ (in most applications x_t is replaced by S_t).

The famous Ornstein–Uhlenbeck (1930) process is the most typical model with an additive error term: $dx_t = \vartheta(\mu - x_t)dt + \sigma dw_t$.

There are very many stochastic differential equations that could find interesting applications. As it was shown (Skiadas-Katsamaki 1995) even a general stochastic exponential model could give realistic paths especially during the first stages of a diffusion process: $dx_t = \mu(x_t)^b dt + \sigma dw_t$. In the same paper three methods for estimating the parameter σ are given.

Modeling using stochastic differential equations has several applications but also faces the problems arising from the introduction of stochastic theory. First of all, a stochastic differential equation gives a solution which may provide several stochastic paths during a simulation. However, one cannot find one final path as it is the case in a deterministic process. In most cases the deterministic solution arises by eliminating the error term. An infinite number of stochastic paths could provide the mean value of the stochastic process as a limit of a summation. When there exists an exact solution of the stochastic differential equation it can be estimated the mean value and if possible the variance. More useful, after estimating the mean value and the variance, is the estimation of the confidence intervals, thus informing regarding the limits of the real life application modeled.

Acknowledgments

For biography *see* the entry ►Chaotic Modelling.

Cross References

►Brownian Motion and Diffusions
►Chaotic Modelling
►Ito Integral
►Numerical Methods for Stochastic Differential Equations
►Optimal Statistical Inference in Financial Engineering
►Probability Theory: An Outline
►Stochastic Differential Equations
►Stochastic Modeling Analysis and Applications
►Stochastic Models of Transport Processes
►Stochastic Processes

References and Further Reading

Kloeden PE, Schurz H, Platten E, Sorensen M (1992). On effects of discretization on estimators of drift parameters for diffusion processes. Research Report no. 249, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81(3):637–654

Brown R (1828) A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Philos Mag 4:161–173

Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik 17:549–560

Fick A (1855) Über Diffusion. Poggendorff's Annalen 94:59–86

Fick A (1855b) On liquid diffusion. Philos Mag J Sci 10:31–39

Fourier J (1822) Theorie analytique de la chaleur. Firmin Didot, Paris

Fourier J (1878) The analytical theory of heat. Cambridge University Press, Cambridge

Gardiner CW (1990) Handbook of stochastic methods for physics, chemistry and natural science, 2nd edn. Springer, Berlin

Giovanis AN, Skiadas CH (1995) Forecasting the electricity consumption by applying stochastic modeling techniques: the case of Greece. In: Janssen J, Skiadas CH, Zopounidis C (eds) Advances in applying stochastic modeling and data analysis. Kluwer Academic, Dordrecht

Itô K (1944) Stochastic integral. In: Proceedings of the imperial academy of Tokyo, vol 20, pp 519–524

Itô K (1951) On stochastic differential equations. Mem Am Math Soc 4:1–51

Katsamaki A, Skiadas CH (1995) Analytic solution and estimation of parameters on a stochastic exponential model for a technological diffusion process. Appl Stoch Model Data Anal 11(1):59-75
Kloeden PE, Platen E (1999) Numerical solution of stochastic differential equations. Springer, Berlin
Kloeden PE, Platen E, Schurz H (2003) Numerical solution of SDE through computer experiments. Springer, Berlin
Oksendal B (1989) Stochastic differential equations: an introduction with applications, 2nd edn. Springer, New York
Skiadas CH (1985) Two generalized rational models for forecasting innovation diffusion. Technol Forecast Soc Change 27: 39–61
Skiadas CH (1986) Innovation diffusion models expressing asymmetry and/or positively or negatively influencing forces. Technol Forecast Soc Change 30:313–330
Skiadas CH (1987) Two simple models for the early and middle stage prediction of innovation diffusion. IEEE Trans Eng Manag 34:79–84
Skiadas CH (2010) Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential. Meth Comput Appl Probab 12(2):261–270
Skiadas CH, Giovanis AN (1997) A stochastic bass innovation diffusion model studying the growth of electricity consumption in Greece. Appl Stoch Model Data Anal 13:85–101
Skiadas CH, Giovanis AN, Dimoticalis J (1993) A sigmoid stochastic growth model derived from the revised exponential. In: Janssen J, Skiadas CH (eds) Applied stochastic models and data analysis. World scientific, Singapore, pp 864–870
Skiadas CH, Giovanis AN, Dimoticalis J (1994) Investigation of stochastic differential models: the Gompertzian case. In: Gutierez R, Valderama Bonnet MJ (eds) Selected topics on stochastic modeling. World Scientific, Singapore, pp 296–310
Smoluchowski M (1906) Zur kinetischen theorie der Brownschen molekularbewegung und der suspensionen. Ann D Phys 21: 756–780
Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian motion. Phys Rev 36:823–41

Stochastic Modeling Analysis and Applications

Anil G. Ladde[1], Gangaram S. Ladde[2]
[1]Chesapeake Capital Corporation, Richmond, VA, USA
[2]Professor
University of South Florida, Tampa, FL, USA

The classical random flow and Newtonian mechanics are two theoretical approaches to analyze dynamic processes in biological, engineering, physical and social sciences under random perturbations. Historically, in the classical approach (Bartlett; 1969, Ross; 1971), one considers a dynamic system as a random flow or process with a certain probabilistic laws such as: diffusion, Markovian, nonmarkovian and etc. From this type consideration, one attempts to determine the state transition probability distributions/density functions (STPDF) of the random process. The determination of the unknown STPDF leads to the study of deterministic problems in the theory of ordinary or partial or integro-differential equations (Lakshmikantham and Leela 1969a, b). For example, a random flow that obeys a Markovian probabilistic law leads to

$$\frac{\partial}{\partial s}P(s,x,t,B) = q(s,x)P(s,x,t,B) - \int_{R^n-\{x\}} P(s,y,t,B)Q(s,x,dy), \tag{1}$$

that is, Kolmogorov's backward equation, where, $P(s,x,t,B)$ is STPDF; $Q(s,x,dy)$ is the state transition intensity function (STIF) and $q(s,x) = -Q(s,x,\{x\})$. In particular, in the case of Markov chain (see ►Markov Chains) with finite number of states r, equation (1) reduces to:

$$\frac{\partial}{\partial s}P(s,t) = Q(s)P(s,t),\ P(t,t) = I, \tag{2}$$

where, $P_{ij}(s,t) = P(s,i,t,\{j\})$; $P(s,t) = (P_{ij}(s,t))_{r\times r}$; an intensity matrix $Q(s)$ and the identity I are $r \times r$ matrices. These types of equations are referred as master equations in the literature (Arnold 1974; Bartlett 1960; Gīhman 1972; Gikhman and Skorokhod 1969; Goel and Richter-Dyn 1974; Kimura and Ohta 1971; Kloeden and Platen 1992; Ladde 1991; Ladde and Sambandham 2004; Ricciardi 1977; Soong 1973). The solution processes of such differential equations are used to find the higher moments and other statistical properties of dynamic processes described by random flows or processes in sciences. We remark that in general, Kolmogorov's backward or forward (master equations) are nonlinear and non stationary deterministic differential equations (Arnold 1974; Gīhman 1972; Gikhman and Skorokhod 1969; Goel and Richter-Dyn 1974; Ricciardi 1977; Soong 1973). As a result of this, the close form STPDF are not feasible.

A modern approach (Arnold 1974; Gīhman 1972; Ito 1951, Kloeden and Platen 1992; Ladde and Ladde 2009; Ladde 1991; Ladde and Lakshmikantham 1980; Ladde and Sambandham 2004; Nelson 1967; Øksendal 1985; Ricciardi 1977; Soong 1973; Wong 1971) of stochastic modeling of dynamic processes in sciences and engineering sciences is based on fundamental theoretical information, a practical experimental setup and basic laws in science and engineering sciences. Depending on the nature of stochastic disturbances, there are several probabilistic models, namely, ►Random walk, Poisson, Brownian motion (see

►Brownian Motion and Diffusions), Colored Noise processes. In the following, we very briefly outline the salient features of Random Walk and Colored Noise dynamic modeling approaches (Kloeden and Platen 1992; Ladde and Ladde 2009; Wong 1971).

Random Walk Modeling Approach (Ladde and Ladde 2009)

Let $x(t)$ be a state of a system at a time t. The state of the system is observed over an interval of $[t, t+\Delta t]$, where Δt is a small increment in t. Without loss in generality, it is assumed that $x(t)$ is 1-dimensional state and Δt is positive. The state is under the influence of random perturbations. We experimentally observe the data-set of the state: $x(t_0) = x(t), x(t_1), x(t_2), \ldots, x(t_i), \ldots, x(t_n) = x(t+\Delta t)$ of a system at $t_0 = t, t_1 = t+\tau, t_2 = t+2\tau, \ldots, t_i = t+i\tau, \ldots, t_n = t+\Delta t = t+n\tau$ over the interval $[t, t+\Delta t]$, where n belongs to $\{1, 2, 3, \ldots\}$ and $\tau = \frac{\Delta t}{n}$. These observations are made under the following conditions:

RWM 1 The system is under the influence of independent and identical random impulses that are taking place at $t_1, t_2, \ldots, t_i, \ldots, t_n$.

RWM 2 The influence of a random impact on the state of the system is observed on every time subinterval of length τ.

RWM 3 For each $i \in I(1,n) = \{1, 2, \ldots, k, \ldots, n\}$, it is assumed that the state is either increased by $\Delta x(t_i)$ ("success"-the positive increment $(\Delta x(t_i) > 0)$) or decreased by $\Delta x(t_i)$ ("failure"-the negative increment $(\Delta x(t_i) < 0)$). We refer $\Delta x(t_i)$ as a microscopic/local experimentally or knowledge-base observed increment to the state of the system at the ith impact on the subinterval of length τ.

RWM 4 It is assumed that $\Delta x(t_i)$ is constant for $i \in I(1,n)$ and is denoted by $\Delta x(t_i) \equiv Z_i = Z$ with $|Z_i| = \Delta x > 0$. Thus, for each $i \in I(1,n)$, there is a constant random increment Z of magnitude Δx to the state of the system per impact on the subinterval of length τ.

RWM 5 For each random impact and any real number p satisfying $0 < p < 1$, it is assumed that

$$P(\{Z_i = \Delta x > 0\}) = p \text{ and } P(\{Z_i = -\Delta x < 0\}) = 1 - p = q. \tag{3}$$

From RWM1, RWM2 and RWM3, under n independent and identical random impacts, the initial state and n experimental or knowledge-base observed random increments Z_i of constant magnitude Δx in the state, the aggregate change of the state of the system $x(t+\Delta t) - x(t)$ under n observations of the system over the given interval $[t, t+\Delta t]$ of length Δt is described by

$$x(t+\Delta t) - x(t) = n\frac{\left[\sum_{i=1}^{n} Z_i\right]}{n} = \frac{\Delta t}{\tau} S_n, \tag{4}$$

where $S_n = \frac{1}{n}\left[\sum_{i=1}^{n} Z_i\right]$ and $Z_i = x(t_i) - x(t_{i-1})$. S_n is the sample average of the state aggregate incremental data. It is clear that $x(t+\Delta t) - x(t) = x(t_n) - x(t)$ is a discrete-time-real-valued stochastic process which is the sum of n independent Bernoulli random variables Z_i $(Z_i = Z)$, $i = 1, 2, \ldots, n$. We also note that for each n, $x(t_n) - x(t_0)$ is a binomial random variable with parameters (n, p). Moreover, the random variable $x(t_n) - x(t)$ takes values from the set $\{-n\Delta x, (2-n)\Delta x, \ldots, (2m-n)\Delta x, \ldots, n\Delta x\}$. The stochastic process $x(t_n) - x(t)$ is referred to as a *Random Walk process*. Let m be a number of positive increments Δx to the state of the system out of total n changes. $(n-m)$ is the number of negative increments $-\Delta x$ to the state of the system out of total n changes. Furthermore, $m \in I(0,n)$, we further note that

$$S_n = \frac{1}{n}\left[(2m-n)S_n^+\right], \tag{5}$$

where $S_n^+ = \frac{1}{n}\left[\sum_{i=1}^{n} |Z_i|\right]$.

Therefore, the *aggregate change* of state, $x(t+\Delta t) - x(t)$ under n identical random impacts on the system over the given interval $[t, t+\Delta t]$ of time is described by

$$x(t+\Delta t) - x(t) = \frac{1}{n}(2m-n)\frac{S_n^+}{\tau}\Delta t. \tag{6}$$

Moreover, from (6), we have:

$$E[x(t+\Delta t) - x(t)] = (p-q)\frac{S_n^+}{\tau}\Delta t \tag{7}$$

and

$$\mathrm{Var}(x(t+\Delta t) - x(t)) = 4pq\frac{\left(S_n^+\right)^2}{\tau}\Delta t. \tag{8}$$

$\frac{S_n^+}{\tau}$ and $\frac{(S_n^+)^2}{\tau}$ are *sample microscopic or local average increment* and *sample microscopic or local average square increment* per unit time over the uniform length of sample subintervals $[t_{k-1}, t_k]$, $k = 1, 2, \ldots, n$ of interval $[t, t+\Delta t]$, respectively.

We note that the physical nature of the problem imposes certain restrictions on Δx and τ. Similarly, the parameter p cannot be taken arbitrary. In fact, the following conditions seem to be natural for sufficiently large n:

For $x(t+\Delta t)-x(t)=n\Delta x$, $\Delta t=n\tau$, $4pq=(p+q)^2-(p-q)^2=1-(p-q)^2$, and

$$\lim_{\tau\to 0}\left[\frac{\left(S_n^+\right)^2}{\tau}\right]=2D,\ \lim_{\Delta x\to 0}\lim_{\tau\to 0}\left[(p-q)\frac{S_n^+}{\tau}\right]=C \text{ and } \lim_{\Delta x\to 0}\lim_{\tau\to 0}4pq=1, \tag{9}$$

where C and D are certain constants, the former is called a *drift* coefficient, and the latter is called a *diffusion* coefficient. Moreover, C can be interpreted as the *average/mean/expected rate of change of state* of the system per unit time, and D can be interpreted as the mean square rate of change of the system per unit time over an interval of length Δt. From (7), (8) and (9), we obtain

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}E[x(t+\Delta t)-x(t)]=C\Delta t, \tag{10}$$

and

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}\mathrm{Var}(x(t+\Delta t)-x(t))=2D\Delta t. \tag{11}$$

Now, we define

$$y(t,n,\Delta t)=\frac{x(t+\Delta t)-x(t)-n(p-q)S_n^+}{\sqrt{4npq\left(S_n^+\right)^2}}. \tag{12}$$

By the application of the DeMoivre–Laplace Central Limit Theorem, we conclude that the process $y(t,n,\Delta t)$ is approximated by standard normal random variable for each t (zero mean and variance one). Moreover,

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)=\frac{x(t+\Delta t)-x(t)-C\Delta t}{\sqrt{2D\Delta t}}. \tag{13}$$

For fixed Δt, the random variable $\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)$ has standard normal distribution (zero mean and variance one). Now, by rearranging the expressions in (13), we get

$$x(t+\Delta t)-x(t)=C\Delta t+\sqrt{2D}\,\Delta w(t) \tag{14}$$

where $\sqrt{\Delta t}\left[\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)\right]=\Delta w(t)=w(t+\Delta t)-w(t)$, $w(t)$ is a Wiener process. Thus the aggregate change of state of the system $x(t+\Delta t)-x(t)$ in (14) under independent and identical random impacts over the given interval $[t,t+\Delta t]$ is interpreted as the sum of the average/expected/mean change $(C\Delta t)$ and the mean square change $\left(\sqrt{2D}\,\Delta w(t)\right)$ of state of the system due to the random environmental perturbations.

If Δt is very small, then its differential $dt=\Delta t$, and from (14) the Itô–Doob differential dx is defined by

$$dx(t)=C\,dt+\sqrt{2D}\,dw(t), \tag{15}$$

where C and D are as defined before. The equation in (15) is called the Itô–Doob type stochastic differential equation (Arnold 1974; Gīhman and Skorohod 1972; Ito 1951; Kloeden and Platen 1992; Laddle and Laddle 2009; Laddle and Lakshmikantham 1980; Øksendal 1985; Soong 1973; Wong 1971).

Observation (1) We recall that the experimental or knowledge base observed constant random variables: $x(t_0)=x(t),Z_1,Z_2,\ldots,Z_k,\ldots,Z_n$ in (4) are mutually independent. Therefore, expectations

$$E[x(t+\Delta t)-x(t)] \text{ and } E\left[(x(t+\Delta t)-x(t))^2\right]=\mathrm{Var}(x(t+\Delta t)-x(t))$$

in (7) and (8) can be replaced by the conditional expectations as:

$$E[x(t+\Delta t)-x(t)]=E[x(t+\Delta t)-x(t)\mid x(t)=x] \tag{16}$$

and

$$\mathrm{Var}(x(t+\Delta t)-x(t))=E\left[(x(t+\Delta t)-x(t))^2\mid x(t)=x\right]. \tag{17}$$

(2) We further note that based on experimental observations, information and basic scientific laws/principles in biological, chemical, engineering, medical, physical and social sciences, we infer that in general the magnitude of the microscopic or local increment depends on both the initial time t and the initial state $x(t)\equiv x$ of a system. As a result of this, in general, the drift (C) and the diffusion (D) coefficients defined in (9) need not be absolute constants. They may depend on both the initial time t and the initial state $x(t)\equiv x$ of the system, as long as their dependence on t and x is very smooth. From this discussion, (16) and (17), one can incorporate both time and state dependent random environmental perturbation effects. As a result of this, (14) reduces to:

$$x(t+\Delta t)-x(t)=C(t,x)\Delta t+\sigma(t,x)\Delta w(t), \tag{18}$$

where $C(t,x)$ and $\sigma^2(t,x)=2D(t,x)$ are also referred to as the average/expected/mean rate and the mean square rate of the state of the system on the interval of length Δt. Moreover, the Itô–Doob type stochastic differential equation (15) becomes:

$$dx(t)=C(t,x)\,dt+\sigma(t,x)dw(t). \tag{19}$$

(3) From (16), (17) and (19), we have

$$\frac{d}{dt}E[x(t)\mid x(t)=x]=C(t,x), \tag{20}$$

$$dx=C(t,x)dt+\sigma(t,x)\xi(t)\,dt, \tag{21}$$

$$dx=C(t,x)\,dt \tag{22}$$

where $w(t)$ is the Wiener process and $\xi(t)$ is the white noise process. We further remark that either (19) or (21) is considered as a stochastic perturbation of deterministic

differential equation (22). The random terms $\sigma(t,x)\,dw(t)$ and $\sigma(t,x)\xi(t)$ in the right-hand side of (19) and (21), respectively, can be, normally, interpreted as random perturbations caused by the presence of microscopic and/or the imperfectness of the controlled conditions, either known or unknown and/or either environmental or internal fluctuations in the parameters in $C(t,x)$. It is this idea that motivates us to build a more general and feasible stochastic mathematical model for dynamic processes in biological, chemical, engineering, medical, physical and social sciences.

Sequential Colored Noise Modeling Approach (Ladde and Ladde 2009; Wong 1971)

The idea is to start with a deterministic mathematical model (22) that is based on phenomenological or biological or chemical/medical/physical social laws and the knowledge of system or environmental parameter(s). From Observation (3), one can identify parameter(s) and the source of random internal or environmental perturbations of parameter(s) of the mathematical model (22), and formulate a stochastic mathematical model in general form as:

$$dx = F(t,x,\xi(t))\,dt, \quad x(t_0) = x_0, \tag{23}$$

and, in particular,

$$dx = C(t,x)\,dx + \sigma(t,x)\xi(t)\,dt, \quad x(t_0) = x_0, \tag{24}$$

where ξ is a stochastic process that belongs to $R[[a,b], R[\Omega,R]]$; rate functions F, $C(t,x)$ and $\sigma(t,x)$ are sufficiently smooth, and are defined on $[a,b]\times R$ into R, $x_0 \in R$ and $t_0 \in [a,b]$. If the sample paths $\xi(t,\omega)$ of $\xi(t)$ are smooth functions (sample continuous), then one can utilize the usual deterministic calculus, and can look for the solution process determined by (23) and (24). We note that such a solution process is a random function with all sample paths starting at x_0. In general this is not feasible, for example, if $\xi(t)$ in (23) or (24) is a Gaussian process. The sequential colored noise modeling (CNM) approach alleviates the limitations of a one-shot modeling approach. The basic ideas are as follows:

CNM 1 Let us start with a sequence $\{\xi_n(t)\}_{n=1}^{\infty}$ of sufficiently smooth (sample path wise continuous) Gaussian processes which converges in some sense to a Gaussian white noise process $\xi(t)$ in (24). For each n, we associate a stochastic differential equation with a smooth random process as follows:

$$dx_n = C(t,x_n) + \sigma(t,x_n)\,\xi_n(t)\,dt, \quad x_n(t_0) = x_0 \tag{25}$$

where $C(t,x)$ and $\sigma(t,x)$ are described in (24).

CNM 2 We assume that the IVP (25) has a unique solution process. The IVP (25) generates a sequence $\{x_n(t)\}_{n=1}^{\infty}$ of solution processes corresponding to the chosen Gaussian sequence $\{\xi_n(t)\}_{n=1}^{\infty}$ in CNM1.

CNM 3 Under reasonable conditions on rate functions $C(t,x)$, $\sigma(t,x)$ in (24) and a suitable convergent sequence of Gaussian processes $\{\xi_n(t)\}_{n=1}^{\infty}$ in CNM1, it is shown that the sequence of solution processes $\{x_n(t)\}_{n=1}^{\infty}$ determined by (25) converges in almost surely or in quadratic mean or even in probability to a process $x(t)$. Moreover, $x(t)$ is the solution process of (24).

CNM 4 The above described ideas CNM1, CNM2 and CNM3 make a precise mathematical interpretation of (24). However, we still need to show that (24) can be modeled by an Itô–Doob form of stochastic differential equation (19). Moreover, one needs to highlight on the concept of convergence of $\{\xi_n(t)\}_{n=1}^{\infty}$ to the white noise process in (24). For this purpose, we define

$$w_n(t) - w_n(t_0) = \int_{t_0}^{t} \xi_n(s)\,ds, \tag{26}$$

and rewrite the IVP (25) into its equivalent integral form:

$$\begin{aligned} x_n(t) &= x_n(t_0) + \int_{t_0}^{t} C(s,x_n(s))\,ds \\ &\quad + \int_{t_0}^{t} \sigma(s,x_n(s))\,\xi_n(s)\,ds \\ &= x_n(t_0) + \int_{t_0}^{t} C(s,x_n(s))\,ds \\ &\quad + \int_{t_0}^{t} \sigma(s,x_n(s))\,dw_n(s). \end{aligned} \tag{27}$$

CNM 5 To conclude the convergence of $\{x_n(t)\}_{n=1}^{\infty}$, we need to show the convergence of both terms in the right-hand side of (27). The procedure for showing this convergence generates the following two mathematical steps:

Step 1: This step is to establish the following as in Ladde and Ladde (2009) and Wong (1971):

$$\begin{aligned} \lim_{n\to\infty}[y_n(t)] &= \lim_{n\to\infty}\left[\int_{t_n}^{t} \phi(s,w_n(s))\,dw_n(s)\right] \\ &= \int_{t_0}^{t} \phi(s,w(s))\,dw(s) \\ &\quad + \frac{1}{2}\int_{t_0}^{t} \frac{\partial}{\partial z}\phi(s,w(s))\,ds, \end{aligned} \tag{28}$$

where ϕ is a known smooth function of two variables. This is achieved by considering a deterministic partial indefinite integral of a given smooth deterministic function ϕ:

$$\psi(t,x) = \int_0^x \phi(t,z)\,dz. \tag{29}$$

Step 2: This step deals with the procedure of finding a limit of the sequence of the solution process $\{x_n(t)\}_{n=1}^{\infty}$ determined by (25) or its equivalent stochastic differential equation (27) as in Ladde and Ladde; 2009 and Wong; 1971:

$$dx_n = C(t,x_n)\,dt + \sigma(t,x_n)\,dw_n(t),$$
$$x_n(t_0) = x_0, \tag{30}$$

where $w_n(t)$ is as defined in (26). For this purpose, we assume that $\sigma(t,z)$ in (24) satisfies the conditions: $\sigma(t,z) \neq 0$, and it is continuously differentiable. We set $\phi(t,z) = \frac{1}{\sigma(t,z)}$ in (29). Under the smoothness conditions on rate functions C, σ and imitating the procedure outlined in Step 1, one can conclude that $\{x_n(t)\}_{n=1}^{\infty}$ converges to a process $x(t)$ on $[t_0,b]$. The final conclusion is to show that $x(t)$ satisfies the following Itô–Doob type stochastic differential equation:

$$dx = \left[C(t,x) + \frac{1}{2}\sigma(t,x)\,\frac{\partial}{\partial x}\,\sigma(t,x)\right] dt + \sigma(t,x)\,dw(t), \quad x(t_0) = x_0. \tag{31}$$

This is achieved by the procedure of solving the Itô–Doob type stochastic differential equation in the form of (30). The procedure is to reduce differential equation (30) into the following reduced integrable differential equation as in (Gīhman and Skorohod (1972); Kloeden and Platen (1992); Ladde and Ladde (2009) and Wong (1971)):

$$dm = f(t)\,dt + g(t)\,dw(t), \tag{32}$$

where $f(t)$ and $g(t)$ are suitable stochastic processes determined by rate functions C and σ in (24). The extra term $\frac{1}{2}\sigma(t,x)\frac{\partial}{\partial x}\sigma(t,x)$ in (31) is referred to as the *correction term*.

In summary, it is further detailed as shown in Ladde and Ladde (2009) and Wong (1971) that if we interpret Gaussian white-noise driven differential equation (24) by the limit of a sequence of stochastic differential equations (25) with a sequential colored noise process, then the Gaussian white-noise driven differential equation (24) is equivalent to the Itˆo–Doob type stochastic differential equation (31). Moreover, this material is 1-dimensional state variable, however, it can be easily extended to multi-dimensional state space.

Several dynamic processes are under both internal and external random distributions. The usage of this information coupled with different modes in probabilistic analysis, namely, an approach through sample calculus, L^p-calculus, and Itô–Doob calculus as in (Ladde and Lakshmikantham; 1980, Ladde and Sambandham; 2004, Nelson; 1967, Øksendal; 1985 and Soong; 1973) leads to different dynamic models. The majority of the dynamic models are in the context of Itô–Doob calculus (Arnold; 1974, Gīhman; 1972, Ito; 1951, Kloeden and Platen; 1992, Ladde; 1991, Ladde and Ladde; 2009, Ladde and Lakshmikantham; 1980; Nelson; 1967, Øksendal; 1985, Soong; 1973, Wong; 1971) and are described by systems of stochastic differential equations

$$dx = f(t,x)\,dt + \sigma(t,x)w(t), \quad x(t_0) = x_0, \tag{33}$$

where dx is the Itô–Doob type stochastic differential of x, $x \in R^n$, w is m-dimensional normalized Wiener process defined on a complete probability space $(\Omega, \Im, P)$, $f(t,x)$ is drift rate vector, and $\sigma(t,x)$ is a diffusion rate matrix of size $n \times m$. Various qualitative properties (Arnold; 1974, Ladde; 1991, Ladde and Lakshmikantham; 1980, Ladde and Sambandham; 2004, Soong; 1973, Wong; 1971) have played a very significant role in state estimation and system designing processes since the beginning or middle of the twentieth century.

Acknowledgment

This research was supported by Mathematical Sciences Division, US Army Research Office, Grant No. W911NF-07-1-0283.

About the Author

Dr. Gangaram Ladde is Professor of Mathematics and Statistics, University of South Florida (since 2007). Prior to that he was Professor of Mathematics, University of Texas at Arlington (1980–2007). He received his Ph.D. in Mathematics from University of Rhode Island in 1972. He has published more than 150 papers, has co-authored 4 monographs, and co-edited 6 proceedings of international conferences, including, (1) *Stochastic Versus Deterministic Systems of Differential Equations*, (with M. Sambandham, Marcel Dekker, Inc, New York, 2004) and (2) *Random Differential Inequalities* (with V. Lakshmikantham, Academic Press, New York, 1980). Dr. Ladde is the Founder and Joint Editor-in-Chief (1983–present) of the *Journal of Stochastic Analysis and Applications*. He is also a Member of Editorial Board of several journals in Mathematical Sciences. Dr. Ladde is recipient of several research awards and grants.

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Markov Chains
►Random Walk

►Stochastic Differential Equations
►Stochastic Modeling, Recent Advances in
►Stochastic Models of Transport Processes
►Stochastic Processes: Classification

References and Further Reading

Arnold L (1974) Stochastic differential equations: theory and applications. Wiley-Interscience (Wiley), New York, Translated from the German

Bartlett MS (1960) Stochastic population models in ecology and epidemiology. Methuen's Monographs on Applied Probability and Statistics, Methuen, London

Gĭhman ĬĬ, Skorohod AV (1972) Stochastic differential equations. Springer, New York, Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72

Gikhman II, Skorokhod AV (1969) Introduction to the theory of random processes. Translated from the Russian by Scripta Technica, W.B. Saunders, Philadelphia, PA

Goel NS, Richter-Dyn N (1974) Stochastic models in biology. Academic (A subsidiary of Harcourt Brace Jovanovich), New York–London

Ito K (1951) On stochastic differential equations. Mem Am Math Soc 1951(4):51

Kimura M, Ohta T (1971) Theoretical aspects of population genetics. Princeton University Press, Princeton, NJ

Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Applications of mathematics (New York), vol 23, Springer, Berlin

Ladde GS (1991) Stochastic delay differential systems. World Scientific, Hackensack, NJ, pp 204–212

Ladde AG, Ladde GS (2009) An introduction to differential equations: stochastic modeling, methods and analysis, vol II. In Publication Process

Ladde GS, Lakshmikantham V (1980) Random differential inequalities. Mathematics in Science and Engineering, vol 150, Academic (Harcourt Brace Jovanovich), New York

Ladde GS, Sambandham M (2004) Stochastic versus deterministic systems of differential equations. Monographs and textbooks in pure and applied mathematics, vol 260. Marcel Dekker, New York

Lakshmikantham V, Leela S (1969a) Differential and integral inequalities: theory and applications, volume I: ordinary differential equations. Mathematics in science and engineering, vol 55-I. Academic, New York

Lakshmikantham V, Leela S (1969b) Differential and integral inequalities: theory and applications, vol II: functional, partial, abstract, and complex differential equations. Mathematics in science and engineering, vol 55-II. Academic, New York

Nelson E (1967) Dynamical theories of Brownian motion. Princeton University Press, Princeton, NJ

Oksendal B (1985) Stochastic differential equations. An introduction with applications. Universitext, Springer, Berlin

Ricciardi LM (1977) Diffusion processes and related topics in biology. Springer, Berlin, Notes taken by Charles E. Smith, Lecture Notes in Biomathematics, vol 14

Ross SM (1972) Introduction to probability models. Probability and mathematical statistics, vol 10. Academic, New York

Soong TT (1973) Random differential equations in science and engineering. Mathematics in science and engineering, vol 103. Academic (Harcourt Brace Jovanovich), New York

Wong E (1971) Stochastic processes in information and dynamical systems. McGraw-Hill, New York, NY

Stochastic Models of Transport Processes

Alexander D. Kolesnik
Professor
Institute of Mathematics & Computer Science, Academy of Sciences of Moldova, Kishinev, Moldova

The transport process $\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))$ in the Euclidean space, $\mathbb{R}^m$, $m \geq 1$, is generated by the stochastic motion of a particle that, at the time instant $t = 0$, starts from some initial point (e.g., origin) of $\mathbb{R}^m$ and moves with some finite speed c in random direction. The motion is controlled by some stochastic process $x(t)$, $t \geq 0$, causing, at random time instants, the changes of direction chosen randomly according to some distribution on the unit sphere $S_1^m \subset \mathbb{R}^m$. Such stochastic motions, also called random flights, represent the most important type of random evolutions (for limit and asymptotic theorems for general random evolutions see, for instance, Papanicolaou [1975], Pinsky [1991], Korolyuk and Swishchuk [1994] and the bibliographies therein). While the finiteness of the velocity is the basic feature of such motions, the models differ with respect to the way of choosing the new directions (the scattering function), the type of the governing stochastic process $x(t)$, and the dimension of the space $\mathbb{R}^m$. If the new directions are taken on according to the uniform probability law and the phase space $\mathbb{R}^m$ is isotropic and homogeneous, $\mathbf{X}(t)$ is referred to as the *isotropic* transport process. The most studied model is referred to the case when the speed c is constant and $x(t)$ is the homogeneous Poisson process (see ►Poisson Processes).

The simplest one-dimensional isotropic transport process with constant finite speed c driven by a homogeneous Poisson process of rate $\lambda > 0$ was first studied by Goldstein (1951) and Kac (1956). They have shown that the transition density $f = f(x,t)$, $x \in \mathbb{R}^1$, $t > 0$, of the process satisfies the telegraph equation

$$\frac{\partial^2 f}{\partial t^2} + 2\lambda \frac{\partial f}{\partial t} - c^2 \frac{\partial^2 f}{\partial x^2} = 0, \tag{1}$$

and can be found by solving this equation with the initial conditions $f(x,0) = \delta(x)$, $\frac{\partial f}{\partial t}\Big|_{t=0} = 0$, where $\delta(x)$ is the one-dimensional Dirac delta-function. The explicit form of

the transition density of the process (i.e., the fundamental solution to (1)) is given by the formula

$$f(x,t) = \frac{e^{-\lambda t}}{2}\left[\delta(ct+x)+\delta(ct-x)\right] + \frac{e^{-\lambda t}}{2c}\left[\lambda I_0\left(\frac{\lambda}{c}\sqrt{c^2t^2-x^2}\right) + \frac{\lambda ct}{\sqrt{c^2t^2-x^2}}\, I_1\left(\frac{\lambda}{c}\sqrt{c^2t^2-x^2}\right)\right]\Theta(ct-|x|), \quad x\in\mathbb{R}^1, \quad |x|\le ct, \quad t>0, \tag{2}$$

where $I_0(x)$ and $I_1(x)$ are the Bessel functions of zero and first orders, respectively, with imaginary argument and $\Theta(x)$ is the Heaviside function. The first term in (2) represents the density of the singular component of the distribution (which is concentrated in two terminal points $\pm ct$ of the interval $[-ct, ct]$), while the second one represents the density of the absolutely continuous part of the distribution (which is concentrated in the open interval $(-ct, ct)$).

Let $\mathbf{X}(t)$, $t>0$, be the isotropic transport process in the Euclidean plane, $\mathbb{R}^2$, generated by the random motion of a particle moving with constant speed c and choosing new directions at random Poissonian (λ) instants according to the uniform probability law on the unit circumference. Then the transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^2$, $t>0$, of $\mathbf{X}(t)$ has the form (Stadje 1987; Masoliver et al. 1993; Kolesnik and Orsingher 2005)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{2\pi ct}\,\delta(c^2t^2-\|\mathbf{x}\|^2) + \frac{\lambda}{2\pi c}\,\frac{\exp\left(-\lambda t+\frac{\lambda}{c}\sqrt{c^2t^2-\|\mathbf{x}\|^2}\right)}{\sqrt{c^2t^2-\|\mathbf{x}\|^2}} \times\Theta\left(ct-\|\mathbf{x}\|\right), \quad \mathbf{x}=(x_1,x_2)\in\mathbb{R}^2, \quad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2}\le ct, \quad t>0. \tag{3}$$

Similar to the one-dimensional case, the density (3) is the fundamental solution (the Green's function) to the two-dimensional telegraph equation

$$\frac{\partial^2 f}{\partial t^2}+2\lambda\frac{\partial f}{\partial t} = c^2\left\{\frac{\partial^2 f}{\partial x_1^2}+\frac{\partial^2 f}{\partial x_2^2}\right\}. \tag{4}$$

The transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^3$, $t>0$, of the isotropic transport process $\mathbf{X}(t)$ with unit speed $c=1$ in the three-dimensional Euclidean space, $\mathbb{R}^3$, is given by the formula (Stadje 1989)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{4\pi t^2}\,\delta(t^2-\|\mathbf{x}\|^2) + \frac{\lambda\, e^{-\lambda t}}{4\pi\|\mathbf{x}\|}\left[\lambda\int_{-1}^{-\|\mathbf{x}\|/t}\exp(\lambda(\xi t+\|\mathbf{x}\|)\operatorname{arth}\xi)\,(\operatorname{arth}\xi)^2\,d\xi + \frac{1}{t}\operatorname{arth}\left(\frac{\|\mathbf{x}\|}{t}\right)\right]\Theta\left(t-\|\mathbf{x}\|\right), \quad \mathbf{x}=(x_1,x_2,x_3)\in\mathbb{R}^3, \quad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2+x_3^2}\le t, \quad t>0, \tag{5}$$

where $\operatorname{arth}(x)$ is the hyperbolic inverse tangent function.

In the four-dimensional Euclidean space, $\mathbb{R}^4$, the transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^4$, $t>0$, of the isotropic transport process $\mathbf{X}(t)$ has the following form (Kolesnik 2006)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{2\pi^2(ct)^3}\,\delta(c^2t^2-\|\mathbf{x}\|^2) + \frac{\lambda t}{\pi^2(ct)^4} \times\left[2+\lambda t\left(1-\frac{\|\mathbf{x}\|^2}{c^2t^2}\right)\right]\exp\left(-\frac{\lambda}{c^2t}\|\mathbf{x}\|^2\right) \times\Theta\left(ct-\|\mathbf{x}\|\right), \quad \mathbf{x}=(x_1,x_2,x_3,x_4)\in\mathbb{R}^4, \quad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2+x_3^2+x_4^2}\le ct, \quad t>0. \tag{6}$$

We see that in the spaces $\mathbb{R}^2$ and $\mathbb{R}^4$, the transition densities of $\mathbf{X}(t)$ have very simple analytical forms (3) and (6) expressed in terms of elementary functions. In contrast, the three-dimensional density (5) has the fairly complicated form of an integral with variable limits which, apparently, cannot be explicitly evaluated. This fact shows that the behavior of transport processes in the Euclidean spaces $\mathbb{R}^m$ substantially depends on the dimension m. Moreover, while the transition densities of the processes on the line $\mathbb{R}^1$ and in the plane $\mathbb{R}^2$ are the fundamental solutions (i.e., the Green's functions) to the telegraph equations (1) and (4), respectively, the similar results for other spaces have not been obtained so far.

However, for the integral transforms of the distributions of $\mathbf{X}(t)$, one can give the most general formulas that are valid in any dimensions. Let $H(t)=\mathrm{E}\left\{e^{i(\boldsymbol{\alpha},\mathbf{X}(t))}\right\}$ be the characteristic function (Fourier transform) of the isotropic transport process $\mathbf{X}(t)$ in the Euclidean space $\mathbb{R}^m$ of arbitrary dimension $m\ge 2$. Here, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)\in\mathbb{R}^m$ is the real m-dimensional vector of inversion parameters

and $(\boldsymbol{\alpha}, \mathbf{X}(t))$ means the inner product of the vectors $\boldsymbol{\alpha}$ and $\mathbf{X}(t)$. Introduce the function

$$\varphi(t) = 2^{(m-2)/2}\,\Gamma\left(\frac{m}{2}\right)\frac{J_{(m-2)/2}(ct\|\boldsymbol{\alpha}\|)}{(ct\|\boldsymbol{\alpha}\|)^{(m-2)/2}}, \qquad m \geq 2, \tag{7}$$

where $\|\boldsymbol{\alpha}\| = \sqrt{\alpha_1^2 + \cdots + \alpha_m^2}$, $\Gamma(x)$ is the Euler gamma-function and $J_{(m-2)/2}(x)$ is the Bessel function of order $(m-2)/2$ with real argument. Note that (7) is the characteristic function of the uniform distribution on the surface of the sphere of radius ct in the space $\mathbb{R}^m$, $m \geq 2$. Then the characteristic function $H(t)$, $t \geq 0$, satisfies the following convolution-type Volterra integral equation of second kind (Kolesnik 2008):

$$H(t) = e^{-\lambda t}\varphi(t) + \lambda\int_0^t e^{-\lambda(t-\tau)}\varphi(t-\tau)H(\tau)\,d\tau, \quad t \geq 0. \tag{8}$$

In the class of continuous functions, the integral equation (8) has the unique solution given by the uniformly converging series

$$H(t) = e^{-\lambda t}\sum_{n=0}^{\infty}\lambda^n\,[\varphi(t)]^{*(n+1)}, \tag{9}$$

where $[\varphi(t)]^{*(n+1)}$ means the $(n+1)$-multiple convolution of function (7) with itself. The Laplace transform $\mathcal{L}$ of the characteristic function $H(t)$ has the form (Kolesnik 2008)

$$\mathcal{L}[H(t)](s) = \frac{F\left(\frac{1}{2}, \frac{m-2}{2}; \frac{m}{2}; \frac{(c\|\boldsymbol{\alpha}\|)^2}{(s+\lambda)^2+(c\|\boldsymbol{\alpha}\|)^2}\right)}{\sqrt{(s+\lambda)^2 + (c\|\boldsymbol{\alpha}\|)^2} - \lambda\,F\left(\frac{1}{2}, \frac{m-2}{2}; \frac{m}{2}; \frac{(c\|\boldsymbol{\alpha}\|)^2}{(s+\lambda)^2+(c\|\boldsymbol{\alpha}\|)^2}\right)}, \quad m \geq 2, \tag{10}$$

for $\mathrm{Re}\, s > 0$, where $F(\xi, \eta; \zeta; z)$ is the Gauss hypergeometric function.

One of the most remarkable features of the isotropic transport processes in $\mathbb{R}^m$, $m \geq 2$, is their weak convergence to the Brownian motion (see ►Brownian Motion and Diffusions) as both the speed c and the intensity of switchings λ tend to infinity in such a way that the following Kac condition holds:

$$c \to \infty, \quad \lambda \to \infty, \quad \frac{c^2}{\lambda} \to \rho, \quad \rho > 0. \tag{11}$$

Under this condition (11), the transition density $f = f(\mathbf{x}, t)$, $\mathbf{x} \in \mathbb{R}^m$, $m \geq 2$, $t > 0$, of the isotropic transport process $\mathbf{X}(t)$ converges to the transition density of the homogeneous Brownian motion with zero drift and diffusion coefficient $\sigma^2 = 2\rho/m$ (Kolesnik 2008), i.e.,

$$\lim_{\substack{c,\,\lambda\to\infty \\ (c^2/\lambda)\to\rho}} f(\mathbf{x}, t) = \left(\frac{m}{4\rho\pi t}\right)^{m/2} \times \exp\left(-\frac{m}{4\rho t}\,\|\mathbf{x}\|^2\right), \qquad m \geq 2,$$

where $\|\mathbf{x}\|^2 = x_1^2 + \cdots + x_m^2$.

Some of these results are also valid for the transport processes with arbitrary scattering functions. Suppose that both the initial and each new direction are taken on according to some arbitrary distribution on the unit sphere $S_1^m \subset \mathbb{R}^m$, $m \geq 2$. Let $\chi(\mathbf{x})$, $\mathbf{x} \in S_1^m$ denote the density of this distribution, assumed to exist. Introduce the function

$$\psi(t) = \int_{S_1^m} e^{ict(\boldsymbol{\alpha}, \mathbf{x})}\,\chi(\mathbf{x})\,\mu(d\mathbf{x}),$$

where $\mu(d\mathbf{x})$ is the Lebesgue measure on S_1^m. Then the characteristic function of such a transport process satisfies a Volterra integral equation similar to (8), in which the function $\varphi(t)$ is replaced everywhere by the function $\psi(t)$. The unique continuous solution of such an equation is similar to (9) with the same replacement.

About the Author

Alexander Kolesnik, Ph.D. in Probability and Statistics (1991) and Habilitation (2010), is a Leading Scientific Researcher (Professor). He has published more than 40 articles. He is currently preparing a monograph on the statistical theory of transport processes at finite velocity. He was coeditor (1996–2006) of InterStat (Electronic Journal on Probability and Statistics, USA), and external referee of many international journals on probability and statistics.

Cross References

►Brownian Motion and Diffusions
►Poisson Processes
►Stochastic Modeling Analysis and Applications
►Stochastic Modeling, Recent Advances in

References and Further Reading

Goldstein S (1951) On diffusion by discontinuous movements and on the telegraph equation. Q J Mech Appl Math 4:129–156
Kac M (1956) A stochastic model related to the telegrapher's equation. In: Some stochastic problems in physics and mathematics, Magnolia petroleum company colloquium, lectures in the pure and applied science, No. 2 (Reprinted in: Rocky Mount J Math (1974), 4:497–509)
Kolesnik AD (2006) A four-dimensional random motion at finite speed. J Appl Probab 43:1107–1118
Kolesnik AD (2008) Random motions at finite speed in higher dimensions. J Stat Phys 131:1039–1065

Kolesnik AD, Orsingher E (2005) A planar random motion with an infinite number of directions controlled by the damped wave equation. J Appl Probab 42:1168–1182
Korolyuk VS, Swishchuk AV (1994) Semi-Markov random evolutions. Kluwer, Amsterdam
Masoliver J, Porrá JM, Weiss GH (1993). Some two and three-dimensional persistent random walks. Physica A 193:469–482
Papanicolaou G (1975) Asymptotic analysis of transport processes. Bull Am Math Soc 81:330–392
Pinsky M (1991) Lectures on random evolution. World Scientific, River Edge
Stadje W (1987) The exact probability distribution of a two-dimensional random walk. J Stat Phys 46:207–216
Stadje W (1989) Exact probability distributions for non-correlated random walk models. J Stat Phys 56:415–435

Stochastic Processes

Rolando Rebolledo
Professor, Head of the Center for Stochastic Analysis,
Facultad de Matemáticas
Universidad Catolica de Chile, Santiago, Chile

The word "stochastic process" is derived from the Greek noun "stokhos" which means "aim." Another related Greek word "stokhastikos," "the dart game," provides an alternative image for randomness or chance. Although the concept of Probability is often associated with dice games, the dart game seems to be more adapted to the modern approach to both Probability Theory and Stochastic Processes. Indeed, the fundamental difference between a dice game and darts is that while in the first, one cannot control the issue of the game, in the dart game, one tries to attain an objective with different degrees of success, thus, the player increases his knowledge of the game at each trial. As a result, time is crucial in the dart game, the longer you play, the better you increase your skills.

Definition of a Stochastic Process

The mathematical definition of a stochastic process, in the Kolmogorov model of Probability Theory, is given as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, Ω is a non empty set called *sample space*, $\mathcal{F}$ is a sigma field of subsets of Ω, which represents the family of *events*, and $\mathbb{P}$ is a *probability measure* defined on $\mathcal{F}$. T is another non empty set, and $(E, \mathcal{E})$ a measurable space to represent all possible *states*. Then, a *stochastic process with states in* E is a map $X : T \times \Omega \to E$ such that for all $t \in T$, $\omega \mapsto X(t, \omega)$ is a measurable function. In other words, a primary interpretation of a stochastic process X is as a collection of random variables, and as such, notations like $(X_t)_{t\in T}$ are used to refer to X, that is $X_t(\omega) = X(t, \omega)$, for all $(t, \omega) \in T \times \Omega$. If T is an ordered number set, (e.g., $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}^+$, $\mathbb{R}$), it is often referred as the set of *time variables* and taken as a subset of integers or real numbers. For each $\omega \in \Omega$, the map $X(\cdot, \omega) : t \mapsto X(t, \omega)$ is called the *trajectory* of the process. Thus, each trajectory is an element of E^T, the set of all E-valued functions defined on T. Particularly, if T is a countable set, the process is said to be indexed by *discrete times* (the expression *Time Series* is also in use in this case). Discrete time stochastic processes were the first studied in Probability Theory under the name of *chains* (see ▶Markov Chains).

Example 1

1. Consider a sequence $(\xi_n)_{n\geq 1}$ of real random variables. According to the definition, this is a stochastic process. New stochastic processes can be defined on this basis. For instance, take $(S_n)_{n\geq 1}$, defined as, $S_n = \xi_1 + \ldots + \xi_n$, for each $n \geq 1$.

   Suppose now that the random variables $(\xi_n)_{n\geq 1}$ are independent and identically distributed on $\{-1, 1\}$ with $\mathbb{P}(\xi = \pm 1) = 1/2$. Then, $(S_n)_{n\geq 1}$ becomes a *Simple Symmetric Random Walk*.
2. Consider a real function $x : [0, \infty[\to \mathbb{R}$, this is also a stochastic process. It suffices to consider any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define $X(\omega, t) = x(t)$, for all $\omega \in \Omega$, $t \geq 0$. This is a trivial stochastic process.
3. Consider an initial value problem given by

$$\begin{cases} x' = f(t, x); \\ x(0) = x, \end{cases} \tag{1}$$

   where f is a continuous function on the two variables (t, x). Newtonian Mechanics can be written within this framework, which is usually referred as a mathematical model for a *closed dynamical system* in Physics. That is, the system has no interaction with the environment, and time is reversible. Now define Ω as the set of all continuous functions from $[0, \infty[$ into $\mathbb{R}$. Endow Ω with the topology of uniform convergence on compact subsets of the positive real line and call $\mathcal{F}$ the corresponding Borel σ-field. Thus, any $\omega \in \Omega$ is a function $\omega = (\omega(t); t \geq 0)$. Define the stochastic process $X(\omega, t) = \omega(t)$, known as the *canonical process*. The initial value problem is then written as

$$X(\omega, t) = x + \int_0^t f(s, X(\omega, s))ds. \tag{2}$$

   This can be phrased as an example of a *Stochastic Differential Equation*, without noise term. The solution

is a deterministic process which provides a description of the given closed dynamical system. Apparently, there is no great novelty and one can wonder whether the introduction of Ω is useful. However, this framework includes processes describing open dynamical systems too, embracing the interaction of the main system with the environment, and that is an important merit of the stochastic approach. Typically, the interaction of a given system with the environment is described through the action of so-called noises interfering with the main dynamics. Let us complete our example adding a noise term to the closed dynamics.

To consider the action of a *noise*, take a sequence $(\xi_n)_{n\geq 1}$ of random variables defined on Ω, such that $\xi_n(\omega) \in \{-1,1\}$. Let be given a probability $\mathbb{P}$ on the measurable space $(\Omega,\mathcal{F})$ such that $\mathbb{P}(\xi_n = \pm 1) = 1/2$. Call $S_n = \xi_1 + \ldots + \xi_n$ and denote $[t]$ the greatest integer $\leq t$. The equation

$$X(\omega,t) = x + \int_0^t f(s,X(\omega,s))ds + S_{[t]}(\omega), \qquad (3)$$

is an example of a stochastic differential equation driven by a ▶random walk. The stochastic process obtained as a solution is no longer deterministic and describes an open system dynamics. ▽

Distributions

The space of trajectories E^T is usually endowed with the product σ-field $\mathcal{E}^{\otimes T}$ generated by all projections $\pi_t : E^T \to E$, which associate to each function $x \in E^T$ its value $x(t) \in E$, $t \in T$. Thus, a stochastic process is, equivalently, a random variable $X : \Omega \to E^T$, $\omega \mapsto X(\cdot,\omega)$. The *Law* or *Probability Distribution* P_X of a stochastic process X is the image of the probability $\mathbb{P}$ on the measurable space $(E^T,\mathcal{E}^{\otimes T})$ of all trajectories. Given a probability measure P on the space $(E^T,\mathcal{E}^{\otimes T})$, one may construct a *Canonical Process* X whose distribution P_X coincides with P. Indeed, it suffices to consider $\Omega = E^T$, $\mathcal{F} = \mathcal{E}^{\otimes T}$, $\mathbb{P} = P$, $X(t,\omega) = \omega(t)$, for each $\omega = (\omega(s);\ s \in T) \in E^T$, $t \in T$.

Let a finite set $I = \{t_1,\ldots,t_n\} \subset T$ be given, and denote π_I the canonical projection defined on E^T with values in E^I, such that $x \mapsto (x(t_1),\ldots,x(t_n))$. Call $\mathcal{P}_f(T)$ the family of all finite subsets of T. The *Finite Dimensional Distributions* or *Marginal Probability Distributions* of an E-valued stochastic process is the family $(P_{X,I})_{I\in\mathcal{P}_f(T)}$ of distributions, where $P_{X,I}$ is defined as

$$P_{X,I}(A) = P_X\left(\pi_i^{-1}(A)\right) = \mathbb{P}\left((X(t_1,\cdot),\ldots,X(t_n,\cdot)) \in A\right), \qquad (4)$$

for all $A \in \mathcal{E}^{\otimes I}$.

Example 2

1. A *Poisson Process* $(N_t)_{t\geq 0}$ is defined as a stochastic process with values in $\mathbb{N}$ such that
   (a) $N_0(\omega) = 0$ and $t \mapsto N_t(\omega)$ is increasing, for all $\omega \in \mathbb{N}$.
   (b) For all $0 \leq s \leq t < \infty$, $N_t - N_s$ is independent of $(N_u;\ u \leq s)$.
   (c) For all $0 \leq s \leq t < \infty$, the distribution of $N_t - N_s$ is Poisson with parameter $t - s$, that is

$$\mathbb{P}(N_t - N_s = k) = \frac{(t-s)^k}{k!}e^{-(t-s)}.$$

2. A d-dimensional *Brownian Motion* (see also ▶Brownian Motion and Diffusions) is a stochastic process $(B_t)_{t\geq 0}$, taking values in $\mathbb{R}^d$ such that:
   (a) If $0 \leq s < t < \infty$, then $B_t - B_s$ is independent of $(B_u;\ u \leq s)$.
   (b) If $0 \leq s < t < \infty$, then

$$\mathbb{P}(B_t - B_s \in A) = (2\pi(t-s))^{-d/2}\int_A e^{-|x|^2/2(t-s)}dx,$$

   where dx represents the Lebesgue measure on $\mathbb{R}^d$ and $|x|$ is the euclidian norm in that space.

   The Brownian Motion starts at x if $\mathbb{P}(B_0 = x) = 1$. ▽

Construction of Canonical Processes

An important problem in the construction of a canonical stochastic process given the family of its finite dimensional distributions was solved by Kolmogorov in the case of a countable set T and extended to continuous time later by several authors. At present, a particular case, general enough for applications, is the following version of the Daniell–Kolmogorov Theorem. Suppose that E is a Polish space (complete separable metric space) and let $\mathcal{E}$ be its Borel σ-field. Let T be a subset of $\mathbb{R}^+$. Suppose that for each $I \in \mathcal{P}_f(T)$ a probability P_I is given on the space $(E,\mathcal{E}^{\otimes I})$. Then, there exists a probability P on $(E^T,\mathcal{E}^{\otimes T})$ such that for all $I \in \mathcal{P}_f(T)$,

$$P_I(A) = P \circ \pi_I^{-1}(A) = P\left(\pi_I^{-1}(A)\right), \qquad (5)$$

for all $A \in \mathcal{E}^{\otimes I}$, if and only if the following *Consistency Condition* is satisfied:

$$P_I = P_J \circ \pi_{J,I}^{-1}, \qquad (6)$$

for all $I, J \in \mathcal{P}_f(T)$ such that $I \subset J$, where $\pi_{J,I}$ denotes the canonical projection from the space E^J onto E^I.

Example 3 Consider $J = \{t_1, \ldots, t_n\}$ and let Φ_t be the normal distribution of mean zero and variance $t \geq 0$, that is,

$$\Phi_t(A) = (2\pi t)^{-1/2} \int_A e^{-x^2/2t} dx.$$

Let $P_J = \Phi_{t_1} \otimes \Phi_{t_2 - t_1} \otimes \ldots \Phi_{t_n - t_{n-1}}$, that is for all Borel sets $A_1, \ldots, A_n$,

$$P_J(A_1 \times A_2 \times \ldots \times A_n) = \Phi_{t_1}(A_1)\Phi_{t_2-t_1}(A_2) \ldots \Phi_{t_n - t_{n-1}}(A_n).$$

This is a probability on $\mathbb{R}^n$. Take $I = \{t_1, \ldots, t_{n-1}\}$. Notice that $\pi_{J,I}^{-1}(A_1 \times \ldots \times A_{n-1}) = A_1 \times \ldots \times A_{n-1} \times \mathbb{R}$, thus

$$\begin{aligned} P_I(A_1 \times A_2 \times \ldots \times A_{n-1}) &= \Phi_{t_1}(A_1)\Phi_{t_2-t_1}(A_2) \ldots \\ &\quad \Phi_{t_{n-1}-t_{n-2}}(A_{n-1}) \\ &= P_J(A_1 \times A_2 \times \ldots \times \mathbb{R}). \end{aligned}$$ ▽

Regularity of Trajectories

Another interpretation of a stochastic process is based on regularity properties of trajectories. Indeed, if one knows that each trajectory belongs almost surely to a function space $S \subset E^T$, endowed with a σ-field $\mathcal{S}$, one may provide another characterization of the stochastic process X as an S-valued random variable, $\omega \mapsto X(\cdot, \omega)$ defined on Ω.

Regarding the regularity, Kolmogorov first proved one of the most useful criteria on continuity of trajectories. Suppose that $X = (X(t,\omega);\ t \in [0,1],\ \omega \in \Omega)$ is a real-valued stochastic process and assume that there exist $\alpha, \delta > 0$ and $0 < C < \infty$ such that

$$\mathbb{E}\left(|X(t+h) - X(t)|^\alpha\right) < C|h|^{1+\delta}, \tag{7}$$

for all $t \in [0,1]$ and all sufficiently small $h > 0$, then X has continuous trajectories with probability 1. Therefore, if X satisfies (7), then there exists a random variable $\widetilde{X} : \Omega \to C[0,1]$, where $C[0,1]$ is the metric space of real continuous functions defined on $[0,1]$, endowed with the metric of uniform distance, such that $\mathbb{P}\left(\{\omega \in \Omega :\ X(\cdot,\omega) = \widetilde{X}(\omega)\}\right) = 1$.

Wiener Measure, Brownian Motion

The above result is crucial to construct the *Wiener Measure* on the space $C[0,1]$ or, more generally, on $C(\mathbb{R}^+)$, which is the law of the *Brownian Motion* (see also ►Brownian Motion and Diffusions). Indeed, by means of Kolmogorv's Consistency Theorem, one first constructs a probability measure P on the product space $(\mathbb{R}^{\mathbb{R}^+}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+})$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-field of $\mathbb{R}$, considering the consistent family of probability distributions

$$P_I = \Phi_{t_1} \otimes \Phi_{t_2 - t_1} \otimes \ldots \otimes \Phi_{t_n - t_{n-1}}, \tag{8}$$

where $I = \{t_1, \ldots, t_n\}$, and Φ_t denotes the normal distribution with mean 0 and variance t. Since the family $(P_I)_{I \in \mathcal{P}_f(\mathbb{R}^+)}$ is consistent, there exists a unique P probability measure on $(\mathbb{R}^{\mathbb{R}^+}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+})$ such that $P_I = P \circ \pi_I^{-1}$. One can construct the canonical process with law P which should correspond to the Brownian Motion. Unfortunately, the set of real-valued continuous functions defined on $\mathbb{R}^+$ is not an element of $\mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+}$. However, thanks to (7) one proves that the exterior probability measure P^* defined by P is concentrated on the subset $C(\mathbb{R}^+)$ of $\mathbb{R}^{\mathbb{R}^+}$ thus, the restriction P_W of P^* to $C(\mathbb{R}^+)$ gives the good definition of Wiener Measure. Thus, a canonical version of the Brownian Motion is given by the canonical process on the space $C(\mathbb{R}^+)$.

Series Expansion in L^2

In the early years of the Theory of Stochastic Processes, a number of authors, among them Karhunen and Loève, explored other regularity properties of trajectories, deriving some useful representations by means of series expansions in an L^2 space. More precisely, let $T \in \mathcal{B}(\mathbb{R}^+)$ be given and call $\mathfrak{h} = L^2(T)$ the Hilbert space of all real-valued Lebesgue-square integrable functions defined on T. Suppose that all trajectories $X(\cdot,\omega)$ belong to $\mathfrak{h}$ for all $\omega \in \Omega$, and denote $(e_n)_{n \in \mathbb{N}}$ an orthonormal basis of $\mathfrak{h}$. Therefore, $x_n(\omega) = \langle X(\cdot,\omega), e_n \rangle$ satisfies $\sum_{n \in \mathbb{N}} |x_n(\omega)|^2 < \infty$, for all $\omega \in \Omega$. And the series

$$\sum_{n \in \mathbb{N}} x_n(\omega) e_n, \tag{9}$$

converges in $\mathfrak{h}$, providing a representation of $X(\cdot,\omega)$. So that, by an abuse of language one can represent $X(t,\omega)$ by $\sum_{n \in \mathbb{N}} x_n(\omega) e_n(t)$.

Example 4 Consider $T = [0,1]$ and the Haar orthonormal basis on the space $\mathfrak{h} = L^2([0,1])$ constructed by induction as follows: $e_1(t) = 1$ for all $t \in [0,1]$;

$$e_{2^m+1} = \begin{cases} 2^{m/2}, & \text{if } 0 \leq t < 2^{-m-1}, \\ -2^{m/2}, & \text{if } 2^{-m-1} \leq t < 2^{-m}, \\ 0, & \text{otherwise.} \end{cases}$$

And finally, define $e_{2^m+j}(t) = e_{2^m+1}\left(t - 2^{-m}(j-1)\right)$, for $j = 1, \ldots, 2^m$, $m = 0, 1, \ldots$. Given a sequence $(b_n)_{n \geq 1}$ of independent standard normal random variables (that is, with distribution $\mathcal{N}(0,1)$), the $L^2(\Omega \times [0,1])$-convergent series $\sum_{n \geq 1} b_n(\omega) f_n(t)$ provides a representation of the Brownian Motion $(B_t)_{t \in [0,1]}$, where $f_n(t) = \int_0^t e_n(s) ds$, $(t \in [0,1],\ n \in \mathbb{N})$. ▽

The General Theory of Processes

The General Theory of Processes emerged in the seventies as a contribution of the Strasbourg School initiated by Paul André Meyer. This Theory uses the concept of a *History* or *Filtration*, which consists of an increasing family of σ-fields $\mathbb{F} = (F_t)_{t\in T}$, where T is an ordered set, $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for all $s \le t$. Thus, a stochastic process X is *adapted* to $\mathbb{F}$ if for all $t \in T$, the variable $X(t,\cdot)$ is $\mathcal{F}_t/\mathcal{E}$–measurable. Stronger measurability conditions mixing regularity conditions have been introduced motivated by the construction of stochastic integrals and the modern theory of Stochastic Differential Equations. Let $T = \mathbb{R}^+$ and assume E to be a Polish space endowed with the σ-field of its Borel sets. Denote $C_E = C(\mathbb{R}^+, E)$ (respectively $D_E = D(\mathbb{R}^+, E)$) the space of all E-valued continuous functions defined on $\mathbb{R}^+$ to E (resp. the space of all E-valued functions which have left hand limit at each point $t > 0$ and are right-continuous at $t > 0$, endowed with the Skorokhod's topology). Consider now the family $\mathcal{C}_E$ (resp. $\mathcal{D}_E$) of all $\mathbb{F}$-adapted stochastic processes $X : \mathbb{R}^+ \times \Omega \to E$ such that their trajectories belong to C_E (resp. to D_E). The *Predictable* (resp. *Optional*) σ-field on the product set $\mathbb{R}^+ \times \Omega$ is the one generated by $\mathcal{C}_E$ (resp. $\mathcal{D}_E$), that is $\mathcal{P} = \sigma(\mathcal{C}_E)$, (resp. $\mathcal{O} = \sigma(\mathcal{D}_E)$). Then, a process X is *predictable* (resp. *optional*) if $(t,\omega) \mapsto X(t,\omega)$ is measurable with respect to $\mathcal{P}$, (resp. $\mathcal{O}$). A crucial notion in the development of this theory is that of *Stopping Time*: a function $\tau : \Omega \to [0,\infty]$ is a stopping time if for all $t > 0$, $\{\omega \in \Omega : \tau(\omega) \le t\}$ is an element of the σ-field $\mathcal{F}_t$. This definition is equivalent to say that τ is a stopping time if and only if $(t,\omega) \mapsto 1_{[0,\tau(\omega)[}(t)$ is an optional process, where the notation 1_A is used for the indicator or characteristic function of a set A.

The development of the General Theory of Processes encountered at least two serious difficulties which could not be solved in the framework of Measure Theory and required a use of Capacity Theory. They are the *Section Theorem* and the *Projection Theorem*. The Section Theorem asserts that if the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete (that is $\mathcal{F}$ contains all $\mathbb{P}$-null sets) and $A \in \mathcal{O}$, then there exists a stopping time τ such that its graph is included in A. And the Projection Theorem states that given an optional set $A \subset \mathbb{R}^+ \times \Omega$, the projection $\pi(A)$ on Ω belongs to the complete σ-field $\mathcal{F}$. For instance, this result allows to prove that given a Borel set B of the real line, the random variable $\tau_B(\omega) = \inf\{t \ge 0 : X(t,\omega) \in B\}$ ($\inf \emptyset = \infty$), defines a stopping time for an $\mathbb{F}$-adapted process X with trajectories in D almost surely, provided the filtration $\mathbb{F}$ is right-continuous, that is, for all $t \ge 0$, $\mathcal{F}_t = \mathcal{F}_{t+} := \bigcap_{s>t} \mathcal{F}_s$, and in addition each σ-field contains all $\mathbb{P}$-null sets. Within this theory, the system $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P})$ is usually called a *Stochastic Basis* and a system $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, E, \mathcal{E}, \mathbb{P}, (X_t)_{t\in T})$ provides the whole structure needed to define an E-valued adapted stochastic process.

Attending to measurability properties only, stochastic processes may be classified as optional or predictable, as mentioned before, for which no probability is needed. However, richer properties of processes strongly depend on the probability considered in the stochastic basis. For instance, the definitions of *martingales*, *submartingales*, *supermartingales*, *semimartingales* depend on a specific probability measure, through the concept of *conditional expectation*. Let us mention that *semimartingales* form the most general class of possible integrands to give a rigorous meaning to *Stochastic Integrals* and *Stochastic Differential Equations*.

Probability is moreover fundamental for introducing concepts as *Markov Process* (see ▶Markov Processes), *Gaussian Process*, *Stationary Sequence and Stationary Process*.

Extensions of the Theory

Extensions to the theory have included changing either the nature of T to consider *Random Fields*, where $t \in T$ may have the meaning of a space label (T is no more a subset of the real line), or the state space E, to deal for instance with measure-valued processes, or random distributions.

Example 5 Let $(T, \mathcal{T}, \nu)$ be a σ-finite measure space, and $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. Call $\mathcal{T}_\nu$ the family of all sets $A \in \mathcal{T}$ such that $\nu(A) < \infty$. A *Gaussian white noise* based on ν is a random set function W defined on $\mathcal{T}_\nu$ and values in $\mathbb{R}$ such that

(a) $W(A)$ is centered Gaussian and $\mathbb{E}\left(W(A)^2\right) = \nu(A)$, for all $A \in \mathcal{T}_\nu$;
(b) If $A \cap B = \emptyset$, then $W(A)$ and $W(B)$ are independent.

In particular, if $T = \mathbb{R}^{+2}$, $\mathcal{T}$ the corresponding Borel σ-field, and $\nu = \lambda$ the product Lebesgue measure, define $B_{t_1,t_2} = W(]0,t_1] \times]0,t_2])$, for all $(t_1,t_2) \in T$. The process $(B_{t_1,t_2})_{(t_1,t_2)\in T}$ is called the *Brownian sheet*. ▽

Going further, on the state space E consider the algebra $\mathfrak{E}$ of all bounded $\mathcal{E}$-measurable complex-valued functions. Then, to each E-valued stochastic process X one associates a family of maps $j_t : \mathfrak{E} \to L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$, where $j_t(f)(\omega) = f(X(t,\omega))$, for all $t \ge 0$, $\omega \in \Omega$. The family $(j_t)_{t\in\mathbb{R}^+}$, known as the *Algebraic Flow* can be viewed as a family of complex random measures (each j_t is a Dirac measure supported by $X(t,\omega)$) or, better, as a *-homomorphism between the two *-algebras $\mathfrak{E}$, $L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$, the * operation being here the

complex conjugation. The stochastic process is completely determined by the algebraic flow $(j_t)_{t\in\mathbb{R}^+}$.

Example 6 Consider a Brownian motion B defined on a stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P})$, with states in $\mathbb{R}$, and call $\mathfrak{B}$ the algebra of bounded complex valued Borel function defined on the real line. $\mathfrak{B}$ is a *-algebra of functions, that is, there exists an involution * (the conjugation), such that $f \mapsto f^*$ is antilinear and $(fg)^* = g^* f^*$, for all $f, g \in \mathfrak{B}$. The algebraic flow associated to B is given by $j_t(f) = f(B_t)$, for all $t \geq 0$, and any $f \in \mathfrak{B}$, that is $j_t : \mathfrak{B} \to L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathbb{P}(B_0 = x) = 1$, then $j_0(f) = f(x)$ almost surely. Moreover, notice that Itô's formula implies that for all bounded f of class C^2, it holds

$$j_t(f) = f(x) + \int_0^t j_s\left(\frac{d}{dx}f\right) dB_s + \int_0^t j_s\left(\frac{1}{2}\frac{d^2}{dx^2}f\right) ds. \quad \triangledown$$

Algebraic flows provide a suitable framework to deal with more generalized evolutions, like those arising in the description of *Open Quantum System Dynamics*, where the algebras are non commutative. Thus, given two unital *-algebras (possibly non commutative) $\mathfrak{A}, \mathfrak{B}$, a notion of *Algebraic Stochastic Process* is given by a flow $(j_t)_{t\in\mathbb{R}^+}$, where $j_t : \mathfrak{B} \to \mathfrak{A}$ is a *-homomorphisms, for all $t \geq 0$. That is, each j_t is a linear map, which satisfies $(j_t(b))^* = j_t(b^*)$, $j_t(a^*b) = j_t(a)^* j_t(b)$, for all $a, b \in \mathfrak{B}$, and $j_t(\mathbf{1}_{\mathfrak{B}}) = \mathbf{1}_{\mathfrak{A}}$, where $\mathbf{1}_{\mathfrak{A}}$ (resp. $\mathbf{1}_{\mathfrak{B}}$) is the unit of $\mathfrak{A}$ (resp. $\mathfrak{B}$).

The Dawning of Stochastic Analysis as a Pillar of Modern Mathematics

These days, Stochastic Processes provide the better description of complex evolutionary phenomena in Nature. Coming from our understanding of the macro world, through our everyday life, exploring matter at its smallest component, stochastic modeling has become fundamental. In other words, stochastic processes have become influential in all sciences, namely, in biology (population dynamics, ecology, neurosciences), computer science, engineering (especially electric and operation research), economics (via finance), physics, among others. The new branch of Mathematics, known as Stochastic Analysis, is founded on stochastic processes. Stochastics is invading all branches of Mathematics: Combinatorics, Graph Theory, Partial and Ordinary Differential Equations, Group Theory, Dynamical Systems, Geometry, Functional Analysis, among many other specific subjects. The dawning of Stochastic Analysis era is a fundamental step in the evolution of human understanding of Chance as a natural interconnection and interaction of matter in Nature. This has been a long historical process which started centuries ago with the dart game.

Acknowledgments

The author is gratefully indebted with a number of anonymous referees for heartening support. Their comments were fundamental to improve the first version of this contribution. Also, no symphony orchestra could sound appropriately with no experimented conductor. The hearted conductor of this encyclopedia has been Professor Miodrag Lovric to whom I express my deep gratitude for his efficient and courageous work.

This work received partial support of grant PBCT-ADI13 of the Chilean Science and Technology Bicentennial Foundation.

About the Author

The following bibliography is nothing but a very small sample of references on stochastic processes, which could be termed classic, as well as more recent textbooks. General references as well as specialized books on the field are fast increasing, following the success of stochastic modeling, and one can be involuntarily and easily unfair by omitting outstanding authors.

Rolando Rebolledo obtained his "Doctorat d'État" at the Université Pierre et Marie Curie (Paris VI), France, in 1979. He is Professor at the Faculty of Mathematics, and Head of the Center for Stochastic Analysis, Pontificia Universidad Católica de Chile. He was President of the Sociedad de Matemática de Chile during five periods (1982–1985, 1994–1995, 1995–1998). He was Chairman of the Latin American Regional Committee of the Bernoulli Society (1989–1993), member of the Council of that Society and Scientific Secretary of the Committee for the Year 2000 of the Bernoulli Society. He chaired the Commission on Development and Exchanges of the International Mathematical Union (1994–1998, 1998–2002). Professor Rebolledo is a member of the American Mathematical Society, Bernoulli Society, Fellow of the International Statistics Institute since 1994. He has been twice awarded with the "Presidential Chair" in Chile (1995–1998, 1999–2002), and with the Medal of the Catholic University for outstanding research achievements (1996 and 1999). Dr. Rebolledo has published over 80 research papers, and edited five Proceedings of the International ANESTOC Workshops. Rolando Rebolledo has been Visiting Professor at many universities all over the world, including Denmark, Germany, Brazil, Venezuela, Italy, France, Russia, Australia, USA, and Portugal.

Cross References

►Brownian Motion and Diffusions
►Extremes of Gaussian Processes
►Gaussian Processes

►Lévy Processes
►Markov Chains
►Markov Processes
►Martingales
►Point Processes
►Poisson Processes
►Random Walk
►Renewal Processes
►Sampling Problems for Stochastic Processes
►Statistical Inference for Stochastic Processes
►Stochastic Differential Equations
►Stochastic Processes: Applications in Finance and Insurance
►Stochastic Processes: Classification

References and Further Reading

Accardi L, Lu YG, Volovich I (2002) Quantum theory and its stochastic limit. Springer, Berlin
Bhattacharya R, Waymire EC (2007) A basic course in probability theory. Springer Universitext, New York
Bhattacharya R, Waymire EC (2009) Stochastic processes with applications. SIAM Classics in Applied Mathematics, Philadelphia
Dellacherie C (1972) Capacités et processus stochastiques. Springer, New York
Dellacherie C, Meyer PA (1978–1987) Probabilités et potentiel, vols 1–4. Hermann, Paris
Doob JL (1953) Stochastic processes. Wiley, New York
Dynkin EB (1965) Markov processes. Springer, Berlin (Translated from Russian)
Ethier K, Kurtz TG (1986) Markov processes: characterization and convergence. Wiley, New York
Feller W (1966) An introduction to probability theory and its applications, vol 2. Wiley, New York
Gikhman II, Skorokhod AV (1974–1979) Theory of stochastic processes, vol 1–3. Springer, Berlin (Translated from Russian)
Itô K (2006) Essentials of stochastic processes. American Mathematical Society, Providence
Karatzas I, Shreve SE (1991) Brownian motion and stochastic calculus. Springer, New York
Lévy P (1965) Processus stochastiques et mouvement Brownien. Gauthier-Villars, Paris
Meyer PA (1966) Probability and potentials. Ginn-Blaisdell, Boston
Meyer PA (1993) Quantum probability for probabilists. Lecture notes in mathematics, vol 1538, Springer, Berlin
Neveu J (1975) Discrete-parameter martingales. North-Holland, Amsterdam; American Elsevier, New York
Parthasarathy KR (1992) An introduction to quantum stochastic calculus. Birkhaüser, Basel
Protter P (1990) Stochastic integration and differential equations: a new approach. Springer, Berlin
Rebolledo R (2006) Complete positivity and the Markov structure of open quantum systems, in open quantum systems II. Lecture notes in mathematics, 1882, pp 149–182
Varadhan SRS (2007) Stochastic processes. Courant lectures notes in mathematics, vol 16. American Mathematical Society, New York

Stochastic Processes: Applications in Finance and Insurance

Leda D. Minkova
Associate Professor, Faculty of Mathematics and Informatics
Sofia University "St. Kl. Ohridski", Sofia, Bulgaria

The applications of ►stochastic processes and martingale methods (see ►Martingales) in finance and insurance have attracted much attention in recent years.

Martingales in Finance

Let us consider a continuous time arbitrage free financial market with one risk-free investment (bond) and one risky asset (stock). All processes are assumed to be defined on the complete probability space $(\Omega, \mathcal{F}_T, (\mathcal{F}_t), P)$ and adapted to the filtration $(\mathcal{F}_t)$, $t \le T$. The bond yields a constant rate of return $r \ge 0$ over each time period. The risk-free bond represents an accumulation factor and its price process B equals

$$dB_t = rB_t dt,\ t \in [0, T],\ B_0 = 1, \tag{1}$$

or $B_t = e^{rt}$. The evolution of the stock price S_t is described by the linear stochastic differential equation

$$dS_t = S_t(\mu dt + \sigma dW_t),\ t \in [0, T],\ S_0 = S, \tag{2}$$

where the expected rate of return μ and the volatility coefficient σ are constants. The stochastic process W_t, $t \ge 0$ is a one-dimensional Brownian motion. The solution of Eq. 2 is given by

$$S_t = S \exp\left(\sigma W_t + \left(\mu - \frac{\sigma^2}{2}\right)t\right),\ t \in [0, T]. \tag{3}$$

The process (3) is considered by Samuelson (1965) and is called a geometric Brownian motion. The market with two securities is called a standard diffusion (B, S) market and is suggested by F. Black and M. Scholes (1973). The references are given in Shiryaev (1999) and Rolski et al. (1999).

A European call (put) option, written on risky security gives its holder the right, but not obligation to buy (sell) a given number of shares of a stock for a fixed price at a future date T. The exercise date T is called maturity date and the price K is called a strike price. The problem of option pricing is to determine the value to assign to the option at a time $t \in [0, T]$. The writer of the option has to calculate the fair price as the smallest initial investment that would

allow him to replicate the value of the option throughout the time T. The replication portfolio can be used to hedge the risk inherent in writing the option.

Definition 1 **(Martingale measure)** *A probability measure $\overline{P}$ defined on $(\Omega, \mathcal{F}_T)$ is called a martingale measure if it is equivalent to P ($\overline{P} \sim P$) and the discounted process $\overline{S}_t = S_t B_t^{-1}$ is a $\overline{P}$– local martingale.*

For the Black–Scholes model, the martingale measure is unique and is defined by the following theorem of Girsanov type.

Theorem 1 *The unique martingale measure $\overline{P}$ is given by the Radon–Nikodym derivative*

$$\frac{d\overline{P}}{dP} = \exp\left(-\frac{\mu - r}{\sigma} W_T - \frac{1}{2}\left(\frac{\mu - r}{\sigma}\right)^2 T\right), \; P - a.s.$$

Under the martingale measure, $\overline{P}$, the discounted stock price $\overline{S}_t$ satisfies the equation

$$d\overline{S}_t = \sigma \overline{S}_t d\overline{W}_t, \; t \geq 0,$$

where

$$\overline{W}_t = W_t + \frac{\mu - r}{\sigma} t, \; t \leq T$$

is a standard Brownian motion (see ►Brownian Motion and Diffusions) with respect to the measure $\overline{P}$.

The new probability measure $\overline{P}$ is called also a *risk-neutral measure*. The ratio $\frac{\mu - r}{\sigma}$ is called a *market price of risk*.

Consider a European call option written on a stock S_t, with exercise date T and strike price K. If we assume that the price of a stock is described by (2) and the payoff function is $f_T = \max(S_T - K, 0)$, then the fair price C_t of the European call option at time t is given by the famous Black–Scholes formula Black F, Scholes M (1973).

Theorem 2 **(Black–Scholes formula)** *The value C_t at time t of the European call option is given by*

$$C_t = S_t \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2), \; t \leq T$$

where

$$d_1 = \frac{\log\left(\frac{S_t}{K}\right) + (T-t)\left(r + \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T-t}},$$

$$d_2 = \frac{\log\left(\frac{S_t}{K}\right) + (T-t)\left(r - \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}$$

and Φ is the standard Gaussian cumulative distribution function.

Insurance Risk Model

The standard model of an insurance company, called *risk process* $\{X(t), t \geq 0\}$ is given by

$$X(t) = ct - \sum_{k=1}^{N(t)} Z_k, \quad \left(\sum_{1}^{0} = 0\right). \tag{4}$$

Here c is a positive real constant representing the *risk premium* rate. The sequence $\{Z_k\}_{k=1}^{\infty}$ of mutually independent and identically distributed random variables, with common distribution function F, $F(0) = 0$, and mean value μ, is independent of the *counting process* $N(t)$, $t \geq 0$. The process $N(t)$ is interpreted as the number of claims on the company during the interval $[0, t]$. In the classical risk model, also called the Cramér–Lundberg model, the process $N(t)$ is a homogeneous Poisson process (see ►Poisson Processes), see for instance Grandell (1991). The ruin probability of a company with initial capital $u \geq 0$ is given by

$$\Psi(u) = P\left(u + X(t) < 0 \text{ for some } t > 0\right).$$

The martingale techniques have been introduced by H. Gerber in 1973 (see Gerber 1979). Since then, the martingale approach is a basic tool in risk theory (see the References in Schmidli (1996), Rolski et al. (1999), and Embrechts et al. (1997)).

Under the net profit condition $\theta = \frac{c}{\lambda\mu} - 1 > 0$, the following fundamental result holds (Embrechts et al. 1997).

Theorem 3 **(Cramér–Lundberg theorem)** *Assume that there exists $R > 0$ such that*

$$\int_0^{\infty} e^{Rx} dF_I(x) = 1 + \theta, \tag{5}$$

where $F_I(x) = \int_0^x (1 - F(y))dy$ is the integrated tail distribution of F.

a) *For all $u \geq 0$,*

$$\Psi(u) \leq e^{-Ru}; \tag{6}$$

b) $\lim_{u\to\infty} e^{Ru}\Psi(u) = \left[\frac{R}{\theta\mu}\int_0^{\infty} xe^{Rx}(1-F(x))dx\right]^{-1} < \infty$, *provided that*

$$\int_0^{\infty} xe^{Rx}(1 - F(x))dx < \infty.$$

c)

$$1 - \Psi(u) = \frac{\theta}{1+\theta}\sum_{n=0}^{\infty}\left(\frac{1}{1+\theta}\right)^n F_I^{*n}(u). \tag{7}$$

The condition (5) is known as the *Cramér condition*. Inequality (6) is called the *Lundberg inequality* and the constant R is the adjustment coefficient or *Lundberg exponent* (see Grandell 1991). Formula (7) is known as Pollaczek–Khinchin formula.

Example 1 **(Exponentially Distributed Claims)** *Suppose that the claim sizes are exponentially distributed with parameter* μ, *that is* $F(z) = 1 - e^{-\frac{z}{\mu}}$, $z \geq 0, \mu > 0$.

In this case, $F_I(z)$ *is also an exponential distribution function and the solution of equation (5) is*

$$R = \frac{1}{\mu}\frac{\theta}{1+\theta}.$$

The Pollaczek–Khinchin formula (7) gives the ruin probability

$$\Psi(u) = \frac{1}{1+\theta}e^{-\frac{1}{\mu}\frac{\theta}{1+\theta}u}, \quad u \geq 0.$$

Cross References

►Brownian Motion and Diffusions
►Insurance, Statistics in
►Martingales
►Optimal Statistical Inference in Financial Engineering
►Radon–Nikodým Theorem
►Stochastic Processes
►Stochastic Processes: Classification
►Testing Exponentiality of Distribution

References and Further Reading

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–657

Embrechts P, Klüppelberg C, Mikosh T (1997) Modelling extremal events for insurance and finance. Springer, Berlin

Gerber HU (1979) An introduction to mathematical risk theory. S.S. Huebner Foundation, Wharton School, Philadelphia

Grandell J (1991) Aspects of risk theory. Springer, New York

Pliska SR (1997) Introduction to mathematical finance. Blackwell, Oxford

Rolski T, Schmidli H, Schmidt V, Teugels J (1999) Stochastic processes for insurance and finance. Wiley, Chichester

Samuelson PA (1965) Rational theory of warrant pricing. Ind Manag Rev 6:13–31

Schmidli H (1996) Martingales and Insurance Risk. In Eighth International Summer School on Probability Theory and Mathematical Statistics, pp 155–188

Shiryaev AN (1999) Essentials of stochastic finance: facts, models, theory. World Scientific, Singapore

Stochastic Processes: Classification

Venkatarama Krishnan
Professor Emeritus ECE
UMass Lowell, Lowell, MA, USA

Definitions

Let $\{\Omega, \mathcal{F}, P\}$ be a complete probability space where Ω is the *sample space*, $\mathcal{F}$ is the σ-field associated with the sample space containing all the null sets of Ω, and P is the probability measure defined on the field $\mathcal{F}$. Let $\{\mathbb{R}, \mathcal{R})$ be a measurable range space called the *state space*, where $\mathbb{R} \equiv (-\infty, \infty)$ is the real line and $\mathcal{R}$ is the σ-field associated with the real line $\mathbb{R}$. A *random variable* X is a function that assigns a rule of correspondence between each $\omega \in \Omega$ and each $x \in \mathbb{R}$. This correspondence will induce a probability measure P_X defined on the field $\mathcal{R}$. Thus, X maps the probability space $\{\Omega, \mathcal{F}, P\}$ to the probability range space $\{\mathbb{R}, \mathcal{R}, P_X\}$

$$X : \{\Omega, \mathcal{F}, P\} \longrightarrow \{\mathbb{R}, \mathcal{R}, P_X\}. \tag{1}$$

The distribution function $F_X(x)$ of X is given by

$$P\{\omega : X(\omega) \leq x\} = P\{X \leq x\} = F_X(x), \; x \in \mathbb{R} \tag{2}$$

and the density function $f_X(x)$, which may include impulse functions of x, is the derivative of $F_X(x)$.

The definition (see, e.g., Gikhman and Skorokhod 1996, p. 1 and 144) of a *stochastic* (or random) process requires a parameter set Θ and an increasing sequence of sub σ-fields $\{\mathcal{F}_\theta \subset \mathcal{F}, \theta \in \Theta\}$ called the *filtration* σ*-field* such that $\mathcal{F}_\zeta \subset \mathcal{F}_\theta$ for each $\{\theta, \zeta \in \Theta, \zeta < \theta\}$. The filtration σ-field is a consequence of the distinction between the uncertainty of the future and the knowledge of the past. The family $\{X(\theta), \mathcal{F}_\theta\}$ of random variables defined on the probability space $\{\Omega, \mathcal{F}, P\}$ will be called a *random function* if the parameter set Θ is arbitrary and a *stochastic process* if the parameter set Θ is the time set $\mathbb{T} \equiv (-\infty, \infty)$, and θ is interpreted as time t. Thus, $X(t) \in \mathcal{F}_t$ is a stochastic process that maps the probability space $\{\Omega, \mathcal{F}, P\}$ to the range space $\{\mathbb{R}, \mathcal{R}, P_X\}$ for every point $\omega \in \Omega$ and $t \in \mathbb{T}$. $X(t)$ is said to be *adapted* to the filtration field $\{\mathcal{F}_t, t \in \mathbb{T}\}$ if $X(t)$ is $\mathcal{F}_t$-measurable in the sense the inverse image set $\{X(t)^{-1}[\mathbb{B}]\} \in \mathcal{F}_t$ for every subset $\mathbb{B}$ of the real line $\mathbb{R} \in \mathcal{R}$.

The important point to emphasize is that a stochastic process is not a single time function but an ensemble of time functions. If the time parameter t belongs to a set of integers $\mathbb{Z} \equiv \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ then $X(n)$ or X_n denotes a *discrete-time* stochastic process.

A non-negative real line will be represented by $\mathbb{R}^+ \equiv [0,\infty)$ and non-negative time set by $\mathbb{T}^+ \equiv [0,\infty)$. A set of non-negative integers will be denoted by $\mathbb{N} \equiv \{0,1,\ldots\}$ and a set of positive integers by $\mathbb{N}^+ \equiv \{1,2,\ldots,N\}$.

Since $X(t)$ is a random variable for every $t \in \mathbb{T}$, the distribution function $F_X(x:t)$ will be given by

$$P\{X(\omega,t) \le x\} = P\{X(t) \le x\} \equiv F_X(x:t),\ x \in \mathbb{R},\ t \in \mathbb{T} \tag{3}$$

and the density function $f_X(x:t)$, which again may include impulse functions of x, is the partial derivative of $F_X(x;t)$ with respect to x.

Autocorrelation and autocovariance functions for a stochastic process $X(t)$ for $\{t_1,t_2 \in \mathbb{T}\}$ are defined by:

$$\begin{aligned} R_X(t_1,t_2) &= \big[X(t_1)X(t_2)\big] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1,x_2:t_1,t_2)dx_1dx_2, \end{aligned} \tag{4}$$

$$\begin{aligned} C_X(t_1,t_2) &= E\big\{[X(t_1)-\mu_x(t_1)][X(t_2)-\mu_x(t_2)]\big\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1-\mu_x(t_1)][x_2-\mu_x(t_2)] \\ &\quad f(x_1,x_2:t_1,t_2)dx_1dx_2, \end{aligned} \tag{5}$$

where $\mu_x(t_1)$ and $\mu_x(t_2)$ are the mean values of $X(t)$ at times t_1 and t_2 respectively.

Stochastic processes can be classified in different categories but many of them straddle categories.

Stationary and Ergodic Process

A stochastic process $X(t)$ is nth order *stationary* if the nth order distribution function satisfies

$$\begin{aligned} &F_X(x_1,\ldots,x_n:t_1,\ldots,t_n) = F_X(x_1,\ldots,x_n: \\ &\quad t_1+\tau,\ldots,t_n+\tau) \text{for any } \tau \in \mathbb{T}. \end{aligned} \tag{6}$$

It is *strictly stationary* if Eq. (6) is true for all $n \in \mathbb{Z}$. However, the most useful concepts of stationarity are the first order stationarity defined by

$$F_X(x:t) = F_X(x:t+\tau) = F_X(x), \tag{7}$$

and the second order stationarity called *wide sense stationary* defined by

$$F_X(x_1,x_2:t_1,t_2) = F_X(x_1,x_n:t_1+\tau,t_2+\tau) = F_X(x_1,x_2:\tau). \tag{8}$$

Wide sense stationarity can be determined from the following two criteria:

1. The expected value $E[X(t)] = \mu_X$ = a constant.
2. The autocorrelation function $R_X(t_1,t_2) = R_X(t_2-t_1) = R_X(\tau)$ is a function of the time difference τ.

A stationary process $X(t)$ is *mean ergodic* if the ensemble average is equal to the time average of the sample function $X(t)$.

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)dt = \int_{-\infty}^{\infty}xf_X(t)dt = \mu_X, \tag{9}$$

or, equivalently the covariance $C_X(\tau)$ satisfies the condition $\int_{-\infty}^{\infty}|C_X(\tau)|d\tau < \infty$.

A stationary process is *correlation ergodic* if

$$\begin{aligned} &\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)X(t+\tau)dt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1x_2f_X(x_1,x_2:\tau)dx_1dx_2 = R_X(\tau), \end{aligned} \tag{10}$$

which is equivalent to the condition $\int_{-\infty}^{\infty}\big|E\{[X(t)X(t+\tau)]^2\} - E\{[X(t)]^2\}\big|d\tau < \infty$.

State and Time Discretized Process

The stochastic process $X(t)$ can be classified into four broad categories depending upon whether the state space is discretized with $\mathbb{R} \equiv \mathbb{Z}$ or the time is discretized with $\mathbb{T} \equiv \mathbb{Z}$ or both. As mentioned earlier, discrete-time random processes will be denoted by X_n or $X(n)$ where $n \in \mathbb{Z}$.

1. *Discrete State Discrete Time Process* (DSDT)
   At any given time $i > 0$ a particle takes a positive step from $X_0 = 0$ with probability p and a negative step with probability q with $p+q = 1$. The random variable Z_i representing each step is independent and identically distributed. The position X_n of the particle at time n is a stochastic process $X_n = Z_1 + Z_2 + \cdots + Z_n$. It represents a DSDT process with discrete time set $\mathbb{N}^+ = \{1,\ldots,n,\ldots\}$ and discrete state space $\mathbb{R} = \mathbb{Z} = \{\ldots,-1,0,1,\ldots\}$ representing the position of the particle. This process known as a *simple* ►*random walk* (see, e.g., Cox and Miller 1977, p. 25) is nonstationary. If $p = q$ then the process is called a *symmetric* simple random walk.
2. *Discrete State Continuous Time Process* (DSCT)
   A customer arrives at the service counter of a supermarket at a random time $t \ge 0$ at an average rate of λ per unit time interval. If $N(t)$ is the stochastic process representing the number of customers arriving in the time interval $[0,t]$ then $N(t)$ is a DSCT process with time set $\mathbb{T}^+ = \{0 \le t < \infty\}$ and discrete state space $\mathbb{N} = \{0,1,\ldots\}$ representing the number of customers. This process known as *Poisson process* (see ►Poisson Processes) is nonstationary.
3. *Continuous State Discrete Time Process* (CSDT)
   In the DSDT process of (1), each step of the particle at any time $i > 0$ is a continuous random variable Z instead of a discrete one, governed by a distribution function $F_Z(z)$ with mean μ_Z. If X_n is the position of the particle at time $i = n$ then X_n represents a CSDT

process with discrete time set $\mathbb{N}^+ = \{1, \ldots, n, \ldots\}$, and continuous state space $\mathbb{R}^+ = \{0 \le x < \infty\}$ representing the position of the particle. This process is nonstationary.

4. *Continuous State Continuous Time Process* (CSCT)
In the DSDT process of (1) the particle undergoes a positive or negative step of Δx in a time interval Δt. If certain limiting conditions on Δx and Δt are satisfied then as Δx and Δt tend to 0, a CSCT process results, which is called *Wiener process* (see, e.g., Cox and Miller 1977, p. 205) or *Brownian motion* (see ▶Brownian Motion and Diffusions). Extrusion of plastic shopping bags where the thicknesses of the bags vary constantly with respect to time with the statistics being constant over long periods of time is an example of a CSCT process. These processes are nonstationary.

Gaussian Process

A stochastic process $X(t)$ defined on a complete probability space is a *Gaussian stochastic process* if for any collection of times $\{t_0, t_1, \ldots, t_n\} \in \mathbb{T}$, the random variables $X_0 = X(t_0), X_1 = X(t_1), \ldots, X_n = X(t_n)$ are jointly Gaussian distributed for all $n \in \mathbb{Z}$, with joint probability function

$$f_{X_0 X_1 X_2 \ldots X_n}(x) = \frac{1}{(2)^{n/2}|\mathbf{C}_X|} \exp\left(- \frac{(\mathbf{x}-\boldsymbol{\mu}_X)^T \mathbf{C}_X^{-1} (\mathbf{x}-\boldsymbol{\mu}_X)}{2} \right) \tag{11}$$

where $\boldsymbol{\mu}_X$ is the mean vector and $\mathbf{C}_X$ is the covariance matrix of the random variables $\{X_0, X_1, \ldots, X_n\}$. The Wiener process is also an example of a Gaussian process.

Markov Process

Let the σ-field $\mathcal{F}_t$ generated by $\{X(s), s \le t, t \in \mathbb{T}\}$ represent the past history up to the present and the σ-field $\mathcal{F}_t^c$ generated by $\{X(s), s > t, t \in \mathbb{T}\}$ represent the future evolution. Let a random variable Y be $\mathcal{F}_t$-measurable and another random variable Z be $\mathcal{F}_t^c$-measurable. Then the process $\{X(t), t \in \mathbb{T}\}$ is called a *Markov process* (see Markov Processes) if the following hold:

1. Given the present information $X(t)$, the past Y and the future Z are conditionally independent.

$$E[YZ|X(t)] = E[Y|X(t)]E[Z|X(t)]. \tag{12}$$

2. The future Z, conditioned on the past history up to the present $\mathcal{F}_t$, is equal to the future given the present.

$$E[Z|\mathcal{F}_t] = E[Z|X(t)]. \tag{13}$$

3. The future Z, conditioned on the past value $X(s)$ is the future conditioned on the present value $X(t)$ and again conditioned on the past value $X(s)$.

$$E[Z|X(s)] = E\{E[Z|X(t)]|X(s)\} \text{ for } s < t. \tag{14}$$

This is known as the *Chapman-Kolmogorov equation* (see, e.g., Ross 2000, p. 166).

In terms of probability, with $\tau > 0$ and states x_h, x_i, x_j, Eq. (13) is equivalent to:

$$\begin{aligned} P\{X(t+\tau) = x_j | X(t) = x_i, X(u) \\ = x_h,\ 0 \le u < t\} = P\{X(t+\tau) \\ = x_j | X(t) = x_i\}. \end{aligned} \tag{15}$$

Or, for $t_0 < t_1 < \ldots < t_{n-1} < t_n$, and $\{x_k, k = 0, \ldots, n, \ldots\}$ belonging to some discrete-state space

$$\begin{aligned} P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) \\ = x_{n-1}, \ldots, X(t_0) = x_0\} \\ = P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\}. \end{aligned} \tag{16}$$

A Markov process has an important property that the density $f_{\tau_i}(t)$ of the random time τ_i spent in any given state x_i is an exponential and hence it is called *memoryless*.

Markov Chains

Discrete state Markov processes are called *chains*, and if time is continuous they are called *continuous* ▶*Markov chains*, and if time is discrete they are called *discrete Markov Chains*. The Poisson process is an example of a continuous Markov chain.

A stochastic process $\{X(t), t \in \mathbb{T}^+\}$ is a continuous-time Markov chain if for each of the discrete states h, i, j and any time $\tau > 0$

$$\begin{aligned} P\{X(t+\tau) = j | X(t) = i, X(u) = h, 0 \le u < t\} \\ = P\{X(t+\tau) = j | X(t) = i\}. \end{aligned} \tag{17a}$$

The quantity $P\{X(t+\tau) = j | X(t) = i\}$ is the time dependent transition probability defined by $p_{ij}(t, \tau)$, which is generally a function of times t and τ. If the transition from the state i to the state j is dependent only on the time difference $\tau = (t+\tau) - t$ then the transition probability is stationary and the Markov chain is called *homogeneous*. In this case transition probability becomes $p_{ij}(\tau)$.

The probability density function $f_{\tau_i}(t)$ of the random time τ_i spent in any given state i for a continuous Markov chain is exponential and hence it is called *memoryless*.

A stochastic process $\{X(n), n = 0, 1, \ldots\}$ is a discrete-time Markov chain if for each of the discrete states i, j and $\{i_k, k = 0, 1, \ldots, n-1\}$ and any time $m > 0$,

$$\begin{aligned} P\{X(n+m) = j | X(n) = i, X(n-1) = i_{n-1}, \ldots, X(0) = i_0\} \\ = P\{X(n+m) = j | X(n) = i\}. \end{aligned} \tag{17b}$$

The quantity $P\{X(n+m) = j|X(n) = i\}$ is called the m-step transition probability defined by $p_{ij}^{(m)}(n)$, which is generally a function of time n. If the transition from the state i to the state j is dependent only on the time difference $m = (n+m) - n$ then the transition probability is stationary and the Markov chain is *homogeneous*. In this case the m-step transition probability becomes $p_{ij}^{(m)}$.

The one-step probability from state i to state j of a homogeneous discrete Markov chain is given by:

$$P\{X(n+1) = j|X(n) = i\} = p_{ij}. \tag{18}$$

The probability mass function f_{τ_i} of the random time τ_i spent in any given state i for a discrete Markov chain is geometric and hence it is called *memoryless*.

Semi-Markov Process

In a Markov process the distributions of state transition times are exponential for a continuous process, and geometric for a discrete process and hence they are considered memoryless. While the definition of a *semi-Markov process* $X(t)$ defined on a complete probability space is the same as that of a Markov process (Eqs. 15 and 16), the distributions of transition times $\tau_i \in \mathbb{T}$ between states need not be memoryless but can be arbitrary. For a continuous-time semi-Markov process the state transitions can occur at any instant of time $t \in \mathbb{T}$ with an arbitrary density $f_{\tau_i}(t)$ for the time τ_i spent in state x_i and for a discrete-time semi-Markov process the state transitions can occur at time instants $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ with an arbitrary probability mass f_{τ_i} for the time τ_i spent in state i. If the amount of time spent in each state is 1 then this semi-Markov process is a Markov chain. Markov processes are a subclass of semi-Markov processes.

Independent Increment Process

A stochastic process $\{X(t), t \in \mathbb{T}\}$ is defined on a complete probability space with a sequence of time variables $\{t_0 < t_1 < \ldots < t_n\} \in \mathbb{T}$. If the increments $X(t_0), [X(t_1) - X(t_0)], \ldots, [X(t_n) - X(t_{n-1})]$ of the process $\{X(t), t \in \mathbb{T}\}$ are a sequence of independent random variables then the process is called an *independent increment* process (see, e.g., Krishnan 2006, p. 507). If the distribution of the increments $X_t - X_s, t > s$ depends only on the time difference $t - s = \tau$, then the process is a *stationary independent increment* process.

If the time set is discrete given by $\mathbb{N}^+ = \{1, 2, \ldots\}$ then the independent increment process is a sequence of independent random variables given by $Z_0 = X_0, \{Z_i = X_i - X_{i-1}, i \in \mathbb{N}^+\}$. Independent increment process is a special case of a Markov process. It is not a stationary process because of the following (see, e.g., Krishnan 2005, p. 61):

$$\begin{aligned} E[X(t)] &= \mu_0 + \mu_1 t, \text{ where } \mu_0 = E[X(t_0)] \text{ and} \\ \mu_1 &= E[X(t_1)] - \mu_0; \\ \mathrm{Var}[X(t)] &= \sigma_0^2 + \sigma_1^2 t, \text{ where } \sigma_0^2 = E[X(t_0) - \mu_1]^2 \text{ and} \\ \sigma_1^2 &= E[X(t_1) - \mu_0]^2 - \sigma_0^2. \end{aligned} \tag{19}$$

Poisson and Wiener processes are examples of stationary independent increment processes.

Uncorrelated and Orthogonal Increment Process

A stochastic process $\{X(t), t \in \mathbb{T}\}$ with $s_1 < t_1, s_2 < t_2$ and $t_1 \le t_2$

1. Has *uncorrelated increments* (see, e.g., Krishnan 2006, p. 508) if

$$E[(X_{t_2} - X_{s_2})(X_{t_1} - X_{s_1})] = E[(X_{t_2} - X_{s_2})]E[(X_{t_1} - X_{s_1})]. \tag{20}$$

2. Has *orthogonal increments* (see, e.g., Krishnan 2006, p. 508) if

$$E[(X_{t_2} - X_{s_2})(X_{t_1} - X_{s_1})] = 0. \tag{21}$$

Clearly, independent increments imply uncorrelated increments but the converse is not true.

General Random Walk Process

The simple random walk discussed earlier can be generalized. Starting from $X_0 = 0$ a particle takes independent identically distributed random steps $Z_1, Z_2, \ldots, Z_n$, whose values are drawn from an arbitrary distribution, which do not change with the state of the process. This distribution may be continuous with density function $f_Z(z)$ or discrete with probability of transition from state i to state j being p_{ij}. In the latter case p_{ij} will be dependent on the difference $j-i$, or, $p_{ij} = p_{j-i}$. The position $X_n = Z_1 + Z_2 + \cdots + Z_n$, $n \in \mathbb{N}^+$ of the particle is a stochastic process where n is the number of state transitions, which is always forward from state x_i to x_{i+1}. Depending upon whether the instants of these transitions are taken from the set $\mathbb{T}^+$ or $\mathbb{N}^+$ the process X_n is either a continuous-time or a discrete-time *general random walk* (see, e.g., Cox and Miller 1977, p. 46). In either case the distribution of the time intervals between these transitions is arbitrary and hence it is a special case of a semi-Markov process.

Birth and Death Process

Let $\{X(t), t \ge 0\}$ be a continuous Markov chain. State transitions can occur only from the state $x_i = i$ to $x_{i+1} = i + 1$, or $x_{i-1} = i - 1$, or stays at $x_i = i$. $X(t)$ is called a *birth and*

death process (see, e.g., Kleinrock 1975, p. 53) if in a small interval Δt

$$P\{X(t+\Delta t) - X(t) = j | X(t) = i\}$$
$$= \begin{cases} \lambda_i \Delta t + o(\Delta t), & \text{if} \quad j = 1, \\ \mu_i \Delta t + o(\Delta t), & \text{if} \quad j = -1, \\ o(\Delta t), & \text{if} \quad |j| > 1. \end{cases} \tag{22}$$

$$\text{and } P\{X(t+\Delta t) - X(t) = 0 | X(t) = i\}$$
$$= 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t), \tag{23}$$

where $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. λ_i is the rate at which births occur and μ_i is the rate at which deaths occur when the population size is i. The probability of the population size being i at any time $t > 0$ is given by $P\{X(t) = i\} = P_i(t)$. This is a Markov process with independent increments. If $\lambda_i = i\,\lambda$ and $\mu_i = i\,\mu$ then this process is called a *linear* birth and death process.

The *pure birth process* is a sub-class of birth and death process with $\mu_i \equiv 0$ for all i. State transitions can occur only from the state $x_i = i$ to $x_{i+1} = i + 1$ with rate λ_i or stays in the same state $x_i = i$.

The *Poisson process* is a sub-class of pure birth processes with $\lambda_i \equiv \lambda$ a constant for all i. Here the probability of i events in time t is given by $P_i(t, \lambda) = [(\lambda t)^i / i!] e^{-\lambda t}, t > 0$. This process has stationary independent increments.

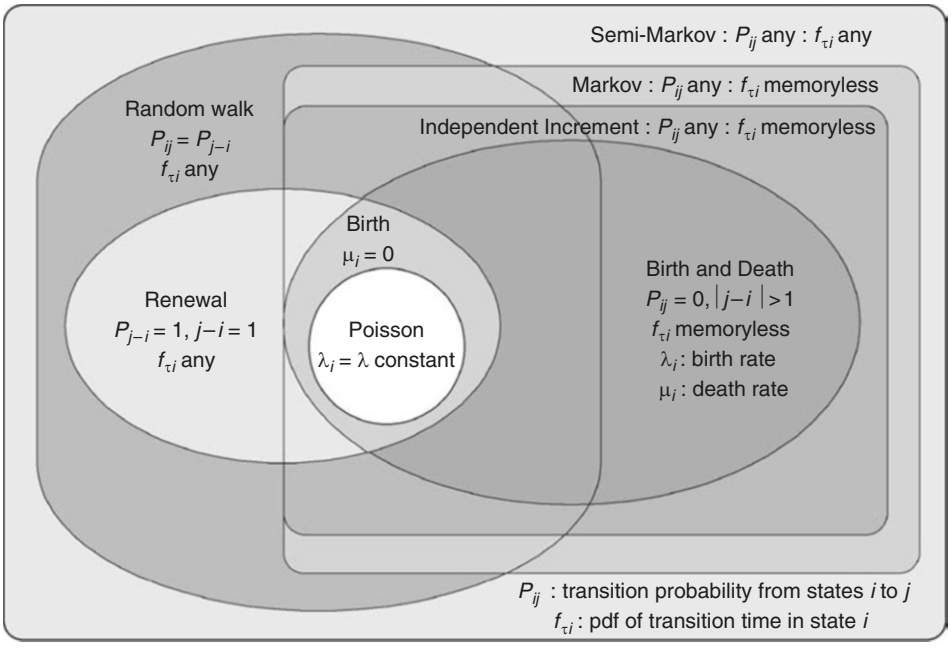
Renewal Process

In the general random walk process X_n discussed in the previous section the interest was in the probability of the state of the particle after n transitions. In ►renewal processes the concern is only in the number of transitions that occur in a time interval $[0, t]$ and not on the state. Starting from $t = 0$ the transitions occur at sequence of times $0 < t_1 < t_2 < \ldots < t_n, n > 0$ with *inter-arrival times* defined by random variables $Y_1 = t_1, Y_2 = (t_2 - t_1), \ldots, Y_n = (t_n - t_{n-1})$. The random variables $Y_i, i \in \mathbb{N}^+$ are independent and identically distributed with an arbitrary density function $f(y)$ with $E[Y_i] = \mu$ for all i.

The stochastic process defined by $X_n = Y_1 + Y_2 + \cdots + Y_n$ is called a *renewal* process (see, e.g., Cox and Miller 1977, p. 340), where a renewal occurs at the epochs at $t_1 < t_2 < \cdots < t_n$. In this process X_n represents the *time* of the *nth* renewal whereas in the random walk X_n represents the *state* of the process at time n. This process is a subclass of semi-Markov processes and also a subclass of random walk processes. If the density function $f(y)$ is either exponential or geometric then this process is Markov. The relationship among the various discrete-state random processes similar to the one in Kleinrock (1975, p. 25) is shown in Fig. 1.

Martingale Process

A *martingale process* (see, e.g., Doob 1990, p. 91 and p. 294; ►Martingales) is a stochastic process where the best estimate of the future value conditioned on the past history



Stochastic Processes: Classification. Fig. 1 Relationships among some discrete state stochastic processes

including the present is the present value. Since there is no trend to the process it is unpredictable. Many problems in engineering and finance can be cast in the martingale framework. Pricing stock options (see, e.g., Ross 2000, p. 556) and bonds has been cast in the martingale framework.

Let $\{\Omega, \mathcal{F}, P\}$ be a complete probability space and let $\{\mathcal{F}_n, n \in \mathbb{N}\}$ be an increasing family of sub σ-fields of $\mathcal{F}$. The real valued sequence of random variables $\{X_n, n \in \mathbb{N}\}$ adapted to the family $\{\mathcal{F}_n, n \in \mathbb{N}\}$ is a discrete $\mathcal{F}_n$-martingale if for all n:

1. $E|X_n| < \infty$
2. $E\{X_n|\mathcal{F}_m\} = X_m$ for $m \le n$

If condition (2) is modified as

3. $E\{X_n|\mathcal{F}_m\} \ge X_m$ for $m \le n$ submartingale
4. $E\{X_n|\mathcal{F}_m\} \le X_m$ for $m \le n$ supermartingale

Analogously, let $\{\mathcal{F}_t, t \in \mathbb{T}^+\}$ be an increasing family of sub σ-fields of $\mathcal{F}$ of a complete probability space. The real valued stochastic process $\{X(t), t \in \mathbb{T}^+\}$ adapted to the family $\{\mathcal{F}_t, t \in \mathbb{T}^+\}$ is a continuous $\mathcal{F}_t$-martingale if for all $t \in \mathbb{T}^+$:

1. $E|X(t)| < \infty$,
2. $E\{X(t)|\mathcal{F}_s\} = X_s$ for $s \le t$.

If condition (2) is modified as

3. $E\{X(t)|\mathcal{F}_s\} \ge X_s$ for $s \le t$ submartingale.
4. $E\{X(t)|\mathcal{F}_s\} \le X_s$ for $s \le t$ supermartingale.

Note that any martingale is both a submartingale and a supermartingale.

In the simple random walk process given in DSDT, if $n(p-q)$ is subtracted from X_n, then $Y_n = [X_n - n(p-q)]$ is an example of a discrete martingale with respect to the sequence $\{Z_k, k = 1, \ldots, n-1\}$ even though X_n is not. The Wiener process $W(t)$ is an example of a continuous $\mathcal{F}_t$-martingale. In the Poisson process $N(t)$, if the mean λt is subtracted then $Y(t) = [N(t) - \lambda t]$ is another example of a continuous $\mathcal{F}_t$-martingale even though $N(t)$ is not. However, both X_n and $N(t)$ are Markov processes leading to the conclusion that a Markov process is not necessarily a martingale. It can also be shown that a martingale is not necessarily a Markov process.

The martingale property captures the notion of a fair game. A fair coin is tossed and a player wins a dollar if the toss is heads and loses a dollar if the toss is tails. At the end of the *mth* toss the player has X_m dollars. The estimated amount of money after the $m+1$st toss is still X_m dollars since the expected value of the $m+1$st toss is zero.

Periodic Process

Let $\{X(t), t \in \mathbb{T}\}$ be a stochastic process defined on a complete probability space taking values in the range space $\{\mathbb{R}, \mathcal{R}\}$. $X(t)$ is *periodic in the wide sense* (see, e.g., Krishnan 2006, p. 558) with period $T_c (T_c > 0)$ if the mean $\mu_X(t)$ and the autocorrelation function $R_X(t,s)$ satisfy

$$\mu_X(t) = \mu_X(t + kT_c) \text{ for all } t \text{ and integer } k \tag{24}$$

$$\begin{aligned} R_X(t,s) &= R_X(t + kT_c, s) \\ &= R_X(t, s + kT_c) \text{ for all } t, s \text{ and integer } k. \end{aligned} \tag{25}$$

Note that $R_X(t,s)$ is periodic in both arguments t and s.

However, for a stationary periodic process $X(t)$ with $\tau = t - s$, Eq. (25) simplifies to

$$R_X(\tau) = R_X(\tau + kT_c) \text{ for all } \tau \text{ and integer } k. \tag{26}$$

Since $R_X(\tau)$ is uniformly continuous, a zero mean stationary periodic stochastic process $X(t)$ with fundamental frequency $\omega_c = 2\pi/T_c$ can be represented in the mean square sense by a Fourier series

$$\begin{aligned} X(t) &= \sum_{n=-\infty}^{\infty} X_n \exp(jn\omega_0 t), X_0 = 0 \\ \text{where } X_n &= \frac{1}{T_c} \int_0^{T_c} X(t) \exp(-jn\omega_0 t) dt. \end{aligned} \tag{27}$$

Cyclostationary process

Allied to the periodic process is the *cyclostationary process* (see, e.g., Krishnan 2006, p. 560). A *strict sense* cyclostationary process $X(t)$ on a complete probability space with period $T_c (T_c > 0)$ is defined by

$$\begin{aligned} &F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) \\ &= F_X(x_1, \ldots, x_n; t_1 + kT_c, \ldots, t_n + kT_c) \end{aligned} \tag{28}$$

for all n and k.

Since the above definition is too restrictive, a *wide sense* cyclostationary $X(t)$ can be defined by

$$\begin{aligned} \mu_X(t) &= \mu_X(t + kT_c) \\ R_X(t_1, t_2) &= R_X(t_1 + kT_c, t_2 + kT_c). \end{aligned} \tag{29}$$

About the Author

Venkatarama Krishnan, Ph D, is Professor Emeritus in the Department of Electrical and Computer Engineering at the University of Massachusetts Lowell. Previously, he has taught at Smith College (2003), the Indian Institute of Science Bangalore (1971–1987), Polytechnic University of New York (1964–1971), University of Pennsylvania (1961–1964), Villanova University (1958–1961), and Princeton

University (1957–1958). In 1956 he was the recipient of an Orson Desaix Munn Scholarship from Princeton University. He was also a co-director (1992–2000) of the Center for Advanced Computation and Telecommunications at University of Massachusetts Lowell. He has taught Probability and Stochastic Processes continuously for over forty years and received the best teaching award from University of Massachusetts Lowell in 2000. He has authored four books in addition to technical papers, the latest book being *Probability and Stochastic Processes* published by Wiley in 2006. Prof. Krishnan is a life senior member of IEEE, and is listed in *Who is Who in America, 2010.*

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Lévy Processes
►Markov Chains
►Markov Processes
►Martingales
►Point Processes
►Poisson Processes
►Random Walk
►Renewal Processes
►Stochastic Processes

References and Further Reading

Doob JL (1990) Stochastic processes, Wiley, New York
Gikhman II, Skorokhod AV (1996) Introduction to the theory of random processes. Dover, New York
Krishnan V (2005) Nonlinear filtering and smoothing. Dover, New York
Krishnan V (2006) Probability and random processes.Wiley, Hoboken, NJ
Cox DR, Miller HD (1977) The theory of stochastic processes. Chapman and Hall/CRC, London
Kleinrock L (1975) Queueing systems, vol 1. Wiley, New York
Ross SM (2000) Introduction to probability models. Harcourt Academic, San Diego

Stratified Sampling

Michael P. Cohen
Adjunct Professor
George Mason University, Fairfax, VA, USA
NORC at the University of Chicago, Washington DC, USA

Stratification refers to dividing a population into groups, called *strata*, such that pairs of population units within the same stratum are deemed more similar (*homogeneous*) than pairs from different strata. The strata are mutually exclusive (non-overlapping) and exhaustive of the population. Clearly sufficient information on each population unit must be available before we can divide the population into strata.

The primary reason for dividing a population into strata is to make use of the strata in drawing a sample. For example, instead of drawing a simple random sample of sample size n from the population, one may draw a ►simple random sample of sample size n_h from stratum h of L strata, where $n = n_1 + \cdots + n_L$. The sample selection for any stratum is done independently of the other strata. The stratum sample sizes n_h are often chosen proportional to the number of population units in stratum h but other allocations of the stratum samples may be preferred in specific situations.

There are two major reasons for drawing a stratified sample instead of an unstratified one:

1. Such samples are generally more efficient (in the sense that estimates have smaller variances) than samples that do not use stratification. There are exceptions, primarily when the strata are far from homogeneous with respect to the variable being estimated.
2. The sample sizes are controlled (rather than random) for the population strata. This means, in particular, that one may guarantee adequate sample size for estimates that depend only on certain strata. For instance, if men and women are in separate strata, one can assure the sample size for estimates for men and for women.

Estimation Under Simple Random Sampling Within Strata

The independence of the sample selection by strata allows for straightforward variance calculation when simple random sampling is employed within strata. Let Y_T denote the population total for a variable Y for which an estimate is sought. Let N_h and n_h denote respectively the population size and sample size for stratum h. Let, moreover, Y_{hj} and y_{hi} denote respectively the Y-value of the jth population element or i^{th} sample element in stratum h. Then, if

$$\overline{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj} \text{ and } \overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi},$$

define

$$S_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(Y_{hj} - \overline{Y}_h)^2 \text{ and } s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)^2.$$

We estimate Y_T by $\hat{y}$ where $\hat{y} = \sum_{h=1}^{L} N_h \bar{y}_h$. The variance of $\hat{y}$ is

$$V(\hat{y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h}(1 - n_h/N_h)S_h^2$$

and the variance is estimated by

$$\hat{V}(\hat{y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h}(1 - n_h/N_h)s_h^2.$$

Similarly, the population mean $\overline{Y} = Y_T/N$, where $N = \sum_{h=1}^{L} N_h$ is the size of the population, is estimated by $\hat{y}/N$ and its variance by $\hat{V}(\hat{y})/N^2$.

Allocation of Sample Sizes to Strata Under Simple Random Sampling within Strata

For a total sample size of n and given values of S_h, the question arises how should one allocate the sample to the strata; that is, how should one choose the n_h, $h = 1, \ldots, L$, so that $n = n_1 + \cdots + n_L$ and $V(\hat{y})$ is minimized? This is a straightforward constrained minimization problem (solved with Lagrange multipliers) that yields the solution:

$$n_h = \frac{nN_hS_h}{\sum_{k=1}^{L} N_kS_k}$$

Note that, as one would expect, the more variability in a stratum (larger S_h), the larger the relative sample size in that stratum. This method of determining the stratum sample sizes is termed *Neyman allocation* in view of the seminal paper on stratified sampling by Neyman (1934).

Sometimes the strata are not equally costly to sample. For example, there may be additional travel costs in sampling a rural geographically-determined stratum over an urban one. If it costs C_h to sample a unit in stratum h, then the allocation

$$n_h = \frac{nN_hS_h/\sqrt{C_h}}{\sum_{k=1}^{L} N_kS_k/\sqrt{C_k}}$$

is best in two senses: It minimizes $V(\hat{y})$ subject to fixed total cost (a fixed budget) $C_T = C_1 + \cdots + C_L$ and it minimizes C_T subject to fixed $V(\hat{y})$.

These allocations assume that the S_h, $h = 1, \ldots, L$, are known. In practice, rough estimates, perhaps based on a similar previous survey, will serve. The same comment applies to the costs for the cost-based allocation.

In the absence of any prior information, even approximate, the simple *proportional allocation* $n_h = nN_h/N$ is often used. In this case, the estimator $\hat{y}$ has a particularly simple form

$$\hat{y} = \sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{L} \frac{N_h}{(nN_h/N)} \sum_{i=1}^{n_h} y_{hi}$$
$$= \frac{N}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}.$$

Therefore $\hat{y}$ is just the sum of the sample values expanded by N/n. In many surveys a wide variety of quantities are estimated and their within-stratum variability may differ so proportional allocation may be employed as a compromise.

Unbiased estimation requires at least one sample selection per stratum. Unbiased variance estimation requires at least two selections per stratum.

Stratum Boundaries

Sometimes stratification is based on small discrete categories like gender or race. Other times, one may have data on a variable that can be regarded as continuous closely related to the variable one wants to estimate from the sample. For example, one may want to estimate the output of factories based on strata defined by the number of workers at the factory. One stratum might be all factories with 75–100 workers. In this case, 75 and 100 are said to be the stratum boundaries. How should these boundaries be chosen?

One method that has been shown to be good is the cumulative square root of frequencies method developed by Dalenius and Hodges (1957): Start by assuming (in our example) that the factories have been divided into a rather large number of categories based on the numbers of workers, numbered from fewest workers to the most workers. If f_k is the number of factories in category k, calculate $Q_k = \sqrt{f_1} + \cdots + \sqrt{f_k}$. Divide the factories into strata so that the differences between the at adjacent stratum boundary points are as equal as possible.

More recently, Lavallée and Hidiroglou (1988) developed an iterative procedure especially designed for skewed populations.

Variance Estimation for Stratified Samples

For simple estimators and stratified sampling, direct formulas are available to calculate variance estimates. These formulas are tailored to the specific estimator whose variance is sought. General purpose variance estimators have

been developed, however, that allow one to estimate variances for a wide class of estimators using a single procedure. See Wolter (2007) and Shao and Tu (1995) for a complete discussion of these procedures.

The procedure *balance half-sample replication* (or *balanced repeated replication*) has been developed as a variance estimation procedure when two primary sampling units (PSUs) are selected from each stratum. There may be additional sampling within each PSU so the sample design may be complex. The variance estimation is based on half sample replicates, each replicate consisting of one PSU from each stratum. The pattern that determines which PSU to choose from each stratum for a particular replicate is based on a special kind of matrix, called a Hadamard matrix.

A form of the *jackknife method* (see ►Jackknife) is also widely employed with two PSU per stratum sample designs (although it can be extended to other designs). This jackknife method is based on forming replicates, but the replicate consists of one PSU selected to be in the replicate from a specific stratum, with both PSUs being in the replicate for all other strata.

Various forms of the *bootstrap method* (see ►Bootstrap Methods) have been employed in recent years as general variance estimation methods for stratified sampling.

Although not as generic, the *Taylor series* (or *linearization*) method is a powerful technique for estimating variances in complex samples.

Stratified Sampling with Maximal Overlap (Keyfitzing)

Sometimes it is worthwhile to select a stratified sample in a manner that maximizes overlap with another stratified sample, subject to the constraint that the probabilities of selection are the ones desired. For example, cost savings may arise if a new stratified sample is similar to a previous one, yet births, deaths, and migration in the population may preclude it being exactly the same. Keyfitz (1951) developed a method to deal with this problem, so it is often called *Keyfitzing*. More recent researchers have extended the method to more general situations.

Stratification in Two Phases

It may be that it is clearly desirable to stratify on a certain characteristic, but that characteristic may not be available on the sampling frame (list of units from which the sample is selected). For example, in travel surveys one would likely want to stratify on household type (e.g., single adult head of household or adult couple with children) but this information is usually not provided on an address list. One solution is to first conduct a large, relatively inexpensive first phase of the survey for the sole purpose of obtaining the information needed to stratify. This information is then employed in the stratification of the second stage of the survey. This process is called *two-phase sampling* or *double sampling*.

Let n_h^I be the size of the first stage sample that lies in stratum h and let $n^I = n_1^I + \cdots + n_L^I$ be the first-stage sample size. At the second stage, n_h^{II} units with Y-values $y_{h1}, \ldots, y_{hn_h^{II}}$ are sampled in stratum h. Then one can estimate Y_T by

$$\tilde{y} = N \sum_{h=1}^{L} \frac{n_h^I}{n^I} \sum_{i=1}^{n_h^{II}} \frac{y_{hi}}{n_h^{II}}$$

Approximate variance formulas can also be given. See, e.g., Raj and Chandhok (1998) or Scheaffer et al. (2006). Because the n_h^I are random, the usual (one-phase) variance formulas would underestimate the variance.

Poststratification

After a sample has been selected and the data collected, sometimes the estimation procedures of stratification can be employed even if the sample selection was for an unstratified design. An important requirement is that the population proportions N_h/N must be known, at least approximately. If so, then

$$\hat{y} = N \sum_{h=1}^{L} \frac{N_h}{N} \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} = N \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h$$

is an improved estimate of the population total. The usual variance estimator $\hat{V}(\hat{y})$, however, is no longer valid as it does not account for the randomness of the n_h. More complicated variance estimators can be developed for this purpose.

Another reason to employ poststratification is to reduce bias due to nonresponse.

Controlled Selection

Controlled selection is a sample selection method that is related to stratified sampling but differs in that independent selections are not made from the cells ("strata"). The method was introduced by Goodman and Kish (1950). For an example of controlled selection, imagine a two-dimensional array of cells of population units, say of industrial classification categories by geographic areas. All population units lie in exactly one cell, analogous to strata. The sample size is not large enough for there to be the two selections per cell needed for unbiased variance estimation if the selections were independent by cell. Under controlled selection, only certain balanced patterns of cell combinations can be selected. When properly carried out, this is a valid probability selection technique.

About the Author

Dr. Michael P. Cohen is Senior Consultant to the National Opinion Research Center and Adjunct Professor, Department of Statistics, George Mason University. He was President of the Washington Statistical Society (2007–2008), and of the Washington Academy of Sciences (2003–2004). He served as Assistant Director for Survey Programs of the U.S. Bureau of Transportation Statistics (2002–2006). He is a Fellow of the American Statistical Association, the American Educational Research Association, and the Washington Academy of Sciences. He is an Elected Member of the International Statistical Institute and Sigma Xi and a Senior Member of the American Society for Quality. Dr. Cohen has over 60 professional publications. He served as an Associate Editor, *Journal of the American Statistical Association*, Applications and Case Studies Section (2004–2006). He has been an Associate Editor of the *Journal of Official Statistics* since 2003. He is the Guest Problem Editor of the *Journal of Recreational Mathematics* for 2009–2010.

Cross References

►Balanced Sampling
►Jackknife
►Multistage Sampling
►Sampling From Finite Populations
►Simple Random Sample

References and Further Reading

Bethlehem J (2009) Applied survey methods: a statistical perspective. Wiley, Hoboken
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Dalenius T, Hodges JL (1957) The choice of stratification points. Skandinavisk Aktuarietidskrift 1–2:203–213
Goodman R, Kish L (1950) Controlled selection – a technique in probability sampling. J Am Stat Assoc 45:350–372
Keyfitz N (1951) Sampling with probabilities proportional to size: adjustment for changes in the probabilities. J Am Stat Assoc 46:105–109
Knottnerus P (2003) Sample survey theory: some Pythagorean perspectives. Springer, New York
Lavallée P, Hidiroglou M (1988) On the stratification of skewed populations. Surv Methodol 14:33–43
Lohr S (1999) Sampling: design and analysis. Brooks/Cole, Pacific Grove
Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc 97:558–606
Raj D, Chandhok P (1998) Sample survey theory. Narosa Publishing House, New Delhi
Scheaffer RL, Mendenhall W, Ott RL (2006) Elementary survey sampling, 6th edn. Duxbury, Belmont
Shao J, Tu D (1995) The jackknife and the bootstrap. Springer, New York
Wolter KM (2007) Introduction to variance estimation, 2nd edn. Springer, New York

Strong Approximations in Probability and Statistics

Murray D. Burke
Professor
University of Calgary, Calgary, AB, Canada

Strong approximations in Probability and Statistics are results that describe the closeness almost surely of random processes such as partial sums and ►empirical processes to certain ►Gaussian processes. As a result, strong laws such as the law of the iterated logarithm and weak laws such as the central limit theorem (see ►Central Limit Theorems) follow.

Let $X_1, X_2, \ldots$ be a sequence of independent random variables with the same distribution function. Put $S_n = X_1 + \cdots + X_n$. If the mean $m = E(X_1)$ exists (finite), then the strong law of large numbers states that $S_n/n \to m$, almost surely, as $n \to \infty$. One can ask the question, at what rate does this convergence take place? This question is answered, in 1941 by Hartman and Wintner, who proved the law of the iterated logarithm (LIL): If, in addition, the variance σ^2 of X_1 is finite, then

$$\begin{aligned} \limsup_{n\to\infty} \frac{S_n - nm}{\sigma\sqrt{2n\log\log n}} &\to_{a.s.} 1, \\ \liminf_{n\to\infty} \frac{S_n - nm}{\sigma\sqrt{2n\log\log n}} &\to_{a.s.} -1. \end{aligned} \tag{1}$$

To gain further insight about the asymptotic behavior of partial sums, we can consider $S_{\lfloor nt \rfloor}$, $0 \le t \le 1$, as a random process. In 1964, Strassen proved that it can be approximated by a standard Brownian motion process (see ►Brownian Motion and Diffusions). A standard Brownian motion (or Wiener process) is a random process $\{W(t); t \ge 0\}$ that has stationary and independent increments, where the distribution of $W(t)$ is normal with mean 0 and variance t, for any fixed $t > 0$ and $W(0) = 0$.

Strassen showed that if $m = E(X_1)$ and $Var(X_1) = \sigma^2 < \infty$, then there exists a common probability space on which one can define a standard Brownian motion process W and a sequence of independent and identically distributed random variables $Y_1, Y_2, \ldots$ such that $\{S_n = \sum_{i=1}^n X_i : n \ge 1\} =_D \{\tilde{S}_n = \sum_{i=1}^n Y_i : n \ge 1\}$ and, as

$n \to \infty$,

$$\sup_{0\le t\le 1} \frac{|\sigma^{-1}(\tilde{S}_{\lfloor nt\rfloor} - m\lfloor nt\rfloor) - W(nt)|}{\sqrt{n\log\log n}} \to_{a.s.} 0, \qquad (2)$$

where $\lfloor nt\rfloor$ is the largest integer less than or equal nt.

Statement (2) is an example of a strong approximation which gives rise to the *strong invariance principle*. From it one can deduce the law of the iterated logarithm for partial sums (1) from that of standard Brownian motion (Khinchin's LIL). Alternately, one can prove it for a specific sequence of random variables, say simple coin tossing, and then, via (2), it is inherited by any independent sequence with a common distribution having finite variance.

If one assumes further conditions on the moments of the random variables (beyond finite variance) then the rate of convergence in (2) can be improved. In particular, if one assumes that X_1 has a finite moment generating function in an open interval containing the origin, then Komlós et al. (1975) have proven a Theorem 1-type result with convergence statement:

$$\limsup_{n\to\infty} \sup_{0\le t\le 1} \frac{|\sigma^{-1}(\tilde{S}_{\lfloor nt\rfloor} - m\lfloor nt\rfloor) - B(nt)|}{\log n} \le C, \ a.s. \quad (3)$$

for some constant $C > 0$.

Many almost-sure results including (3) are proven by first establishing an inequality for the maximal deviations and then applying a Borel-Cantelli lemma (see ▶Borel–Cantelli Lemma and Its Generalizations). The Komlós et al. inequality is:

$$P\left\{\max_{1\le k\le n} |\sigma^{-1}(\tilde{S}_k - mk) - B(k)| > c_1\log n + x\right\} < c_2 e^{-c_3 x},$$

where c_1, c_2, c_3 are positive constants depending only on the distribution of X_1. The Borel-Cantelli lemma to be used is: for any sequence of events $A_n, n \ge 1$, if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n$, infinitely often$) = 0$. Massart (1989) proved a multivariate version of (3).

The rate $\mathcal{O}(\log n)$ in (3) is the best rate possible. This is a consequence of the Erdös- Rényi laws of large numbers:

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with mean $E(X_1) = m$ and where the ▶moment generating function $M(t) = E(e^{t(X_1-m)})$ of $X_1 - m$ is finite in an interval containing $t = 0$. Then, for any $c > 0$,

$$\max_{0\le k\le n-\lfloor c\log n\rfloor} \frac{S_{k+\lfloor c\log n\rfloor} - S_k - m\lfloor c\log n\rfloor}{\lfloor c\log n\rfloor} \to_{a.s.} \alpha(c),$$

where $\alpha(c) = \sup\{x : \varrho(x) \ge e^{-1/c}\}$, with $\varrho(x) = \inf_t e^{-tx}M(t)$, the Chernoff function of $X_1 - m$.

If the left side of (3) converged to 0, almost surely, then $\sigma^{-1}(X_i - m)$ and $B(i) - B(i-1)$ would share the same function α. Since α uniquely determines the distribution function of a random variable, $\sigma^{-1}(X_i - m) =_D B(i) - B(i-1)$, a standard normal distribution.

Empirical process are important in many areas of statistics. If $X_1, X_2, \ldots$ is a sequence of independent k-dimensional random vectors with distribution function F, let $F_n(x) = n^{-1}\sum_{i=1}^{n} I[X_i \le x]$, $x \in R$, is the proportion of $X_1, X_2, \ldots, X_n$ that are less than or equal to the real vector $x = (x_1, \ldots, x_k)$ in the usual partial ordering of R^k. The empirical process is defined as

$$\alpha_n(x) = \sqrt{n}[F_n(x) - F(x)], \ \ x \in R^k.$$

Strong approximation results are available for the empirical process which describe its behavior in terms of both $x \in R^k$ and the sample size. A Kiefer process $K_F(x, y)$ is a Gaussian process defined on $R^k \times [0, \infty)$ that has mean zero and covariance function $E(K(x_1, y_1)K(x', y')) = (\min\{y_1, y'\})(F(x \wedge x') - F(x)F(x'))$, where $x \wedge x' = \left(\min\left\{x_1, x_1'\right\}, \ldots, \min\left\{x_k, x_k'\right\}\right)$.

In 1988, Csörgő and Horváth proved that there exists a common probability space on which one can define a Kiefer process K and a sequence of independent and identically distributed random variables $Y_1, Y_2, \ldots$ such that its empirical process $\{\tilde{\alpha}_n(x);\ x \in R^k, n = 1, 2, \ldots\} =_D \{\alpha_n(x);\ x \in R^k, n = 1, 2, \ldots\}$, the empirical process of the original sequence of X_i, and

$$\limsup_{n\to\infty} \max_{1\le j\le n} \sup_{x\in R^k} \frac{|\tilde{\alpha}_j(x) - j^{-1/2}K(x, j)|}{n^{-1/(4k)}(\log n)^{3/2}} \le C, \ \ a.s. \qquad (4)$$

When the dimension $k = 1$, the denominator in (4) can be improved to $n^{-1/2}(\log n)^2$. Similar to partial sums, the law of the iterated logarithm for the empirical process can be deduced from that of the Kiefer process, that is

$$\limsup_{n\to\infty} \sup_{x\in R^k} \frac{|\alpha_n(x)|}{\sqrt{\frac{1}{2}\log\log n}} =_{a.s.} 1.$$

Other results involve the strong approximation of the empirical process by a sequence of Brownian bridges B_n, where each is a Gaussian process defined on R^k and each has mean zero and covariance function $EB_n(x_1)B_n(x') = F(x \wedge x') - F(x)F(x')$. For general F, Borisov proved an approximation with rate $O(n^{-1/(2(2k-1))}\log n)$, a.s. When F has a density, Rio obtained a rate of $O(n^{-1/12}(\log n)^{(5k+1)/6})$, a.s. Here the exponent of n is independent of the dimension. When F is the uniform distribution on $[0, 1]^k$, Massart, in 1989, proved (4) with a rate of $O(n^{-1/(2(k+1))}(\log n)^2)$ and obtained an approximation in terms of sequences of Brownian bridges with a rate of $O(n^{-1/(2k)}(\log n)^{1/2})$.

About the Author

Murray David Burke is Professor of Mathematics and Statistics at the University of Calgary. He was Chair of the Division of Statistics and Actuarial Science (1988–1991, 1997–2001) and President of the Probability Section of the Statistical Society of Canada (2009–2010). He is an elected member of the International Statistical Institute.

Cross References

►Approximations to Distributions
►Borel–Cantelli Lemma and Its Generalizations
►Brownian Motion and Diffusions
►Convergence of Random Variables
►Empirical Processes
►Laws of Large Numbers
►Limit Theorems of Probability Theory

References and Further Reading

Borisov IS (1982) An approximation of empirical fields. In: Nonparametric statistical inference. Coll Math Soc János Bolyai, Budapest, Hungary, 1980, vol 32. North Holland, Amsterdam, 1982, pp 77–87

Csörgő M, Horváth L (1988) A note on strong approximations of multivariate empirical processes. Stoch Proc Appl 27:101–109

Csörgő M, Révész P (1981) Strong approximations in probability and statistics. Academic, New York

DasGupta A (2008) Asymptotic theory of statistics and probability. Springer, New York

Hartman P, Wintner A (1941) On the law of the iterated logarithm. Am J Math 63:169–176

Komlós J, Major P, Tusnády G (1975) An approximation of partial sums of independent r.v.'s and the sample df. I. Z Wahrcheinlichkeitstheorie verw Gebiete 32:111–131

Massart P (1989) Strong approximations for multivariate empirical and related processes, via KMT constructions. Ann Probab 17:266–291

Rio E (1996) Vitesses de convergence dans le principe d'invariance faible pour la fonction de répartition empirique multivarée. CR Acad Sci Paris t 322(1):169–172

Strassen V (1964) An invariance principle for the law of the iterated logarithm. Z Wahrcheinlichkeitstheorie verw Gebiete 3:211–226

Structural Equation Models

Scott L. Hershberger
Global Director of Survey Design
Harris Interactive, New York, NY, USA

Introduction

A *structural equation model* is a representation of a series of hypothesized relationships between observed variables and *latent variables* into a composite hypothesis concerning patterns of statistical dependencies. The hypothesized relationships are described by parameters that indicate the magnitude of the relationship (direct or indirect) that independent (*exogenous*) variables (either observed or latent) have on dependent (*endogenous*) variables (either observed or latent). By enabling the representation of hypothesized relationships into testable mathematical models, a structural equation model offers a comprehensive method for the quantification and testing of theoretical models. Once a theory has been proposed, it can be tested against empirical data.

The term *structural equation model* was first coined by econometricians and is probably the most appropriate name for the process just briefly sketched. *Path analysis*, developed by Sewall Wright (1921), is an early form of SEM that is restricted to observed variables. The exogenous observed variables are assumed to have been measured without error and have unidirectional (*recursive*) relations with one another. As it turns out, *path analysis rules* are still used today to identify the structural equations underlying the models. Using the path analysis approach, models are presented in the form of a drawing (often called a *path diagram*), and the structural equations of the model are inferred by reading the diagram correctly. However, the term *path analysis* implies too many restrictions on the form of the model. *Structural equation modeling* (SEM), on the other hand, has grown to incorporate latent and observed variables that can be measured with and without error and have bidirectional (*nonrecursive*) relationships among variables. Another term used frequently is *causal analysis*. Unfortunately, this is also a misleading term. Although SEM may appear to imply causality, the structural equations are not causal relations but functional relations. *Covariance structure modeling* is another popular term that is used mostly by psychologists. Unfortunately, it too is restrictive. Although the covariance structure of observed data is the most commonly modeled, SEM can be used to model other moments of the data. For example, mean structures are occasionally modeled, and facilities are provided for this in a number of SEM software programs. Modeling the third (skew) and fourth (kurtosis) moments of the data is also possible.

Mathematical Representation

To date, several mathematical models for SEM have been proposed. Although these mathematical models can translate data equally well into the model parameters, they differ in how parsimoniously this translation process is conducted. Perhaps the most well known of these mathematical models, the Keesling–Wiley–Jöreskog (*LISREL*) model, can require up to nine symbols in order to represent a model. In contrast, the *COSAN* model can generally represent the same model using only two symbols. Striking

a compromise between the LISREL and COSAN models, the Benter-Weeks (*EQS*) model can represent any model using only four symbols. Mathematically, the EQS model is represented by

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\gamma}\boldsymbol{\xi}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are coefficient matrices, and $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are vectors of random variables. The random variables within $\boldsymbol{\eta}$ are endogenous variables and the variables within $\boldsymbol{\xi}$ are exogenous variables. Endogenous and exogenous variables can be either latent or observed. The matrix $\boldsymbol{\beta}$ consists of coefficients (parameters) that describe the relations among the endogenous variables. The matrix $\boldsymbol{\xi}$ consists of coefficients (parameters) that describe the relations between exogenous and endogenous variables.

It is important to note that the primary interest in SEM centers on describing the network of relations among the variables (implying that one is generally interested in the covariance structure among the variables). Although the structural equation model is written in terms of equations lining the variables, the data used to solve the model parameters are actually covariances or correlations. In fact, this approach is no different from how many other multivariate statistical models are evaluated. For example, multiple regression uses a series of equations that link dependent to independent variables, but it is the correlational structure of the data that is used to solve for the regression coefficients. Similarly, in the EQS model, the sample covariance structure (C) among a set of variables x, y is defined as

$$C = (x + y)(x + y)' = \mathbf{J}(\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \boldsymbol{\beta})^{-1}\mathbf{J}'$$

where $\boldsymbol{\Gamma}$ is a matrix of coefficients linking exogenous ξ with endogenous η variables, $\boldsymbol{\beta}$ is a matrix of coefficients linking endogenous variables, and $\boldsymbol{\Phi}$ represents the covariances among the exogenous variables. The $\mathbf{J}$ matrix serves as a "filter" for selecting the observed variables from the total number of variables to be included in the model.

The Confirmatory Factor Analysis Model

A popular type of structural equation model is the *confirmatory factor analysis model*. In contrast to *exploratory factor analysis* (EFA), where all loadings are free to vary, confirmatory factor analysis (CFA) allows for the explicit constraint of certain loadings to be zero. As traditionally given, the confirmatory factor model in matrix notation is

$$\mathbf{Y} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a vector of scores on the observed variables, $\boldsymbol{\Lambda}$ is a *factor pattern loading matrix*, $\boldsymbol{\xi}$ is a matrix of *common factors*, and $\boldsymbol{\varepsilon}$ is a matrix of measurement errors in the observed variables. As such, the covariance structure implied by the confirmatory factor model is defined as

$$C = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

where C is the sample variance-covariance matrix, $\boldsymbol{\Phi}$ is a matrix of the factor variance-covariances, and $\boldsymbol{\Psi}$ is a variance-covariance matrix among the measurement errors.

In the EQS representation, the confirmatory factor model is generally expressed as

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\gamma}\boldsymbol{\xi} \text{ with } \boldsymbol{\beta} = 0$$

and the covariance structure implied by the model is given as

$$C(\eta\eta') = (0\eta + \gamma\xi)(0\eta + \gamma\xi)' = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$$

where the asymmetric relations in the model (the effects of the common and error factors on the observed variables) are in $\boldsymbol{\Gamma}$ and the symmetric relations (the factor and error variances and covariances) are in $\boldsymbol{\Phi}$. Note that for the confirmatory factor model the matrix $\boldsymbol{\beta}$ is dropped from the EQS model because in CFA there are no regression relations between endogenous variables.

Model Estimation

Model estimation proceeds by rewriting the structural equations so that each of the parameters of the equations is a function of the elements of the sample covariance matrix C. Subsequently, after obtaining values for the parameters, it one were to substitute these values back into the expression for the covariance structure implied by the model, the resulting sample matrix C can be represented as $\widehat{C}$. Clearly, $\widehat{C}$ should be very close to C because it was the elements of C that assisted in solving for the model parameters: The difference should be small if the model is consistent with the data.

The evaluation of $C - \widehat{C}$ depends on the estimation method used to solve for the model parameters. The most commonly used estimation methods for solving the parameters are *unweighted least squares* (ULS), *generalized* (*weighted*) *least squares* (GLS), and *maximum likelihood* (ML). With each estimation method, the structural equations are solved iteratively, until optimal estimates of the parameters are obtained. Optimal parameter values are values that imply covariances $(\widehat{C})$ close to the observed covariances (C). The difference $C - \widehat{C}$ is known as a *discrepancy function* (F). In order to minimize this discrepancy function, the partial derivatives of F are taken with respect to the elements of $C - \widehat{C}$. The form of the discrepancy function varies across the different estimation methods. However, the general form of this discrepancy function is

$$F = \sum_{ij} (C - \widehat{C})' W (C - \widehat{C})$$

in which a weighted sum of differences between the IJ elements of C and $\widehat{C}$ is calculated. As C and $\widehat{C}$ become more different, the discrepancy function becomes larger implying less correspondence between the model-implied covariances and the observed covariances. Most currently available SEM programs (e.g., SPSS' AMOS, EQS, LISREL, Mplus, Mx, the SEM package in R, SAS PROC CALIS) include ULS, GLS, and ML as standard estimation methods.

Model Assessment and Fit

For a model with positive *df* degrees of freedom, it is very unlikely that the discrepancy function will equal 0, implying a model with perfect fit to the data. Thus, there must be some measure of how large the discrepancy function must be in order to determine that the model does not fit the data. If multivariate normality is present, a *chi-square goodness-of-fit test* for the model is available using the sample size and the value of the discrepancy function

$$\chi^2 = (N-1)(F)$$

with *df* = (the number of unique elements of C) − (the number of parameters solved). If chi-square is not significant, then no significant discrepancy exists between the model-implied and observed covariance matrices. As such, the model fits the data and is confirmed. However, the chi-square test suffers from several weaknesses, including a dependence on sample size, and vulnerability to departures from multivariate normality. Thus, it is recommended that other descriptive fit criteria (e.g., ratio of χ^2 to *df*) and fit indices (e.g., the *comparative fit index*, the *root mean square error of approximation*) be examined in addition to the χ^2 value to assess the fit of the proposed model. Quite a few fit criteria and indices have been developed, each with its own strengths and weaknesses, and is it usually advisable to report a range of them.

Model Identification

Only identified models should be estimated. The process of *model identification* involves confirming that a unique numerical solution exists for each of the parameters of the model. Model identification should be distinguished from *empirical identification*, which involves assessing whether the rank of the *information matrix* is not deficit. Most SEM programs automatically check for empirical identification. On the other, model identification is not as easily or automatically assessed. For structural equation models in general, the most frequently invoked identification rules are the t-rule and the rank and order conditions. The *t*-rule is a simple rule to apply, but is only a necessary not a sufficient condition of identification. The *t*-rule is that the number of nonredundant elements in the covariance matrix of the observed variables (p) must be greater than or equal to the number of unknown parameters in the proposed model. Thus, if $t \leq p(p+1)/2$ the necessary condition of identification is met. Unfortunately, although the *t*-rule is simple to apply, it is only good for determining *underidentified* models. The order condition requires that for the model to be identified, the number of p variables excluded from each structural equation must equal $p - 1$. Unfortunately, the order condition is also a necessary bit not sufficient condition for identification. Only the rank condition is a necessary and sufficient condition for identification; however, it is not easy to apply. In general terms, the rank condition requires that the rank of any model matrices (e.g., $\mathbf{\Phi}, \boldsymbol{\beta}, \mathbf{\Gamma}$) be of at least rank $p-1$ for all submatrices formed by removing the parameter of interest. However, the usefulness of these criteria is doubtful because a failure to meet them does not necessarily mean the model is not identified. As it turns out, the only sure way to assess the identification status of a model prior to model fitting is to show through algebraic manipulation that each of the model parameters can be solved in terms of the p variances and $p(p-1)/2$ covariances.

Equivalent Structural Equation Models

Equivalent structural equation models may be defined as the set of models that, regardless of the data, yield identical (a) implied covariance, correlation, and other moment matrices when fit to the same data, which in turn imply identical (b) residuals and fitted moment matrices, (c) fit functions and chi-square values, and (d) goodness-of-fit indices based on fit functions and chi-square. One most frequently thinks of equivalent models as described in (a) above. To be precise, consider two alternative models, denoted $M1$ and $M2$, each of which is associated with a set of estimated parameters and a covariance implied by those parameter estimates (denoted as $\widehat{C}_{M1}$ and $\widehat{C}_{M2}$). Models $M1$ and $M2$ are considered equivalent if, for any sample covariance matrix C, the implied matrices $\widehat{C}_{M1} = \widehat{C}_{M2}$ or alternatively, $(C - \widehat{C}_{M1}) = (C - \widehat{C}_{M2})$. Because of this equivalence, the values of statistical tests of fit that are based on the discrepancy between the sample covariance matrix and the model-implied covariance matrix will be identical. Thus, even when a hypothesized model fits well according to multiple fit indices, there may be equivalent models with identical fit – even if the theoretical implications of those models are very different. However, model equivalence is not unique to SEM. For example, in *exploratory factor analysis*, without the arbitrary constraint of extracting orthogonal factors in decreasing order of magnitude,

there would potentially be an infinite number of equivalent initial solutions.

About the Author

Scott L. Hershberger, Ph.D. is formerly Quantitative Professor of Psychology at the California State University, Long Beach and is now Global Director of Survey Design at Harris Interactive. He is a past Associate editor of the journal, *Structural Equation Modeling*, and is an elected member of the Royal Statistical Society and the International Statistical Institute. He has authored or co-authored numerous articles and several books on multivariate analysis and psychometrics.

Cross References

►Causal Diagrams
►Causation and Causal Inference
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Tests
►Factor Analysis and Latent Variable Modelling
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Psychiatry, Statistics in
►Sociology, Statistics in

References and Further Reading

Bentler P (1995) EQS program manual. Multivariate Software, Encino
Bollen KA (1989) Structural equation models with latent variables. Wiley, New York
Bollen KA, Long JS (eds) (1993) Testing structural equation models. Sage, Newbury Park
Hoyle RH (ed) (1995) Structural equation modeling: concepts, issues, and applications. Sage, Thousand Oaks
Raykov T, Marcoulides GA(2006) A first course in structural equation modeling, 2nd edn. Lawrence Erlbaum Associates, Mahwah
Wright S (1921) Correlation and causation. J Agr Res 20:557–585

Structural Time Series Models

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán, San Miguel de Tucumán, Argentina

Introduction

The basic idea of structural time series models is that they are set up as regression models in which the explanatory variables are functions of time with coefficients which change over time. Thus within a regression framework a simple trend would be modeled in terms of a constant and a time with a random disturbance added on, that is

$$y_t = \alpha + \beta t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{1}$$

This model is easy to estimate using ordinary ►least squares, but suffers from the disadvantage that the trend is deterministic. In general, this is too restrictive, however, the necessary flexibility is introduced by letting the coefficients α and β evolve over time as stochastic processes. In this way the trend can adapt to underlying changes. The current, or *filtered*, estimate of the trend is estimated by putting the model in state space form and applying the Kalman filter. Related algorithms are used for making *predictions* and for *smoothing*, which means computing the best estimate of the trend at all points in the sample using the full set of observations. The extent to which the parameters are allowed to change is governed by *hyperparameters*. These can be estimated by maximum likelihood but, again, the key to this is the state space form and the Kalman filter. The STAMP package of Koopman et al. (2000) carries out all the calculations and is set up so as to leave the user free to concentrate on choosing a suitable model.

An excellent general presentation of the Kalman filter is given in this Encyclopedia by M. S. Grewal under the title *Kalman Filtering*. We give below a set of particular results about the filter that are for application within the areas covered by Time Series and Econometric. Similarly, a general presentation of smoothing is given as well in this Encyclopedia by A.W. Bowman under the title *Smoothing Techniques*. We recall that in our context smoothing means computing the best estimates based on the full sample, therefore we give below a set of particular results that are for application within the areas covered by Time Series and Econometric.

The classical approach to time series modeling is based on the fact that a general model for any indeterministic stationary series is the autoregressive-moving average of order (p, q). This is usually referred to as ARMA(p, q). The modeling strategy consists of first specifying suitable values of p and q on the basis of an analysis of the correlogram and other relevant statistics. The model is then estimated, usually under the assumption that the disturbance is Gaussian. The residuals are then examined to see if they appear to be random, and various test statistics are computed. In particular, the Box–Ljung Q-statistic, which is based on the first P residual autocorrelations, is used to test for residual serial correlation. Box and Jenkins (1976) refer to these stages as identification, estimation and diagnostic checking. If the diagnostic checks are satisfactory, the model is ready to be used for forecasting. If they are not, another specification must be tried. Box and Jenkins stress the role

of parsimony in selecting p and q to be small. However, it is sometimes argued, particularly in econometrics, that a less parsimonious pure autoregressive (AR) model is often to be preferred as it is easier to handle.

Many series are not stationary. In order to handle such situations Box and Jenkins proposed that a series be differenced to make it stationary. After fitting an ARMA model to the differenced series, the corresponding integrated model is used for forecasting. If the series is differenced d times, the overall model is called ARIMA(p, d, q). Seasonal effects can be captured by seasonal differencing.

The model selection methodology for structural models is somewhat different in that there is less emphasis on looking at the correlograms of various transformations of the series in order to get an initial specification. This is not to say that correlograms should never be examined, but the experience is that they can be difficult to interpret without prior knowledge of the nature of the series and in small samples and/or with messy data they can be misleading. Instead the emphasis is on formulating the model in terms of components which knowledge of the application or an inspection of the graph suggests might be present. For example, with monthly observations, one would probably wish to build a seasonal pattern into the model at the outset and only drop it if it proved to be insignificant. Once a model has been estimated, the same type of diagnostics tests as are used for ARIMA models can be performed on the residuals. In particular the Box–Ljung statistic can be computed, with the number of relative hyperparameters subtracted from the number of residual autocorrelations to allow for the loss of degrees of freedom. Standard tests for non-normality and heteroscedasticity can also be carried out, as can tests of predictive performance in a post-sample period. Plots of residuals should be examined, a point which Box and Jenkins stress for ARIMA model building. In a structural time series model, such plots can be augmented by graphs of the smoothed components. These can often be very informative since it enables the model builder to check whether the movements in the components correspond to what might be expected on the basis of prior knowledge.

State Space Form, Kalman Filtering and Smoothing

As we say before, a structural time series model is one in which the trend, seasonal and error terms in the basic model, plus other relevant components, are modeled explicitly. This is in sharp contrast to the philosophy underlying ARIMA models where trend and seasonal are removed by differencing prior to detailed analysis.

The statistical treatment of the structural time series models is based on the state space form, the Kalman filter and the associated smoother. The likelihood is constructed from the Kalman filter in terms of the one-step ahead prediction errors and maximized with respect to the hyperparameters by numerical optimization. The score vector for the parameters can be obtained via a smoothing algorithm which is associated with the Kalman filter. Once the hyperparameters have been estimated, the filter is used to produced one-step ahead predictions residuals which enables us to compute diagnostic statistics for normality, serial correlation and goodness of fit. The smoother is used to estimate unobserved components, such as trends and seasonals, and to compute diagnostic statistics for detecting ►outliers and structural breaks. ARIMA models can also be handled using the Kalman filter. The state space approach becomes particularly attractive when the data are subject to missing values or temporal aggregation.

State Space Form

All linear time series have a state space representation. This representation relates the disturbance vector $\{\boldsymbol{\varepsilon}_t\}$ to the observation vector $\{\mathbf{y}_t\}$ via a Markov process (see ►Markov Processes) $\{\boldsymbol{\alpha}_t\}$. A convenient expression of the state space form is

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim N(\mathbf{0}, \mathbf{H}_t), \\ \boldsymbol{\alpha}_t &= \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{R}_t\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(\mathbf{0}, \mathbf{Q}_t), \qquad t = 1, \ldots n, \end{aligned} \tag{2}$$

where $\mathbf{y}_t$ is a $p \times 1$ vector of observations and $\boldsymbol{\alpha}_t$ is an unobserved $m \times 1$ vector called the *state vector*. The idea underlying the model is that the development of the system over time is determined by $\boldsymbol{\alpha}_t$ according to the second equation of (2), but because $\boldsymbol{\alpha}_t$ cannot be observed directly we must base the analysis on observations $\mathbf{y}_t$. The first equation of (2) is called the *measurement equation*, and the second one, the *transition equation*. The system matrices $\mathbf{Z}_t$, $\mathbf{T}_t$ and $\mathbf{R}_t$ have dimensions $p \times m$, $m \times m$ and $m \times g$ respectively. The disturbance terms $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are assume to be serially independent and independent of each other at all time points. The matrix $\mathbf{H}_t$ has dimension $p \times p$ with rank p, and the matrix $\mathbf{Q}_t$ has dimension $g \times g$ with rank $g \leq m$. The matrices $\mathbf{Z}_t, \mathbf{T}_t, \mathbf{R}_t, \mathbf{H}_t$ and $\mathbf{Q}_t$ are fixed and their unknown elements, if any, are place in the hyperparameter vector $\boldsymbol{\psi}$ which can be estimated by maximum likelihood. In univariate time series $p = 1$, so $\mathbf{Z}_t$ is a row vector.

The initial state vector $\boldsymbol{\alpha}_0$ is assumed to be $N(\mathbf{a}_0, \mathbf{P}_0)$ where $\mathbf{a}_0$ and $\mathbf{P}_0$ are known. When $\mathbf{a}_0$ and $\mathbf{P}_0$ are unknown, $\boldsymbol{\alpha}_0$ is taken as diffuse. An adequate approximation can

often be achieved numerically by taking $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \kappa \mathbf{I}_m$, where κ is a scalar which tends to infinity.

Kalman Filter

In the Gaussian state space model (2), the Kalman filter evaluate the minimum mean squared error estimator of the state vector $\boldsymbol{\alpha}_{t+1}$ using the set of observations $\mathbf{Y}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$, denoted $\mathbf{a}_{t+1} = E(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)$, and the corresponding variance matrix $\mathbf{P}_{t+1} = \mathrm{Var}(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)$, for all t. This means that the Kalman filter allows to continuously update the estimation of the state vector whenever a new observation is available. Since all distributions are normal, conditional distributions are also normal. Let $\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_t$, then $\mathbf{v}_t$ is the one-step ahead forecast error $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{Y}_{t-1})$. Demote its variance matrix by $\mathbf{F}_t$. Then

$$\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_t\mathbf{Z}_t' + \mathbf{H}_t, \quad t = 1, \ldots n. \tag{3}$$

It is possible to show that the updating recursion is given by

$$\mathbf{a}_{t+1} = \mathbf{T}_{t+1}\mathbf{a}_t + \mathbf{K}_t\mathbf{v}_t, \tag{4}$$

where

$$\mathbf{K}_t = \mathbf{T}_{t+1}\mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}, \tag{5}$$

and

$$\mathbf{P}_{t+1} = \mathbf{T}_{t+1}\mathbf{P}_t\left(\mathbf{T}_{t+1}' - \mathbf{Z}_t'\mathbf{K}_t'\right) + \mathbf{R}_{t+1}\mathbf{Q}_{t+1}\mathbf{R}_{t+1}', \tag{6}$$

for $t = 0, 1, \ldots, n-1$, with $\mathbf{K}_0 = \mathbf{0}$.

The set (3) to (6) constitute the Kalman filter for model (2). The derivation of the Kalman recursions can be found in Anderson and Moore (1979), Harvey (1989), Abril (1999) and Durbin and Koopman (2001).

The output of the Kalman filter is used to compute the log-likelihood function $\log L(\mathbf{y}_t, \boldsymbol{\psi})$, conditional on the hyperparameter vector $\boldsymbol{\psi}$, as given by

$$\log L(\mathbf{y}_t, \boldsymbol{\psi}) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{n}\mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t, \tag{7}$$

apart from a possible constant. Numerical maximization of (7) with respect to the hyperparameter vector $\boldsymbol{\psi}$ yields the maximum likelihood estimator $\widetilde{\boldsymbol{\psi}}$. Usually (7) is called the *prediction error decomposition* of the likelihood.

Smoothing

The work of de Jong (1988, 1989), Kohn and Ansley (1989) and Koopman (1993) leads to a smoothing algorithm from which different estimators can be computed based on the full sample $\mathbf{Y}_n$. Smoothing takes the form of a backwards recursion

$$\mathbf{u}_t = \mathbf{F}_t^{-1}\mathbf{v}_t - \mathbf{K}_t'\mathbf{r}_t, \qquad \mathbf{M}_t = \mathbf{F}_t^{-1} + \mathbf{K}_t'\mathbf{N}_t\mathbf{K}_t,$$
$$\mathbf{r}_{t-1} = \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{L}_t'\mathbf{r}_t, \quad \mathbf{N}_{t-1} = \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t + \mathbf{L}_t'\mathbf{N}_t\mathbf{L}_t, \tag{8}$$

for $t = n, n-1, \ldots, 1$, where $\mathbf{L}_t = \mathbf{T}_{t+1} - \mathbf{K}_t\mathbf{Z}_t$, $\mathbf{r}_n = \mathbf{0}$ and $\mathbf{N}_n = \mathbf{0}$. The recursions require memory space for storing the Kalman output $\mathbf{v}_t$, $\mathbf{F}_t$ and $\mathbf{K}_t$ for $t = 1, \ldots, n$. The series $\{\mathbf{u}_t\}$ will be referred to as *smoothing errors*. The smoothing quantities $\mathbf{u}_t$ and $\mathbf{r}_t$ play a pivotal role in the construction of diagnostic tests for outliers and structural breaks. The smoother can be used to compute the smoothed estimator of the disturbance vector $\widehat{\boldsymbol{\varepsilon}}_t = E(\boldsymbol{\varepsilon}_t|\mathbf{Y}_n)$. The smoothed estimator of the state vector $\widehat{\boldsymbol{\alpha}}_t = E(\boldsymbol{\alpha}_t|\mathbf{Y}_n)$ is constructed as follows

$$\widehat{\boldsymbol{\alpha}}_t = \mathbf{a}_t + \mathbf{P}_t\mathbf{r}_{t-1}, \tag{9}$$

for $t = 1, \ldots, n$, where $\mathbf{r}_t$ satisfies the backwards recursions given in (8).

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadística* (Journal of the Argentinean Statistical Society).

Cross References

- ▶Autocorrelation in Regression
- ▶Box–Jenkins Time Series Models
- ▶Forecasting with ARIMA Processes
- ▶Kalman Filtering
- ▶Markov Processes
- ▶Model Selection
- ▶Residuals
- ▶Time Series
- ▶Trend Estimation

References and Further Reading

Abril JC (1999) Análisis de Series de Tiempo Basado en Modelos de Espacio de Estado. EUDEBA, Buenos Aires

Anderson BDO, Moore JB (1979) Optimal filtering. Prentice-Hall, Englewood Cliffs, New Jersey

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control (revised edition), Holden-Day, San Francisco

de Jong P (1988) A cross-validation filter for time series models. Biometrika 75:594–600

de Jong P (1989) Smoothing and interpolation with the state-space model. J Am Stat Assoc 84:1085–1088

Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford

Harvey AC (1989) Forecasting, structural time series models and the kalman filter. Cambridge University Press, Cambridge

Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME, J Basic Eng 83D:35–45

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction problems. Trans ASME, J Basic Eng 83D:95–108
Kohn R, Ansley CF (1989) A fast algorithm for signal extraction, influence and cross-validation in state space models. Biometrika 76:65–79
Koopman SJ (1993) Disturbance smoother for state space models. Biometrika 80:117–126
Koopman SJ, Harvey AC, Doornik JA, Shephard N (2000) STAMP: structural time series analyser, modeller and predictor. Timberlake Consultant Ltd, London

Student's *t*-Distribution

Bronius Grigelionis
Professor, Head of the Mathematical Statistics Department
Institute of Mathematics and Informatics, Vilnius, Lithuania

We say that a random variable X has a Student t distribution with $\nu > 0$ degrees of freedom, a scaling parameter $\delta > 0$ and a location parameter $\mu \in R^1$, denoted $T(\nu, \delta, \mu)$, if its probability density function (pdf) is

$$f_X(x) = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\pi}\delta\Gamma\left(\frac{1}{2}\nu\right)}\left[1+\left(\frac{x-\mu}{\delta}\right)^2\right]^{-\frac{\nu+1}{2}}, \quad x \in R^1,$$

where $\Gamma(z)$ is the Euler's gamma function. $T(1, \delta, \mu)$ is the Cauchy distribution. $T(\nu, \delta, \mu)$ is heavy tailed and for an integer r

$$E(X-\mu)^{2r} = \begin{cases} \frac{\delta^{2r-1}\nu^r\Gamma\left(\frac{r}{2}+1\right)\Gamma\left(\frac{\nu}{2}-r\right)}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\nu\right)}, & \text{if } 2r < \nu, \\ +\infty, & \text{if } 2r \geq \nu. \end{cases}$$

Because

$$f_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi y}} e^{-\frac{(x-\mu)^2}{2y}} g(y)\,dy, \quad x \in R^1,$$

where

$$g(y) = \frac{\left(\frac{1}{2}\delta^2\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{1}{2}\nu\right)} y^{-\frac{\nu}{2}-1} e^{-\frac{\delta^2}{2y}}\,dy, \quad y > 0$$

is pdf of the inverse (reciprocal) gamma distribution, which is a member of the Thorin class, the Student t distribution is a marginal distribution of a Thorin subordinated Gaussian Lévy process (see, e.g., Grigelionis, 2007 and references therein). This property implies that $T(\nu, \delta, \mu)$ is self-decomposable, i.e., for every $c \in (0,1)$, there exists a random variable X_c, independent of X, such that $X \overset{\text{law}}{=} cX + X_c$, and therefore $T(\nu, \delta, \mu)$ is infinitely divisible. Self-decomposability of $T(\nu, \delta, \mu)$ permits to construct several classes of stationary stochastic processes with marginal Student t distributions and various types of dependence structure, relevant for modeling of economic and financial time series. In the fields of finance Lévy processes with marginal Student t distributions can often be fitted extremely well to model distributions of logarithmic asset returns (see Heyde and Leonenko, 2005).

The classical Student t distribution was introduced in 1908 by W.S. Gosset ("Student"), proving that the distribution law $\mathcal{L}(t_n) = T\left(n-1, \sqrt{n-1}, 0\right)$, where

$$t_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}, \quad n \geq 2,$$

$X_1, \ldots, X_n$ are independent normally distributed random variables, $\mathcal{L}(X_i) = N(\mu, \sigma^2)$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Properties of the classical Student t distributions are surveyed in Johnson, Kotz, 1970.

During last century the theory of Student t statistics has evolved into the theory of general Studentized statistics and self-normalized processes, and the Student t distribution was generalized to the multivariate case, leading to multivariate processes with matrix self-normalization (see de la Peña et al., 2009).

We say that a random d-dimensional vector X has a Student t distribution with $\nu > 0$ degrees of freedom, a symmetric positive definite scaling $d \times d$ matrix Σ and a location vector $\mu \in R^d$, denoted $T_d(\nu, \Sigma, \mu)$, if its pdf is

$$f_X(x) = \frac{\Gamma\left(\frac{1}{2}(\nu+d)\right)}{(\nu\pi)^{d/2}\Gamma\left(\frac{1}{2}\nu\right)|\Sigma|^{1/2}} \times \left(1+\frac{\left((x-\mu)\Sigma^{-1}, x-\mu\right)}{\nu}\right)^{-\frac{\nu+d}{2}}, \quad x \in R^d,$$

where $(x, y) = \sum_{i=1}^d x_i y_i$, $x, y \in R^d$, $|\Sigma| := \det\Sigma$ (see Johnson and Kotz, 1972).

We have that

$$Ee^{i(z,X)} = \frac{e^{i(\mu,z)}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{1}{2}\nu\right)} \times \left(\nu(z\Sigma, z)\right)^{\frac{\nu}{4}} K_{\frac{\nu}{2}}\left(\sqrt{\nu(z\Sigma, z)}\right), \quad z \in R^d,$$

where K_ν is the modified Bessel function of the third kind, i.e.,

$$K_\nu(x) = \frac{1}{2}\int_0^\infty u^{-\nu-1}\exp\left\{-\frac{1}{2}x(u+u^{-1})\right\}du, \quad x > 0,\ \nu \in R^1,$$

implying that for $c \in R^d$, $c \neq 0$, $\mathcal{L}((c,X)) = T\left(\nu, \sqrt{\nu\,(c\Sigma,c)},\ (c,\mu)\right)$, which means that $T_d(\nu,\sum,\mu)$ is marginal self-decomposable (see, Barndorff–Nielsen and Pérez-Abreu, 2002).

If $\nu > d+1$, $EX = \mu$ and $E(c_1,X-\mu)(c_2,X-\mu) = \nu(c_1\Sigma^{-1},c_2)\Gamma\left(\frac{\nu-d-1}{2}\right)$, $c_1,c_2 \in R^d$.

As $\nu \to \infty$, $T_d(\nu,\Sigma,\mu) \Rightarrow N_d(\mu,\Sigma)$ and, in particular, $T\left(\nu,\sqrt{\nu}\sigma,\mu\right) \Rightarrow N(\mu,\sigma^2)$, where "$\Rightarrow$" means weak convergence of probability laws.

Let M_d be an Euclidean space of symmetric $d\times d$ matrices with the scalar product $\langle A_1,A_2\rangle := \mathrm{tr}(A_1A_2)$, $A_1,A_2 \in M_d$, $M_d^+ \subset M_d$ be the cone of non-negative definite matrices, $\mathcal{P}\left(M_d^+\right)$ be the class of probability distributions on M_d^+.

Since

$$Ee^{i(z,X)} = e^{i(z,\mu)} \int\limits_{M_d^+} e^{-\frac{1}{2}(zA,z)}\,U(\,dA),$$

where

$$\phi_U(\Theta) := \int\limits_{M_d^+} e^{-\mathrm{tr}(\Theta A)}\,U(\,dA)$$

$$= \frac{[2\nu\mathrm{tr}(\Sigma\,\Theta)]^{\frac{\nu}{4}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{1}{2}\nu\right)} K_{\frac{\nu}{2}}\left(\sqrt{2\nu\mathrm{tr}(\Sigma\,\Theta)}\right),$$

$$\Theta \in M_d^+,\ U \in \mathcal{P}\left(M_d^+\right),$$

$\mathcal{L}(X-\mu)$ is a U-mixture of centered Gaussian distributions (see Grigelionis, 2009).

If $\nu \geq d$ is an integer, $U = \mathcal{L}\left(\nu W_\nu^{-1}\right)$, where $W_\nu = \sum_{i=1}^{\nu} Y_i^T Y_i$, $Y_1,\ldots,Y_\nu$ are independent d-dimensional centered Gaussian vectors with the covariance matrix Σ, z^T is the transposed vector z, i.e., U is the inverse Wishart distribution.

About the Author

Bronius Grigelionis graduated from the Department of Physics and Mathematics, Vilnius University in 1959. He was a postgraduate student at the Kiev University (1959–1960) and Moscow University (1960–1962) supervised by Prof. B.V. Gnedenko. He earned a doctor's (Ph.D.) in 1963 and the degree of Doctor habilius in 1969 at the Vilnius University. He was a senior Research Fellow at the Institute of Physics and Mathematics and a lecturer of the Vilnius University in 1963–1970. Since 1970 he has been Head of the Mathematical Statistics Department at the Institute of Mathematics and Informatics and Professor of Vilnius University. He is a member of the Lithuanian Mathematics Society, Lithuanian Academy of Sciences, International Statistical Institute, Bernoulli Society, Lithuanian Catholic Academy of Sciences. He has supervised 19 Ph.D. students.

Cross References

►Confidence Interval
►Correlation Coefficient
►Financial Return Distributions
►Heteroscedastic Time Series
►Hotelling's T^2 Statistic
►Multivariate Statistical Distributions
►Regression Models with Symmetrical Errors
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistical Distributions: An Overview
►Student's *t*-Tests

References and Further Reading

Barndorff-Nielsen OE (2002) Pérez-Abreu V (2002). Extensions of type G and marginal infinite divisibility. Teor Veroyatnost i Primenen 47(2):301–319

de la Peña VH, Lai TL, Shao QM (2009) Self-normalized processes: limit theory and statistical applications. Springer, Berlin

Grigelionis B (2007) On subordinated multivariate Gaussian Lévy processes. Acta Appl Math 96:233–246

Grigelionis B (2009) On the Wick theorem for mixtures of centered Gaussian distributions. Lith Math J 49(4):372–380

Heyde CC, Leonenko NN (2005) Student processes. Adv Appl Prob 37:342–365

Johnson NL, Kotz S (1970) Distributions in statistics: continuous univariate distributions vol 2. Wiley, New York

Johnson NL, Kotz S (1972) Distributions in statistics: continuous multivariate distributions. Wiley, New York

Student (1908) On the probable error of mean. Biometrika 6:1–25

Student's *t*-Tests

Damir Kalpić[1], Nikica Hlupić[2], Miodrag Lovrić[3]
[1]Professor and Head, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[3]Professor, Faculty of Economics
University of Kragujevac, Kragujevac, Serbia

Introduction

Student's *t*-tests are parametric tests based on the Student's or *t*-distribution. Student's distribution is named in honor of William Sealy Gosset (1876–1937), who first determined it in 1908. Gosset, "one of the most original minds in contemporary science" (Fisher 1939), was one of the best Oxford graduates in chemistry and mathematics in his generation. In 1899, he took up a job as a brewer

at Arthur Guinness Son & Co, Ltd in Dublin, Ireland. Working for the Guinness brewery, he was interested in quality control based on small samples in various stages of the production process. Since Guinness prohibited its employees from publishing any papers to prevent disclosure of confidential information, Gosset had published his work under the pseudonym "Student" (the other possible pseudonym he was offered by the managing director La Touche was "Pupil," see Box 1987, p. 49), and his identity was not known for some time after the publication of his most famous achievements, so the distribution was named Student's or t-distribution, leaving his name less well known than his important results in statistics. His, now, famous paper "The Probable Error of a Mean" published in *Biometrika* in 1908, where he introduced the t-test (initially he called it the z-test), was essentially ignored by most statisticians for more than 2 decades, since the "statistical community" was not interested in small samples ("only naughty brewers take n so small," Karl Pearson writing to Gosset, September 17, 1912, quoted by E.S. Pearson 1939, p. 218). It was only R. Fisher who appreciated the importance of Gosset's small-sample work, and who reconfigured and extended it to two independent samples, correlation and regression, and provided correct number of degrees of freedom. "It took the genius and drive of a Fisher to give Student's work general currency" (Zabel 2008, p. 6); "The importance of 1908 article is due to what Fisher found there, not what Gosset placed there" (Aldrich 2008, p. 11).

One-Sample t-Test

In the simplest form, also called the one-sample t-test, Student's t-test is used for testing a statistical hypothesis (Miller and Miller 1999) about the mean μ of a normal population whose variance σ^2 is unknown and sample size n is relatively small ($n \leq 30$). For a comparison of means of two independent univariate normal populations with equal (but unknown) variances we use two-sample t-test, and both of these tests have their multivariate counterparts based on multivariate extension of the t-variable called Hotelling's T^2 statistic ▶Hotelling's T^2 statistic (Johnson and Wichern 2007). Student's t-test also serves as the basis for the analysis of dependent samples (populations) in paired difference t-test or repeated measures design, in both univariate (Bhattacharyya and Johnson 1977) and multivariate cases (Johnson and Wichern 2007).

To understand the motivation for Student's t-test, suppose that we have at our disposal a relatively large sample of size $n > 30$ from a normal population with unknown mean μ and known variance σ^2. What we want is to determine the mean μ, i.e., to test our supposition (null hypothesis) $H_0 : \mu = \mu_0$ against one of the alternative hypotheses $\mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$. Maximum likelihood principle (method) (Hogg et al. 2005, or Anderson 2003) leads to the sample mean $\overline{X}$ as the test statistic, and it is known that $\overline{X}$ has Gaussian or normal distribution with mean μ and variance σ^2/n. Hence, we might calculate (provided σ^2) the probability of observing $\overline{x}$ in a certain range under the assumption of the supposed distribution $N(\mu_0, \sigma^2/n)$ and thereby assess our supposition about the unknown μ. Yet, this would require (numerical) evaluation of the integral of normal density for every particular pair (μ_0, σ^2) and, therefore, we construct the universal standard normal variable or z-score

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{1}$$

which, in our example, represents the distance from the observed $\overline{X}$ to the hypothesized population mean μ_0, expressed in terms (units) of standard deviation $\sigma/\sqrt{n}$ of $\overline{X}$. Thus, variable Z is an independent parameter and it has a standard normal distribution that has been extensively tabulated and is readily available in statistical books and software. The test itself is now based on Z as the test statistic and the rationale behind the test is that if the null hypothesis is true, then the larger the distance from $\overline{x}$ to μ_0 (larger $|z|$-value), the smaller the probability of observing such an $\overline{x}$. Therefore, given a level of significance α, we reject H_0 if $|z| \geq z_{\alpha/2}$, $z \geq z_\alpha$ or $z \leq -z_\alpha$, respectively, where z_α is the Z-value corresponding to the probability α for a random variable having standard normal distribution to take a value greater than z_α, i.e., $P(z \geq z_\alpha) = \alpha$. By virtue of the central limit theorem (Anderson 2003) and provided that the sample is large enough ($n > 30$), we apply the same test even though the population distribution cannot be assumed to be normal, the only precondition being that the variance is known. Of course, in real applications we rarely know exact population variance σ^2, so we substitute sample variance S^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2}$$

for σ^2 and likelihood ratio test statistic (1) becomes Students's t-variable

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}. \tag{3}$$

Having at our disposal a sufficiently large sample ($n > 30$), we consider s to be a "faithful" estimate of σ and we might still apply the same test, i.e., compare t with z_α values. This would then be only an approximate large-sample test, but

its result would likely correspond to the real truth. However, when population variance σ^2 is not known and the sample size is relatively small ($n \leq 30$), the test we have been discussing is not reliable anymore because t in (3) is not a faithful approximation of z in (1), as a direct consequence of the fact that sample variance S^2 determined from too small a sample does not approximate σ^2 well. Construction of a reliable test under such conditions requires knowledge of the exact distribution of variable T in (3), and due to Gosset, we know that it is a t-distribution with $n-1$ degrees of freedom. The same as with z-test, the rationale behind the t-test is that if the null hypothesis is true, then observing $\bar{x}$ too much distant from μ_0 is not likely. Specifically, for a given level of significance α and one of the alternatives $\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$, following the Neyman–Pearson approach, we calculate the critical value $t_{n-1}(\alpha/2)$ or $t_{n-1}(\alpha)$ defined by $P(t \geq t_{n-1}(\alpha)) = \alpha$, i.e., $t_{n-1}(\alpha)$ is the value corresponding to probability α for a random variable having t-distribution to take a value greater than $t_{n-1}(\alpha)$, and

$$\begin{aligned}\text{reject } H_0 \text{ if } |t| &\geq t_{n-1}(\alpha/2) \quad \text{with the alternative hypothesis } \mu \neq \mu_0,\\ t &\geq t_{n-1}(\alpha) \quad \text{with the alternative hypothesis } \mu > \mu_0,\\ t &\leq -t_{n-1}(\alpha) \quad \text{with the alternative hypothesis } \mu < \mu_0.\end{aligned} \tag{4}$$

Statistical tests imply *reject–do not reject* results, but it is usually more informative to express conclusions in the form of confidence intervals. In the case of the two-sided t-test ($H_1 : \mu \neq \mu_0$) constructed from a random sample of size n, $(1-\alpha)100\%$ confidence interval for the mean of a normal population is

$$\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}. \tag{5}$$

Two-Sample *t*-Test

When we compare parameters of two populations (means, variances, or proportions), we need to distinguish two cases: samples may be independent or dependent according to how they were selected. Two random samples are *independent* if the sample selected from one population is not related in any way to the sample from the other population. However, if the random samples are chosen in such a way that each measurement in one sample can be naturally or by design paired or matched with a measurement in the other sample, then the samples are called *dependent*. Dependent samples occur in two situations:

(a) *Repeated measures design*, when the same subject or unit is measured twice, *before and after* a treatment (e.g., the blood pressure of each subject in the study is recorded twice, before and after a drug is administered)

(b) *Matched pairs design*, when subjects are *matched* as closely as possible, and then one of each pair is randomly assigned to each of the treatment group and control group (see ▶Research Designs).

Two Independent Samples

(a) *Equal variances* $\sigma_1^2 = \sigma_2^2 = \sigma^2$

This is a simpler situation because variances of considered populations, though unknown, are equal. With the respective sample sizes being n_1 and n_2, maximum likelihood principle yields a test based on test statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{6}$$

where S_p^2 is the pooled estimator of common variance σ^2 given by

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}. \tag{7}$$

The pooled t-test is based on the fact that variable T in (6) has Student's distribution with $n_1 + n_2 - 2$ degrees of freedom, i.e., $P(t \geq t_{n_1+n_2-2}(\alpha)) = \alpha$. Hence, for instance, we reject the null hypothesis that both population means are equal ($H_0 : \mu_1 = \mu_2$) if $|t| \geq t_{n_1+n_2-2}(\alpha/2)$.

(b) *Unequal variances* $\sigma_1^2 \neq \sigma_2^2$

When the assumption of equal variances is untenable, we are confronted with what is known as ▶Behrens–Fisher problem, which is still an open challenge. There are, however, approximate solutions and a commonly accepted technique is Welch's t-test, also referred to as Welch–Aspin, Welch–Satterthwaite, or Smith–Satterthwaite test (Winer 1971; Johnson 2005). The test statistic is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8}$$

and it has approximately t-distribution with degrees of freedom estimated as

$$\nu = \frac{(g_1 + g_2)^2}{g_1^2/(n_1-1) + g_2^2/(n_2-1)}; \quad g_i = \frac{s_i^2}{n_i}. \tag{9}$$

The difference between the denominators in (6) and (8) should be noticed; in (6) we have the *estimate of the common variance*, while in (8) we have the *estimate of variance of the difference.*

The test procedure is to calculate the value t of the test statistics given by (8) and degrees of freedom ν according to (9) (if ν is not an integer we round it down rather than up in order to take a conservative approach). Then, given the level of significance α, we use the obtained ν and Student's distribution to calculate critical value $t_\nu(\alpha)$ and draw conclusions comparing t and $t_\nu(\alpha)$ like in an ordinary one-sample t-test.

Two Dependent Samples

The test procedure is essentially the same as for one-sample t-test, the only difference being that we enter (3) with the mean and standard deviation of paired differences instead of with the original data. Number of degrees of freedom is $n - 1$, where n is the number of the observed differences (number of pairs). This test is based on the assumption that the population of paired differences follows normal distribution.

Robustness of *t*-Test

Since the t-test requires certain assumptions in order to be exact, it is of interest to know how strongly the underlying assumptions can be violated without degrading the test results considerably. In general, a test is said to be robust if it is relatively insensitive to violation of its underlying assumptions. That is, a robust test is one in which the actual value of significance is unaffected by failure to meet assumptions (i.e., it is near the nominal level of significance), and at the same time the test maintains high power.

The one-sample t-test is widely considered reasonably robust against the violation of the normality assumption for large sample sizes, except for extremely skewed populations (see Bartlett 1935 or Bradley 1980). Departure from normality is most severe when sample sizes are small and becomes less serious as sample sizes increase (since the sampling distribution of the mean approaches a normal distribution; see ►Central Limit Theorems). However, for extremely skewed distribution even for quite large samples (e.g., 500), t-test may not be robust (Pocock 1982).

Numerous studies have dealt with the adequacy of the two-sample t-test if at least one assumption is violated. In case of unequal variances, it has been shown that the t-test is only robust if sample sizes are equal (e.g., Scheffé 1970; Posten et al. 1982; Zimmerman 2004). However, if two equal sample sizes are very small, the t-test may not be robust (see Huck 2008, pp. 205–207). If both sample size and variances are unequal, the Welch t-test is preferred to as a better procedure.

If the normality assumption is not met, a researcher can select one of the nonparametric alternatives of the t-test – in one-sample scenario ►Wilcoxon–signed–rank test, in two independent samples case ►Wilcoxon–Mann–Whitney test, and if the samples are dependent Wilcoxon-matched pair rank test (for the asymptotic efficiency comparison, see ►Asymptotic Relative Efficiency in Testing).

Extension to comparison of an arbitrary number of independent samples ends up in a technique called ►analysis of variance, abbreviated ANOVA. Multivariate counterparts of one-sample and two-sample t-tests are based on Hotelling's T^2 statistic (Johnson and Wichern 2007), and ANOVA generalizes to multivariate analysis of variance, abbreviated MANOVA (see ►Multivariate Analysis of Variance (MANOVA)).

Cross References

►Behrens–Fisher Problem
►Chernoff-Savage Theorem
►Density Ratio Model
►Effect Size
►Hotelling's T^2 Statistic
►Parametric Versus Nonparametric Tests
►Presentation of Statistical Testimony
►Rank Transformations
►Research Designs
►Robust Inference
►Scales of Measurement and Choice of Statistical Methods
►Significance Testing: An Overview
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Student's t-Distribution
►Validity of Scales
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Aldrich J (2008) Comment on S. L. Zabell's paper: on Student's 1908 paper. The probable error of a mean. J Am Stat Assoc 103(481): 8–11

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, Hoboken

Bartlett MS (1935) The effect of non-normality on the t distribution. Proc Cambridge Philos Soc 31:223–231

Bhattacharyya GK, Johnson RA (1977) Statistical concepts and methods. Wiley, New York

Box JF (1987) Guinness, Gosset, Fisher, and small samples. Stat Sci 2(1):45–52

Bradley JV (1980) Nonrobustness in Z; t; and F tests at large sample sizes. Bull Psychonom Soc 16(5):333–336

Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or *t*-Test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4:1–39
Fisher RA (1939) Student. Ann Eugenic 9:1–9
Hogg RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics, 6th edn. Prentice-Hall, Pearson
Huck SW (2008) Statistical misconceptions. Routledge Academic, New York
Johnson RA (2005) Miller and Freund's probability and statistics for engineers, 7th edn. Prentice-Hall, Pearson
Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Prentice-Hall, Pearson
Miller I, Miller M (1999) John E. Freund's mathematical statistics, 6th edn. Prentice-Hall, Pearson
Pearson ES (1939) Student as a statistician. Biometrika 30:210–250
Pocock SJ (1982) When not to rely on the central limit theorem - an example from absentee data. Commun Stat Part A - Theory Meth 11(19):2169–2179
Posten HO, Yeh HC, Owen DB (1982) Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. Commun Stat - Theory Meth 11:109–126
Scheffé H (1970) Practical solutions of the Behrens-Fisher problem. J Am Stat Assoc 65(332):1501–1508
Student (1908) The probable error of a mean. Biometrika 6:1–25
Winer BJ (1971) Statistical principles in experimental design. McGraw-Hill, New York
Zabel SL (2008) On Student's 1908 article. The probable error of a mean. J Am Stat Assoc 103(481):1–7
Zimmerman DW (2004) Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank transformation tests. Psicológica 25:103–133

Sturges' and Scott's Rules

David W. Scott
Noah Harding Professor, Associate Chairman
Rice University, Houston, TX, USA

Introduction

The fundamental object of modern statistics is the random variable X and its associated probability law. The probability law may be given by the cumulative probability distribution $F(x)$, or equivalently by the probability density function $f(x) = F'(x)$, assuming the continuous case. In practice, estimation of the probability density may approached either parametrically or nonparametrically. If a parametric model $f(x|\theta)$ is assumed, then the unknown parameter θ may be estimated from a random sample using maximum likelihood methods, for example. If no parametric model is available, then a nonparametric estimator such as the histogram may be chosen. This article describes two different methods of specifying the construction of a histogram from a random sample.

Histogram as Density Estimator

The histogram is a convenient graphical object for representing the shape of an unknown density function. We begin by reviewing the stem-and-leaf diagram, introduced by Tukey (1977). Tukey reanalyzed Lord Rayleigh's 15 measurements of the weight of nitrogen. Using the ▶R language, the stem-and-leaf diagram of the weights is given in Fig. 1. One of the 15 raw numbers is $x_1 = 2.30143$. Where does x_1 appear in the diagram? The three digits to the left of "|" are called the *stem*. The stems correspond to the *bins* of a histogram. Here there are four stems, defined by the five cut points $(2.295, 2.300, 2.305, 2.310, 2.315)$. The bin counts are $(6, 2, 0, 7)$, with x_1 falling in the second bin. Rounding x_1 to 2.301 and removing the stem "230," leaves the leaf value of "1," which is what appears to the right of the second stem in Figure 1. In the fourth stem, all seven measurements rounded to 2.310. This sample was measured to high accuracy to estimate the atomic weight of nitrogen, but instead its highly non-normal shape led to the discovery of the noble gas argon.

The ordinary histogram depicts only the bin counts, which we denote by $\{\nu_k\}$, where the integer k indicates the bin number. Then $\sum_k \nu_k = n$, where n denotes the sample size. Given an ordered set of cut points $\{t_k\}$, the *kth* bin B_k is the half-open interval $[t_k, t_{k+1})$. If all of the bins have the same width, then plotting the bin counts gives an indication of the shape of the underlying density; see the left frame of Figure 2 for an example.

The left frame of Fig. 2 depicts a frequency histogram, since the bin counts $\{\nu_k\}$ are plotted. The density histogram is defined by the formula

$$\hat{f}(x) = \frac{\nu_k}{nh} \quad x \in B_k. \tag{1}$$

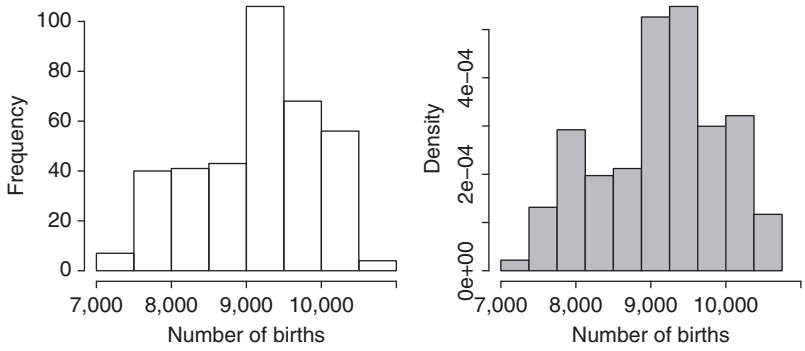
The density histogram estimator is nonnegative and integrates to 1. The right frame of Fig. 2 shows a density histogram with a narrower bin width.

```
> stem (wts)

  The decimal point is 2 digit (s) to the left of the |

229 | 889999
230 | 12
230 |
231 | 0000000
```

Sturges' and Scott's Rules. Fig. 1 Tukey's stem-and-leaf plot of the Raleigh data ($n = 15$)



Sturges' and Scott's Rules. Fig. 2 Histograms of the number of births daily in the USA in 1978. The bin widths are 500 and 375, respectively

Sturges' Rule

The origins of the histogram may be traced back to 1662 and the invention of actuarial tables by John Graunt (1662). But the first practical rule for the construction of histograms took another 260 years. Sturges (1926) essentially developed a normal reference rule, that is, a formula for the number of bins appropriate for normal data. Sturges sought a discrete distribution that was approximately normal to develop his formula. While several come to mind, clearly a binomial random variable $Y \sim B(m,p)$ with $p = \frac{1}{2}$ is suitable. If we imagine appropriately re-scaled normal data, which are continuous, rounded to integer values $(0,1,\ldots,m)$ in the $m+1$ bins (each of width $h=1$)

$$B_0 = \left(-\frac{1}{2},\frac{1}{2}\right] \quad B_1 = \left(\frac{1}{2},\frac{3}{2}\right] \quad \ldots \quad B_m = \left(m-\frac{1}{2},m+\frac{1}{2}\right], \tag{2}$$

then the Binomial probability in the *kth* bin is given by

$$P(Y=k) = \binom{m}{k} p^k (1-p)^{m-k} = \binom{m}{k}\left(\frac{1}{2}\right)^m = \frac{\binom{m}{k}}{2^m \cdot 1}. \tag{3}$$

Comparing the density formulae in Eqs. 1 and 3, we have

$$v_k = \binom{m}{k}, \quad n = 2^m, \quad \text{and} \quad h = 1. \tag{4}$$

If we let K denote the number of bins, then $K = m+1$ for the binomial density, as well as for the appropriately re-scaled normal data. From Eq. 4, we compute

$$n = 2^m = 2^{K-1}; \quad \text{hence} \quad K = 1 + \log_2(n). \tag{5}$$

The formula for K in Eq. 5 is called *Sturges' Rule*.

Scott's Rule

The density histogram $\hat{f}(x) = v_k/nh$ is not difficult to analyze for a random sample of size n from a density $f(x)$. Given a set of equal-width bins, the bin counts $\{v_k\}$ are individually a Binomial random variable $B(n,p_k)$, with probability

$$p_k = \int_{B_k} f(t)\,dt = \int_{t_k}^{t_{k+1}} f(t)\,dt = \int_{t_k}^{t_k+h} f(t)\,dt.$$

So $Ev_k = np_k$. Thus for a fixed point x, the expected value of the density histogram $\hat{f}(x)$ is $(np_k)/nh = p_k/h$. Scott (1979) shows that this is close to the unknown true value $f(x)$ when the bin width h is small.

On the other hand, the variance of v_k is $np_k(1-p_k)$, so that the variance of $\hat{f}(x)$ is $np_k(1-p_k)/(nh)^2 \sim p_k/nh^2$. This variance will be small if h is large. Since h cannot be both small and large, and using the integrated mean squared error as the criterion, Scott (1979) derived the asymptotically optimal bin width to be

$$h_S^* = \left(\frac{6}{n\int f'(t)^2\,dt}\right)^{1/3}. \tag{6}$$

While the formula for h^* in Eq. 6 seems to require knowledge of the unknown density, it is perfectly suitable for deriving Scott's normal-reference bin-width rule. If $f \sim N(\mu,\sigma^2)$, then

$$\int_{-\infty}^{\infty} f'(t)^2\,dt = \frac{1}{4\sqrt{\pi}\sigma^3} \quad \text{and}$$
$$h_S^* = \left(\frac{24\sqrt{\pi}\sigma^3}{n}\right)^{1/3} \approx 3.5\sigma n^{-1/3}. \tag{7}$$

Scott's rule $\hat{h}_S$ replaces σ in the formula for h_S^* by the usual maximum likelihood estimate of the standard deviation.

The Rules in Practice

For the birth count data used in Fig. 2, $n = 365$, $\hat{\sigma} = 817.9$, and the sample range is $(7135, 10711)$; hence, Sturges' and

Scott's rules give

$$K = 9.51\left(\text{or } \hat{h} = \frac{10711 - 7135}{9.51} = 376.0\right) \text{ and } \quad \hat{h}_S = 400.6.$$

Note the density histogram in the right frame of Fig. 2 uses $h = 375$, which has ten bins. Interestingly, the left frame shows the default histogram in R, which implements Sturges' rule as well. However, instead of finding ten bins exactly, R uses the function *pretty* to pick approximately ten bins with "convenient" values for $\{t_k\}$. The result in this case is 8 bins, and $h = 500$. Scott's rule (not shown) is close to $h = 375$.

The Rules with Massive Datasets

While the two rules often give similar results for sample sizes less than a couple hundred, they diverge for larger values of n for any density, including the normal. To see this, let us reconsider the binomial/normal construction at Eq. 2 we used to find Sturges' rule. (The data are basically rounded to one of the $m+1$ integer values $0, 1, \ldots, m$.) Thus we have $K = m+1$ bins, $n = 2^{K-1}, \mu = mp = m/2$, and $\sigma^2 = mp(1-p) = m/4$. Note that the variance of this density increases with the sample size in such a way that Sturges' rule always gives $h = 1$ for any sample size.

By way of contrast, Scott's rule from Eq. 7 is given by

$$h_S^* = 3.5\sqrt{\frac{m}{4}}\, n^{-1/3} = 1.75\sqrt{K-1}\, n^{-1/3}$$
$$= 1.75\sqrt{\log_2(n)}\, n^{-1/3}. \qquad (8)$$

Observe that $h_S^* \to 0$ as the sample size $n \to \infty$. In fact, $h_S^* < 1$ for all $n > 87$ for these data. When $n = 200$, $h_S^* = 0.83$, only 17% less than Sturges' $h = 1$. However, when $n = 10^6$, $h_S^* = 0.0781$. Thus the optimal histogram would have nearly 13 (1/0.0781) times as many bins as when using Sturges' rule.

The bin width given in Eq. 8 is also the *ratio* of Scott's rule to the Sturges bin width (since $h = 1$). If the normal data have any other scale, then the ratio is the same. The trick of using the Binomial model facilitates the conversion of bin counts to bin widths. Otherwise, a more careful analysis of the sample range of normal data would be necessary.

Discussion

Both Sturges' and Scott's rules use the normal-reference principle. However, Sturges makes a deterministic calculation, whereas Scott's rule is based upon a balancing of the global variance and squared bias of the histogram estimator. For normal data, we have seen that Sturges' rule greatly understates the optimal number of bins (according to integrated mean squared error). Thus we say that Sturges' rule tends to oversmooth the resulting histogram. Sturges' rule wastes a large fraction of the information available in large samples.

Why are these rules useful in practice? Terrell and Scott; 1985 show that there exists an "easiest" smooth density, whose optimal bin width is only 1.069 times as wide as Scott's normal reference rule. Terrell concludes that for any other density, the (unknown) optimal bin width will be narrower still. Thus, the normal reference rule is always useful as a first look at the data. Narrower bin widths can be investigated if the sample size is large enough and there is obvious non-normal structure.

Hyndman; 1995 cautions that since both v_k and n in Eqs. 3 and 4 could be multiplied by a constant factor, that K could take the general form $c + \log_2(n)$. The fact that Sturges' rule ($c = 1$) continues to be used is probably due to its simple form and its closeness to the optimal number of bins for textbook-sized problems ($n < 200$). Of course, if you impose the boundary condition that with one sample ($n = 1$) you should choose one bin ($K = 1$), then you would conclude that $c = 1$ is appropriate.

A variation of Scott's rule was independently proposed by Freedman and Diaconis; 1981, who suggested using a multiple of the interquartile range rather than $\hat{\sigma}$ in the normal reference rule. Of course, there are more advanced methods of cross-validation for histograms introduced by Rudemo; 1982. Surveys of these and other ideas may be found in Scott; 1992, Wand; 1997, and Doane 1976. Finally, we note that if the bin widths are not of equal width, then the shape of the frequency histogram can be grossly misleading. The appropriate density histogram has the form v_k/nh_k, but more research is required to successfully construct these generalized histograms in practice.

Acknowledgments

This work was partially supported by NSF award DMS-09-07491, and ONR contract N00014-06-1-0060.

About the Author

Professor Scott was awarded the Founders Award, American Statistical Association (2008), for "superb contributions and leadership in statistical research, particularly in multivariate density estimation and visualization, and in editorship of the Journal of Computational and Graphical Statistics." He has also received the U.S. Army Wilks Award (2004), and was named the Texas Statistician of the Year (1993). He was Editor, *Journal of Computational and Graphical Statistics* (2000–2004).

Cross References

►Exploratory Data Analysis
►Nonparametric Density Estimation
►Nonparametric Estimation
►Stem-and-Leaf Plot

References and Further Reading

Doane DP (1976) Aesthetic frequency classifications. Am Stat 30:181–183
Freedman D, Diaconis P (1981) On the histogram as a density estimator: 12 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57:453–476
Graunt J (1662) Natural and political observations made upon the bills of mortality. Martyn, London
Hyndman RJ (1995) The problem with sturges rule for constructing histograms. Unpublished note, 1995
Rudemo M (1982) Empirical choice of histograms and kernel density estimators. Scand J Stat 9:65–78
Scott DW (1979) On optimal and data-based histograms. Biometrika 66:605–610
Scott DW (1992) Multivariate density estimation: theory, practice, and visualization. Wiley, New York
Sturges HA The choice of a class interval. J Am Stat Assoc 21:65–66
Terrell GR, Scott DW (1985) Oversmoothed nonparametric density estimates. J Am Stat Assoc 80:209–214
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, MA
Wand MP (1997) Data-based choice of histogram bin width. Am Stat 51:59–64

Sufficient Statistical Information

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

In the entry ►Sufficient statistics, it was mentioned that we wished to work with a sufficient or minimal sufficient statistic **T** because such a statistic will summarize data, but preserve all "information" about an unknown parameter $\boldsymbol{\theta}$ contained in the original data. Here, $\boldsymbol{\theta}$ may be real or vector valued. But, how much *(Fisher-)information* do we have in the original data which we attempt to preserve through data summary? Our present concern is to quantify Fisher-information content within some data.

The notion of the information about $\boldsymbol{\theta}$ contained in data was introduced by F. Y. Edgeworth in a series of papers, published in the *J. Roy. Statist. Soc.*, during 1908–1909. Fisher (1922) articulated the systematic development of this concept. The reader is referred to Efron's (1998, p. 101) commentaries on Fisher-information.

Section "►One Parameter Case" introduces a one-parameter situation. Section "►Multi-Parameter Case" discusses the two-parameter case which easily extends to a multi-parameter situation. When one is forced to utilize some less than full information data summary, we discuss in section "►Role in the Recovery of Full Information" how the lost information may be recovered by conditioning on ancillary statistics. Mukhopadhyay (2000, Chap. 6) includes in-depth discussions.

One-Parameter Case

Suppose that X is an observable real valued random variable with the pmf or pdf $f(x;\theta)$ where the unknown parameter $\theta \in \Theta$, an open subinterval of $\Re$, while the $\mathcal{X}$ space is *assumed* not to depend upon θ. We *assume* throughout that the partial derivative $\frac{\partial}{\partial\theta} f(x;\theta)$ is finite for all $x \in \mathcal{X}$, $\theta \in \Theta$. We also *assume* that we can interchange the derivative (with respect to θ) and the integral (with respect to x).

Definition 1 *The Fisher-information or simply the information about θ, contained in the data, is given by*

$$\mathcal{I}_X(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X;\theta)\right\}^2\right].$$

The information $\mathcal{I}_X(\theta)$ measures the square of the sensitivity of $f(x;\theta)$ on an average due to an infinitesimal subtle change in the true parameter value θ. This concept may be understood as follows: Consider

$$\lim_{\Delta\theta\to 0} \frac{f(x;\theta+\Delta\theta) - f(x;\theta)}{\Delta\theta} \div f(x;\theta)$$

which is $\frac{\partial}{\partial\theta}\log f(x;\theta)$. Obviously, $E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] \equiv 0$, and hence one goes on to define $\mathcal{I}_X(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X;\theta)\right\}^2\right]$.

Example 1 Let X be Poisson(λ), $\lambda > 0$. One verifies that $\mathcal{I}_X(\lambda) = \lambda^{-1}$. That is, as we contemplate having larger and larger values of λ, the variability built in X increases, and hence it seems natural that the information about the unknown parameter λ contained in the data X will go down further and further. ▲

Example 2 Let X be $N(\mu,\sigma^2)$ where $\mu \in (-\infty,\infty)$ is an unknown parameter. Here, $\sigma \in (0,\infty)$ is assumed known. One verifies that $\mathcal{I}_X(\mu) = \sigma^{-2}$. Again, as we contemplate having larger and larger values of σ, the variability built in

X increases, and hence it seems natural that the information about the unknown parameter μ contained in the data X will go down further and further. ▲

The following result quantifies the information about an unknown parameter θ contained in a random sample $X_1, \ldots, X_n$ of size n.

Theorem 1 *Let $X_1, \ldots, X_n$ be iid with a common pmf or pdf given by $f(x;\theta)$. We denote $E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right\}^2\right] = \mathcal{I}_{X_1}(\theta)$, the information contained in the observation X_1. Then, the information $\mathcal{I}_X(\theta)$, contained in the random sample $X = (X_1, \ldots, X_n)$, is given by*

$$\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta) \text{ for all } \theta \in \Theta.$$

Next, suppose that we have collected random samples $X_1, \ldots, X_n$ from a population and we have somehow evaluated the information $\mathcal{I}_X(\theta)$ contained in $\mathbf{X} = (X_1, \ldots, X_n)$. Also, suppose that we have a summary *statistic* $T = T(\mathbf{X})$ in mind for which we have evaluated the information $\mathcal{I}_T(\theta)$ contained in T. If it turns out that $\mathcal{I}_T(\theta) = \mathcal{I}_X(\theta)$, can we then claim that the statistic T is indeed sufficient for θ? The answer is yes, we certainly can.

We state the following result by referring to Rao (1973, result (iii), p. 330) for details. In an exchange of personal communications, C.R. Rao had provided a simple way to look at the next Theorem 2. In Mukhopadhyay (2000), the Exercise 6.4.15 gives an outline of Rao's elegant proof whereas in the Examples 6.4.3–6.4.4 of Mukhopadhyay (2000), one finds opportunities to apply this theorem.

Theorem 2 *Suppose that $\mathbf{X}$ is the whole data and $T = T(\mathbf{X})$ is some statistic. Then, $\mathcal{I}_{\mathbf{X}}(\theta) \geq \mathcal{I}_T(\theta)$ for all $\theta \in \Theta$. The two information measures will be equal for all θ if and only if T is a sufficient statistic for θ.*

Multi-Parameter Case

When the unknown parameter $\boldsymbol{\theta}$ is multidimensional, the definition of the Fisher information measure gets more involved. To keep the presentation simple, we only discuss the case of a two-dimensional parameter.

Suppose that X is an observable real valued random variable with the pmf or pdf $f(x;\boldsymbol{\theta})$ where the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta$, an open rectangle $\subseteq \Re^2$, and the $\mathcal{X}$ space does not depend upon $\boldsymbol{\theta}$. We *assume* throughout that $\frac{\partial}{\partial\theta_i}f(x;\boldsymbol{\theta})$ exists, $i = 1, 2$, for all $x \in \mathcal{X}$, $\boldsymbol{\theta} \in \Theta$, and that we can also interchange the partial derivative (with respect to θ_1, θ_2) and the integral (with respect to x).

Definition 2 *Denote $I_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\left\{\frac{\partial}{\partial\theta_i}\log f(X;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta})\right\}\right]$, for $i, j = 1, 2$. The Fisher-information matrix or simply the information matrix about $\boldsymbol{\theta}$ is given by*

$$\mathcal{I}_X(\boldsymbol{\theta}) = \begin{pmatrix} I_{11}(\boldsymbol{\theta}) & I_{12}(\boldsymbol{\theta}) \\ I_{21}(\boldsymbol{\theta}) & I_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

In situations where $\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x;\boldsymbol{\theta})$ exists for all $x \in \mathcal{X}$, for all $i, j = 1, 2$, and for all $\boldsymbol{\theta} \in \Theta$, we can alternatively express

$$I_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\log f(X;\boldsymbol{\theta})\right] \text{ for } i, j = 1, 2,$$

and rewrite $\mathcal{I}_X(\boldsymbol{\theta})$ accordingly.

Having a statistic $T = T(X_1, \ldots, X_n)$, however, the associated information matrix about $\boldsymbol{\theta}$ will simply be calculated as $\mathcal{I}_T(\boldsymbol{\theta})$ where one would replace the original pmf or pdf $f(x;\boldsymbol{\theta})$ by that of T, namely $g(t;\boldsymbol{\theta})$, $t \in \mathcal{T}$. In order to compare two summary statistics T_1 and T_2, we have to consider their individual two-dimensional *information matrices* $\mathcal{I}_{T_1}(\boldsymbol{\theta})$ and $\mathcal{I}_{T_2}(\boldsymbol{\theta})$. It would be tempting to say that T_1 *is more informative* about $\boldsymbol{\theta}$ than T_2 provided that

the matrix $\mathcal{I}_{T_1}(\theta) - \mathcal{I}_{T_2}(\theta)$ is positive semi definite.

A version of Theorem 1. holds in the multiparameter case. One may refer to Rao (1973, Sect. 5a.3).

Example 3 Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ are both unknown parameters. Denote $\boldsymbol{\theta} = (\mu, \sigma^2)$, $\mathbf{X} = (X_1, \ldots, X_n)$. One can verify that the information matrix is given by

$$\mathcal{I}_X(\boldsymbol{\theta}) = n\mathcal{I}_{X_1}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix},$$

for the whole data $\mathbf{X}$. ▲

Example 4 (Example 3. Continued) Let $\overline{X} = n^{-1}\Sigma_{i=1}^n X_i$, the sample mean and $S^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2$, the sample variance, $n \geq 2$. One can check that

$$\mathcal{I}_{\overline{X}}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{pmatrix},$$

$$\mathcal{I}_{S^2}(\boldsymbol{\theta}) = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2}(n-1)\sigma^{-4} \end{pmatrix}.$$

Surely, $\overline{X}$ and S^2 are independent, and hence

$$\mathcal{I}_{\overline{X},S^2}(\boldsymbol{\theta}) = \mathcal{I}_{\overline{X}}(\boldsymbol{\theta}) + \mathcal{I}_{S^2}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix},$$

which coincides with $\mathcal{I}_X(\boldsymbol{\theta})$. ▲

Role in the Recovery of Full Information

In the entry ►Sufficient statistics, we had seen how ancillary statistics could play significant roles in conjunction with non-sufficient statistics. Suppose that T_1 is a non-sufficient statistic for θ and T_2 is ancillary for θ. In other words, in terms of the information content, $\mathcal{I}_{T_1}(\theta) < \mathcal{I}_X(\theta)$ where $\mathbf{X}$ is the whole data and $\mathcal{I}_{T_2}(\theta) = 0$ for all $\theta \in \Theta$. Can we recover all the information contained in $\mathbf{X}$ by reporting T_1 while conditioning on the observed value of T_2? The answer is: we can do so and it is a fairly simple process.

Such a process of conditioning has far reaching implications as emphasized by Fisher (1934, 1956) in his famous "Nile" example. One may also refer to Basu (1964), Hinkley (1980), Ghosh (1988) and Reid (1995) for fuller discussions of *conditional inference*. Also, refer to Mukhopadhyay (2000, Sect. 6.5).

The approach goes through the following steps. One first finds the conditional pdf of T_1 when $T_1 = u$ given that $T_2 = v$, denoted by $g_{T_1|v}(u;\theta)$. Using this conditional pdf, one can obtain the information content:

$$\mathcal{I}_{T_1|v}(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log\{g_{T_1|v}(T_1;\theta)\}\right\}^2\right].$$

In general, the expression of $\mathcal{I}_{T_1|v}(\theta)$ would depend on v, that is, the fixed value of T_2. Next, one averages $\mathcal{I}_{T_1|v}(\theta)$ over all possible values v, that is, evaluates $E_{T_2}[\mathcal{I}_{T_1|T_2}(\theta)]$. Once this last bit of averaging is done, it will coincide with the information content in the joint statistic (T_1, T_2), that is, one can claim:

$$\mathcal{I}_{T_1,T_2}(\theta) = E_{T_2}\left[\mathcal{I}_{T_1|T_2}(\theta)\right].$$

This analysis provides a way to recover the lost information due to reporting T_1 alone via conditioning on an ancillary statistic T_2. Two examples follow that are taken from Mukhopadhyay (2000, pp. 316–318).

Example 5 Let X_1, X_2 be iid $N(\theta, 1)$ where $\theta \in (-\infty, \infty)$ is an unknown parameter. We know that $\overline{X}$ is sufficient for θ. Now, $\overline{X}$ is distributed as $N\left(\theta, \frac{1}{2}\right)$ so that we can immediately write $\mathcal{I}_{\overline{X}}(\theta) = 2$. Now, $T_1 = X_1$ is not sufficient for θ since $\mathcal{I}_{X_1}(\theta) = 1 < \mathcal{I}_{\overline{X}}(\theta)$. That is, if we report only X_1 after the data (X_1, X_2) has been collected, there will be some loss of information. Next, consider an ancillary statistic, $T_2 = X_1 - X_2$ and now the joint distribution of (T_1, T_2) is $N_2\left(\theta, 0, 1, 2, \rho = \frac{1}{\sqrt{2}}\right)$. Hence, we find that the conditional distribution of T_1 given $T_2 = v$ is $N\left(\theta + \frac{1}{2}v, \frac{1}{2}\right), v \in (-\infty, \infty)$. Thus, we first have $\mathcal{I}_{T_1|v}(\theta) = E_{T_1|v}\left[4\left(T_1 - \theta - \frac{1}{2}v\right)^2\right] = 2$ and since this expression does not involve v, we then have $E_{T_2}\left[\mathcal{I}_{T_1|T_2}(\theta)\right] = 2$ which equals $\mathcal{I}_{\overline{X}}(\theta)$. In other words, by conditioning on the ancillary statistic T_2, we have recovered the full information which is $\mathcal{I}_{\overline{X}}(\theta)$. ▲

Example 6 Suppose that (X, Y) is distributed as $N_2(0, 0, 1, 1, \rho)$ where the unknown parameter is the correlation coefficient $\rho \in (-1, 1)$. Now consider the two individual statistics X and Y. Individually, both $T_1 = X$ and $T_2 = Y$ are ancillary for ρ. We note that the conditional distribution of X given $Y = y$ is $N(\rho y, 1-\rho^2)$ for $y \in (-\infty, \infty)$ and accordingly have,

$$\frac{\partial}{\partial\rho}\log f_{X|Y=y}(x;\rho) = \frac{\rho}{1-\rho^2} - \left[\frac{\rho(x-\rho y)^2}{(1-\rho^2)^2} - \frac{y(x-\rho y)}{(1-\rho^2)}\right].$$

In other words, the information about ρ contained in the conditional distribution of $T_1 \mid T_2 = v, v \in \Re$, is given by

$$\frac{2\rho^2}{(1-\rho^2)^2} + \frac{v^2}{(1-\rho^2)},$$

which depends on the value v unlike what we had in Example 5. Then, the information contained in (X, Y) will be given by

$$\mathcal{I}_{X,Y}(\rho) = E_{T_2}\left[E_{T_1|T_2=v}\left(\left\{\frac{\partial}{\partial\rho}\log f_{T_1|T_2=v}(T_1;\rho)\right\}^2\right)\right]$$

$$= E_{T_2}\left[\frac{2\rho^2}{(1-\rho^2)^2} + \frac{T_2^2}{(1-\rho^2)}\right]$$

$$= \frac{2\rho^2}{(1-\rho^2)^2} + \frac{1}{(1-\rho^2)} = \frac{1+\rho^2}{(1-\rho^2)^2}.$$

In other words, even though the statistic X tells us nothing about ρ, by averaging the conditional (on the statistic Y) information in X, we have recovered the full information about ρ contained in the whole data (X, Y). ▲

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Cramér–Rao Inequality
►Estimation
►Statistical Design of Experiments (DOE)
►Statistical Inference for Stochastic Processes
►Statistical View of Information Theory
►Sufficient Statistics

References and Further Reading

Basu D (1964) Recovery of ancillary information. Contributions to statistics, the 70th birthday festschrift volume presented to P. C. Mahalanobis. Pergamon, Oxford

Efron BF (1998) R. A. Fisher in the 21st century (with discussions by Cox DR, Kass R, Barndorff-Nielsen O, Hinkley DV, Fraser DAS, Dempster AP) Stat Sci 13:95–122

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A222:309–368

Fisher RA (1934) Two new properties of mathematical likelihood. Proc R Soc A 144:285–307

Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh/London

Ghosh JK (ed) (1988) Statistical information and likelihood: a collection of critical essays by Dr. D. Basu. Lecture notes in statistics No 45. Springer, New York

Hinkley DV (1980) Fisher's development of conditional inference. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 101–108

Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York

Rao CR (1973) Linear statistical inference and its applications, 2 edn. Wiley, New York

Reid N (1995) The roles of conditioning in inference (with discussions by Casella G, Dawid AP, DiCiccio TJ, Godambe VP, Goutis C, Li B, Lindsay BC, McCullagh P, Ryan LA, Severini TA, Wells MT) Stat Sci 10:138–157

Sufficient Statistics

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

Many fundamental concepts and principles of statistical inference originated in Fisher's work. Perhaps the deepest of all statistical concepts and principles is *sufficiency*. It originated from Fisher (1920) and blossomed further in Fisher (1922). We introduce the notion of sufficiency which helps in summarizing data without any loss of *information*.

Section "►Sufficiency" introduces sufficiency and *Neyman factorization*. Section "►Minimal Sufficiency" discusses *minimal sufficiency*, the *Lehmann-Scheffé approach*, and *completeness*. Section "►Neyman Factorization" shows the importance of *ancillary* statistics including Basu's theorem. Mukhopadhyay (2000, Chap. 6) provides many more details.

Sufficiency

Let $X_1, \ldots, X_n$ be independent real-valued observations having a common probability mass function (pmf) or probability density function (pdf) $f(x;\theta), x \in \mathcal{X}$, the domain space for x. Here, n is known, but $\theta \in \Theta(\subseteq \Re)$ is unknown. In general, however, the X's and θ are allowed to be vector valued. This should be clear from the context. A summary from data $\mathbf{X} \equiv (X_1, \ldots, X_n)$ is provided by some appropriate statistic, $T \equiv T(\mathbf{X})$ which may be vector valued.

Definition 1 *A real valued statistic T is called sufficient for parameter θ if and only if the conditional distribution of the random sample $\mathbf{X} = (X_1, \ldots, X_n)$ given $T = t$ does not involve θ, for all $t \in \mathcal{T}$, the domain space for T.*

In other words, *given* the value t of a *sufficient statistic* T, *conditionally* there is no more information left in the original data regarding θ. That is, once a sufficient summary T becomes available, the original data $\mathbf{X}$ becomes redundant.

Definition 2 *A statistic $\mathbf{T} \equiv (T_1, \ldots, T_k)$ where $T_i \equiv T_i(X_1, \ldots, X_n), i = 1, \ldots, k$, is called jointly sufficient for parameter θ if and only if the conditional distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ given $\mathbf{T} = \mathbf{t}$ does not involve θ, for all $\mathbf{t} \in \mathcal{T} \subseteq \Re^k$.*

Example 1 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (iid) Poisson(λ) where λ is unknown, $0 < \lambda < \infty$. Here, $\mathcal{X} = \{0, 1, 2, \ldots\}$, $\theta = \lambda$, and $\Theta = (0, \infty)$. Then, $T = \Sigma_{i=1}^{n} X_i$ is a sufficient statistic for λ.

Neyman Factorization

Suppose that we have observable real valued iid observations $X_1, \ldots, X_n$ from a population with a common pmf or pdf $f(x;\boldsymbol{\theta})$. Then, the likelihood function is given by $L(\boldsymbol{\theta}) = \Pi_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Fisher (1922) discovered the fundamental idea of factorization whereas Neyman (1935) rediscovered a refined approach to factorize a likelihood function. Halmos and Savage (1949) and Bahadur (1954) introduced measure-theoretic treatments.

Theorem 1 **(Neyman Factorization Theorem).** *A vector valued statistic $\mathbf{T} = \mathbf{T}(X_1, \ldots, X_n)$ is jointly sufficient for $\boldsymbol{\theta}$ if and only if the following factorization holds:*

$$L(\boldsymbol{\theta}) = g(\mathbf{T}(x_1, \ldots, x_n); \boldsymbol{\theta})\, h(x_1, \ldots, x_n),$$
$$\text{for all } x_1, \ldots, x_n \in \mathcal{X},$$

where the functions $g(\mathbf{T};\boldsymbol{\theta})$ and $h(.)$ are both nonnegative, $h(x_1, \ldots, x_n)$ is free from $\boldsymbol{\theta}$, and $g(\mathbf{T};\boldsymbol{\theta})$ depends on $x_1, \ldots, x_n$ only through the observed value $\mathbf{T}(x_1, \ldots, x_n)$ of $\mathbf{T}$.

Example 2 Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\boldsymbol{\theta} = (\mu, \sigma^2) \in \Re \times \Re^+$ is an unknown parameter vector. Let

$\overline{X}, S^2$ respectively be the sample mean and variance. Then, $\mathbf{T} = (\overline{X}, S^2)$ is jointly sufficient for $\boldsymbol{\theta}$. However, this does not imply component-wise sufficiency. To appreciate this fine line, pretend for a moment that one could claim component-wise sufficiency. But, since $(\overline{X}, S^2)$, and hence $(S^2, \overline{X})$, is jointly sufficient for (μ, σ^2). Now, how many would be willing to push an idea that component-wise, S^2 is sufficient for μ or $\overline{X}$ is sufficient for σ^2!

Theorem 2 (Sufficiency in an Exponential Family). *Suppose that $X_1, \ldots, X_n$ are iid with a common pmf or the pdf belonging to a regular k-parameter exponential family, namely*

$$f(x;\boldsymbol{\theta}) = a(\boldsymbol{\theta})g(x)exp\left\{\Sigma_{i=1}^{k} b_i(\boldsymbol{\theta})R_i(x)\right\}$$

with appropriate forms for $g(x) \geq 0$, $a(\boldsymbol{\theta}) \geq 0, b_i(\boldsymbol{\theta})$ and $R_i(x)$, $i = 1, \ldots, k$. Denote $T_j = \Sigma_{i=1}^{n} R_j(X_i), j = 1, \ldots, k$. Then, the statistic $\mathbf{T} = (T_1, \ldots, T_k)$ is jointly sufficient for $\boldsymbol{\theta}$.

Minimal Sufficiency

From the factorization Theorems 1–2, it should be clear that the whole data $\mathbf{X}$ must always be sufficient for the unknown parameter $\boldsymbol{\theta}$. But, we ought to reduce the data by means of summary statistics in lieu of considering $\mathbf{X}$ itself. What is a natural way to define the notion of a "shortest sufficient" or "best sufficient" summary statistic? The other concern should be to get hold of such a summary, if there is one.

Lehmann and Scheffé (1950) gave a mathematical formulation of the concept known as *minimal sufficiency* and proposed a technique to locate minimal sufficient statistics. Lehmann and Scheffé (1955, 1956) included crucial follow-ups.

Definition 3 *A statistic $\mathbf{T}$ is called minimal sufficient for the unknown parameter $\boldsymbol{\theta}$ if and only if*

1. *$\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$, and*
2. *$\mathbf{T}$ is minimal or "shortest" in the sense that $\mathbf{T}$ is a function of any other sufficient statistic.*

Lehmann–Scheffé Approach

The following result was proved in Lehmann and Scheffé (1950). Its proof requires some understanding of the correspondence between a statistic and so called *partitions* it induces on a sample space.

Theorem 3 (Minimal Sufficient Statistics). *Let us denote $h(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \Pi_{i=1}^{n} f(x_i;\boldsymbol{\theta}) / \Pi_{i=1}^{n} f(y_i;\boldsymbol{\theta})$, the ratio of the likelihood functions at $\mathbf{x}$ and $\mathbf{y}$, for $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$. Let $\mathbf{T} \equiv \mathbf{T}(X_1, \ldots, X_n) = (T_1, \ldots, T_k)$ be a statistic such that the following holds:*

- *With any two arbitrary but fixed data points $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ from $\mathcal{X}^n$, $h(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ does not involve $\boldsymbol{\theta}$ if and only if $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$.*

Then, $\mathbf{T}$ is minimal sufficient for $\boldsymbol{\theta}$.

In Examples 1–2 and Theorem 2, the reported sufficient statistics also happen to be the minimal sufficient statistics. It should be noted, however, that a minimal sufficient statistic may exist for some distributions from outside a regular exponential family. For example, let $X_1, \ldots, X_n$ be iid Uniform$(0, \theta)$ where $\theta \in \Re^+$ is an unknown parameter. Here, $X_{n:n}$, the largest order statistic, is a minimal sufficient statistic for θ.

Theorem 4 (Distribution of a Minimal Sufficient Statistic in an Exponential Family). *Under the conditions of Theorem 2, the pmf or the pdf of the minimal sufficient statistic $(T_1, \ldots, T_k)$ also belongs to a k-parameter exponential family.*

In the case of population distributions not belonging to a regular exponential family, however, sometimes one may not achieve any substantial data reduction by invoking the concept of minimal sufficiency. For example, suppose that we have iid observations $X_1, \ldots, X_n$ having the following Cauchy pdf:

$$\frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, -\infty < x, \theta < \infty.$$

Here, $-\infty < \theta < \infty$ is an unknown location parameter. Now, let $\mathbf{T} = (X_{n:1}, \ldots, X_{n:n})$ where $X_{n:1} \leq \ldots \leq X_{n:n}$ are the sample order statistics. One can verify that $\mathbf{T}$ is a minimal sufficient statistic for θ.

A Complete Sufficient Statistic

Consider a real valued random variable X whose pmf or pdf is $f(x;\boldsymbol{\theta})$ for $x \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$. Let $T = T(X)$ be a statistic and suppose that its pmf or pdf is denoted by $g(t;\boldsymbol{\theta})$ for $t \in \mathcal{T}$ and $\boldsymbol{\theta} \in \Theta$. Then, $\{g(t;\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta\}$ is called the family of distributions induced by T.

Definition 4 *The family $\{g(t;\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta\}$ is called complete if and only if the following condition holds. Consider any real valued function $h(t)$ defined for $t \in \mathcal{T}$, having a finite expectation, such that*

$$E_{\boldsymbol{\theta}}[h(T)] = 0 \text{ for all } \boldsymbol{\theta} \in \Theta \text{ implies } h(t) \equiv 0 \text{ w.p.1.}$$

A statistic T is said to be complete if and only if $\{g(t;\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta\}$ is complete. A statistic $\mathbf{T}$ is called complete sufficient for $\boldsymbol{\theta}$ if and only if (1) *$\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$ and* (2) *$\mathbf{T}$ is complete.*

A complete sufficient statistic, if it exists, is also a minimal sufficient statistic. For example, let $X_1, \ldots, X_n$ be iid

Uniform$(0, \theta)$ where $\theta \in \Re^+$ is unknown. Here, $X_{n:n}$, the largest order statistic, is a complete sufficient statistic for θ. Hence, $X_{n:n}$ is also minimal sufficient for θ. This proof bypasses Theorem 3. Now, we state a remarkably general result (Theorem 5) in the case of a regular exponential family of distributions. One may refer to Lehmann (1986, pp. 142–143) for a proof of this result.

Theorem 5 (**Completeness of a Minimal Sufficient Statistic in an Exponential Family**). *Under the conditions of Theorem 2, the minimal sufficient statistic* $(T_1, \ldots, T_k)$ *is complete.*

Ancillary Statistics

The concept called *ancillarity* of a statistic is perhaps the furthest away from the notion of sufficiency. A sufficient statistic $\mathbf{T}$ preserves all the information about $\boldsymbol{\theta}$ contained in the data $\mathbf{X}$. In contrast, an ancillary statistic $\mathbf{T}$ *by itself* provides *no information* about $\boldsymbol{\theta}$. This concept evolved from Fisher (1925) and later it blossomed into the vast area of *conditional inference*. In his 1956 book, Fisher emphasized many positive aspects of ancillarity in analyzing real data. For fuller discussions of *conditional inference* one may look at Basu (1964), Hinkley (1980) and Ghosh (1988). Reid (1995) provides an assessment of conditional inference procedures.

Consider the real valued observable random variables $X_1, \ldots, X_n$ from some population having the common pmf or pdf $f(x; \boldsymbol{\theta})$, where the unknown parameter vector $\boldsymbol{\theta} \in \Theta \subseteq \Re^p$. Let us continue writing $\mathbf{X}$ for the full data and $\mathbf{T} = \mathbf{T}(\mathbf{X})$ for a vector valued statistic.

Definition 5 *A statistic* $\mathbf{T}$ *is called ancillary for* $\boldsymbol{\theta}$ *or simply ancillary provided that the pmf or the pdf of* $\mathbf{T}$ *does not involve* $\boldsymbol{\theta}$.

Here is an important result that ties the notions of complete sufficiency and ancillarity. Basu (1955) came up with this elegant result which we state here under full generality.

Theorem 6 (**Basu's Theorem**). *Suppose that we have two vector valued statistics,* $\mathbf{U} = \mathbf{U}(\mathbf{X})$ *which is complete sufficient for* $\boldsymbol{\theta}$ *and* $\mathbf{W} = \mathbf{W}(\mathbf{X})$ *which is ancillary for* $\boldsymbol{\theta}$. *Then,* $\mathbf{U}$ *and* $\mathbf{W}$ *are independently distributed.*

An ancillary statistic by itself tells one nothing about $\boldsymbol{\theta}$! Hence, one may think that an ancillary statistic may not play a role to come up with a sufficient summary statistic. But, that may not be the case. The following examples will highlight the fundamental importance of ancillary statistics.

Example 3 Suppose that (X, Y) has a curved exponential family of distributions with the joint pdf given by

$$f(x, y; \theta) = \begin{cases} exp\{-\theta x - \theta^{-1} y\} & \text{if } 0 < x, y < \infty \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta (> 0)$ is an unknown parameter. This distribution was discussed by Fisher (1934, 1956) in the context of his famous "Nile" example. Denote $U = XY, V = X/Y$. One can show that U is ancillary for θ, V does not provide the full information about θ, but (U, V) is minimal sufficient for θ. Note that $V^{1/2}$ is the maximum likelihood estimator of θ, but it is not minimal sufficient for θ.

Example 4 This example was due to D. Basu. Let (X, Y) be distributed as bivariate normal with zero means, unit variances, and an unknown correlation coefficient ρ, $-1 < \rho < 1$. Then, marginally, both X and Y are distributed as standard normal variables. Clearly, X by itself is an ancillary statistic, Y by itself is an ancillary statistic, but X and Y combined has all the information about ρ.

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Approximations for Densities of Sufficient Estimators
►Exponential Family Models
►Minimum Variance Unbiased
►Optimal Shrinkage Estimation
►Properties of Estimators
►Rao–Blackwell Theorem
►Statistical View of Information Theory
►Sufficient Statistical Information
►Unbiased Estimators and Their Applications

References and Further Reading

Bahadur RR (1954) Sufficiency and statistical decision functions. Ann Math Stat 25:423–462
Basu D (1955) On statistics independent of a complete sufficient statistic. Sankhyā 15:377–380
Basu D (1964) Recovery of ancillary information. Contributions to statistics, the 70th birthday festschrift volume presented to P. C. Mahalanobis. Pergamon, Oxford
Fisher RA (1920) A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Mon Not R Astron Soc 80:758–770
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A222:309–368
Fisher RA (1925) Theory of statistical estimation. Proc Camb Philos Soc 22:700–725
Fisher RA (1934) Two new properties of mathematical likelihood. Proc R Soc A 144:285–307
Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh/London

Ghosh JK (ed) (1988) Statistical information and likelihood: a collection of critical essays by Dr. D. Basu. Lecture notes in statistics No. 45. Springer, New York
Halmos PR, Savage LJ (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann Math Stat 20:225–241
Hinkley DV (1980) Fisher's development of conditional inference. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 101–108
Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York
Lehmann EL, Scheffé H (1950) Completeness, similar regions and unbiased estimation-Part I. Sankhyā 10:305–340
Lehmann EL, Scheffé H (1955) Completeness, similar regions and unbiased estimation-Part II. Sankhyā 15:219–236
Lehmann EL, Scheffé H (1956) Corrigenda: completeness, similar regions and unbiased estimation-Part I. Sankhyā 17:250
Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York
Neyman J (1935) Sur un teorema concernente le cosidette statistiche sufficienti. Giorn 1st Ital Att 6:320–334
Reid N (1995) The roles of conditioning in inference (with discussions by Casella G, Dawid AP, DiCiccio TJ, Godambe VP, Goutis C, Li B, Lindsay BC, McCullagh P, Ryan LA, Severini TA, Wells MT) Stat Sci 10:138–157

Summarizing Data with Boxplots

Boris Iglewicz
Professor
Temple University, Philadelphia, PA, USA

Introduction

Statisticians have created a variety of techniques for summarizing data graphically. For continuous univariate data the most commonly used graphical display is the histogram. Once the interval width is carefully determined, the histogram provides a visual summary of the data center, spread, ▶skewness, and unusual observations, which may be ▶outliers. While these features are visible, there are no specific numeric summary measures that are part of the histogram display.

Tukey (1977) introduced a simple alternative to the histogram that contains similar features as the histogram, is easier to graph, and includes measures of location, spread, skewness, and a rule for flagging outliers. He called this graphic summary the boxplot. The key components of the boxplot consist of Tukey's five number summary. These are: the median = Q_2; upper quartile = Q_3; lower quartile = Q_1; largest value = $X_{(n)}$; and the smallest value = $X_{(1)}$. This information is all that is needed to graph the simplest version of the boxplot, called the box-and-whisker plot. Such a plot is illustrated as the left plot of Fig. 1. The data consists of daily percent changes in the Dow Jones industrial average closing values for days when the market is open. Thus, if Y_t is the closing Dow Jones Industrial Average at day t, then the data for the boxplots in Fig. 1 consists of $X_t = 100(Y_t - Y_{t-1})/Y_{t-1}$.

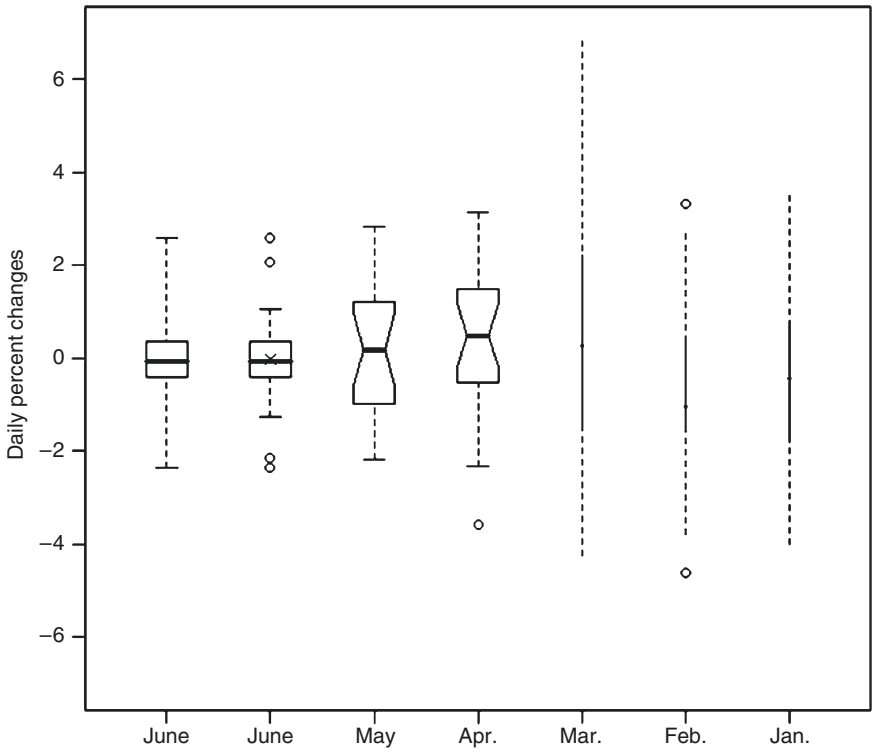
The box-and-whisker plot has a box at the center that contains approximately 50% of the middle observations. The horizontal line inside the plot is the median, Q_2, which provides a nice summary measure for the data center. The upper and lower horizontal lines enclosing the box are the values of Q_3, and Q_1, respectively. From these one can obtain the interquartile range, $IQR = Q_3 - Q_1$, which is a common robust measure of spread. Skewness can also be observed by comparing $Q_3 - Q_2$ with $Q_2 - Q_1$ or $X_{(n)} - Q_2$ with $Q_2 - X_{(1)}$.

Tukey (1977) also added a simple rule for flagging observations as potential outliers. That rule flags observations as outliers if they fall outside the interval $(Q_1 - k(IQR), Q_3 + k(IQR))$. Tukey suggested using $k = 1.5$ for a liberal interval with out values so designated. He also suggested using $k = 3$ to designate far out values. The box-and-whisker plot that incorporates the rule for flagging outliers is called a boxplot. The second from left plot in Fig. 1 illustrates such a boxplot for the June 2009 data. In addition, this boxplot contains an X in the middle designating the location of the sample mean. The inclusion of the sample mean is a useful added feature that some statistical computer packages incorporate.

Although the boxplot is a simple graphic summary procedure, a number of modifications have been suggested and properties studied. In section "▶Varied Versions of the Basic Boxplot" we will briefly review other variants of the basic boxplot. In section "▶Outlier Rule" we will consider further the properties of the outlier identification rule and suggest modified versions. In section "▶Quartiles" we will consider the computation of quartiles. A brief summary will be provided in section "▶Summary".

Varied Versions of the Basic Boxplot

A fair number of alternative versions of the basic boxplot have been introduced and used. Tufte (1983) suggested a slight modification that is useful in summarizing a large number of parallel boxplots that can be especially useful when dealing with data collected over many time periods. Tufte suggested removing the box, as in the three right most graphs in Fig. 1, representing the data for January, February, and March 2009, respectively. The box can now be represented by either a solid line or empty space.



Summarizing Data with Boxplots. Fig. 1 Graph contains several versions of boxplot construction based on daily percent changes of the Dow Jones industrial average grouped by month. The two left hand boxplots consist of June 2009 data with the right one including potential outliers. The next two to the right represent notched boxplots. The three right side boxplots are based on a version suggested by Tufte

The point inside the solid line designates the location of the median. The dashed lines go towards the largest and smallest observations excluding flagged outliers, which are individually plotted on the boxplot graph.

Another avenue of innovation is the thickness of the box. The simplest suggestion, given by McGill et al. (1978), is to make the width proportional to the square root of the sample size, thus showing precision. This is further refined by Benjamini (1988), who suggested replacing the two outer vertical lines of the box by density plots. Such density plots depend on the kernel and window width and are thus not unique. He called these plots histplots. Benjamini also introduced density plots for the entire vertical length of the boxplots. These plots he called vaseplots. Both the hisplots and vaseplots consist of lines. The vaseplot is further refined by Hintze and Nelson (1998) who used a curved density plot as a replacement. As the resulting plot often looks like a violin, they called their modification a violin plot.

A further refinement is the notched boxplot introduced by McGill et al. (1978). The goal is to provide a visual significance test comparing the medians of two adjacent boxplots. If the two medians lie within the two notches, then we can say that the two population medians are not significantly different. Two notched boxplots are shown as the April and May data in Fig. 1, where we can see that the two population medians are not significantly different. The intervals are based on asymptotic results from the normal distribution. These are refined in common statistical packages by using sign test type intervals. Benjamini (1988) suggested using the standard boxplot, but represent the notches by a shaded horizontal region.

Outlier Rule

Tukey's simple outlier labeling rule is heavily used, typically with $k = 1.5$, where observations are labeled as outliers if they lie outside the interval $(Q_1 - k(IQR),\ Q_3 + k(IQR))$. Hoaglin et al. (1986) studied the performance of this rule for random normal data. They found that the rule with $k = 1.5$ is very liberal for moderate to large data sets. For example, for $n = 300$ random normal observations there is an 85% chance that at least one observation will be falsely

labeled as an outlier. Even with $n = 100$ that percentage stays at 53%. For the conservative $k = 3.0$ rule, these out probabilities drop drastically to 0.2 percent for $n = 100$. The problem is that this $k = 1.5$ rule does not take sample-size into account. Consequently, the chances of labeling regular observations as outliers increase with increasing sample-size.

Let $B(k, n)$ = probability that all observations of a random normal sample lie inside $(Q_1 - k(IQR),\ Q_3 + k(IQR))$. Hoaglin and Iglewicz (1987) obtained values of k as functions of n to keep $B(k, n) = 0.95$ or $B(k, n) = 0.90$. That is, all n observations are inside the outlier labeling interval. Thus, for $n = 100$, $B(k, n) = 0.95$, they obtained $k = 2.2$, while for $n = 300, k = 2.4$. Iglewicz and Banerjee (2001) extended this procedure to random samples from a variety of both symmetric and skewed distributions in addition to the normal. Their work was further extended by Sim et al. (2005) and Banerjee and Iglewicz (2007).

Quartiles

Although the computation of quartiles seems to be quite simple on the surface, there are actually a number of choices for computing quartiles. As an example, Frigge et al. (1989) discuss eight options for computing quartiles. Although these choices will have limited effect for large samples, they can differ noticeably for small samples. That can lead to different boundaries for the box part of the boxplot and different values of k to maintain $B(k, n) = 0.95$.

Consider the non-negative number $f = j + g$, where j is the integer part of f and g the fractional part. For example, if $f = 12.8$, then $j = 12$ and $g = 0.8$. Consider the ordered observations $X_{(1)} \le X_{(2)} \le X_{(3)} \le \cdots \le X_{(n)}$, then $X_{(f)} = (1 - g)X_{(j)} + gX_{(j+1)}$. The median is typically obtained as $Q_{(2)} = X_{(f)}$, where $f = (n + 1)/2$. Letting $n = 2N + 1$ for n odd, $Q_{(2)} = X_{(N+1)}$. For $n = 2N$, n even, $Q_{(2)} = (X_{(N)} + X_{(N+1)})/2$. Tukey (1977) suggested a very simple rule for obtaining $Q_{(1)}$ and $Q_{(3)}$, as $Q_{(1)} = X_{(f)}$, where $f = (j + 1)/2$ and j = the integer part of $(n + 1)/2$. Then $Q_{(3)} = X_{(n+1-f)}$. An alternative popular choice for f in $X_{(f)} = Q_{(1)}$ is $f = (n + 1)/4$.

Summary

The boxplot is a heavily used graphical tool for summarizing univariate continuous data. Although the boxplot option shown on the second from the left plot of Fig. 1 is by far the most popular version, a variety of other useful choices have been discussed. These include the notched boxplots that are useful in comparing two population medians, the Tufte version useful when comparing many samples, and plots that incorporate density information. On some occasions, professionals are content with the simpler box-and-whisker plot illustrated as the leftmost plot of Fig. 1. While the illustrations of Fig. 1 consist of vertical boxplots, these could have just as effectively been drawn horizontally.

While this write-up has been devoted to discussion of the popular univariate boxplot, there have been a number of successful introductions of bivariate boxplots. These again use robust measures, but incorporate information on the correlation between the variables. Two bivariate boxplot versions worthy of note are by Goldberg and Iglewicz (1992) and Rousseeuw et al. (1999).

Acknowledgment

The author wishes to thank Alicia Stranberg for help with generating the graph. This article was written while on a Study Leave from Temple University.

About the Author

Boris Iglewicz serves as Professor of Statistics and Director of Biostatistics Research Center, Temple University. He received his Ph.D in statistics from Virginia Tech. At Temple University Dr. Iglewicz has served as the founding director of the graduate programs in statistics and as department chair. He also received the school's Musser Leadership Award for Excellence in Research and chosen as a Senior Research Fellow. Dr. Iglewicz has published about 70 professional journal articles, books, and chapters in books. From the American Statistical Association (ASA) he was chosen as a Fellow and received the following awards and recognitions: Chapter Recognition Award; W. J. Youden Award; Don Owen Award; and SPAIG award. He also served as President of the Philadelphia Chapter of ASA. He is also a Fellow of the Royal Statistical Society, Elected member of the International Statistical Institute, and serves as Associate Editor of *Statistics in Biopharmaceutical Research*. Dr. Iglewicz is listed in American Men and Women of Science, Who's Who in America, and Who's Who in the World.

Cross References

►Data Analysis
►Exploratory Data Analysis
►Five-Number Summaries
►Outliers

References and Further Reading

Banerjee S, Iglewicz B (2007) A simple univariate outlier identification procedure designed for large samples. Commun Stat Simul Comput 36:249–263

Benjamini Y (1988) Opening the box of a boxplot. Am Stat 42: 257–262

Frigge M, Hoaglin DC, Iglewicz B (1989) Some Implementations of the boxplot. Am Stat 43:50–54
Goldberg KM, Iglewicz B (1992) Bivariate extensions of the boxplot. Technometrics 34:307–320
Hintze J, Nelson RD (1998) Violin plots: a boxplot–density trace synergism. Am Stat 52:181–184
Hoaglin DC, Iglewicz B (1987) Fine-tuning some resistant rules for outlier labeling. J Am Stat Assoc 81:1147–1149
Hoaglin DC, Iglewicz B, Tukey JW (1986) Performance of some resistant rules for outlier labeling. J Am Stat Assoc 81:991–999
Iglewicz B, Banerjee S (2001) A simple univariate outlier identification procedure. In: Proceedings of the annual meeting of the american statistical association
McGill R, Tukey JW, Larson WA (1978) Variations of the box plots. Am Stat 32:12–16
Rousseeuw PJ, Ruts I, Tukey JW (1999) The Bagplot: a bivariate boxplot. Am Stat 53:382–387
Sim CH, Gan FF, Chang TC (2005) Outlier labeling with boxplot procedures. J Am Stat Assoc 100:642–652
Tufte E (1983) The visual display of quantitative information. Graphic Press, Cheshire
Tukey JW (1977) Introductory data analysis. Addison-Wesley, Reading

Superpopulation Models in Survey Sampling

Gad Nathan
Professor Emeritus
Hebrew University of Jerusalem, Jerusalem, Israel

Classical sampling theory considers a finite population, $U = \{1, \ldots, N\}$, of known size, N, with a vector of fixed unknown values of a variable of interest, $\boldsymbol{y} = (y_1, \ldots, y_N)$. A sample of size n, $s = \{s_{i_1}, \ldots, s_{i_n}\}$, is selected by a sample design, which assigns to each possible sub-set of U a known probability – $p(s)$. The objective is to estimate some function of $\boldsymbol{y}$, which can be assumed, without loss of generality, to be the population total, $\boldsymbol{y} = \sum_{i=1}^{N} y_i$, on the basis of the sample observations, $\{y_{i_1}, \ldots, y_{i_n}\}$, and the sample probabilities – $p(s)$. Inference based only on the sample selection probabilities is known as *design based* (or ►randomization) inference and the properties of estimators are considered in this framework solely with respect to the known sample selection probabilities. Although design-based inference is widely applied in practice for the estimation of finite population parameters, it suffers from several drawbacks:

1. It can be shown that there is no unbiased estimator, say of the total, which is optimal, in the sense that its randomization variance is minimal for all sets of possible values of the population variables (Godambe 1955).
2. While the use of auxiliary data, e.g., known values of an auxiliary variable for all population units, $\boldsymbol{X} = (X_1, \ldots, X_N)$, for sample design (e.g., stratification) or for estimation (e.g., ratio estimation) is widely applied, in practice, it cannot strictly be justified under the design-based paradigm, unless some model relating the values of X and Y is assumed. For instance the efficiency of ratio estimation is based on the premise that there is a linear relationship between the values of X and Y (without an intercept) – Cochran (1977).
3. The use of sample survey data for analytical purposes, which has developed extensively over the past few decades, cannot be treated on a solid theoretical basis solely under design-based inference -see e.g., Kish and Frankel (1974). Thus, although a regression analysis can formally be carried out on sample data, the results cannot be interpreted easily when the dependent and the independent variables are considered as fixed values, rather than as realizations of random variables, i.e., unless a linear model with random errors is assumed - Brewer and Mellor (1973).

This has led sample survey theoreticians and practitioners to consider a *model based*, or *superpopulation* approach, which assumes that each population unit is associated with a random variable for which a stochastic structure is specified and the actual value associated with a population unit is considered as the realization of the random variable, rather than a fixed unknown value - Cassel et al. 1976. Thus the vector of population values, $\boldsymbol{y}$, is assumed to be the realization of a random vector variable: $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$. The form of the joint distribution of $Y_1, \ldots, Y_N$, often denoted by ξ, is usually assumed to be known, except for unknown parameters. Thus, if we assume a regression model between $\boldsymbol{Y}$ and $\boldsymbol{X}$, we might consider ξ as multivariate normal, i.e., $\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \Sigma)$, where $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown parameters.

There are several different possible interpretations of the superpopulation concept, such as the following - see also Särndal et al. 1992:

1. The finite population may be considered as actually selected from a larger universe by a real world random mechanism or process. This would be the interpretation of a statistical model in the social sciences, such as econometric models. This is the approach usually used by practitioners who wish to analyze sample survey data created by complex sample designs – see, for

instance, Nathan and Holt (1980), Skinner et al. (1989), Pfeffermann (1993) and Chambers and Skinner (2003).
2. The superpopulation joint distribution, ξ, may be considered under a Bayesian approach, as a prior distribution, which reflects the subjective belief in the unknown values of $Y_1, \ldots, Y_N$, so that we consider the problem of finding the posterior distribution of the finite population parameter, given the sample values.
3. The superpopulation distribution may be considered as reflecting nonsampling errors, such as measurement errors, which account for differences between observed values of the variables and their 'true' values.
4. The superpopulation distribution, ξ, may be considered as a purely mathematical device, not associated with any physical process or subjective belief, in order to make explicit theoretical derivations. Thus different estimators or sample designs may be considered and compared, with respect to their performance and characteristics (e.g., bias and variance), under different models. Since in most cases our certainty about the true models is very limited, this can provide a useful tool for checking the robustness of estimators and sample designs to departures from assumed models.

The rapid development of sample survey theory and practice over the past 50 years has occurred in all aspects of sample surveys. However the rapid integration of the superpopulation concept and model-based ideas in mainstream theory and practice of sample survey inference has been one of the major developments. Thirty five years ago, the fundamental divide between advocates of classical design-based inference and design, and those who preferred basing both the sample design and inference only on superpopulation models was still at its zenith and the controversies of the two previous decades, exemplified by Brewer and Mellor (1973), were still raging. The early randomization-based approach, developed by the pioneers of classical design-based sampling theory, was challenged by the study of the logical foundations of estimation theory in survey sampling, for example, Godambe (1955), and by early advocates of pure superpopulation model-based design and prediction approach to inference, for example, Royall (1970). These controversies continued to be fiercely discussed well into the 1980s, see, for example, Hansen et al. (1983), and pure superpopulation based prediction approaches are still being advocated – see Valliant (2000). However the extreme views, relating to both approaches, have mellowed considerably over the past 2 decades, and sample survey theory and practice are currently, by and large, based on a variety of combined approaches, such as model-assisted methods, which integrate superpopulation models with a randomization-based approach – see for example the variety of approaches, many of them based on superpopulation models, used in the latest *Handbook of Statistics* (volume 29), devoted to sample surveys – Rao and Pfeffermann (2009).

The superpopulation concept has served and continues to serve as an extremely important and useful tool for the development of the theory and practice of sample surveys – in their design, estimation and analysis.

About the Author

Dr. Gad Nathan is Professor Emeritus, Department of Statistics, Hebrew University, Jerusalem (since 2002). He is Past President of the Israel Statistical Association (1991–1993). Professor Nathan was Chair, Department of Statistics, Hebrew University, Jerusalem (1974–1977, and 1988–1991). He was also Director, Statistical Methods Division, Central Bureau of Statistics, Jerusalem, (1964–1969), Chief Scientist, Central Bureau of Statistics (Part-time, 1995–2001), Vice-President, International Statistical Institute, (1981–1983), and Vice-President, International Association of Survey Statisticians (1999–2001). He is Elected Member of the International Statistical Institute (1977) and Elected Fellow of the American Statistical Association (1978). He has (co-)authored about 75 publications.

Cross References

►Model Selection
►Multivariate Normal Distributions
►Nonsampling Errors in Surveys
►Random Variable
►Randomization
►Sample Survey Methods
►Sampling From Finite Populations
►Small Area Estimation

References and Further Reading

Brewer KRW, Mellor RW (1973) The effect of sample structure on analytical surveys. Aust J Stat 15:145–152
Chambers RL, Skinner CJ (eds) (2003) Analysis of survey data. Wiley, New York
Cassel CM, Särndal CE, Wretman JH (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63:615–620
Cochran WG (1977) Sampling techniques 3rd edn. Wiley, New York
Godambe VP (1955) A unified theory of sampling from finite populations. J R Stat Soc B 17:269–278
Hansen MH, Madow WG, Tepping BJ (1983) An evaluation of model-dependent and probability-sampling inferences in sample surveys. J Am Stat Assoc 78:776–793
Kish L, Frankel M (1974) Inference from complex samples. J R Stat Soc B 36:1–37

Nathan G, Holt D (1980) The effect of survey design on regression analysis. J R Stat Soc B 43:377–386
Pfeffermann D (1993) The role of sampling weights when modeling survey data. Int Stat Rev 61:317–337
Rao CR, Pfeffermann D (eds) (2009) Handbook of statistics, 29: sample surveys: theory, methods and inference. Elsevier, Amsterdam
Royall RM (1970) On finite population sampling theory under certain linear regression models. Biometrika 57:377–387
Särndal CE, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer, New York
Skinner CJ, Holt D, Smith TMF (eds) (1989) Analysis of complex surveys. Wiley, Chichester
Valliant R, Dorfman AH, Royall RM (2000) Finite population sampling and inference: a prediction approach. Wiley, Chichester/New York

Surveillance

Marianne Frisén
Professor
University of Gothenburg, Gothenburg, Sweden

The Need for Statistical Surveillance

The aim of statistical surveillance is the timely detection of important changes in the process that generates the data. Already at birth surveillance is used, as described by Frisén (1992). The baby might get the umbilical cord around the neck at any time during labour. This will cause a lack of oxygen, and a Caesarean section is urgent. The electrical signal of the heart of the baby during labour is the base for the surveillance system. Detection has to be made as soon as possible to ensure that the baby is delivered without brain damage.

Around 1930, Walter A. Shewhart developed the first versions of sequential surveillance by introducing control charts for industrial applications (see ►Control Charts). Although industrial applications are still important, many new applications have come into focus.

In finance, transaction strategies are of great interest and the timeliness of transactions is important. Most theory of stochastic finance is based on the assumption of an efficient market. When the stochastic model is assumed to be completely known, we can use probability theory to calculate the optimal transaction conditions. When the information about the process is incomplete, as for example when a change can occur in the process, there may be an arbitrage opportunity, as demonstrated by Shiryaev (2002). In these situations, observations should be analysed continuously to decide whether a transaction at that time is profitable as measured either by return or by risk. Statistical inference is needed for the decision. Different aspects of the subject of financial surveillance are described in the book edited by Frisén (2007). There are also other applications in the field of economics. The *detection of turning points in business cycles* is important for both government and industry.

In public health surveillance, the timely detection of various types of adverse health events is crucial. The monitoring of incidences of different diseases and symptoms is carried out by international, national and local authorities to detect outbreaks of infectious diseases. Epidemics, such as influenza, are for several reasons very costly to society, and it is therefore of great value to monitor influenza data, both for the outbreak detection and during the epidemic period in order to allocate medical resources. Methods for surveillance for common diseases also serve as models for the detection of new diseases as well as for detecting bioterrorism. Surveillance for the onset of an outbreak is described in Frisén et al. (2009). Reviews of methods for the surveillance of public health are given by Sonesson and Bock (2003) and Woodall et al. (2008).

The Statistical Surveillance Problem

Terminology

The terminology is diverse. "Optimal stopping rules" (see ►Optimal Stopping Rules) is most often used in probability theory, especially in connection with financial problems. Literature on "change-point problems" does not always treat the case of sequentially obtained observations but often refers to the retrospective analysis of a fixed number of observations. The term "early warning system" is sometimes used in economic and medical literature. "Monitoring" is most often used in medical literature and with a broad meaning. The notations "statistical process control" and "quality control" are used in the literature on industrial production.

Overviews

Surveys and bibliographies on statistical surveillance are given for example by Lai (1995), who gives a full treatment of the field but concentrates on the minimax properties of stopping rules, by Woodall and Montgomery (1999) and Ryan (2000), who concentrate on control charts, and by Frisén (2003), who characterises methods by different optimality properties. The overview by Frisén (2009) and the adjoining discussion takes up many recent issues.

Differences between hypothesis testing and surveillance

In the initial example, the decision concerning whether the baby is at risk has to be made sequentially, based on the data collected so far. Each new time demands a new decision. There is no fixed data set but an increasing number of observations. In sequential hypothesis testing, we have sequentially obtained observations and repeated decisions, but the hypotheses are fixed. In contrast, there are no fixed hypotheses in surveillance. We can never accept any null hypotheses and turn our backs on the mother, since the baby might get the umbilical cord around the neck in the next minute.

Statistical specifications

We denote the process by $X = \{X(t) : t = 1, 2, \ldots\}$, where $X(t)$ is the observation (vector) made at time t, which is usually discrete. The purpose of the monitoring is to detect a possible change, for example the change in distribution of the observations due to the baby's lack of oxygen. The time of the change is denoted by τ. Before the change, the distribution belongs to the family f^D, and after the time τ, the distribution belongs to the family f^C. At each decision time s, we want to discriminate between two events, $C(s)$ and $D(s)$. For most applications, these can be further specified as $C(s) = \{\tau \leq s\}$ (a change has occurred) and $D(s) = \{\tau > s\}$ (no change has occurred yet), respectively.

We use the observations $X_s = \{X(t); t \leq s\}$ to form an alarm criterion which, when fulfilled, is an indication that the process is in state $C(s)$, and an alarm is triggered. We use an alarm statistic, $p(X_s)$, and a control limit, $G(s)$, and the alarm time, t_A, is $t_A = \min\{s; p(X_s) > G(s)\}$. The change point τ can be regarded either as a random variable or as a deterministic but unknown value, depending on what is most relevant for the application.

Evaluation and Optimality

Quick detection and few false alarms are desired properties of methods for surveillance. Different error rates and their implications for active and passive surveillance were discussed by Frisén and de Maré (1991).

Evaluation by significance level, power, specificity, sensitivity, or other well-known metrics may seem convenient. However, these are not easily interpreted in a surveillance situation. For example, when the surveillance continues, the specificity will tend to zero for most surveillance methods. Thus, there is not one unique specificity value in a surveillance situation.

Special metrics such as the expected time to a false alarm ARL^0 and the expected delay of a warranted alarm are used (see Frisén (1992)). The expected delay is different for early changes as compared with late ones. The most commonly used delay measure is ARL^1, the expected delay for a change that appears at the start of the surveillance.

In addition, the optimality criteria are different in surveillance as compared with hypothesis testing. The minimax optimality and the expected delay over the distribution of the change point are frequently used.

Methods

In surveillance, it is important to aggregate the sequentially obtained information in order to take advantage of all information. Different ways of aggregation meet different optimality criteria. Expressing methods for surveillance through likelihood functions makes it possible to link the methods to various optimality criteria. Many methods for surveillance can be expressed by a combination of partial likelihood ratios (Frisén (2003)). The likelihood ratio for a fixed value of τ is $L(s,t) = f_{X_s}(x_s|\tau = t)/f_{X_s}(x_s|D)$. The exact formula for these likelihood components will vary between situations.

The full likelihood ratio method (LR) can be expressed as a weighted sum of the partial likelihoods $L(s,t)$. It is optimal with respect to the criterion of minimal expected delay, as demonstrated by Shiryaev (1963).

The simplest way to aggregate the likelihood components is to add them. Shiryaev (1963) and Roberts (1966) suggested what is now called the Shiryaev-Roberts method. This means that all possible change times, up to the decision time s, are given equal weight.

The method by Shewhart (1931) is simple and the most commonly used method for surveillance. An alarm is given as soon as an observation deviates too much from the target. Thus, only the last observation is considered. The alarm criterion can be expressed by the condition $L(s,s) > G$, where G is a constant.

The CUSUM method was first suggested by Page (1954). The alarm condition of the method can be expressed by the partial likelihood ratios as $t_A = \min\{s; \max(L(s,t); t = 1, 2, \ldots, s) > G\}$, where G is a constant. The CUSUM method satisfies the minimax criterion of optimality, as proved by Moustakides (1986).

The alarm statistic of the EWMA method is an exponentially weighted moving average, $Z_s = (1 - \lambda)Z_{s-1} + \lambda X(s)$, $s = 1, 2, \ldots$ where $0 < \lambda < 1$ and Z_0 is the target value. The EWMA method was described by Roberts (1959).

Complex Situations

Applications contain complexities such as autocorrelations, complex distributions, complex types of changes and

spatial as well as other multivariate settings. Thus, the basic surveillance theory has to be adapted to special cases.

Time series with special dependencies have been treated for example by Basseville and Nikiforov (1993), Schmid (1997) and Lai (1998). Surveillance for special distributions such as, for example, discrete ones were discussed for example by Woodall (1997). Complex changes such as gradual ones from an unknown baseline are of interest at the outbreak of influenza or other diseases. The maximal partial maximum likelihood will give a CUSUM variant. This was used for semiparametric surveillance by Frisén et al. (2009).

Multivariate surveillance is of interest in many areas. In industry, the monitoring of several components in an assembly process requires multivariate surveillance. An example in finance is the on-line decisions on the optimal portfolio of stocks, as described by Okhrin and Schmid (2007). The surveillance of several distribution parameters, such as the mean and the variance (see e.g., Knoth and Schmid (2002)), is another example of multivariate surveillance.

In spatial surveillance, observations are made at different locations. Most methods for spatial surveillance are aimed at detecting spatial clusters, but other relations between the variables can also be of interest. The surveillance of a set of variables for different locations is a special case of multivariate surveillance, as discussed by Sonesson and Frisén (2005) and Sonesson (2007).

About the Author

Dr. Marianne Frisén is Professor of Statistics at University of Gothenburg. She is an Elected member of the ISI. Professor Frisén has has been working in the area of surveillance for over 25 years. She has organized symposiums on financial surveillance, written numerous publications on surveillance, including the text *Financial Surveillance* (Wiley, 2007), the first book-length treatment of statistical surveillance methods used in financial analysis.

Cross References

►Detection of Turning Points in Business Cycles
►Optimal Stopping Rules
►Relationship Between Statistical and Engineering Process Control
►Sequential Probability Ratio Test
►Sequential Sampling
►Significance Testing: An Overview

References and Further Reading

Basseville M, Nikiforov I (1993) Detection of abrupt changes: theory and application. Prentice Hall, Englewood Cliffs

Frisén M (1992) Evaluations of methods for statistical surveillance. Stat Med 11:1489–1502

Frisén M (2003) Statistical surveillance. Optimality and methods. Int Stat Rev 71:403–434

Frisén M (2007) Financial surveillance, edited volume. Wiley, Chichester

Frisén M (2009) Optimal sequential surveillance for finance, public health and other areas. Editor's special invited paper. Sequential Anal 28:310–337, discussion 338–393

Frisén M, de Maré J (1991) Optimal surveillance. Biometrika 78: 271–280

Frisén M, Andersson E, Schiöler L (2009) Robust outbreak surveillance of epidemics in Sweden. Stat Med 28:476–493

Knoth S, Schmid W (2002) Monitoring the mean and the variance of a stationary process. Stat Neerl 56:77–100

Lai TL (1998) Information bounds and quick detection of parameters in stochastic systems. IEEE Trans Inform Theor 44: 2917–2929

Moustakides GV (1986) Optimal stopping times for detecting changes in distributions. Ann Stat 14:1379–1387

Okhrin Y, Schmid W (2007) Surveillance of univariate and multivariate nonlinear time series. In: Frisén M (ed) Financial surveillance. Wiley, Chichester, pp 153–177

Page ES (1954) Continuous inspection schemes. Biometrika 41: 100–114

Roberts SW (1959) Control chart tests based on geometric moving averages. Technometrics 1:239–250

Roberts SW (1966) A Comparison of some control chart procedures. Technometrics 8:411–430

Ryan TP (2000) Statistical methods for quality improvement. Wiley, New York

Schmid W (1997) Cusum control schemes for Gaussian processes. Stat Pap 38:191–217

Shewhart WA (1931) Economic control of quality of manufactured product. MacMillan, London

Shiryaev AN (1963) On optimum methods in quickest detection problems. Theor Probab Appl 8:22–46

Shiryaev AN (2002) Quickest detection problems in the technical analysis of financial data. In: Geman H, Madan D, Pliska S, Vorst T (eds) Mathematical finance – bachelier congress 2000. Springer, Berlin, pp 487–521

Sonesson C, Bock D (2003) A review and discussion of Prospective statistical surveillance in public health, JR Stat Soc A 166:5–21

Sonesson C (2007) A cusum framework for detection of space-time disease clusters using scan statistics. Stat Med 26: 4770–4789

Sonesson C, Frisén M (2005) Multivariate surveillance. In: Lawson A, Kleinman K (eds) Spatial surveillance for public health. Wiley, New York, pp 169–186

Woodall WH (1997) Control charts based on attribute data: bibliography and review. J Qual Technol 29:172–183

Woodall WH, Montgomery DC (1999) Research issues and ideas in statistical process control. J Qual Technol 31: 376–386

Woodall WH, Marshall JB, Joner JMD, Fraker SE, Abdel-Salam ASG (2008) On the use and evaluation of prospective scan methods for health-related surveillance. J R Stat Soc A 171: 223–237

Sonesson, C. and Bock, D. (2003) A review and discussion of prospective statistical surveillance in public health. *J R Stat Soc A 166*: 5–21

Survival Data

D. R. Cox
Honorary Fellow
Nuffield College, Oxford, UK

Preliminaries

The most immediate examples of survival data come from ▶demography and actuarial science and concern the duration of human life. The issues of statistical analysis that arise are similar to those in many fields. Thus survival time may be the length of time before a piece of industrial equipment fails, the length of time before a firm becomes bankrupt, the duration of a period of employment or, particularly in a medical or epidemiological context, the time between diagnosis of a specific condition and death from that condition.

Depending on the perspective involved the term failure time may be used instead of survival time.

Central requirements are that for each study individual we have a clear time origin and a clear end point. For example, time may be measured from the instant an individual enters the study population and the end point may be death from a specific cause, or death (all causes) or cure. Normally the passage of time is clearly defined in the natural way. There may be other possibilities, for example the investigation of tire life in terms of km driven. In applications considerable care is needed over these definitions, ensuring that they are precise and relevant.

A common characteristic of such data is that the frequency distributions are widely dispersed with positive skewness. Another is the presence of right censoring. That is for some, or in some cases, for many individuals, all that is known is that by the end of the study the critical event in question has not occurred, implying that the survival time in question exceeds some given value. In industrial life testing censoring may be by design but more commonly it is just a feature of the data acquisition process.

Formalization

We represent survival time by a random variable T, treated for simplicity as continuously distributed; there is a closely parallel discussion for discrete random variables.

For a given population of individuals the distribution of T can be described in several mutually equivalent ways, for example by

- the survivor function

$$S(t) = P(T > t), \tag{1}$$

- the probability density function

$$f(t) = -S'(t) \tag{2}$$

- the hazard or age-specific failure rate

$$h(t) = f(t)/S(t) = -\frac{d}{dt}\log S(t). \tag{3}$$

A more interpretable specification of the hazard at time t is as a failure rate conditional on survival to time t, that is as

$$\lim P(T < t + \delta \mid t < T)/\delta$$

as δ tends to zero through positive values.

These three specifications are mathematically equivalent; all have their uses in applications.

A central role is played in some parts of the subject by the exponential distribution of rate ρ and mean $1/\rho$, namely the special case

$$S(t) = e^{-\rho t}, \quad f(t) = \rho e^{-\rho t}, \quad h(t) = \rho. \tag{4}$$

The last property shows that failure occurs at random with respect to "age." If $h(t)$ increases with t there is ageing whereas if $h(t)$ decreases with t then in a certain sense old is better than new. There are other possibilities, in particular a bath-tub effect in which high initial values are followed by a decrease followed in turn by a gradual increase.

Many other forms may be used in applications, notably the ▶Weibull distribution with $h(t) = \rho(\rho t)^{\gamma}$.

Statistical Analysis

For n independent individuals from a homogenous population it is convenient to write the data in the form

$$(t_1, d_1), \ldots, (t_n, d_n). \tag{5}$$

Here for individual j, t_j is a time and if $d_j = 1$ this is the relevant value of T whereas if $d_j = 0$ the individual is right censored. This is interpreted to mean that all we know about the value of T for that individual exceeds t_j, a non-trivial assumption implying what is rather misleadingly called uninformative censoring. It excludes for example the deliberate or unwitting withdrawal of individuals from a study because of a presumption of imminent failure.

There are two broad approaches to analysis, parametric based on an assumed form for the distribution, and non-parametric.

The former is typically tackled by the method of maximum likelihood. Let θ denote the parameter specifying the distribution, for example ρ for the exponential distribution and (ρ, γ) for the Weibull distribution. Then the

likelihood is

$$\Pi\{f(t_j;\theta\}^{d_j}\{S(t_j;\theta)\}^{1-d_j}. \quad (6)$$

That is, each failed individual contributes a term depending on the density whereas each censored individual contributes a term depending on the survivor function. The method of maximum likelihood may now be applied (or a Bayesian posterior density calculated).

For the exponential distribution the likelihood takes the form

$$\rho^{\Sigma d_j}\exp(-\rho\Sigma t_j). \quad (7)$$

where Σd_j is the total number of failures. It follows that the maximum likelihood estimate of ρ, obtained by maximizing this expression with respect to ρ, is

$$\frac{\Sigma d_j}{\Sigma t_j}, \quad (8)$$

that is, the total number of failures divided by the total time at risk calculated from all individuals those who fail and those who are censored. This is sometimes called the fundamental theorem of epidemiology.

For a nonparametric analysis a limiting form of a life-table approach is used called the Kaplan-Meier method. Essentially the hazard is estimated as zero at all times at which failures do not occur and as the number of failures divided by the number at risk of failure at times at which failure does occur. The estimated survivor function is reconstructed from this by a discrete version of (3). If required, estimates of, say, the median survival time can be found by interpolation, assuming that sufficient failures have occurred to allow this part of the distribution to be estimated effectively.

Dependencies and Comparisons

Often there are more than a single group of observations and comparisons are required, say between groups of individuals treated differently. In simple cases this can be achieved either by comparing parameters in parametric models fitted separately to the different groups or by graphical comparison of the Kaplan-Meier estimates (see ►Kaplan-Meier Estimator).

In more complicated cases, for example when several explanatory variables are addressed simultaneously, models analogous to regression models are helpful. The most widely used of these is the proportional hazards model. For each individual we suppose available a vector z of explanatory variables and that the corresponding hazard function is

$$h_0(t)\exp(\beta^T z). \quad (9)$$

Here $h_0(t)$, called the baseline hazard, specifies the hazard for a reference individual with $z = 0$.

A typical example with critical event death from cardio-vascular causes might have z_1, age at entry, z_2, systolic blood pressure at entry, both typically measured from some reference level, z_3, zero for men, one for women and z_4, zero for control and one for a new drug under test. A component of β, say the first component β_1, specifies the increase in hazard per unit increase in the component z_1 of z, with all other components of z held fixed. That is for fixed gender, treatment and blood pressure the hazard increases by a factor e^{β_1} per extra year of age.

If the baseline hazard is constant or specified parametrically maximum likelihood estimation is possible, essentially generalizing (3). For example if $h_0(t)$ is an unknown constant, a baseline individual has an exponential distribution. If $h_0(t)$ is left arbitrary a modified form of likelihood-based inference is used called partial likelihood. Problems of interpretation, model choice, etc., are essentially the same as in multiple linear regression. An important possibility is that some components of z may be functions of time.

Generalizations and Literature

There are many generalizations of these ideas of which the most notable is to event-history analysis in which a sequence of events, possibly of different types, may occur on each individual.

There is a very extensive literature, some of it specific to application fields. Cox and Oakes (1984) give a broad introduction and Kalbfleisch and Prentice (2002) a more specialized and thorough account. For a discussion with attention to mathematical detail, see Andersen et al. (1993) and Aalen et al. (2008).

About the Author

Sir David Cox is among the most important statisticians of the past half-century. He has made major contributions to statistical theory, methods, and applications. He has written or co-authored 18 books and more than 300 papers, many of which are seminal works. He was editor of *Biometrika* for 25 years (1966–1991). Professor Cox was elected a Fellow of the Royal Society (F.R.S.) in 1973, knighted by Queen Elizabeth II in 1985 and became an Honorary Fellow of the British Academy in 1997. He has served as President of the Bernoulli Society (1979–1981), of the Royal Statistical Society (1980–1982), and of the International Statistical Institute (1995–1997). He is a Fellow of the Royal Danish Academy of Sciences, of the Indian Academy of Sciences and of the Royal Society of Canada and a Foreign Associate of the US National Academy of

Sciences. He has been awarded the Guy Medal in Silver, Royal Statistical Society (1961), Guy Medal in Gold, Royal Statistical Society (1973), Weldon Memorial Prize, University of Oxford (1984), Kettering Prize and Gold Medal for Cancer Research (1990), Marvin Zelen Leadership Award, Harvard University (1998). In 2010 he received the Copley Medal of the Royal Society. He has supervised or been associated with more than 60 doctoral students, many of whom have become leading researchers themselves (including a number of authors of this Encyclopedia: Anthony Atkinson, Adelchi Azzalini, Gauss Cordeiro, Vern Farewell, Roderick Little, Francisco Louzada-Neto, Peter McCullagh, and Basilio Pereira). Professor Cox holds 21 honorary doctorates, the last one from the University of Gothenburg (2007).

"His outstanding contributions to the theory and applications of statistics have a pervasive influence. For instance, the introduction of what is nowadays called 'Cox regression' in survival analysis has started a research area with numerous books and thousands of papers on statistical theory and on statistical practice. It has changed the way in which survival studies in medicine and technology are performed and evaluated." (Inauguration of Doctors Ceremony, University of Gothenburg, October 20, 2007).

Cross References

►Bayesian Semiparametric Regression
►Censoring Methodology
►Degradation Models in Reliability and Survival Analysis
►Demographic Analysis: A Stochastic Approach
►Event History Analysis
►First-Hitting-Time Based Threshold Regression
►Frailty Model
►Generalized Weibull Distributions
►Hazard Ratio Estimator
►Hazard Regression Models
►Kaplan-Meier Estimator
►Life Table
►Logistic Distribution
►Medical Research, Statistics in
►Modeling Survival Data
►Population Projections
►Statistical Inference in Ecology
►Testing Exponentiality of Distribution
►Weibull Distribution

References and Further Reading

Aalen OO, Borgan O, Gjessing HK (2008) Survival and event history analysis. Springer, New York

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data. 2nd edn. Wiley, New York