M

Margin of Error

Judith M. Tanur
Distinguished Teaching Professor Emerita
Stony Brook University, Stony Brook, NY, USA

Margin of error is a term that probably originated in the popular reporting of results of ▶public opinion polls but has made its way into more professional usage. It usually represents half of the length of a confidence interval (most usually a 95% confidence interval, though it could in theory be any confidence interval) for a proportion or percentage, calculated under the assumption of simple random sampling. The sample value of the proportion, $\hat{p}$, is used as an estimate of the population proportion π, and the standard error (se) is estimated as $\sqrt{\hat{p}(1-\hat{p})/n}$. Then a 95% confidence interval is given as $\hat{p} \pm 1.96\times$ se and the margin of error is $1.96\times$ se. For example, if an opinion poll gives a result of 40% of 900 respondents in favor of a proposition (a proportion of .40), then the estimated se of the proportion is $\sqrt{(0.4 \times 0.6)/900} = .016$ and that is expressed as 1.6 percentage points. Then the margin of error would be presented as $1.96 \times 1.6 = 3.2$ percentage points.

The fact that the margin of error is often reported in the popular press represents progress from a time when sample results were not qualified at all by notions of sample-to-sample variability. Such reporting, however, is frequently subject to misinterpretation, though reporters often caution against such misinterpretation. First, like the confidence interval, the margin of error does not represent anything about the probability that the results are close to truth. A 95% confidence interval merely says that, with the procedure as carried out repeatedly by drawing a sample from this population, 95% of the time the stated interval would cover the true population parameter. There is no information whether this current interval does or does not cover the population parameter and similarly the margin of error gives no information whether it covers the true population percentage. Second, the procedure assumes simple random sampling, but frequently the sampling for a survey is more complicated than that and hence the standard error calculated under the assumption of simple random sampling is an underestimate. Third, the margin of error is frequently calculated for the sample as a whole, but when interest centers on a subgroup of respondents (e.g., the percentage of females who prefer a particular candidate) the sample size is smaller and a fresh margin of error should be calculated for the subgroup, though it frequently is not. And finally, and perhaps most importantly, there is a tendency to assume that the margin of error takes into account all possible "errors" when in fact it deals only with sampling error. Nonsampling errors, such as noncoverage, nonresponse, or inaccurate responses are not taken into account via a confidence interval or the margin of error and may indeed be of much larger magnitude than the sampling error measured by the standard error.

About the Author

For biography *see* the entry ▶Nonsampling Errors in Surveys.

Cross References

▶Confidence Interval
▶Estimation
▶Estimation: An Overview
▶Public Opinion Polls

Marginal Probability: Its Use in Bayesian Statistics as Model Evidence

Luis Raúl Pericchi
Professor
University of Puerto Rico, San Juan, Puerto Rico

Definition

Suppose that we have vectors of random variables $[\mathbf{v},\mathbf{w}] = [v_1, v_2, \ldots, v_I, w_1, \ldots, w_J]$ in $\Re^{(I+J)}$. Denote as the **joint** density function: $f_{\mathbf{v},\mathbf{w}}$, which obeys: $f_{\mathbf{v},\mathbf{w}}(v, w) \geq 0$ and

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

$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{v},\mathbf{w}}(v,w)dv_1 \dots dv_I dw_1 \dots dw_I = 1$. Then the probability of the set $[A_v, B_w]$ is given by

$$P(A_v, B_w) = \int \dots \int_{A_v,B_w} f_{\mathbf{v},\mathbf{w}}(v,w)\mathbf{dvdw}.$$

The *marginal* density $f_{\mathbf{v}}$ is obtained as

$$f_{\mathbf{v}}(v) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{v},\mathbf{w}}(v,w)dw_1 \dots dw_I.$$

The *marginal probability* of the set A_v is then obtained as,

$$P(A_v) = \int \dots \int_{A_v} f_{\mathbf{v}}(v)dv.$$

We have assumed that the random variables are continuous. When they are discrete, integrals are substituted by sums. We proceed to present an important application of marginal probabilities for measuring the probability of a model.

Measuring the Evidence in Favor of a Model

In Statistics, a parametric model, is denoted as $f(x_1, \dots, x_n|\theta_1, \dots, \theta_k)$, where $\mathbf{x} = (x_1, \dots, x_n)$ is the vector of n observations and $\theta = (\theta_1, \dots, \theta_k)$ is the vector of k parameters. For instance we may have $n = 15$ observations normally distributed and the vector of parameters is (θ_1, θ_2) the location and scale respectively, denoted by $f_{Normal}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta_2} \exp\left(-\frac{1}{2\theta_2^2}(\mathbf{x}_i - \theta_1)^2\right)$.

Assume now that there is reason to suspect that the location is zero. As a second example, it may be suspected that the sampling model which usually has been assumed Normally distributed, is instead a Cauchy, $f_{Cauchy}(X|\theta) = \prod_1^n \frac{1}{\pi\theta_2} \frac{1}{\left(1+\left(\frac{x_i-\theta_1}{\theta_2}\right)^2\right)}$. The first problem is a *hypothesis test* denoted by

$$H_0 : \theta_1 = 0 \text{ VS } H_1 : \theta_1 \neq 0,$$

and the second problem is a *model selection* problem:

$$M_0 : f_{Normal} \text{ VS } M_1 : f_{Cauchy}.$$

How to measure the evidence in favor of H_0 or M_0? Instead of maximized likelihoods as it is done in traditional statistics, in ►Bayesian statistics the central concept is *the evidence* or *marginal probability density*

$$m_j(\mathbf{x}) = \int f_j(\mathbf{x}|\theta_j)\pi(\theta_j)\mathbf{d}\theta_j,$$

where j denotes either model or hypothesis j and $\pi(\theta)$ denotes the prior for the parameters under model or hypothesis j.

Marginal probabilities embodies the likelihood of a model or hypothesis in great generality and can be claimed it is the natural probabilistic quantity to compare models.

Marginal Probability of a Model

Once the marginal densities of the model j, for $j = 1, \dots, J$ models have been calculated and assuming the prior model probabilities $P(M_j), j = 1, \dots, J$ with $\sum_{j=1}^{J} P(M_j) = 1$ then, using Bayes Theorem, *the marginal probability of a model* $P(M_j|\mathbf{x})$ can be calculated as,

$$P(M_j|\mathbf{x}) = \frac{\mathbf{m_j}(\mathbf{x}) \cdot \mathbf{P(M_j)}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{m_i}(\mathbf{x}) \cdot \mathbf{P(M_i)}}.$$

We have then the following formula for any two models or hypotheses:

$$\frac{P(M_j|\mathbf{x})}{P(M_i|\mathbf{x})} = \frac{P(M_j)}{P(M_i)} \times \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

or in words: Posterior Odds equals Prior Odds times Bayes Factor, where the Bayes Factor of M_j over M_i is

$$B_{j,i} = \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

Jeffreys (1961).

In contrast to ►*p-values*, which have interpretations heavily dependent on the sample size n, and its definition is not the same as the scientific question, the posterior probabilities and Bayes Factors address the scientific question: "how probable is model or hypothesis j as compared with model or hypothesis i?," and the interpretation is the same for any sample size, Berger and Pericchi (2001). Bayes Factors and Marginal Posterior Model Probabilities have several advantages, like for example large sample consistency, that is as the sample size grows the Posterior Model Probability of the sampling model tends to one. Furthermore, if the goal is to predict future observations y_f it is **not** necessary to select one model as *the* predicting model since we may predict by the so called Bayesian Model Averaging, which if quadratic loss is assumed, the optimal predictor takes the form,

$$E[Y_f|\mathbf{x}] = \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{E}[\mathbf{Y_f}|\mathbf{x}, \mathbf{M_j}] \times \mathbf{P(M_j}|\mathbf{x)},$$

where $E[Y_f|\mathbf{x}, \mathbf{M_j}]$ is the expected value of a future observation under the model or hypothesis M_j.

Intrinsic Priors for Model Selection and Hypothesis Testing

Having said some of the advantages of the marginal probabilities of models, the question arises: how to assign the conditional priors $\pi(\theta_j)$? In the two examples above which priors are sensible to use? The problem is *not* a simple one since it is not possible to use the usual Uniform priors since then the Bayes Factors are undetermined. To solve this problem with some generality, Berger and Pericchi (1996)

introduced the concepts of Intrinsic Bayes Factors and Intrinsic Priors. Start by splitting the sample in two subsamples $\mathbf{x} = [\mathbf{x}(\mathbf{l}), \mathbf{x}(-\mathbf{l})]$ where the training sample $\mathbf{x}(\mathbf{l})$ is as small as possible such that for $j = 1, \ldots, J : 0 < m_j(\mathbf{x}(\mathbf{l})) < \infty$. Thus starting with an improper prior $\pi^N(\theta_j)$, which does not integrate to one (for example the Uniform), by using the minimal training sample $\mathbf{x}(\mathbf{l})$, all the conditional prior densities $\pi(\theta_j|\mathbf{x}(\mathbf{l}))$ *become* proper. So we may form the Bayes Factor using the training sample $\mathbf{x}(\mathbf{l})$ as

$$B_{ji}(\mathbf{x}(\mathbf{l})) = \frac{\mathbf{m_j}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}{\mathbf{m_i}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}.$$

This however depends on the particular training sample $\mathbf{x}(\mathbf{l})$. So some sort of average of Bayes Factor is necessary. In Berger and Pericchi (1996) it is shown that the average should be the arithmetic average. It is also found a theoretical prior that is an approximation to the procedure just described as the sample size grows. This is called an *Intrinsic Prior*. In the examples above: (i) in the normal case, assuming for simplicity that the variance is known and $\theta_2^2 = 1$ then it turns out that the Intrinsic Prior is Normal centered at the null hypothesis $\theta_1 = 0$ and with variance 2. On the other hand in the Normal versus Cauchy example, it turns out that the improper prior $\pi(\theta_1, \theta_2) = 1/\theta_2$ is the appropriate prior for comparing the models. For other examples of Intrinsic Priors see for instance, Berger and Pericchi (1996a,b, 2001), and Moreno et al. (1998).

About the Author

Luis Raúl Pericchi is Full Professor Department of Mathematics, College of Natural Sciences, University of Puerto Rico, Rio Piedras Campus, San Juan, and Director of the Biostatistics and Bioinformatics Core of the Comprehensive Cancer Center of the University of Puerto Rico. He received his Ph.D. in 1981, Imperial College, London (his supervisor was Professor A.C. Atkinson). He was Founder Coordinator of the Graduate Studies in Statistics (1997–2000) and Director of the Department of Mathematics (2001–2006). Professor Pericchi is Elected Member of the International Statistical Institute (1989) and Past President of the Latin American Chapter of the Bernoulli Society for Probability and Mathematical Statistics (1997–2000). Dr Pericchi was Associate Editor, *International Statistical Review* (1988–1991), Associate Editor of *Bayesian Analysis* (2006–2009). He is currently Associate Editor of the Brazilean Journal of Bayesian Analysis. He has (co)-authored more than 70 scientific articles.

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Inversion of Bayes' Formula for Events
►Model Selection
►Statistical Evidence

References and Further Reading

Berger JO, Pericchi LR (1996a) The intrinsic Bayes factor for model selection and Prediction. J Am Stat Assoc 91:109–122
Berger JO, Pericchi LR (1996b) The intrinsic Bayes factors for linear models. In: Bernardo JM et al (eds) Bayesian statistics 5. Oxford University Press, London, pp 23–42
Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison. IMS LectureNotes-Monograph Series 38:135–207
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, London
Moreno E, Bertolino F, Racugno W (1998) An intrinsic limiting procedure for model selection and hypothesis testing. J Am Stat Assoc 93(444):1451–1460

Marine Research, Statistics in

Gunnar Stefansson
Professor, Director of the Statistical Center
University of Iceland, Reykjavik, Iceland

Marine science is a wide field of research, including hydrography, chemistry, biological oceanography and fishery science. One may consider that the longer-term aspects of global warming and issues with pollution monitoring are the most critical statistical modeling issues. Somewhat subjectively, the next in line are probably issues which relate to the sustainable use of marine resources, commonly called fishery science. Statistics enters all of the above subfields but the most elaborate models have been developed for fishery science and aspects of these will mainly be described here. Within marine research it was quite common up through about 1980 to use models of the biological processes set up using differential equations, but had no error component and basically transformed observed data through an arbitrary computational mechanism into desired measures of population size, growth, yield potential and so forth (Baranov 1918; Beverton and Holt 1957; Gulland 1965).

Data in fishery science are quite noisy for several reasons. One source of variation is measurement error and one should expect considerable variability in data which

are almost always collected indirectly. Thus one cannot observe the marine community through simple population measurements but only with surveys (bottom trawl, divers etc) or sampling of catch, both of which will provide measures which only relate indirectly to the corresponding stock parameters, are often biased and always quite variable. The second source of variation is due to the biological processes themselves, all of which have natural variation. A typical such process is the recruitment process, i.e., the production of a new yearclass by the mature component of the stock in question. Even for biology, this process is incredibly variable and it is quite hard to extract meaningful signals out of the noise. Unfortunately this process is the single most important process with regard to sustainable utilization (Beverton and Holt 1957, 1993).

As is to be expected, noisy input data will lead to variation in estimates of stock sizes, productivity and predictions (Patterson et al. 2001). As is well-known to statisticians, it is therefore important not only to obtain point estimates but also estimates of variability. In addition to the general noise issue, fisheries data are almost never i.i.d. and examples show how ignoring this can easily lead to incorrect estimates of stock size, state of utilization and predictions (Myers and Cadigan 1995).

Bayesian approaches have been used to estimate stock sizes (Patterson 1999). A particular virtue of Bayesian analysis in this context is the potential to treat natural mortality more sensibly than in other models. The natural mortality rate, M, is traditionally treated as a constant in parametric models and it turns out that this is very hard to estimate unless data are quite exceptional. Thus, M is commonly assumed to be a known constant and different values are tested to evaluate the effect of different assumptions. The Bayesian approach simply sets a prior on the natural mortality like all other parameters and the resulting computations extend all the way into predictions. Other methods typically encounter problems in the prediction phase where it is difficult to encompass the uncertainty in M in the estimate of prediction uncertainty.

One approach to extracting general information on difficult biological parameters is to consider several stocks and even several species. For the stock-recruit question it is clear when many stocks are considered that the typical behavior is such that the stock tend to produce less at low stock sizes, but this signal can rarely be seen for individual stocks. Formalizing such analyses needs to include parameters (as random effects) for each stock and combining them reduces the noise enough to provide patterns which otherwise could not be seen (see e.g., Myers et al. 1999).

In addition to the overall view of sustainable use of resources, many smaller statistical models are commonly considered. For example, one can model growth alone, typically using a nonlinear model, sometimes incorporating environmental effects and/or random effects (Miller 1992; Taylor and Stefansson 1999; Brandão et al. 2004; Gudmundsson 2005).

Special efforts have been undertaken to make the use of nonlinear and/or random effects models easier for the user (Skaug 2002; Skaug and Fournier 2006). Although developed for fishery science, these are generic C++-based model-building languages which undertake automatic differentiation transparently to the user (Fournier 1996).

Most of the above models have been developed for "data-rich" scenarios but models designed for less informative data sets abound. Traditionally these include simple models which were non-statistical and were simply a static model of equilibrium catch but a more time-series orientated approach was set up by Collie and Sissenwine (1983). In some cases these simple population models have been extended to formal random effects models (Conser 1991; Trenkel 2008).

At the other extreme of the complexity scale, several multispecies models have been developed, some of which are formal statistical models (Taylor et al. 2007), though most are somewhat ad-hoc and do not take a statistical approach (Helgason and Gislason 1979; Fulton et al. 2005; Pauly et al. 2000). Simple mathematical descriptions of species interactions are not sufficient here since it is almost always essential to take into account spatial variation in species overlap, different nursery and spawning areas and so forth. For these reasons a useful multispecies model needs to take into account multiple areas, migration and maturation along with several other processes (Stefansson and Palsson 1998). To become statistical models, these need to be set up in the usual statistical manner with likelihood functions, parameters to be formally estimated, methods to estimate uncertainty and take into account the large number of different data sources available through appropriate weighting or comparisons (Richards 1991; Stefansson 1998, 2003).

In the year 2010, the single most promising venue of further research concerns the use of random effects in nonlinear fisheries models. Several of these have been described by Venables and Dichmont (2004) and some examples go a few decades back in time as seen above, often in debated implementations (de Valpine and Hilborn 2005). How this can be implemented in the context of complex multispecies models remains to be seen.

Cross References

►Adaptive Sampling
►Bayesian Statistics

►Mathematical and Statistical Modeling of Global Warming
►Statistical Inference in Ecology

References and Further Reading

Baranov FI (1918) On the question of the biological basis of fisheries. Proc Inst Icht Invest 1(1):81–128
Beverton RJH, Holt SJ (1957) On the dynamics of exploited fish populations, vol 19. Marine Fisheries, Great Britain Ministry of Agriculture, Fisheries and Food
Beverton RJH, Holt SJ (1993) On the dynamics of exploited fish populations, vol 11. Chapman and Hall, London
Brandão A, Butterworth DS, Johnston SJ, Glazer JP (2004) Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster, Jasus lalandii. Fish Res 70(2–3):339–349, 2004
Collie JS, Sissenwine MP (1983) Estimating population size from relative abundance data measured with error. Can J Fish Aquat Sci 40:1871–1879
Conser RJ (1991) A delury model for scallops incorporating length-based selectivity of the recruiting year-class to the survey gear and partial recruitment to the commercial fishery. Northeast Regional Stock Assessment Workshop Report, Woods Hole, MA, Res. Doc. SAW12/2, Appendix to CRD-91-03, 18pp
de Valpine P, Hilborn R (2005) State-space likelihoods for nonlinear fisheries timeseries. Can J Fish Aquat Sci 62(9):1937–1952
Fournier DA (1996) AUTODIF. A C++ array language extension with automatic differentiation for use in nonlinear modeling and statistic. Otter Research, Nanaimo, BC, 1996
Fulton EA, Smith ADM, Punt AE (2005) Which ecological indicators can robustly detect effects of fishing? ICES J Marine Sci 62(3):540
Gudmundsson G (2005) Stochastic growth. Can J Fish Aquat Sci 62(8):1746–1755
Gulland JA (1965) Estimation of mortality rates. Annex to Arctic Fisheries Working Group Report. ICES (Int. Counc. Explor. Sea) Document C.M. D:3 (mimeo), 1965
Helgason T, Gislason H (1979) VPA-analysis with species interaction due to predation. ICES C.M. 1979/G:52
Millar RB (1992) Modelling environmental effects on growth of cod: fitting to growth increment data versus fitting to size-at-age data. ICES J Marine Sci 49(3):289
Myers RA, Cadigan NG (1995) Statistical analysis of catch-at-age data with correlated errors. Can J Fish Aquat Sci (Print) 52(6):1265–1273
Myers RA, Bowen KG, Barrowman NJ (1999) Maximum reproductive rate of fish at low population sizes. Can J Fish Aquat Sci 56(12):2404–2419
Patterson KR (1999) Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Can J Fish Aquat Sci 56(2):208–221
Patterson K, Cook R, Darby C, Gavaris S, Kell L, Lewy P, Mesnil B, Punt A, Restrepo V, Skagen DW, Stefansson G (2001) Estimating uncertainty in fish stock assessment and forecasting. Fish Fish 2(2):125–157
Pauly D, Christensen V, Walters C (2000) Ecopath, Ecosim, and Ecospace as tools for evaluating ecosystem impact of fisheries. ICES J Marine Sci 57(3):697
Richards LJ (1991) Use of contradictory data sources in stock assessments. Fish Res 11(3–4):225–238
Skaug HJ (2002) Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. J Comput Gr Stat pp 458–470
Skaug HJ, Fournier DA (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput Stat Data Anal 51(2):699–709
Stefansson G (1998) Comparing different information sources in a multispecies context. In Funk F, Quinn II TJ, Heifetz J, Ianelli JN, Powers JE, Schweigert JF, Sullivan PJ, Zhang CI (eds.), Fishery Stock Assessment Models: Proceedings of the international symposium; Anchorage 1997, 15th Lowell Wakefield Fisheries Symposium, pp 741–758
Stefansson G (2003) Issues in multispecies models. Natural Res Model 16(4):415–437
Stefansson G, Palsson OK (1998) A framework for multispecies modelling of boreal systems. Rev Fish Biol Fish 8:101–104
Taylor L, Stefansson G (1999) Growth and maturation of haddock (Melanogrammus aeglefinus) in icelandic waters. J Northwest Atlantic Fish Sci 25:101–114
Taylor L, Begley J, Kupca V, Stefansson G (2007) A simple implementation of the statistical modelling framework Gadget for cod in Icelandic waters. African J Marine Sci 29(2):223–245, AUG 2007. ISSN 1814-232X. doi: 10.2989/AJMS.2007.29.2.7.190
Trenkel VM (2008) A two-stage biomass random effects model for stock assessment without catches: what can be estimated using only biomass survey indices? Can J Fish Aquat Sci 65(6):1024–1035
Venables WN, Dichmont CM (2004) GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fish Res 70(2–3):319–337

Markov Chain Monte Carlo

Siddhartha Chib
Harry C. Hartkopf Professor of Econometrics and Statistics
Washington University in St. Louis, St. Louis, MO, USA

Introduction

Suppose that π is a probability measure on the probability space $(S, \mathcal{A})$, h is a measurable function from $S \to \mathrm{R}$, and one is interested in the calculation of the expectation

$$\bar{h} = \int h d\pi$$

assuming that the integral exists. In many problems, especially when the sample space S is multivariate or when the normalizing constant of π is not easily calculable, finding the value of this integral is not feasible either by numerial methods of integration (such as the method of quadrature) or by classical Monte Carlo methods (such as the method of rejection sampling). In such instances, it is usually possible to find $\bar{h}$ by Markov chain Monte Carlo, or MCMC for short, a method that stems from Metropolis et al. (1953)

in connection with work related to the hydrogen bomb project. It found early and wide use in computational statistical mechanics and quantum field theory where it was used to sample the coordinates of a point in phase space. Applications and developments of this method in statistics, in particular for problems arising in ►Bayesian statistics, can be traced to Hastings (1970), Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990).

The idea behind MCMC is to generate a sequence of draws $\{\boldsymbol{\psi}^{(g)}, g \geq 0\}$ that follow a Markov chain (see ►Markov Chains) with the property that the unique invariant distribution of this Markov chain is the target distribution π. Then, after ignoring the first n_0 draws to remove the effect of the initial value $\boldsymbol{\psi}^{(0)}$, the sample

$$\{\boldsymbol{\psi}^{(n_0+1)}, ..., \boldsymbol{\psi}^{(n_0+M)}\}$$

for M large, is taken as an approximate sample from π and $\bar{h}$ estimated by the sample average

$$M^{-1}\sum_{g=1}^{M} h(\boldsymbol{\psi}^{(n_0+g)})$$

Laws of large numbers for Markov chains show that

$$M^{-1}\sum_{g=1}^{M} h(\boldsymbol{\psi}^{(n_0+g)}) \rightarrow \int h d\pi$$

as the simulation sample size M goes to infinity (Tierney 1994; Chib and Greenberg 1995; Chen et al. 2000; Liu 2001; Robert and Casella 2004).

A key reason for the interest in MCMC methods is that, somewhat surprisingly, it is straightforward to construct one or more Markov chains whose limiting invariant distribution is the desired target distribution. A leading method is the Metropolis–Hasting (M-H) method.

Metropolis–Hastings method

In the Metropolis–Hastings method, as the Hastings (1970) extension of the Metropolis et al. (1953) method is called, the Markov chain simulation is constructed by a recursive two step process.

Let $\pi(\boldsymbol{\psi})$ be a probabiliy measure that is dominated by a sigma-finite measure μ. Let the density of π with respect to μ be denoted by $p(\cdot)$. Let $q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger})$ denote a conditional density for $\boldsymbol{\psi}^{\dagger}$ given $\boldsymbol{\psi}$ with respect to μ. This density $q(\boldsymbol{\psi}, \cdot)$ is referred to as the proposal or candidate generating density. Then, the Markov chain in the M-H algorithm is constructed in two steps as follows.

Step 1 Sample a proposal value $\boldsymbol{\psi}^{\dagger}$ from $q(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi})$ and calculate the quantity (the *acceptance probability* or the *probability of move*)

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger}) = \begin{cases} \min\left[\frac{p(\boldsymbol{\psi}^{\dagger})q(\boldsymbol{\psi}^{\dagger}, \boldsymbol{\psi})}{p(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger})}, 1\right] & \text{if } p(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger}) > 0; \\ 1 & \text{otherwise}. \end{cases}$$

Step 2 Set

$$\boldsymbol{\psi}^{(g+1)} = \begin{cases} \boldsymbol{\psi}^{\dagger} & \text{with prob } \alpha(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi}^{\dagger}) \\ \boldsymbol{\psi}^{(g)} & \text{with prob } 1 - \alpha(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi}^{\dagger}) \end{cases}$$

If the proposal value is rejected then the next sampled value is taken to be the current value which means that when a rejection occurs the current value is repeated and the chain stays at the current value. Given the new value, the same two step process is repeated and the whole process iterated a large number of times.

Given the form of the acceptance probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ it is clear that the M-H algorithm does not require knowledge of the normalizing constant of $p(\cdot)$. Furthermore, if the proposal density satisfies the symmetry condition $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$, the acceptance probability reduces to $p(\boldsymbol{\psi}')/p(\boldsymbol{\psi})$; hence, if $p(\boldsymbol{\psi}') \geq p(\boldsymbol{\psi})$, the chain moves to $\boldsymbol{\psi}'$, otherwise it moves to $\boldsymbol{\psi}'$ with probability given by $p(\boldsymbol{\psi}')/p(\boldsymbol{\psi})$. The latter is the algorithm originally proposed by Metropolis et al. (1953).

A full expository discussion of this algorithm, along with a derivation of the method from the logic of reversibility, is provided by Chib and Greenberg (1995).

The M-H method delivers variates from π under quite general conditions. A weak requirement for a law of large numbers for sample averages based on the M-H output involve positivity and continuity of $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ for $(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and connectedness of the support of the target distribution. In addition, if π is bounded then conditions for ergodicity, required to establish the central limit theorem (see ►Central Limit Theorems), are satisfied (Tierney 1994).

It is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the invariant distribution without staying at one place for many iterations. Generally, the empirical behavior of the M-H output is monitored by the autocorrelation time of each component of $\boldsymbol{\psi}$ defined as

$$\left\{1 + 2\sum_{s=1}^{M} \rho_{ks}\right\},$$

where ρ_{ks} is the sample autocorrelation at lag s for the kth component of $\boldsymbol{\psi}$, and by the acceptance rate which is the proportion of times a move is made as the sampling proceeds. Because independence sampling produces an autocorrelation time that is theoretically equal to one, one tries to tune the M-H algorithm to get values close to one, if possible.

Different proposal densities give rise to specific versions of the M-H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\psi, \psi') = q(\psi' - \psi)$. The candidate ψ' is thus drawn according to the process $\psi' = \psi + z$, where z follows the distribution q, and is referred to as the random walk M-H chain. The random walk M-H chain is perhaps the simplest version of the M-H algorithm and is quite popular in applications. One has to be careful, however, in setting the variance of z because if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\psi, \psi') = q(\psi')$. Proposal values are thus drawn independently of the current location ψ.

Multiple-Block M-H

In applications when the dimension of ψ is large it is usually necessary to construct the Markov chain simulation by first grouping the variables ψ into smaller blocks. Suppose that two blocks are adequate and that ψ is written as (ψ_1, ψ_2), with $\psi_k \in \Omega_k \subseteq \Re^{d_k}$. In that case the M-H algorithm requires the specification of two proposal densities,

$$q_1\left(\psi_1, \psi_1^{\dagger} | \psi_2\right) \; ; \; q_2\left(\psi_2, \psi_2^{\dagger} | \psi_1\right),$$

one for each block ψ_k, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha\left(\psi_1, \psi_1^{\dagger} | \psi_2\right) = \min\left\{\frac{p\left(\psi_1^{\dagger}, \psi_2\right) q_1\left(\psi_1^{\dagger}, \psi_1 | \psi_2\right)}{p(\psi_1, \psi_2) q_1\left(\psi_1, \psi_1^{\dagger} | \psi_2\right)}, 1\right\}$$

and

$$\alpha\left(\psi_2, \psi_2^{\dagger} | \psi_1\right) = \min\left\{\frac{p\left(\psi_1, \psi_2^{\dagger}\right) q_2\left(\psi_2^{\dagger}, \psi_2 | \psi_1\right)}{p(\psi_1, \psi_2) q_2\left(\psi_2, \psi_2^{\dagger} | \psi_1\right)}, 1\right\},$$

as the probability of move for block ψ_k conditioned on the other block. Then, one cycle of the algorithm is completed by updating each block using a M-H step with the above probability of move, given the most current value of the other block.

Gibbs Sampling

A special case of the multiple-block M-H method is the Gibbs sampling method which was introduced by Geman and Geman (1984) in the context of image-processing and broadened for use in Bayesian problems by Gelfand and Smith (1990). To describe this algorithm, suppose that the parameters are grouped into two blocks (ψ_1, ψ_2) and each block is sampled according to the full conditional distribution of block ψ_k,

$$p(\psi_1 | \psi_2) \; ; \; p(\psi_2 | \psi_1)$$

defined as the conditional distribution under π of ψ_k given the other block. In parallel with the multiple-block M-H algorithm, the most current value of the other block is used in sampling the full conditional distribution. Derivation of these full conditional distributions is usually quite simple since, by ►Bayes' theorem, each full conditional is proportional to $p(\psi_1, \psi_2)$, the joint distribution of the two blocks. In addition, the introduction of latent or auxiliary variables can sometimes simplify the calculation and sampling of the full conditional distributions. Albert and Chib (1993) develop such an approach for the Bayesian analysis of categorical response data.

Concluding Remarks

Some of the recent theoretical work on MCMC methods is related to the question of the rates of convergence (Cai 2000; Fort et al. 2003; Jarner and Tweedie 2003; Douc et al. 2007) and in the development of adaptive MCMC methods (Atchade and Rosenthal; Andrieu and Moulines 2005; 2006).

The importance of MCMC methods in statistics and in particular Bayesian statistics cannot be overstated. The remarkable growth of Bayesian thinking over the last 20 years was made possible largely by the innovative use of MCMC methods. Software programs such as WINBUGS and the various MCMC packages in R have contributed to the use of MCMC methods in applications across the sciences and social sciences (Congdon 2006) and these applications are likely to continue unabated.

About the Author

Siddhartha Chib is the Harry Hartkopf Professor of Econometrics and Statistics at the Olin Business School, Washington University in St. Louis. He is a Fellow of the American Statistical Association and the Director of the NBER-NSF Seminar in Bayesian Inference in Econometrics and Statistics. Professor Chib has made several contributions in the areas of binary, categorical and censored response models, the Metropolis-Hastings algorithm and MCMC methods, the estimation of the marginal likelihood and Bayes factors, and in the treatment of hidden Markov and change-point models, and stochastic volatility and diffusion models. He has served as an Associate Editor

of the *Journal of the American Statistical Association* (Theory and Methods), *Journal of Econometrics*, the *Journal of Business and Economics Statistics*, and others. Currently he is an Associate Editor of the *Journal of Computational and Graphical Statistics*, and *Statistics and Computing*.

Cross References

►Bayesian Reliability Modeling
►Bayesian Statistics
►Bootstrap Methods
►Markov Chains
►Model Selection
►Model-Based Geostatistics
►Monte Carlo Methods in Statistics
►Non-Uniform Random Variate Generations
►Rubin Causal Model
►Small Area Estimation
►Social Network Analysis
►Statistics: An Overview

References and Further Reading

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679
Andrieu C, Moulines E (2006) On the ergodicity properties of some adaptive MCMC algorithms. Ann Appl Probab 16:1462–1505
Atchade YF, Rosenthal JS (2005) On adaptive Markov Chain Monte Carlo algorithms. Bernoulli 11:815–828
Cai HY (2000) Exact bound for the convergence of Metropolis chains. Stoch Anal Appl 18:63–71
Chen MH, Shao QM, Ibrahim JG (2000) Monte Carlo methods in Bayesian computation. Springer, New York
Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. Am Stat 49(4):327–335
Congdon P (2006) Bayesian statistical modelling, 2nd edn. Wiley, Chichester
Douc R, Moulines E, Soulier P (2007) Computable convergence rates for subgeometric ergodic Markov chains. Bernoulli 13:831–848
Fort G, Moulines E, Roberts GO, Rosenthal JS (2003) On the geometric ergodicity of hybrid samplers. J Appl Probab 40:123–146
Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans PAMI 6: 721–741
Hastings WK (1970) Monte-Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109
Jarner SF, Tweedie RL (2003) Necessary conditions for geometric and polynomial ergodicity of random-walk-type markov chains. Bernoulli 9:559–578
Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–550 (with discussion)
Tierney L (1994) Markov-chains for exploring posterior distributions. Ann Stat 22:1701–1728

Markov Chains

Arnoldo Frigessi[1,2], Bernd Heidergott[3]
[1]Director
Norwegian Centre for Research-Based Innovation
"Statistics for Innovation," Oslo, Norway
[2]Professor
University of Oslo & Norwegian Computing Centre, Oslo, Norway
[3]Associate Professor
Vrije Universiteit, Amsterdam, The Netherlands

Introduction

Markov chains, which comprise Markov chains and ►Markov processes, have been successfully applied in areas as divers as biology, finance, manufacturing, telecommunications, physics and transport planning, and even for experts it is impossible to have an overview on the full richness of Markovian theory. Roughly speaking, Markov chains are used for modeling how a system moves from one state to another at each time point. Transitions are random and governed by a conditional probability distribution which assigns a probability to the move into a new state, given the current state of the system. This dependence represents the memory of the system. A basic example of a Markov chain is the so-called random walk defined as follows. Let $X_t \in N$, for $t \in \mathbb{N}$, be a sequence of random variables with initial value $X_0 = 0$. Furthermore assume that $P(X_{t+1} = X_t+1|X_t \geq 1) = p = 1-P(X_{t+1} = X_t-1|X_t \geq 1)$. The sequence $X = \{X_t : t \in \mathbb{N}\}$ is an example of a Markov chain (for a detailed definition see below) and the aspects of X one is usually interested in in Markov chain theory is (i) whether X returns to 0 in a finite number of steps (this holds for $0 \leq p \leq 1/2$), (ii) the expected number of steps until the chain returns to 0 (which is finite for $0 \leq p < 1/2$), and (iii) the limiting behavior of X_t.

In the following we present some realistic examples. A useful model in modeling infectious diseases assumes that there are four possible states: Susceptible (S), Infected (I), Immune (A), Dead (R). Possible transitions are from S to I, S or R; from I to A or R; from A to A or R; from R to R only. The transitions probabilities, from S to I, S to R

and the loop S to S, must sum to one and can depend on characteristics of the individuals modeled, like age, gender, life style, etc. All individuals start in S, and move at each time unit (say a day). Given observations of the sequence of visited states (called trajectory) for a sample of individuals, with their personal characteristics, one can estimate the transition probabilities, by ►logistic regression, for example. This model assumes that the transition probability at time t from one state A to state B, only depends on the state A, and not on the trajectory that lead to A. This might not be realistic, as for example a perdurance in the diseased state I over many days, could increase the probability of transition to R. It is possible to model a system with longer memory, and thus leave the simplest setting of a Markov Chain (though one can formulate such a model still as a Markov Chain over a more complex state space which includes the length of stay in the current state). A second example refers to finance. Here we follow the daily value in Euro of a stock. The state space is continuous, and one can model the transitions from state x Euro to y Euro with an appropriate Normal density with mean $x - y$. The time series of the value of the stock might well show a longer memory, which one would typically model with some autoregressive terms, leading to more complex process again. As a further example, consider the set of all web pages on the Internet as the state space of a giant Markov chain, where the user clicks from one page to the next, according to a transition probability. A Markov Chain has been used to model such a process. The transitions from the current web page to the next web page can be modeled as a mixture of two terms: with probability λ the user follows one of the links present in the current web page and among these uniformly; with probability $1 - \lambda$ the user chooses another web page at random among all other ones. Typically $\lambda = 0.85$. Again, one could discuss how correct the assumption is, that only the current web page determines the transition probability to the next one. The modeler has to critically validate such hypothesis before trusting results based on the Markov Chain model, or chains with higher order of memory. In general a stochastic process has the Markov property if the probability to enter a state in the future is independent of the states visited in the past given the current state. Finally, Markov Chain Monte Carlo (MCMC) algorithms (see ►Markov Chain Monte Carlo) are Markov chains, where at each iteration, a new state is visited according to a transition probability that depends on the current state. These stochastic algorithm are used to sample from a distribution on the state space, which is the marginal distribution of the chain in the limit, when enough iterations have been performed.

In the literature the term Markov processes is used for Markov chains for both discrete- and continuous time cases, which is the setting of this paper. Standard textbooks on Markov chains are Kijima (1997), Meyn and Tweedie (1993), Nummelin (1984), Revuz (1984). In this paper we follow (Iosifescu 1980) and use the term 'Markov chain' for the discrete time case and the term 'Markov process' for the continuous time case. General references on Markov chains are Feller (1968), Gilks et al. (1995), Haeggstroem (2002), Kemeny and Snell (1960), Seneta (1973).

Discrete Time Markov Chains

Consider a sequence of random variables $X = \{X_t : t \in \mathbb{N}\}$ defined on a common underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with state discrete space $(S, \mathcal{S})$, i.e., X_t is $\mathcal{F} - \mathcal{S}$-measurable for $t \in \mathbb{N}$. The defining property of a Markov chain is that the distribution of X_{t+1} depends on the past only through the immediate predecessor X_t, i.e., given $X_0, X_1, \ldots, X_t$ it holds that

$$\mathbb{P}(X_{t+1} = x | X_0 = x_0, X_1 = x_1, \ldots X_{t-1} = x_{t-1}, X_t = y) = \mathbb{P}(X_{t+1} = x | X_t = y),$$

where x, y and all other x_i are element of the given state space S. If $\mathbb{P}(X_{t+1} = x | X_t = y)$ does not depend on t, the chain is called *homogenous* and it is called inhomogeneous otherwise. Provided that S is at most countable, the transition probabilities of a homogeneous Markov Chain are given by $P = (p_{x,y})_{S \times S}$, where $p_{x,y} = \mathbb{P}(X_{t+1} = y | X_t = x)$ is the probability of a transition from x to y. The matrix P is called the *one-step transition probability matrix* of the Markov chain. For the introductory ►random walk example the transition matrix is given by $p_{i,i+1} = p$, $p_{i,i-1} = p - 1$, for $i \geq 1, p_{o,1} = 1$ and otherwise zero, for $i \in \mathbb{Z}$. The row sums are one and the k-th power of the transition matrix represent the probability to move between states in k time units.

In order to fully define a Markov Chain it is necessary to assign an initial distribution $\mu = (\mathbb{P}(X_0 = s) : s \in S)$. The marginal distribution at time t can then be computed, for example, as

$$\mathbb{P}(X_t = x) = \sum_{s \in S} p_{s,x}^{(t)} \mathbb{P}(X_0 = s),$$

where $p_{s,x}^{(t)}$ denotes the s, x element of the t-th power of the transition matrix. Note that given an initial distribution μ and a transition matrix P, the distribution of the Markov chain X is uniquely defined.

A Markov chain is said to be *aperiodic* if for each pair of states i, j the greatest common divisor of the set of all t such that $p_{ij}^{(t)} > 0$ is one. Note that the random walk in

our introductory example fails to be aperiodic as any path from starting in 0 and returning there has a length that is a multiple of 2.

A distribution $(\pi_i : i \in S)$ is called a *stationary distribution* of P if

$$\pi P = \pi.$$

A key topic in Markov chain theory is the study of the limiting behavior of X. Again, with initial distribution μ, X has limiting distribution ν for initial distribution μ if

$$\lim_{t\to\infty} \mu P^t = \nu. \tag{1}$$

Note that any limiting distribution is a stationary distribution. A case of particular interest is that when X has a unique stationary distribution, which is then also the unique limiting distribution and thus describes the limit behavior of the Markov chain. If P fails to be aperiodic, then the limit in (1) may not exists and should be replaced by the Cesaro limit

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \mu P^k = \nu,$$

which always exists for finite Markov chains.

A Markov chain is called *ergodic* if the limit in (1) is independent of the initial distribution. Consequently, an ergodic Markov chain has a unique limiting distribution and this limiting distribution is also a stationary distribution, and since any stationary distribution is a limiting distribution it is also unique.

A Markov chain is called *irreducible* if for any pair of states $i, j \in S$, there exists a path from i to j that X will follow with positive probability. In words, any state can be reached from any other state with positive probability. An irreducible Markov chain is called *recurrent* if the number of steps from a state i to the first visit of a state j, denoted by $\tau_{i,j}$, is almost surely finite for all $i, j \in S$, and it is called *positive recurrent* if $\mathbb{E}[\tau_{i,i}] < \infty$ for at least one $i \in S$. Note that for $p = 1/2$ the random walk is recurrent and for $p < 1/2$ it is positive recurrent.

The terminology developed so far allows to present the main result of Markov chain theory: Any aperiodic, irreducible and positive recurrent Markov chain P possesses a unique stationary distribution π which is the unique probability vector solving $\pi P = \pi$ (and which is also the unique limiting distribution). This ▶ergodic theorem is one of the central results and it has been established in many variations and extensions, see the references. Also, efficient algorithms for computing π have been a focus of research as for Markov chains on large state-spaces computing π is a non-trivial task.

An important topic of the statistics of Markov chains is to estimate the (one-step) transition probabilities. Consider a discrete time, homogeneous Markov chain with finite state space $S = \{1, 2, \ldots, m\}$, observed at time points $0, 1, 2, \ldots, T$ on the trajectory $s_0, s_2, s_2, \ldots, s_T$. We wish to estimate the transition probabilities $p_{i,j}$ by maximum likelihood. The likelihood is

$$\mathbb{P}(X_0 = s_0) \prod_{t=0}^{'T-1} \mathbb{P}(X_t = s_{t+1} | X_y = s_t)$$
$$= \mathbb{P}(X_0 = s_0) \prod_{i=1}^{m}\prod_{j=1}^{m} p_{i,j}^{k(i,j)}$$

where $k(i, j)$ is the number of transitions from i to j in the observed trajectory. Ignoring the initial factor, the maximum likelihood estimator of $p_{i,j}$ is found to be equal to $\hat{p}_{i,j} = \frac{k(i,j)}{k(i,\cdot)}$, where $k(i, \cdot)$ is the number of transitions out from state i. Standard likelihood asymptotic applies, despite the data are dependent, as $k(i, \cdot) \to \infty$, which will happen if the chain is ergodic. The asymptotic variance of the maximum likelihood estimates can be approximated as $\mathrm{var}(\hat{p}_{i,j}) \sim \hat{p}_{i,j}(1 - \hat{p}_{i,j})/k(i, \cdot)$. The covariances are zero, except $\mathrm{cov}(\hat{p}_{i,j}, \hat{p}_{i,j'}) \sim -\hat{p}_{i,j}\hat{p}_{i,j'}/k(i, \cdot)$ for $j \neq j'$. If the trajectory is short, the initial distribution should be considered. A possible model is to use the stationary distribution $\pi(s_0)$, which depend on the unknown transition probabilities. Hence numerical maximization is needed to obtain the maximum likelihood estimates. In certain medical applications, an alternative asymptotic regime can be of interest, when many (k) short trajectories are observed, and $k \to \infty$. In this case the initial distribution cannot be neglected.

Markov Chains and Markov Processes

Let $\{X_t : t \geq 0\}$ denote the (continuous time) Markov process on state space $(S, \mathcal{S})$ with transition matrix $P(t)$, i.e.,

$$(P(t))_{ij} = \mathbb{P}(X_{t+s} = j | X_s = i), \quad s \geq 0, \quad i, j \in S.$$

Under some mild regularity conditions is holds that the *generator matrix* Q, defined as

$$\left.\frac{d}{dt}\right|_{t=0} P(t) = Q,$$

exists for $P(t)$. The stationary distribution of a Markov process can be found as the unique probability π that solves $\pi Q = 0$, see Anderson (1991). A generator matrix Q is called uniformizable with rate μ if $\mu = \sup_j |q_{jj}| < \infty$. While any finite dimensional generator matrix is uniformizable a classical example of a Markov process on denumerable state space that fails to have this property is the M/M/∞

queue. Note that if Q is uniformizable with rate μ, then Q is uniformizable with rate η for any $\eta > \mu$. Let Q be uniformizable with rate μ and introduce the Markov chain P_μ as follows

$$[P_\mu]_{ij} = \begin{cases} q_{ij}/\mu & i \neq j \\ 1 + q_{ii}/\mu & i = j, \end{cases} \tag{2}$$

for $i, j \in S$, or, in shorthand notation,

$$P_\mu = I + \frac{1}{\mu} Q,$$

then it holds that

$$P(t) = e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} (P_\mu)^n, \quad t \geq 0. \tag{3}$$

Moreover, the stationary distribution of P_μ and $P(t)$ coincide. The Markov chain $\mathcal{X}_\mu = \left\{X_n^\mu : n \geq 0\right\}$ with transition probability matrix P_μ is called the *sampled chain*. The relationship between $\mathcal{X}$ and $\mathcal{X}_\mu$ can be expressed as follows. Let $N_\mu(t)$ denote a Poisson process (see ►Poisson Processes) with rate μ, then $X^\mu_{N_\mu(t)}$ and X_t are equal in distribution for all $t \geq 0$. From the above it becomes clear that the analysis of the stationary behavior of a (uniformizable) continuous time Markov chain reduces to that of a discrete time Markov chain.

About the Authors

Arnoldo Frigessi is Professor in statistics, University of Oslo. He is director of the center for research based innovation Statistics for Innovation (sfi)2 and holds a position at the Norwegian Computing Center. Previously he hold positions at the University of Roma Tre and University of Venice. He is an Elected member of the Royal Norwegian Academy of Science and Letters. He is past scientific secretary of the Bernoulli Society for Mathematical Statistics and Probability. His research is mainly in the area of Bayesian statistics and MCMC, both methodological and applied.

Dr Bernd Heidergott is Associate Professor at the Department of Econometrics, Vrije Universiteit Amsterdam, the Netherlands. He is also research fellow at the Tinbergen Institute and at EURANDOM, both situated in the Netherlands. He has authored and co-authored more than 30 papers and two books, *Max-Plus linear Systems and Perturbation Analysis* (Springer, 2007), and *Max Plus at Work* (with Jacob van der Woude and Geert Jan Olsder, Princeton, 2006.)

Cross References

►Box–Jenkins Time Series Models
►Ergodic Theorem
►Graphical Markov Models
►Markov Processes
►Nonlinear Time Series Analysis
►Optimal Stopping Rules
►Record Statistics
►Statistical Inference for Stochastic Processes
►Stochastic Global Optimization
►Stochastic Modeling Analysis and Applications
►Stochastic Processes: Classification

References and Further Reading

Anderson W (1991) Continuous-time Markov chains: an applications oriented approach. Springer, New York
Feller W (1968) An Introduction to Probability Theory and its Applications, vol 1, 3rd edn. Wiley, New York
Gilks W, Richardson S, Spiegeihalter D (eds) (1995) Markov Chain Monte Carlo in practice. Chapman & Hall, London
Haeggstroem O (2002) Finite Markov chains and algorithmic applications, London Mathematical Society Student Texts (No. 52)
Iosifescu M (1980) Finite Markov processes and their applictions. Wiley, New York
Kemeny J, Snell J (1960) Finite Markov chains, (originally published by Van Nostrand Publishing Company Springer Verlag, 3rd printing, 1983)
Kijima M (1997) Markov processes for stochastic modelling. Chapman & Hall, London
Meyn S, Tweedie R (1993) Markov chains and stochastic stability. Springer, London
ummelin E (1984) General irreducible Markov chains and non-negative operators. Cambridge Univesity Press, Cambridge
Revuz D (1984) Markov chains, 2nd edn. Noth-Holland, Amsterdam
Seneta E (1973) Non-negative matrices and Markov chains (originally published by Allen & Unwin Ltd., London, Springer Series in Statistics, 2nd revised edition, 2006)

Markov Processes

Zoran R. Pop-Stojanović
Professor Emeritus
University of Florida, Gainesville, FL, USA

The class of Markov Processes is characterized by a special stochastic dependence known as the *Markov Dependence* that was introduced in 1907 by A.A. Markov while extending in a natural way the concept of stochastic independence that will preserve, for example, the asymptotic properties of sums of random variables such as the law of large numbers. One of his first applications of this dependence was in investigation of the way the vowels and consonants alternate in literary works in the Russian literature. This dependence that Markov introduced, dealt with what we

call today *a discrete-parameter Markov Chain with a finite number of states,* and it can be stated as follows: a sequence $\{X_n;\ n = 1, 2, \dots\}$ of real-valued random variables given on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, each taking on a finite number of values, satisfies

$$\mathrm{P}[X_{n+1} = x_{n+1} | X_1, X_2, \dots, X_n] = \mathrm{P}[X_{n+1} = x_{n+1} | X_n]. \quad (1)$$

Roughly speaking, (1) states that *any prediction of* X_{n+1} *knowing*

$$X_1, X_2, \dots, X_n,$$

can be achieved by using X_n *alone.*

This concept was further extended (as shown in what follows), for the *continuous-parameter Markov processes* by A.N. Kolmogorov in 1931. Further essential developments in the theory of continuous-parameter Markov Processes were due to W. Feller, J.L. Doob, G.A. Hunt, and E.B. Dynkin.

In order to introduce a continuous-parameter Markov Process, one needs the following setting. Let $\mathbf{T} \equiv [0, +\infty) \subset \mathbb{R}$ be the parameter set of the process, referred to in the sequel as *time*, where $\mathbb{R}$ denotes the one-dimensional Euclidean space; let $X = \{X_t,\ \mathcal{F}_t,\ t \in \mathbf{T}\}$ be the process given on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ that takes values in a topological space $(\mathcal{S}, \mathcal{E})$, where $\mathcal{E}$ is a Borel field of $\mathcal{S}$, that is, a σ-field generated by open sets in $\mathcal{S}$. The process X is adapted to the increasing family $\{\mathcal{F}_t,\ t \in \mathbf{T}\}$ of σ-fields of $\mathcal{F}$, where $\mathcal{F}_0$ contains all P-null sets. All X_t's are $\mathcal{E}$-measurable. Here, X_t *is adapted to* $\mathcal{F}_t$ *means that all random events related to* X_t *are contained in* $\mathcal{F}_t$ *for every value* t *of the parameter of the process, that is,* X_t *is* $\mathcal{F}_t$*-measurable in addition of being* $\mathcal{E}$*-measurable.* In order to describe the Markov dependence for the process X, the following two σ-fields are needed: $\forall t,\ t \in \mathbf{T},\ \mathcal{F}_t^{past} = \sigma(\{X_s,\ s \in [0, t]\})$ and $\mathcal{F}_t^{future} = \sigma(\{X_s,\ s \in [t, +\infty)\})$. Here, the *past* and the *future* are relative to the instant t that is considered as the *present.* Now the process $X = \{X_t, \mathcal{F}_t,\ t \in \mathbf{T}\}$ is called a *Markov Process* if and only if one of the following equivalent conditions is satisfied:

$$\begin{aligned}
&(i)\ \forall t,\ t \in \mathbf{T},\ A \in \mathcal{F}_t,\ B \in \mathcal{F}_t^{future}:\\
&\quad \mathrm{P}(A \cap B | X_t) = \mathrm{P}(A | X_t)\mathrm{P}(B | X_t).\\
&(ii)\ \forall t,\ t \in \mathbf{T},\ B \in \mathcal{F}_t^{future}:\\
&\quad \mathrm{P}(B | \mathcal{F}_t) = \mathrm{P}(B | X_t).\\
&(iii)\ \forall t,\ t \in \mathbf{T},\ A \in \mathcal{F}_t:\\
&\quad \mathrm{P}\left(A | \mathcal{F}_t^{future}\right) = \mathrm{P}(A | X_t).
\end{aligned} \quad (2)$$

Observe that (ii) in (2) is the analog of (1) stating that *the probability of an event in the future of the Markov process* X *depends only on the probability of the present state of the process and it is independent of the past history of the process.* There are numerous phenomena occurring in physical sciences, social sciences, econometrics, the world of finance, to name just a few, that can all be modelled by Markov processes. Among Markov processes there is a very important subclass of the so-called *strong Markov processes*. This proper subclass of Markov processes is obtained by *randomizing* the parameter of the process. This randomization of the parameter leads to the so-called *optional times of the process* and the Markov property (2) is replaced by the *strong Markov property*, where in (2) deterministic time t is replaced by an *optional time* of the process. The most important example of a strong Markov process is the *Brownian Motion Process* (see ►Brownian Motion and Diffusions) that models the physical phenomenon known as the *Brownian Movement of particles*. Another important class of processes – *Diffusion processes,* are *strong Markov Processes with continuous paths.*

One of the most important properties of Markov processes is *that times between transitions from one state to another, are random variables that are conditionally independent of each other given the successive states being visited, and each such sojourn time has an exponential distribution with the parameter dependent on the state being visited.* This property coupled with the property that successive states visited by the process form a Markov chain (see ►Markov Chains), clearly describe the structure of a Markov process. Other important examples of Markov processes are ►Poisson processes, Compound Poisson processes, The ►Random Walk, Birth and Death processes, to mention just a few. The last mentioned class of Markov processes has many applications in biology, ►demography, and ►queueing theory.

For further details and proofs of all facts mentioned here, a reader may consult the enclosed list of references.

Cross References

►Brownian Motion and Diffusions
►Markov Chains
►Martingale Central Limit Theorem
►Optimal Stopping Rules
►Poisson Processes
►Random Permutations and Partition Models
►Random Walk
►Statistical Inference for Stochastic Processes
►Stochastic Processes
►Stochastic Processes: Classification
►Structural Time Series Models

References and Further Reading

Blumenthal RM, Getoor RK (1968) Markov processes and potential theory. Academic Press, New York
Chung KL (1982) Lectures from Markov processes to Brownian motion. Springer, New York
Çinlar E (1975) Introduction to stochastic processes. Prentice Hall, New Jersey
Doob JL (1953) Stochastic processes. Wiley, New York
Dynkin EB (1965) Markov process, 2 Volumes. Springer, New York
Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York

Martingale Central Limit Theorem

Petra Posedel
Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

The martingale central limit theorem (MCLT) links the notions of martingales and the Lindeberg–Feller classical central limit theorem (CLT, see ▶Central Limit Theorems) for independent summands.

Perhaps the greatest achievement of modern probability is the unified theory of limit results for sums of independent random variables, such as the law of large numbers, the central limit theorem, and the law of the iterated logarithm. In comparison to the classical strong law of large numbers, the classical CLT says something also about the rate of this convergence. We recall the CLT for the case of independent, but not necessarily identically distributed random variables. Suppose that $\{X_i,\ i \geq 1\}$ is a sequence of zero-mean independent random variables such that $Var[X_n] = \sigma_n^2 < \infty,\ n \geq 1$. Let $S_n = \sum_{i=1}^n X_i,\ n \geq 1$ and set $Var[S_n] = s_n^2$. If the Lindeberg condition holds, i.e., $\frac{\sum_{i=1}^n E\left[X_i \mathbb{1}_{\{|X_i| \geq \epsilon s_n\}}\right]}{s_n^2} \longrightarrow 0$ as $n \to \infty$, for all $\epsilon > 0$, and $\mathbb{1}_{\{.\}}$ denoting the indicator function, then $\frac{S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1)$, where $N(0,1)$ denotes the standard normal random variable.

Limit theorems have applicability far beyond the corresponding results for sums of independent random variables. Namely, since sums of independent random variables centered at their expectations have a specific dependence structure (i.e., are martingales), there is interest in extending the results to sums of dependent random variables.

In order to define martingales and state the MCLT attributed to Brown (1971), one needs the following setting. Let $(\Omega,\ \mathcal{F},\ \mathbb{P})$ be a probability space and let $\{\mathcal{F}_n,\ n \geq 0\}$ be an increasing sequence of σ-fields of $\mathcal{F}$ sets.

Definition 1 *A sequence $\{Y_n,\ n \geq 0\}$ of random variables on Ω is said to be a martingale with respect to $\{\mathcal{F}_n,\ n \geq 0\}$ if (1) Y_n is measurable with respect to $\mathcal{F}_n$, (2) $E|Y_n| < \infty$, and (3) $E[Y_n|\mathcal{F}_m] = Y_m$ a.s. for all $m < n,\ m,\ n \geq 0$.*

In order to highlight the dependence structure of the underlying random variables, one should note that condition (3) is weaker than independence since it cannot be deduced which structure conditional higher-order moments may have given the past. The mathematical theory of martingales may be regarded as an extension of the independence theory, and it too has its origins in limit results, beginning with Bernstein (1927) and Lévy's (1935) early central limit theorems. These authors introduced the martingale in the form of consecutive sums with a view to generalizing limit results for sums of independent random variables. However, it was the subsequent work of Doob, including the proof of the celebrated martingale convergence theorem, that completely changed the direction of the subject, and his book (Doob 1953), popularly called in academia the *Holy Bible for stochastic processes*, has remained a major influence for nearly three decades.

The main result that follows applies the CLT to sequences of random variables that are martingales. If $\{S_n,\ \mathcal{F}_n\}$ is a martingale, it seems natural to replace $Var[S_n]$ in the CLT by the sum of conditional variances. Secondly, the norming by $1/n$ is very restrictive. For a sequence of independent, but not identically distributed random variables, it seems appropriate to norm by a different constant, and for a sequence of dependent random variables norming by another random variable should be considered. The limit theory for martingales essentially covers that for the categories of processes with independent increments and ▶Markov processes. Using stochastic processes that are martingales for analyzing limit results, one has at their disposal all the machinery from martingale theory. This reason makes martingales considerably attractive for inference purposes. A standard reference on martingales is Williams (1991).

Theorem 1 *Let $\{S_n,\ \mathcal{F}_n,\ n \geq 1\}$ be a zero-mean martingale with $S_0 = 0$, whose increments have finite variance. Write*

$$S_n = \sum_{i=1}^n X_i, \quad V_n^2 = \sum_{i=1}^n E\left[X_i^2 | \mathcal{F}_{i-1}\right], \quad \text{and}$$
$$s_n^2 = E\left[V_n^2\right] = E\left[S_n^2\right]. \tag{1}$$

If

$$\frac{V_n^2}{s_n^2} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{\sum_{i=1}^{n} E\left[X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon s_n\}}\right]}{s_n^2} \xrightarrow{\mathbb{P}} 0 \qquad (2)$$

as $n \to \infty$, *for all* $\epsilon > 0$, *and* $\mathbb{1}_{\{.\}}$ *denoting the indicator function, then*

$$\frac{S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1), \qquad (3)$$

where $N(0,1)$ *denotes the standard normal random variable.*

Roughly speaking, (3) says that the sum of martingale differences, when scaled appropriately, is approximately normally distributed provided the conditional variances are sufficiently well behaved. The theorem seems relevant in any context in which conditional expectations, given the past, have a simple and possibly explicit form. Various results on sums of independent random variables in fact require only orthogonality of the increments, i.e., $E[X_i X_j] = 0$, $i \neq j$, and this property holds for martingales whose increments have finite variance. The MCLT reduces to the sufficiency part of the standard Lindeberg–Feller result in the case of independent random variables.

The interpretation of V_n^2 is highlighted and particularly interesting for inference purposes. Let $X_1, X_2, \ldots$ be a sequence of observations of a stochastic process whose distribution depends on a (single) parameter θ, and let $L_n(\theta)$ be the likelihood function associated with $X_1, X_2, \ldots$. Under very mild conditions, score functions $S_n = \partial \log L_n(\theta)/\partial\theta$ form a martingale whose conditional variance $V_n^2 = I_n(\theta)$ is a generalized form of the standard Fisher information, as shown in Hall and Heyde (1980). Namely, suppose that the likelihood function $L(\theta)$ is differentiable with respect to θ and that $E_\theta[\partial \log L(\theta)/\partial\theta]^2 < \infty$.
Let θ be a true parameter vector. We have

$$S_n = \frac{\partial \log L_n(\theta)}{\partial\theta} = \sum_{i=1}^{n} x_i(\theta),$$

$$x_i(\theta) = \frac{\partial}{\partial\theta}[\log L_i(\theta) - \log L_{i-1}(\theta)],$$

and thus $E_\theta[x_i(\theta)|\mathcal{F}_{i-1}] = 0$ a.s., so that $\{S_n, \mathcal{F}_n, n \geq 1\}$ is a square-integrable martingale. Set $V_n^2 = \sum_{i=1}^{n} E_\theta\left[x_i^2(\theta)\middle|\mathcal{F}_{i-1}\right]$. The quantity V_n^2 reduces to the standard Fisher information $I_n(\theta)$ in the case where the observations $\{X_i, i \geq 1\}$ are independent random variables. If the behavior of V_n^2 is very erratic, then so is that of S_n, and it may not be possible to obtain a CLT.

So, if we have a reasonably large sample, we can assume that estimators obtained from estimating functions that are martingales, have an approximately normal distribution, which can be used for testing and constructing confidence intervals. A standard reference for the more general theory of martingale estimating functions is Sørensen (1999).

Billingsley (1961), and independently Ibragimov (1963), proved the central limit theorem for martingales with stationary and ergodic differences. For such martingales the conditional variance V_n^2 is asymptotically constant, i.e., $\frac{V_n^2}{s_n^2} \xrightarrow{P} 1$. Brown (1971) showed that the first part of condition (2) and not stationarity or ergodicity is crucial for such a result to hold. Further extensions in view of other central limit theorems for double arrays are based on Dvoretzky (1970) and McLeish (1974), where limit results employ a double sequence schema $\{X_{n,j}, 1 \leq j \leq k_n < \infty, n \geq 1\}$ and furnish conditions for the row sums $S_n = \sum_{j=1}^{k_n} X_{n,j}$ to converge in distributions to a mixture of normal distributions with means zero. A large variety of negligibility assumptions have been made about differences $X_{n,j}$ during the formulation of martingale central limit theorems. The classic condition of negligibility in the theory of sums of independent random variables asks the $X_{n,j}$ to be uniformly asymptotically negligible.

A comprehensive review on mainly one-dimensional martingales can be found in Helland (1982). Multivariate versions of the central limit theorem for martingales satisfying different conditions or applicable to different frameworks, can be found in Hutton and Nelson (1984), Sørensen (1991), Küchler and Sørensen (1999), Crimaldi and Pratelli (2005), and Hubalek and Posedel (2007).

Cross References

►Central Limit Theorems
►Markov Processes
►Martingales
►Statistical Inference for Stochastic Processes

References and Further Reading

Bernstein S (1927) Sur l'extension du théorèmelimite du calcul des probabilitiés aux sommes de quantités dépendantes. Math Ann 85:1–59
Billingsley P (1961) The Lindeberg-Lévy theorem for martingales. Proc Am Math Soc 12:788–792
Brown BM (1971) Martingale central limit theorems. Ann Math Stat 42:59–66
Chow YS, Teicher H (1997) Probability theory, 3rd edn. Springer, New York
Crimaldi I, Pratelli L (2005) Convergence results for multivariate martingales. Stoch Proc Appl 115(4):571–577
Doob JL (1953) Stochastic processes. Wiley, New York

Dvoretzky A (1970) Asymptotic normality for sums of dependent random variables. Proceedings of the Sixth Berkeley Symposium on Statistics and Probability. pp 513–535
Hall P, Heyde CC (1980) Martingale limit theory and its application. Academic, New York
Helland IS (1982) Central limit theorems for martingales with discrete or continuous time. Scand J Stat 9:79–94
Hubalek F, Posedel P (2007) Asymptotic analysis for a simple explicit estimator in Barndorff-Nielsen and Shephard stochastic volatility models. Thiele Research Report 2007–2005
Hutton JE, Nelson PI (1984) A mixing and stable central limit theorem for continuous time martingales. Technical Report No. 42, Kansas State University, Kansas
Ibragimov IA (1963) A central limit theorem for a class of dependent random variables. Theor Probab Appl 8:83–89
Küchler U, Sørensen M (1999) A note on limit theorems for multivariate martingales. Bernoulli 5(3):483–493
Lévy P (1935) Propriétés asymptotiques des sommes de variables aléatoires enchainées. Bull Sci Math 59(series 2):84–96, 109–128
McLeish DL (1974) Dependent Central Limit Theorems and invariance principles. Ann Probab 2:620–628
Sørensen M (1991) Likelihood methods for diffusions with jumps. In: Prabhu NU, Basawa IV (eds) Statistical inference in stochastic processes. Marcel Dekker, New York, pp 67–105
Sørensen M (1999) On asymptotics of estimating functions. Brazilian J Probab Stat 13:111–136
Williams D (1991) Probability with martingales. Cambridge University Press, Cambridge

Martingales

Rüdiger Kiesel
Professor, Chair for energy trading and financial services
Universität Duisburg-Essen, Duisburg, Germany

The fundamental theorem of asset pricing (The term *fundamental theorem of asset pricing* was introduced in Dybvig and Ross [1987]. It is used for theorems establishing the equivalence of an economic modeling condition such as no-arbitrage to the existence of the mathematical modeling condition existence of equivalent martingale measures.) links the martingale property of (discounted) asset price processes under a particular class of probability measures to the 'fairness' (in this context no arbitrage condition) of financial markets. In elementary models one such result is *In an arbitrage-free complete financial market model, there exists a unique equivalent martingale measure*, see e.g., Bingham and Kiesel (2004).

So despite martingales have been around for more than three and a half centuries they are still at the forefront of applied mathematics and have not lost their original motivation of describing the notion of fairness in games of chance. The *Oxford English Dictionary* lists under the word *martingale* (we refer to Mansuy [2009] for a interesting account of the etymology of the word): A system of gambling which consists in doubling the stake when losing in order to recoup oneself (1815).

Indeed, the archetype of a martingale is the capital of a player during a fair gambling game, where the capital stays "constant on average"; a supermartingale is "decreasing on average," and models an unfavourable game; a submartingale is "increasing on average," and models a favorable game.

Gambling games have been studied since time immemorial – indeed, the Pascal–Fermat correspondence of 1654 which started the subject was on a problem (de Méré's problem) related to gambling. The doubling strategy above has been known at least since 1815. The term "martingale" in our sense is due to J. Ville (1910–1989) in his thesis in 1939. Martingales were studied by Paul Lévy (1886–1971) from 1934 on (see obituary Loève (1973)) and by J.L. Doob (1910–2004) from 1940 on. The first systematic exposition was Doob (1953). Nowadays many very readable accounts exist, see Neveu (1975), Williams (1991) and Williams (2001).

Martingales are of central importance in any modelling framework which uses ▶stochastic processes, be it in discrete or continuous time. The concept has been central to the theory of stochastic processes, stochastic analysis, in mathematical statistics, information theory, and in parts of mathematical physics, see Kallenberg (1997) and Meyer (2009) for further details. The Martingale gambling insight 'You can't beat the system' establishes properties of martingale transforms and lays the foundation of stochastic integrals, Øksendal (1998). Martingale stopping results establish optimality criteria which help develop optimal strategies for decision problems (and exercising financial options), see Chow (1971) and Shiryaev (2007).

We can here only give a few fundamental definitions and results and point to the vast literature for many more exiting results.

For the definition, let I be a suitable (discrete or continuous) index set and assume that an index t is always taken from I. Given a stochastic basis $(\Omega, \mathcal{F}, I\!P, I\!F = \{\mathcal{F}_t\})$ (where the filtration $I\!F$ models the flow of information) we call a process $X = (X_t)$ a *martingale* relative to $(\{\mathcal{F}_t\}, I\!P)$ if

(i) X is adapted (to $\{\mathcal{F}_t\}$).
(ii) $I\!E|X_t| < \infty$ for all t.
(iii) For $s \le t$ we have $I\!E[X_t|\mathcal{F}_s] = X_s \quad I\!P - a.s.$.

X is a *supermartingale* if in place of (ii)

$$\mathbb{E}[X_t|\mathcal{F}_s] \le X_s \quad \mathbb{P}-a.s.;$$

X is a *submartingale* if in place of (iii)

$$\mathbb{E}[X_t|\mathcal{F}_s] \ge X_s \quad \mathbb{P}-a.s..$$

Basic examples are the mean-zero ►random walk: $S_n = \sum X_i$, with X_i independent, where for $\mathbb{E}(X_i) = 0$ S_n is a martingale (submartingales: positive mean; supermartingale: negative mean) and stock prices: $S_n = S_0\zeta_1\cdots\zeta_n$ with ζ_i independent positive r.vs with existing first moment. (See Williams (1991) and Williams (2001) for many more examples). In continuous time the central example is that of Brownian motion, see Revuz and Yor (1991), Karatzas and Shreve (1991), which of course is a central process for many branches of probability (see also ►Brownian Motion and Diffusions).

Now think of a gambling game, or series of speculative investments, in discrete time. There is no play at time 0; there are plays at times $n = 1, 2, \ldots$, and

$$\Delta X_n := X_n - X_{n-1}$$

represents our net winnings per unit stake at play n. Thus if X_n is a martingale, the game is "fair on average."

Call a process $C = (C_n)_{n=1}^{\infty}$ *predictable* if C_n is $\mathcal{F}_{n-1}$-measurable for all $n \ge 1$. Think of C_n as your stake on play n (C_0 is not defined, as there is no play at time 0). Predictability says that you have to decide how much to stake on play n based on the history *before* time n (i.e., up to and including play $n-1$). Your winnings on game n are $C_n\Delta X_n = C_n(X_n - X_{n-1})$. Your total (net) winnings up to time n are

$$Y_n = \sum_{k=1}^{n} C_k\Delta X_k = \sum_{k=1}^{n} C_k(X_k - X_{k-1}).$$

This constitutes the *Martingale transform* of X by C.

The central theorem for betting and applications in finance says that "You can't beat the system!," i.e., if X is a martingale then the martingale transform is a martingale (under some mild regularity conditions on C). So in the martingale case, predictability of C means we can't foresee the future (which is realistic and fair). So we expect to gain nothing – as we should, see e.g., Neveu (1975). Likewise one can analyze different strategies to stop the game, then Doob's stopping time principle reassures that it is not possible to beat the system, see e.g., Williams (2001).

Martingale transforms were introduced and studied by Burkholder (1966). They are the discrete analogs of stochastic integrals and dominate the mathematical theory of finance in discrete time, see Shreve (2004), just as stochastic integrals dominate the theory in continuous time, see Harrison and Pliska (1981). The various links between mathematical finance and martingale theory are discussed in Musiela and Rutkowski (2004) and Karatzas and Shreve (1998).

Martingale-convergence results are among the most important results in probability (arguably in mathematics). Hall and Heyde (1980) and Chow (1988) are excellent sources, but Doob (1953) lays the foundations. Martingale techniques play a central role in many parts of probability, consult Rogers (1994), Revuz and Yor (1991), Karatzas and Shreve (1991) or Kallenberg (1997) for excellent accounts. Martingales appear in time series theory and sequential analysis, see Lai (2009) and Hamilton (1994).

About the Author

Rüdiger Kiesel holds the chair of energy trading and financial services (sponsored by the Stifterverband für die Deutsche Wissenschaft and RWE Supply & Trading; the first such chair in Europe). Previously, he was Professor and Head of the Institute of Financial Mathematics at Ulm University. Kiesel also holds guest professorships at the London School of Economics and the Centre of Mathematical Applications at the University of Oslo. His main research areas are currently risk management for power utility companies, design and analysis of credit risk models, valuation and hedging of derivatives (interest-rate, credit- and energy-related), methods of risk transfer and structuring of risk (securitization), and the stochastic modelling of financial markets using Lévy-type processes. He is on the editorial board of the *Journal of Energy Markets* and co-author (with Nicholas H. Bingham) of the Springer Finance monograph *Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives* (2nd edition, 2004).

Cross References

►Brownian Motion and Diffusions
►Central Limit Theorems
►Khmaladze Transformation
►Martingale Central Limit Theorem
►Point Processes
►Radon–Nikodým Theorem
►Statistical Inference for Stochastic Processes
►Statistics and Gambling
►Stochastic Processes
►Stochastic Processes: Applications in Finance and Insurance
►Stochastic Processes: Classification

References and Further Reading

Bingham N, Kiesel R (2004) Risk-Neutral valuation: pricing and hedging of financial derivatives, 2nd edn. Springer, London
Burkholder DL (1966) Martingale transforms. Ann Math Stat 37:1494–1504

Chow YS, Teicher H (1988) Probability theory: independence, interchangeability, martingales, 2nd edn. Springer, New York
Chow YS, Robbins H, Siegmaund D (1971) Great expectations: the theory of optimal stopping. Houghton Mifflin, Boston
Doob JL (1953) Stochastic processes. Wiley, New York
Dybvig PH, Ross SA (1987) Arbitrage. In: Milgate M, Eatwell J, Newman P (eds) The new palgrave: dictionary of economics. Macmillan, London
Hall P, Heyde CC (1980) Martingale limit theory and applications. Academic, New York
Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
Harrison JM, Pliska SR (1981) Martingales and stochastic integrals in the theory of continuous trading. Stoch Proc Appl 11: 215–260
Kallenberg O (1997) Foundations of probability. Springer, New York
Karatzas I, Shreve S (1991) Brownian motion and stochastic calculus, 2nd edn, 1st edn 1988. Springer, Berlin
Karatzas I, Shreve S (1998) Methods of mathematical finance. Springer, New York
Lai TL (2009) Martingales in sequential analysis and time series, 1945–1985. Electron J Hist Probab Stat 5
Loève M (1973) Paul Lévy (1886–1971), obituary. Ann Probab 1:1–18
Mansuy R (2009) The origins of the word 'martingale'. Electron J Hist Probab Stat 5
Meyer P-A (2009) Stochastic processes from 1950 to the present. Electron J Hist Probab Stat 5
Musiela M, Rutkowski M (2004) Martingale methods in financial modelling, 2nd edn. Springer, Heidelberg
Neveu J (1975) Discrete-parameter martingales. North-Holland, Amsterdam
Øksendal B (1998) Stochastic differential equations: an introduction with applications, 5th edn. Springer, Berlin
Revuz D, Yor M (1991) Continuous martingales and Brownian motion. Springer, New York
Rogers L, Williams D (1994) Diffusions, Markov processes and martingales. Volume 1: foundations, 2nd edn. Wiley, Chichester
Shiryaev AN (2007) Optimal stopping rules, 3rd edn. Springer, Berlin
Shreve S (2004) Stochastic calculus for finance I: the binomial asset pricing model. Springer, New York
Williams D (1991) Probability with martingales. Cambridge University Press, Cambridge
Williams D (2001) Weighing the odds. Cambridge University Press, Cambridge

Mathematical and Statistical Modeling of Global Warming

Chris P. Tsokos
Distinguished University Professor
University of South Florida, Tampa, FL, USA

Introduction

Do we scientifically understand the concept of "Global Warming"? A very basic definition of "Global Warming" is an increase in temperature at the surface of the earth supposedly caused by the greenhouse effect, carbon dioxide, CO_2 (greenhouse gas). The online encyclopedia, Wikipedia, defines the phenomenon of "GLOBAL WARMING" as the increase in the average temperature of the earth's near surface air and oceans in the recent decades and its projected continuation.

For the past 3 years this has been a media chaos: pro and concerned skeptics. The Intergovernmental Panel of the United States on Climate Change (IPCC) – "Climate Change 2007" claimed that the following are some of the causes of Global Warming:

- Increase in temperature – Increase in sea level
- Unpredictable pattern in rainfall
- Increase in extreme weather events
- Increase in river flows
- Etc.

Furthermore, the award winning documentary narrated by Vice President Gore strongly supports the IPCC findings. However, the ABC news program 20/20 "Give Me a Break," raises several questions and disputes the process by which IPCC stated their findings. A number of professional organizations, the American Meteorological Society, American Geographical Union, AAAS, supported the subject matter. The U.S. National Academics blame global warming on human activities.

The concerned skeptics raise several points of interest concerning Global Warming. Great Britain's Channel 4 Documentary entitled "*The Great Global Warming Swindle*" disputes several of the aspects of Vice President former documentary. NASA scientists reveal through their scientific experiments and studies that the increase in atmospheric temperature is due to the fact that sea spots are hotter than previously thought. Their findings are also reported by the *Danish National Space Center*, DNSC, on similar investigations conducted by NASA. DNSC stated that there is absolutely nothing we can do to correct this situation. *Times Washington Bureau Chief*, Bill Adair, states that "Global Warming has been called the most dire issue facing the planet and yet, if you are not a scientist, it can be difficult to sort out the truth." The Wall Street Journal in a leading article "Global Warming is 300-year-old News," stated that "the various kind of evidence examined by the *National Research Council*, NRC, led it to conclude that the observed disparity between the surface and atmospheric temperature trends during the 20-year period is probably at least partially real." It further stated that "uncertainties in all aspects exist- cannot draw any conclusion concerning *Global Warming*." However, the NRC study concluded with an important statement that "major advances in scientific

methods will be necessary before these questions on *Global Warming* can be resolved."

Furthermore, the temperature increase that we are experiencing are infinitesimal, during the past 100 years – the mean global surface air temperature increased by approximately $1.3^{\circ}F$ $(0.32^{\circ}F)$. Dr. Thomas G. Moore, Senior Fellow at the Hoover Institute at Stanford University, in his article entitled "Climate of Fear: Why We Shouldn't Worry About Global Warming" is not concerned with such small changes in temperatures. Furthermore, in his interview with *Newsweek*, he said more people die from cold than from warmth and an increase of a few degrees could prevent thousands of deaths.

It is well known that carbon dioxide, CO_2, and surface/atmospheric temperatures are the primary cause of "GLOBAL WARMING." Jim Verhult, Perspective Editor, *St. Petersburg Times*, writes, "carbon dioxide is invisible – no color, no odor, no taste. It puts out fires, puts the fizz in seltzer and it is to plants what oxygen is to us. It's hard to think of it as a poison." The U.S.A. is emitting approximately 5.91221 billion metric tons of CO_2 in the atmosphere, which makes us the world leader; however, by the end of 2007, the Republic of China became the new leader. Temperatures and CO_2 are related in that as CO_2 emissions increase, the gasses start to absorb too much sunlight and this interaction warms up the globe. Thus, the rise in temperature and the debate of "GLOBAL WARMING."

While working on the subject matter, an article appeared on the front page of the *St. Petersburg Times* on January 23, 2007. This article, entitled "Global Warming: Meet your New Adversary," was written by David Adams. The highlight of this article was a section called "By the Numbers," which stated some information concerning the continental United States: 2006 hottest year; U.S. top global warming polluter; 20% increase of CO_2 since 1990; 15% of CO_2 emissions by 2020; 78 number of days U.S. fire season has increased; and 200 million people that will be displaced due to global warming. Our data for the continental U.S. does not support the first four statistics, we have no data for the fifth, and the sixth is quite hypothetical. The final assertion, with "0" representing the number of federal bills passed by the Congress to cap America's global warming pollution. Thus, it is very important that we perform sophisticated statistical analysis and modeling to fully understand the subject matter. Also, very recently, the Supreme Court of the U.S., in one of its most important environmental decisions, ruled that the Environmental Protection Agency (EPA) has the authority to regulate the greenhouse gases that contribute to global climate changes unless it can provide a scientific basis for its refusal.

We believe that a contributing factor in creating these controversies among scientists (and this is passed onto the policymakers and the media) is a lack of precise and accurate statistical analysis and modeling of historical data with an appropriate degree of confidence. The problem of "GLOBAL WARMING" is very complex with a very large number of contributing entities with significant interactions. The complexity of the subject matter can be seen in the attached diagram "A Schematic View" (Fig. 1). We believe that statisticians/mathematicians can help to create a better understanding of the subject problem that hopefully will lead to the formulation of legislative policies.

Thus, to scientifically make an effort to understand "Global Warming," we must study the marriage of CO_2 and atmosphere temperature, individually and together, using available historical data. Here we shall briefly present some parametric statistical analysis, forecasting models for CO_2 and atmospheric temperature, T_a along with a differential equation, that give the rate of change of CO_2 as a function of time. Scientists can utilize these preliminary analysis and models to further the study of Global Warming. Additional information can be found in Tsokos (2007a, b), and Tsokos 2008b.

Atmospheric Temperature, T_a

Here we shall utilize historical temperature data recorded in the Continental United States from 1895 to 2007, to parametrically identify the probability density of the subject data and to develop a forecasting model to predict short and long term values of T_a.

The probability density function, pdf, of T_a is the three-parameter lognormal pdf. It is given by

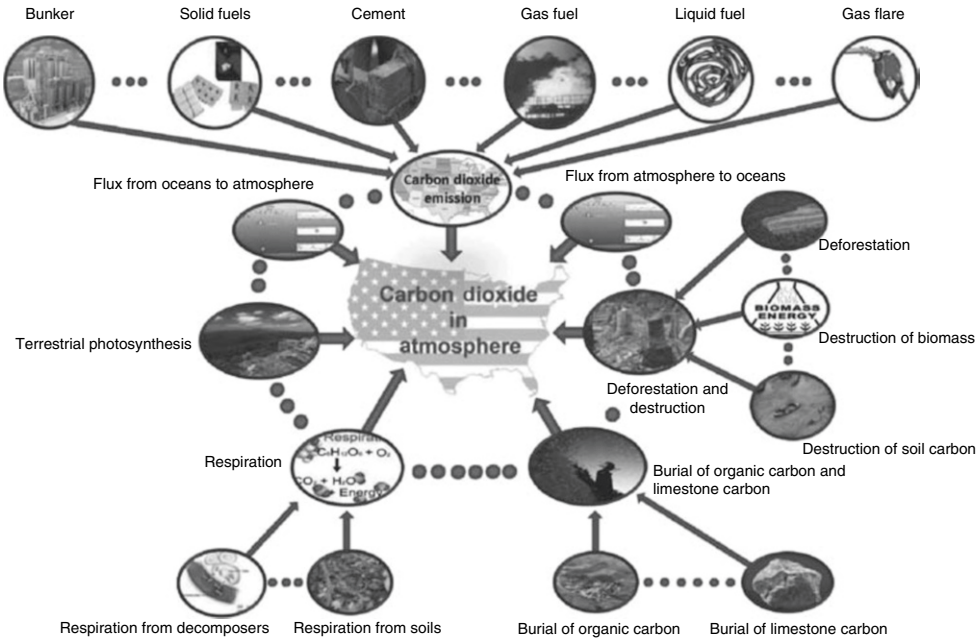
$$f(t;\mu,\theta,\sigma) = \frac{\exp\left\{-\frac{1}{2}\left[\ln(t-\theta)-\mu\right]^2\right\}}{(t-\theta)\sigma\sqrt{2\pi}},\ t \geq \theta, \sigma, \mu > 0, \tag{1}$$

where μ, σ and θ, are the scale, shape and location parameters, respectively.

For the given T_a data the maximum likelihood estimation of population parameter, μ, σ and θ are $\hat{\mu} = 3.59, \hat{\sigma} = 0.019$ and $\hat{\theta} = 0.195$. Thus, the actual pdf that we will be working with is given by

$$f(t;\hat{\mu},\hat{\theta},\hat{\sigma}) = \frac{\exp\left\{-\frac{1}{2}\left[\ln(t-0.195)-2.59\right]^2\right\}}{(t-0.195)\cdot 0.019\sqrt{2\pi}},\ t \geq 0.195. \tag{2}$$

Having identified the pdf that probabilistically characterizes the behavior of the atmospheric T_a, we can obtain the expected value of T_a, all the useful basic statistics along with being able to obtain confidence limits on the true T_a.



Mathematical and Statistical Modeling of Global Warming. Fig. 1 Carbon dioxide (CO_2) in the atmosphere in USA "A Schematic View"

Such a pdf should be applicable in other countries around the world.

The subject data, T_a, is actually a stochastic realization and is given as nonstationary time series. The development of the multiplicative seasonal autoregressive integrated moving average, ARIMA model is defined by

$$\Phi_p(B^s)\phi(1-B)^d(1-B^s)^D x_t = \theta_q(B)\Gamma_Q(B^s)\varepsilon_t, \qquad (3)$$

where p is the order of the autoregressive process; d is the order of regular differencing; q is the order of the moving average process; P is the order of the seasonal autoregressive process; D is the order of the seasonal differencing; Q is the order of the seasonably moving average process; and s refers to the seasonal period, and

$$\begin{aligned}\phi_p(B) &= (1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)\\ \theta_q(B) &= (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)\\ \Phi_P(B^s) &= 1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps}\\ \Gamma_Q(B^s) &= 1-\Gamma_1 B^s-\Gamma_2 B^{2s}-\cdots-\Gamma_Q B^{Qs}.\end{aligned}$$

The developing process of (3) using the actual data is complicated and here we present the final useful form of the model. The reader is referred to Shih and Tsokos (2007, 2009) for details.

The estimated forecasting model for the atmospheric data is given by

$$\begin{aligned}\hat{x}_t &= 1.0941x_{t-1} - 0.057x_{t-2} - 0.0371x_{t-3} + 0.9954x_{t-12}\\ &\quad - 1.0891x_{t-13} + 0.0567x_{t-14} + 0.0369x_{t-15}\\ &\quad + 0.0046x_{t-24} + 0.0895x_{t-25} - 0.0004x_{t-26}\\ &\quad + 0.0017x_{t-27} - 0.9861\varepsilon_{t-1} - 0.9742\Gamma_1\varepsilon_{t-12}\\ &\quad + 0.9607\varepsilon_{t-13}. \qquad (4)\end{aligned}$$

The mean of the residuals, $\bar{r}$, the variance, S_r^2, the standard deviation, S_r, standard error, SE, and the mean square error, MSE, are presented below for one unit of time ahead forecasting.

$\bar{r}$	S_r^2	S_r	SE	MSE
−0.008512476	4.331902	2.081322	0.05673052	4.328756

These numerical results give an indication of the quality of the developed model.

Carbon Dioxide, CO_2

Parametric Analysis

The other most important entity in Global Warming is CO_2. The complexity of CO_2 in the atmosphere is illustrated by the schematic diagram that was introduced. To

better understand CO_2, we need to probabilistically determine the best probability distribution, pdf, that characterizes its behavior. Presently, scientists working on the subject matter make the assumption that CO_2 in the atmosphere follows the classical Gaussian pdf and that is not the best possible fit of the actual data and could lead to misleading decisions. The actual data that we are using was collected in the Island of Hawaii/Mauna Loa from 1990 to 2004. Through goodness-of-fit statistical testing, the best fit of the CO_2 data that we can study its behavior probabilistically is the three-parameter Weibull pdf. The cumulative three-parameter Weibull probability distribution is given by

$$F(x) = 1 - \exp\left\{-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right\},\ \gamma \le x < \infty, \delta > 0, \beta > 0 \quad (5)$$

where α, β, and γ are the shape, scale, and location parameter. The nth moment, mean and variance are given by

$$m_n = \beta^n \Gamma\left(1+\frac{n}{\alpha}\right),\ \mu = \beta\Gamma\left(1+\frac{1}{\alpha}\right) \text{ and } \sigma^2 = \beta^2\Gamma\left(1+\frac{2}{\alpha}\right) - \mu,$$

respectively, where Γ is the gamma function. The approximate maximum likelihood estimates of the true parameters, α, β and γ for the Hawaii data are given by

$$\hat{\alpha} = 2.108,\ \hat{\beta} = 17.092,\ \text{and } \hat{\gamma} = 349.6.$$

Thus, the cumulative pdf that we can use to probabilistically characterize the CO_2 behavior and answer related questions is given by:

$$F(x) = 1 - \exp\left\{-\left(\frac{x-349.6}{17.092}\right)^{2.108}\right\}. \quad (6)$$

For additional details of the subject area see Shih and Tsokos (2009).

Forecasting Model of CO_2

Here we present a forecasting model of CO_2 in the atmosphere. Having such a model will allow us to accurately predict the amount of CO_2 in the atmosphere, and make appropriate decisions as needed. The actual CO_2 data as a function of time results in a nonstationary time series. For details in the development of this model, see Shih and Tsokos (2009). The best forecasting model that we developed is an ARIMA model with second order autoregressive process, with a first order moving average process and a 12-month seasonal effect. Its final form is given by

$$\begin{aligned}\hat{CO}_{2_A} = {} & 0.6887x_{t-1} + 0.1989x_{t-2} + 0.1124x_{t-3} + 1.0759x_{t-12} \\ & - 0.74097x_{t-13} - 0.213997x_{t-14} - 0.12093x_{t-15} \\ & - 0.0683x_{t-24} + 0.047038x_{t-25} + 0.013585x_{t-26} \\ & + 0.00768x_{t-27} - 0.00076x_{t-36} + 0.005234x_{t-37} \\ & + 0.0015116x_{t-38} + 0.00085x_{t-39} - 0.8787\varepsilon_{t-12}.\end{aligned}$$

A similar statistical model can be developed for CO_2 emission, Shih and Tsokos (2009).

A Differential Equation of CO_2 in the Atmosphere

The main attributable variables in CO_2 in the atmosphere are:

E: CO_2 emission (fossil fuel combination)
D: Deforestation and destruction
R: Terrestrial plant respiration
S: Respiration
O: the flux from oceans to atmosphere
P: terrestrial photosynthesis
A: the flux from atmosphere to oceans
B: Burial of organic carbon and limestone carbon

One important question that we would like to know is the rate of change of CO_2 as a function of time. The general form of the differential equation of the subject matter is of the form:

$$\frac{d(CO_2)}{dt} = f(E, D, R, S, O, P, A, B)$$

or

$$CO_{2_A} = \int (E + D + R + S + (O - A) - P - B)dt.$$

Here, B, P and R are constants, thus

$$\begin{aligned}CO_{2_A} = {} & \int (k_E E + k_D D + k_R R + k_S S + k_{O-A}(O - A) \\ & + k_P P - k_B B)dt.\end{aligned}$$

Using the available data we can estimate the functional analytical form of all the attributable variables that appear

in the integrand. Thus, the final working form of CO_2 in the atmosphere is given by

$$CO_2 = \begin{cases} k_E\{-593503t + 2.4755 \times 10^9 e^{-\frac{1}{1200}}\} \\ +k_D(10730.5t + 0.01625t^2) \\ +k_S\{-0.132(1995 + \frac{t}{12})^4 + 1054.4(1995 + \frac{t}{12})^3 \\ -315462(1995 + \frac{t}{12})^2 + 3 \times 10^8 t\} \\ +K_{A-O}\{42.814t - 4.2665t^2 \\ +0.0967t^3\} - k_P \int Pdt - k_B \int Bdt \end{cases}.$$

Having a workable form of the differential equation, we can develop the necessary algorithm to track the influence the attributable variables will have in estimating the change of rate of CO_2 as a function of time.

Conclusion

Finally, is the "Global Warming" phenomenon real? Yes. However, it is not as urgent as some environmentalists claim. For example, our statistical analytical models predict that in the next 10 years, 2019, we will have an increase of carbon dioxide in the atmosphere in the continental U.S. of approximately 7%. In developing a strategic legislative plan, we must address the economic impact it will have in our society. In our present global economic crisis, introducing legislation to address Global Warming issues will present additional critical economic problems. In a global context we must consider about 155 economic developing countries that have minimal to no strategic plans in effect that collect the necessary information that addresses the subject matter in their country. Furthermore, we have approximately 50 undeveloped countries that have minimum understanding about the concept of global warming. Thus, talking about developing global strategies and policies about "Global Warming" is quite premature.

Acknowledgments

This article is a revised and extended version of the paper published in *Hellenic News of America*, 23, 3, November 2009.

About the Author

Chris P. Tsokos is Distinguished University Professor of Mathematics and Statistics and Director of the Graduate Program in Statistics at the University of South Florida. He is the author/co-author of more than 285 research journal publications and more than 20 books plus special volumes. He has also directed more than 37 Ph.D. theses as a major professor. Dr. Tsokos is the recipient of many distinguished awards and honors, including Fellow of the American Statistical Association, USF Distinguished Scholar Award, Sigma Xi Outstanding Research Award, USF Outstanding Undergraduate Teaching Award, USF Professional Excellence Award, URI Alumni Excellence Award in Science and Technology, Pi Mu Epsilon, election to the International Statistical Institute, Sigma Pi Sigma, USF Teaching Incentive Program, and several humanitarian and philanthropic recognitions and awards. Professor Tsokos is an Editor/Chief-Editor/Co-Chief Editor of a number of journals including *International Journal of Environmental Sciences*, *International Journal of Mathematical Sciences*, *International Journal of Business Systems*, *International Journal of Nonlinear Studies*, and *Nonlinear Mathematics, Theory, Methods and Applications*. He also serves as an Associate Editor for a number of international journals.

"Professor Chris P. Tsokos' contributions to statistics, mathematical sciences, engineering and international education over a period of almost a half century are well-known, well-recognized and well-documented in the literature. In particular, his most notable work in the Bayesian reliability, stochastic dynamic systems and statistical modeling in a nonlinear and nonstationary world is well-recognized and well-established." (G. S. Ladde and M. Sambandham (2008). Professor Chris P. Tsokos: a brief review of statistical, mathematical and professional contributions and legacies, *Neural, Parallel & Scientific Computations*, 16 (1), Special issue in honor of Dr. Chris P. Tsokos.)

Cross References

►Environmental Monitoring, Statistics Role in
►Forecasting with ARIMA Processes
►Marine Research, Statistics in
►Statistics and Climate Change
►Time Series

References and Further Reading

Hachett K, Tsokos CP (2009) A new method for obtaining a more effective estimate of atmospheric temperature in the continental United States. Nonlinear Anal-Theor 71(12):e1153–e1159

Shih SH, Tsokos CP (2007) A weighted moving average procedure for forecasting. J Mod Appl Stat Meth 6(2):619–629

Shih SH, Tsokos CP (2008a) A temperature forecasting model for the continental United States. J Neu Par Sci Comp 16:59–72

Shih SH, Tsokos CP (2008b) Prediction model for carbon dioxide emission in the atmosphere (2008). J Neu Par Sci Comp 16: 165–178
Shih SH, Tsokos CP (2009) A new forecasting model for nonstationary environmental data. Nonlinear Anal-Theor 71(12):e1209–e1214
Tsokos CP (2007a) St. Petersburg Times, Response to "Global Warming: Meet Your News Adversary"
Tsokos CP (2007b) Global warming: MEDIA CHAOS: can mathematics/statistics help? International Conference on Dynamical Systems and Applications, Atlanta, GA
Tsokos CP (2008a) Statistical modeling of global warming. Proc Dyn Syst Appl 5:460–465
Tsokos CP (2008b) Global warming (2008). The Fifth World Congress of IFNA (July 2–9, Orlando, Florida)
Tsokos CP, Xu Y (2009) Modeling carbon dioxide emission with a system of differential equations. Nonlinear Anal-Theor 71(12):e1182–e1197
Wooten R, Tsokos CP (2010) Parametric analysis of carbon dioxide in the atmosphere. J Appl Sci 10:440–450

Maximum Entropy Method for Estimation of Missing Data

D. S. Hooda
Professor and Dean (Research)
Jaypee University of Engineering and Technology, Guna, India

In field experiments we design the field plots. In case we find one or more observations missing due to natural calamity or destroyed by a pest or eaten by animals, it is cumbersome to estimate the missing value or values as in field trials it is practically impossible to repeat the experiment under identical conditions. So we have no option except to make best use of the data available. Yates (1933) suggested a method: "Substitute x for the missing value and then choose x so as to minimize the error sum of squares."

Actually, the substituted value does not recover the best information, however, it gives the best estimate according to a criterion based on the least square method. For the randomized block experiment

$$x = \frac{pP + qQ - T}{(p-1)(q-1)}, \tag{1}$$

where

p = number of treatments;
q = number of blocks;
P = total of all plots receiving the same treatment as the missing plot;
Q = total of all plots in the same block as the missing plot; and
T = total of all plots.

For the Latin Square Design, the corresponding formula is

$$x = \frac{p(P_r + P_c + P_t) - 2T}{(p-1)(q-1)}, \tag{2}$$

where

p = number of rows or columns of treatments;
P_r = total of row containing the missing plot;
P_c = total of column containing the missing plot;
P_t = total of treatment contained in the missing plot; and
T = grand total.

In case more than one plot yields are missing, we substitute the average yield of available plots in all except one of these and substitute x in this plot. We estimate x by Yate's method and use this value to estimate the yields of other plots one by one.

Next we discuss the maximum entropy method. If $x_1, x_2, \ldots, x_n$ are known yields and x is the missing yield. We obtain the maximum entropy estimate refer to Kapur and Kesavan (1992) for x by maximizing:

$$-\sum_{i=0}^{n} \frac{x_i}{T+x} \log \frac{x_i}{T+x} - \frac{x}{T+x} \log \frac{x}{T+x}. \tag{3}$$

Thus we get

$$\hat{x} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T}}, \tag{4}$$

where $T = \sum_{i=1}^{n} x_i$.

The value given by (4) is called maximum entropy mean of $x_1, x_2, \ldots, x_n$.

Similarly, if two values x and y are missing, x and y are determined from

$$\hat{x} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T+y}}, \tag{5}$$

$$\hat{y} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T+x}}. \tag{6}$$

The solution of (5) and (6) is

$$\hat{x} = \hat{y} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T}}. \tag{7}$$

Hence all the missing values have the same estimate and this does not change if the missing values are estimated one by one.

There are three following drawbacks of the estimate given by (4)

(1) $\hat{x}$ is rather unnatural. In fact $\hat{x}$ is always greater than arithmetic mean of $x_1, x_2, \ldots, x_n$.
(2) If two values are missing, the maximum entropy estimated for each is the same as given by (7).
(3) This is not very useful for estimating missing values in design of experiments.

The first drawback can be overcome by using generalized measure of entropy instead of Shannon entropy. If we use Burg's measure given by

$$B(P) = \sum_{i=1}^{n} \log p_i. \tag{8}$$

Then we get the estimate

$$\hat{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}. \tag{9}$$

In fact we choose a value $\hat{x}$, which is as equal to $x_1, x_2, \ldots, x_n$ as possible and so we maximize a measure of equality. Since there are many measures of equality, therefore our estimate will also depend on the measure of equality we choose.

The second drawback can be understood by considering the fact that the information theoretic estimate for a missing value depends on:

(a) The information available to us
(b) The purpose for which missing value is to be used.

As for the third drawback, according to the principle of maximum entropy, we should use all the information given to us and avoid scrupulously using any information not given to us. In design of experiments, we are given information about the structure of the design, which we are not using this knowledge in estimating the missing values. Consequently, the estimate is not accurate; however, information theoretic model defined and studied by Hooda and Kumar (2005) can be applied to estimate the missing value x_{ij} in contingency tables. Accordingly, the value x_{ij} is to be chosen to minimize the measure of dependence D.

About the Author

Professor D. S. Hooda is Vice President of the International Forum of Interdisciplinary Mathematics. He is General Secretary of Indian Society of Information Theory and Applications. He is an Elected member of the International Statistical Institute. American Biographical Institute, USA, chose him in 2004 for his outstanding research and conferred with honorary appointment to Research Board of Advisors of the institute. Indian Society of Information Theory has bestowed on him a prestigious award in 2005 for his outstanding contribution and research in information theory. He was Pro-Vice-Chancellor of Kurukshetra University. He has published about 80 papers in various journals and four books in mathematics and statistics. Presently, Professor Hooda is Dean (Research) Jaypee Institute of Engineering and Technology, Raghogarh, Guna.

Cross References

►Entropy
►Estimation
►Estimation: An Overview
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sampling From Finite Populations

References and Further Reading

Hooda DS, Kumar P (2005) Information theoretic model for analyzing independence of attributes in contingency table. Paper presented at the international conference held at Kuala Lumpur, Malayasia, 27–31 Dec 2005

Kapur JN, Kesavan HK (1992) Entropy optimization principles with applications. Academic, San Diego

Yates F (1933) The analysis of replicated experiments when the field experiments are incomplete. Empire J Exp Agr 1:129–142

Mean, Median and Mode

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Mean, median and mode indicate central point of distribution or data set. Let P_X denotes distribution of a random variable X. Any reasonable rule $\mathcal{O} = \mathcal{O}(P_X)$ indicating a point $\mathcal{O}$ to be the center of P_X should satisfy the following postulates:

A1 If $P(a \le X \le b) = 1$ then $a \le \mathcal{O}(P_X) \le b$
A2 $\mathcal{O}(P_{X+c}) = \mathcal{O}(P_X) + c$ for any constant c [transitivity]
A3 $\mathcal{O}(P_{cX}) = c\mathcal{O}(P_X)$ for any constant c [homogeneity]

The *mean* is a synonym of the first moment, i.e. the expected value EX. For a continuous random variable X it may be expressed in terms of density function $f(x)$, as the integral $EX = \int_{-\infty}^{+\infty} xf(x)dx$. In discrete case it is defined as the sum of type $EX = \sum_i x_i p_i$, where x_i is a possible value of X, $i \in I$, while $p_i = P(X = x_i)$ is its probability. The mean fulfils all the above postulates and, moreover, an extra condition

AM $E(X-EX)^2 \le E(X-c)^2$ for any $c \in R$

It is worth to add that mean may not exist.

The *median* $Me = Me(X)$ is a scalar α defined by conditions $P_X(X \le \alpha) \ge \frac{1}{2}$ and $P_X(X \ge \alpha) \ge \frac{1}{2}$. In terms of the cumulative distribution function $F = F_X$ it means that $F(\alpha) \ge \frac{1}{2}$ and $\lim_{x \uparrow \alpha} F(\alpha) \le \frac{1}{2}$. In particular, if X is continuous with density f, then the desired conditions reduces to $\int_{-\infty}^{\alpha} f(x)dx \ge \frac{1}{2}$ and $\int_{\alpha}^{\infty} f(x)dx \ge \frac{1}{2}$. In discrete case it can be expressed in the form $\sum_{\{i:x_i \le \alpha\}} p_i \ge \frac{1}{2}$ and $\sum_{\{i:x_i \ge \alpha\}} p_i \ge \frac{1}{2}$. The median also satisfies the conditions $A1 - A3$ and, moreover

AMe $E|X - MeX| \le E|X - c|$ for any $c \in R$.

The mode $Mo = Mo(X)$ of a random variable X is defined in terms of its density function f (continuos case) or its probability mass function $p_i = P(X = x_i)$ (discrete case). Namely, $Me(X) = \arg\max f(x)$, or is an element x in the set of possible values $\{x_i : i \in I\}$ that $P(X = x) = \max\{p_i : i \in I\}$. The mode also satisfies the conditions $A1 - A3$. It is worth to add that mode may not be unique. There exist bimodal and multimodal distributions. Moreover the set of possible modes may be interval.

In the context of data set, represented by a sequence $x = (x_1, \ldots, x_n)$ of observations, the postulates $A1 - A3$ may be reformulated as follows:

S1 $\mathcal{O}(x_{i_1}, \ldots, x_{i_n}) = \mathcal{O}(x_1, \ldots, x_n)$ for any permutation $i_1, .., i_n$ of the indices $1, \ldots, n$
S2 $\min\{x_1, \ldots, x_n\} \le \mathcal{O}(x_1, \ldots, x_n) \le \max\{x_1, \ldots, x_n\}$
S3 $\mathcal{O}(x_1 + c, \ldots, x_n + c) = \mathcal{O}(x_1, \ldots, x_n) + c$
S4 $\mathcal{O}(cx_1, \ldots, cx_n) = c\mathcal{O}(x_1, \ldots, x_n)$.

In this case the mean, median and mode are defined as follows.

The mean of the data $x = (x_1, \ldots, x_n)$, denoted usually by $\overline{x}$, is the usual arithmetic average $\overline{x} = \frac{1}{n}\sum x_i$. The mean not only satisfies all conditions $S1 - S4$ but also possesses the property

SM $\sum_{i=1}^n (x_i - \overline{x})^2 \le \sum_{i=1}^n (x_i - c)^2$ for all $c \in R$.

Now let us arrange the elements of the sequence $x = (x_1, \ldots, x_n)$ in the not decreasing order $x_{[1]} \le x_{[2]} \le \ldots \le x_{[n]}$. The median of the data set $x = (x_1, \ldots, x_n)$ is defined by the formula

$$Me(x) = \begin{cases} x_{\left[\frac{n+1}{2}\right]}, & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & \text{if } n \text{ is even.} \end{cases}$$

The median satisfies the conditions $S1 - S4$ and, moreover,

SMe $\sum_{i=1}^n |x_i - Me(x)| \le \sum_{i=1}^n |x_i - c|$ for all $c \in R$.

The mode of the data $x = (x_1, \ldots, x_n)$, denoted by $Mo(x)$, is the value in the set that occurs most often. For instance if $x = (7, 3, 18, 24, 9, 3, 18)$ then $x \uparrow = (3, 7, 9, 13, 18, 18, 24)$. For such data $Me(x) = x_{[4]} = 13$ and $Mo(x) = 18$.

It is worth to add that the mean is very sensitive for outlying observations.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Expected Value
►Geometric Mean
►Harmonic Mean
►Mean, Median, Mode: An Introduction
►Random Variable
►Robust Statistical Methods
►Sampling Distribution
►Skewness

References and Further Reading

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton
Joag-Dev K (1989) MAD property of median. A simple proof. Am Stat 43:26–27
Prokhorov AW (1982a) Expected value. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 3. Soviet Encyclopedia, Moscow, pp 600–601 (in Russian)
Prokhorov AW (1982b) Mode. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 3. Soviet Encyclopedia, Moscow p 763 (in Russian)

Mean, Median, Mode: An Introduction

S. N. Gupta
University of South Pacific, Suva, Fiji

Introduction

Mean, median and mode are three statistical measures commonly used to summarize data sets. They are known by the common name *average*. In its broadest sense, an *average* is simply any single value that is representative of

many numbers. Averages are also called *measures of central tendency* because an average is usually located near the center of the data set. Some examples: average age of the players of a cricket team, average reaction time of a particular chemical, average amount spent by a customer in a shopping mall, etc.

The Mean

The *mean*, also known as *arithmetic mean*, is the most widely used average and is defined as the sum of the observations divided by the number of observations. The formula for computing mean is: $\bar{x} = (\sum x)/n$, where $\bar{x}$ is the symbol for mean (pronounced "x-bar"), x is the *symbol* for variable, $\sum x$ is the *sum* of observations (i.e., the sum of the values of the variable x) and n is the *number* of observations.

Although, there are also other kinds of means (such as the ▸harmonic mean and the ▸geometric mean), the arithmetic mean is by far the most popular. For this reason, the word arithmetic is rarely used in practice and we simply refer to the "mean."

Example 1 The ages (in weeks) of five babies are 5, 9, 8, 6 and 10. Find the mean.

Solution: The mean of the set is given by $\bar{x} = \frac{1}{n}\sum x = \frac{5+9+8+6+10}{5} = \frac{38}{5} = 7.6$ weeks.

Calculation of Mean for Discrete Frequency Distribution

Sometimes, it is convenient to represent the data in form of a frequency distribution. In such cases the formula for mean is: $\bar{x} = \frac{\sum fx}{\sum f}$, where f is the frequency, $\sum f$ is the sum of the frequencies, $\sum fx$ is the sum of each observation multiplied by its frequency.

Example 2 Data for numbers of children in 35 families are given below. Find the mean.

No. of children (x):	0	1	2	3	4
Frequency (f):	2	9	11	8	5

Solution:

x	0	1	2	3	4	
f	2	9	11	8	5	$\sum f = 35$
fx	0	9	22	24	20	$\sum fx = 75$

The mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{75}{35} = 2.1$ children per family.

Calculation of Mean for Grouped Frequency Distribution

It is not possible to calculate exact mean in grouped frequency distribution, because some information is lost when the data are grouped. So, only an approximate value of mean is obtained based on the assumption that all observations in a class interval occur at the *midpoint* (x_m) of that interval. Thus, the formula of Example 2 can be used after replacing x by x_m.

Example 3 The following is the distribution of the number of fish caught by 50 fishermen in a village. Find the mean number of fish caught by a fisherman.

No. of fish caught:	11–15	16–20	21–25	26–30
No. of fishermen:	12	14	13	11

Solution:

No. of fish caught	Midpoint (x_m)	f	fx_m
11–15	13	12	156
16–20	18	14	252
21–25	23	13	299
26–30	28	11	308
		$\sum f = 50$	$\sum fx_m = 1015$

Therefore, the mean is $\bar{x} = \frac{\sum fx_m}{\sum f} = \frac{1015}{50} = 20.3$ fish per fisherman.

Weighted Mean

When *weights* (measures of relative importance) are assigned to observations, weighted means are used. If an observation x is assigned a weight w, the weighted mean is given by $\bar{x} = \sum wx / \sum w$.

The Median

The *median* is another kind of average. It is defined as the centre value when the data are arranged in order of magnitude. Thus, the median is a value such that 50% of the data are below median and 50% are above median.

Calculation of Median for Raw Data

The observations are first arranged in ascending order of magnitude. If there are n observations, the median is

1. The value of the $[(n+1)/2]$th observation, *when n is odd.*
2. The mean of the $[n/2]$th and $[(n/2)+1]$th observations, *when n is even.*

Example 4 Find the median for the following data set:

16, 32, 20, 13, 13, 24, 10.

Solution: Arranging the data in ascending order we have

10, 13, 13, 16, 20, 24, 32.

Here, $n = 7$, which is odd. Therefore, median = $\frac{n+1}{2}$ th score $= \frac{7+1}{2}$ th score $=$ 4th score $=$ 16.

Example 5 Find the median for the data:

17, 18, 26, 30, 19, 24, 20, 22, 29, 25.

Solution: Here, n = 10, which is even. Arranging the data in ascending order we have

17, 18, 19, 20, 22, 24, 25, 26, 29, 30.

$$\text{Therefore, median} = \frac{1}{2}\left[\frac{n}{2}\text{th score} + \left(\frac{n}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}\left[\frac{10}{2}\text{th score} + \left(\frac{10}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}[5\text{th score} + 6\text{th score}]$$
$$= \frac{1}{2}[22 + 24] = 23.$$

Calculation of Median for Discrete Frequency Distribution
The same basic formulae as used for raw data are used, but cumulative frequencies are calculated for convenience of locating the observations at specific numbers.

Example 6 Data for the number of books purchased by 28 customers are given below. Find the median.

No. of books (x):	1	2	3	4
No. of customers (f) :	5	9	8	6

Solution:

No. of books (x)	1	2	3	4
No. of customers (f)	5	9	8	6
Cumulative frequency ($c.f.$)	5	14	22	28

Here $n = \sum f = 28$ (even). Therefore,

$$\textit{median} = \frac{1}{2}\left[\frac{28}{2}\textit{th score} + \left(\frac{28}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}[14\text{th score} + 15\text{th score}] = \frac{1}{2}[2+3] = 2.5$$

Calculation of Median for Grouped Frequency Distribution
In a grouped distribution, exact median cannot be obtained because some information is lost in grouping. Here, we first locate the *median class* and then obtain an estimate of the *median* by the formula:

$$\textit{median} = l_1 + \frac{\left(\frac{n}{2} - c\right)}{f}(l_2 - l_1)$$

where, l_1, l_2 are the lower and upper boundaries of the median class, f is the frequency of the median class, n is the sum of all frequencies and c is the cumulative frequency of the class immediately preceding the median class.

Example 7 Find the median for the data of Example 3 above.

Solution: Construct a table for class boundaries and cumulative frequencies:

Class	Class boundaries	f	$c.f.$
11–15	10.5–15.5	12	12
16–20	15.5–20.5	14	26
21–25	20.5–25.5	13	39
26–30	25.5–30.5	11	50
		$n = 50$	

Here, $n/2 = 25$. The median will lie in the class having cumulative frequency ($c.f.$) just larger than 25. The median class is 16–20. Thus, $l_1 = 15.5$, $l_2 = 20.5$, $c = 12$, $f = 14$.

Hence, $\textit{median} = 15.5 + \left(\frac{25-12}{14}\right) \times 5 = 15.5 + 4.64 = 20.14$.

The Mode

The *mode* is the most *frequent* value i.e., the value that has the largest frequency. A major drawback of mode is that a data set may have more than one mode or no mode at all. Also the mode may not always be a central value as in the Example 8(a) below.

Example 8 Find mode in the following data sets:

(a) 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 9.

(b) 12, 14, 15, 15, 15, 19, 19, 19, 20, 20.

(c) 11, 15, 16, 19, 21, 23, 26, 27, 29, 30.

Solution
(a) One mode at 9, (b) Two modes at 15 and 19, (c) No mode as each value occurs only once. For grouped frequency distribution, the mode can be estimated by taking the mid-point of the *modal class* corresponding to the

largest frequency. One advantage of mode is that it can be calculated for both kinds of data, qualitative and quantitative, whereas mean and median can be calculated for only quantitative data. E.g., A group consists of five Hindus, six Muslims and nine Christians. Here, Christianity is most frequent and so it is the mode of this data set.

Remarks If a distribution is symmetrical then mean = median = mode. For skewed distributions a thumb rule (though not without exceptions) is that if the distribution is skewed to the right then mean > median > mode and the inequalities are reversed if the distribution is skewed to the left.

To sum up, there is no general rule to determine which average is most appropriate for a given situation. Each of them may be better under different situations. Mean is the most widely used average followed by median. The median is better when the data set includes ▶outliers or is open ended. Mode is simple to locate and is preferred for finding the most popular item e.g. most popular drink or the most common size of shoes etc.

Cross References

▶Geometric Mean
▶Harmonic Mean
▶Mean Median and Mode
▶Skewness

References and Further Reading

Bluman AG (2007) Elementary statistics: a step by step approach, 6th edn. McGraw Hill, New York
Croucher JS (2002) Statistics: making business decisions. McGraw Hill/Irwin, New York
Mann PS (2006) Introductory statistics, 6th edn. Wiley, New York

Mean Residual Life

Jonathan C. Steele[1], Frank M. Guess[2], Timothy M. Young[2], David J. Edwards[3]
[1]Minitab, Inc., State College, PA, USA
[2]Professor
University of Tennessee, Knoxville, TN, USA
[3]Assistant Professor
Virginia Commonwealth University, Richmond, VA, USA

Theories and applications that use Mean Residual Life (MRL) extend across a myriad of helpful fields, while the methods differ considerably from one application to the next. Accelerated stress testing, fuzzy set engineering modeling, mixtures, insurance assessment of human life expectancy, maintenance and replacement of bridges, replacement of safety significant components in power plants, and evaluation of degradation signals in systems are just a few examples of applications of MRL function analysis. Note that MRL is also called "expected remaining life," plus other phrase variations. For a random lifetime X, the MRL is the conditional expectation $E(X - t|X > t)$, where $t \geq 0$. The MRL function can be simply represented with the reliability function $R(t) = P(X > t) = 1 - F(t)$ as:

$$e(t) = E(X - t|X > t) = \frac{\int_t^\infty R(x)dx}{R(t)}$$

where $R(t) > 0$ for $e(t)$ to be well defined. When $R(0) = 1$ and $t = 0$, the MRL equals the average lifetime. When $R(t) = 0$, then $e(t)$ is defined to be 0. The empirical MRL is calculated by substituting either the standard empirical estimate of $R(t)$ or, when censoring occurs, by substituting the Kaplan-Meier estimate of $R(t)$ (see ▶Kaplan-Meier Estimator). To use the Kaplan-Meier estimate when the final observation is censored requires a modification to define the empirical reliability function as eventually 0.

The reliability function can also be represented as a function of the MRL as:

$$R(t) = \left(\frac{e(0)}{e(t)}\right) \exp^{-\int_0^t \left[\frac{1}{e(x)}\right]dx}.$$

Note that the MRL function can exist, while the hazard rate function might not exist, or vice versa, the hazard rate function can exist while the MRL function might not. Compare Guess and Proschan (1988) plus Hall and Wellner (1981) for comments. When both functions exist, and the MRL function is differentiable, the hazard rate function is a function of the MRL:

$$h(t) = \frac{1 + e'(t)}{e(t)}$$

where $e'(t)$ is the first derivative of the MRL function.

The breadth of applications for the MRL function is astounding. As examples, Chiang (1968) and Deevey (1947) cite the use of the MRL for annuities via expected life tables (see ▶Life Table) in ancient Roman culture. Bhattacharjee (1982) suggests how to use the MRL to decide when to sell an item that has maintenance costs, which has copious natural applications, such as to real estate. Steele (2006) and Guess et al. (2006) illustrate a confidence interval for the range of values where one MRL function dominates

another and use it to reveal an opportunity to increase the profitability of a process that manufactures engineered medium density fiberboard. See also the insightful results on MRL functions of mixtures, ►order statistics, and coherent systems from Navarro and Hernandez (2008). Another topic of extensive research over the years is testing classes of MRL functions. For more on those tests, see references in Hollander and Proschan (1984), Hollander and Wolfe (1999) or Anis et al. (2004), for example. A brief list of other MRL papers, among many wide-ranging papers available, includes Peiravi and Dehqanmongabadi (2008), Zhao and Elsayed (2005), Bradley and Gupta (2003), Asadi and Ebrahimi (2000), Oakes and Dasu (1990), Berger et al. (1988), Guess and Park (1988), and Guess et al. (1986). We would recommend many other useful papers, but space severely limits our list.

While we do not give a complete inventory, note that R packages like *evd*, *ismev*, and *locfit* possess capabilities such as MRL plotting and/or computing the MRL for censored data; compare Shaffer et al. (2008). Another freeware, Dataplot, the software for the NIST website, does a MRL plot, but calls it a "conditional mean exceedance" plot, see Heckert and Filliben (2003). For-profit statistical software, such as JMP, MINITAB, PASW (formerly SPSS), SAS, etc., can be appropriately utilized for computing the MRL, using the basic formulas above (PASW and others use the phrase "life tables," which often contain a column for MRL). Pathak et al. (2009) illustrate the use of MATLAB for computing several different lifetime data functions including the MRL. Steele (2006) computes MRL via Maple.

Cross References

►Conditional Expectation and Probability
►Hazard Ratio Estimator
►Kaplan-Meier Estimator
►Life Expectancy
►Life Table

References and Further Reading

Anis MZ, Basu SK, Mitra M (2004) Change point detection in MRL function. Indian Soc Probab Stat 8:57–71

Asadi M, Ebrahimi N (2000) Residual entropy and its characterizations in terms of hazard function and mean residual life function. Stat Probab Lett 49(3):263–269

Berger RL, Boos DD, Guess FM (1988) Tests and confidence sets for comparing two mean residual life functions. Biometrics 44(1):103–115

Bhattacharjee MC (1982) The class of mean residual lives and some consequences. J Algebra Discr 3(1):56–65

Bradley DM, Gupta RC (2003) Limiting behaviour of the mean residual life. Ann I Stat Math 55(1):217–226

Chiang CL (1968) Introduction to stochastic processes in biostatistics. Wiley, New York

Deevey ES (1947) Life tables for natural populations of animals. Q Rev Biol 22:283–314

Guess FM, Hollander M, Proschan F (1986) Testing exponentiality versus a trend change in mean residual life. Ann Stat 14(4): 1388–1398

Guess FM, Park DH (1988) Modeling discrete bathtub and upside-down bathtub mean residual-life functions. IEEE T Reliab 37(5):545–549

Guess FM, Proschan F (1988) MRL: theory and applications. In: Krishnaiah PR, Rao CR (eds) Handbook of statistics 7: quality control and reliability. North Holland, Amsterdam, pp 215–224

Guess FM, Steele JC, Young TM, León RV (2006) Applying novel mean residual life confidence intervals. Int J Reliab Appl 7(2):177–186

Hall WJ, Wellner JA (1981) Mean residual life. In: Csörgö ZM et al (eds) Statistics and related topics. North Holland, Amsterdan, pp 169–184

Heckert NA, Filliben JJ (2003) CME plot. In: NIST handbook 148: DATAPLOT reference manual, volume I: commands, National Institute of Standards and Technology Handbook Series, pp 2-45–2-47. For more details see link: http://www.itl.nist.gov/div898/software/dataplot/document.htm

Hollander M, Proschan F (1984) Nonparametric concepts and methods in reliability. In: Krishnaiah PR, Sen PK (eds) Handbook of statistics 4: nonparametric methods. North Holland, Amsterdam, pp 613–655

Hollander M, Wolfe D (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York

Navarro J, Hernandez PJ (2008) Mean residual life functions of finite mixtures, order statistics and coherent systems. Metrika 67(3):277–298

Oakes D, Dasu T (1990) A note on residual life. Biometrika 77(2):409–410

Pathak R, Joshi S, Mishra DK (2009) Distributive computing for reliability analysis of MEMS devices using MATLAB. In: Proceedings of the international conference on advances in computing, communication and control (Mumbai, India, January 23–24, 2009). ACM, New York, pp 246–250

Peiravi A, Dehqanmongabadi N (2008) Accelerated life testing based on proportional mean residual life model for multiple failure modes. J Appl Sci 8(22):4166–4172

Shaffer LB, Young TM, Guess FM, Bensmail H, León RV (2008) Using R software for reliability data analysis. Int J Reliab Appl 9(1):53–70

Steele JC (2006) "Function domain sets" confidence intervals for the mean residual life function with applications in production of medium density fiberboard. Thesis at University of Tennessee, Knoxville, TN. Available at link: http://etd.utk.edu/2006/SteeleJonathanCody.pdf

Zhao WB, Elsayed EA (2005) Modelling accelerated life testing based on mean residual life. Int J Syst Sci 36(11): 689–696

Measure Theory in Probability

Milan Merkle
Professor, Faculty of Electrical Engineering
University of Belgrade, Belgrade, Serbia

Foundations of Probability: Fields and Sigma-Fields

Since Kolmogorov's axioms, Probability theory is a legitimate part of Mathematics, with foundations that belong to Measure theory. Although a traditional probabilist works solely with countably additive measures on sigma fields, the concepts of countable additivity and infinite models are by no means natural. As Kolmogorov [1956 p. 15] points out, "... in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes."

To build a probability model, we need first to have a non-empty set Ω which is interpreted as a set of all possible outcomes of a statistical experiment. Then we define which subsets of Ω will be assigned a probability. The family $\mathcal{F}$ of all such subsets has to satisfy

(1) $\Omega \in \mathcal{F}$,
(2) $B \in \mathcal{F} \implies B' \in \mathcal{F}$,
(3) $B_1, B_2 \in \mathcal{F} \implies B_1 \cup B_2 \in \mathcal{F}$,

and then we say that $\mathcal{F}$ is a field. If (3) is replaced by stronger requirement

(3') $B_1, B_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} B_i \in \mathcal{F}$

then we say that $\mathcal{F}$ is a sigma field.

The family $\mathcal{P}(\Omega)$ of all subsets of Ω is a field, and it is the largest field that can be made of subsets of Ω – it clearly contains all other possible fields. The smallest such field is $\mathcal{F}_0 = \{\emptyset, \Omega\}$; it is a subset of any other field.

The intersection of any family of fields is again a field. The union of a family of fields need not be a field. Both statements hold for sigma-fields, too.

Given a collection $\mathcal{A}$ of subsets of Ω, the intersection of all fields (sigma-fields) that contain $\mathcal{A}$ is called a field (sigma-field) *generated by* $\mathcal{A}$.

Having a non-empty set Ω and a field $\mathcal{F}$ of its subsets, a finitely additive probability measure is a function $P: \mathcal{F} \to \mathbb{R}_+$ such that

(a) $P(\Omega) = 1$.
(b) $P(A) \geq 0$ for every $A \in \mathcal{F}$.
(c) $P(A \cup B) = P(A) + P(B)$ whenever $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$ (*finite additivity*).

If (c) is replaced by the condition of *countable additivity*

(c') For any countable collection $A_1, A_2, \ldots$ of sets in $\mathcal{F}$, such that $A_i \cap A_j = \emptyset$ for any $A_i \neq A_j$ and such that $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$ (the latter condition is needless if $\mathcal{F}$ is a sigma-field):

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i)$$

then P is called (a countably additive) *probability measure*, or just *probability*. The triplet $(\Omega, \mathcal{F}, P)$ is called a *probability space*. By Carathéodory extension theorem, any countably additive probability measure P defined on a field $\mathcal{F}$ extends uniquely to a countably additive probability measure on the sigma field generated by $\mathcal{F}$; hence, if P is countably additive, we may always assume that $\mathcal{F}$ is a sigma-field.

A set $B \subset \Omega$ is called a *null set* if $B \subset A$ for some $A \in \mathcal{F}$ with $P(A) = 0$. Let $\mathcal{N}$ be a collection of all null sets in $(\Omega, \mathcal{F}, P)$. If $\mathcal{N} \subset \mathcal{F}$, the sigma-field $\mathcal{F}$ is called *complete*. For any sigma-field $\mathcal{F}$ there exists a complete sigma-field $\bar{\mathcal{F}}$, called a *completion* of $\mathcal{F}$, and defined as the sigma field generated by $\mathcal{F} \cup \mathcal{N}$.

A general positive measure μ is a set function defined on $(\Omega, \mathcal{F})$ with values in $\mathbb{R}_+ \cup \{+\infty\}$, which satisfies (b), (c) or (c'), and $\mu(\emptyset) = 0$. If $\mu(\Omega) < +\infty$, the measure is called *finite* and can be normalized to a probability measure by $P(A) = \mu(A)/\mu(\Omega)$ for all $A \in \mathcal{F}$. If Ω can be represented as a countable union of measurable sets of finite measure, then a measure is called *sigma-finite*. The most commonly used measure in Mathematics is the Lebesgue measure λ on $\mathbb{R}$, with the property that $\lambda([a, b]) = b - a$ for any $a < b$. This measure is not finite, as $\lambda(\mathbb{R}) = +\infty$, but it is sigma-finite.

If there exists a countable set $S \subset \Omega$ such that $\mu(S') = 0$, the measure μ is called *discrete*. Unless the measure is discrete, the sigma-field $\mathcal{F}$ is usually taken to be strictly smaller than $\mathcal{P}(\Omega)$, to ensure that it will be possible to assign some value of the measure to each set in $\mathcal{F}$. This is motivated by existence of non-measurable sets in $\mathbb{R}$ (sets that cannot be assigned any value of Lebesgue measure). Non-measurable sets cannot be effectively constructed and their existence is a consequence of Axiom of Choice [see Solovay (1970)]. The described construction of a probability space ensures that a probability can be assigned to all sets of interest.

The countable (vs. finite) additivity has a role to exclude from consideration measures that are too complicated, and also to enable applicability of fundamental theorems (for details on finitely additive measures see Yosida and Hewitt (1952)). Within axioms (a)-(b)-(c), the countable additivity is equivalent to *continuity of probability*, a property that can be described in two dual (equivalent) forms:

1. If $A_1 \subset A_2 \subset \cdots \subset \ldots$, then $P\left(\bigcup_{n=1}^{+\infty} A_n\right) = \lim_{n\to+\infty} P(A_n)$;
2. If $A_1 \supset A_2 \supset \cdots \supset \ldots$, then $P\left(\bigcap_{n=1}^{+\infty} A_n\right) = \lim_{n\to+\infty} P(A_n)$;

Random Variables and Their Distributions

Let $(\Omega, \mathcal{F}, P)$ be a probability space (usually called *abstract probability space*). Let X be a mapping from Ω to some other space S. A purpose of introducing such mappings can be twofold. First, in some simple models like tossing a coin, we prefer to have a numerical model that can also serve as a model for any experiment with two outcomes. Hence, instead of $\Omega = \{H, T\}$, we can think of $S = \{0, 1\}$ as a set of possible outcomes, which are in fact labels for any two outcomes in a real world experiment. Second, in large scale models, we think of Ω as being a set of possible states of a system, but to study the whole system can be too difficult task, so by mapping we wish to isolate one or several characteristics of Ω.

While Ω can be a set without any mathematical structure, S is usually a set of real numbers, a set in $\mathbb{R}^d$, or a set of functions. To be able to assign probabilities to events of the form $\{\omega \in \Omega \mid X(\omega) \in B\} = X^{-1}(B)$, we have to define a sigma-field $\mathcal{B}$ on S, that will accommodate all sets B of interest. If S is a topological space, usual choices are for $\mathcal{B}$ to be generated by open sets in S (Borel sigma-field), or to be generated by all sets of the form $f^{-1}(U)$, where $U \subset S$ is an open set and f is a continuous function $S \mapsto \mathbb{R}$ (Baire sigma-field). Since for any continuous f and open U, the set $f^{-1}(U)$ is open, the Baire field is a subset of corresponding Borel field. In metric spaces (and, in particular, in $\mathbb{R}^d$, $d \geq 1$) the two sigma fields coincide.

A mapping $X : \Omega \mapsto S$ is called $(\Omega, \mathcal{F}) - (S, \mathcal{B})$ – measurable if $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}$. The term *random variable* is reserved for such a mapping in the case when S is a subset of $\mathbb{R}$. Otherwise, X can have values in $\mathbb{R}^d$, when it is called a *random vector*, or in some functional space, when it is called a *random process*, where trajectories $X(\omega) = f(\omega, \cdot)$ depend on a numerical argument usually interpreted as time, or a *random field* if trajectories are functions of arguments that are not numbers. In general, X can be called *a random element*.

The central issue in a study of random elements is the probability measure $\mu = \mu_X$ induced by X on the space $(S, \mathcal{B})$ by $\mu_X(B) = P(X^{-1}(B))$, $B \in \mathcal{B}$, which is called the *probability distribution of* X. In fact, X is considered to be defined by its distribution; the mapping by itself is not of interest in Probability. In this way, each random element X is associated with two probability triplets: $(\Omega, \mathcal{F}, P)$ and $(S, \mathcal{B}, \mu)$. If a model considers only random variables that map Ω into S, then the first triplet can be discarded, or more formally, $(\Omega, \mathcal{F}, P)$ can be identified with $(S, \mathcal{B}, \mu)$.

The collection of sets $\{X^{-1}(B)\}_{B\in\mathcal{B}}$ is a sigma-field contained in $\mathcal{F}$, which is called a *sigma-field generated by* X, in notation $\sigma(X)$. It is considered in applications as a complete information about X, as it contains all relevant events in Ω from whose realizations we may deduce whether or not $X \in B$, for any $B \in \mathcal{B}$. In particular, if $\mathcal{B}$ contains all singletons $\{x\}$, then we know the value of X.

If there is another sigma-field $\mathcal{G}$ such that $\sigma(X) \subset \mathcal{G} \subset \mathcal{F}$, then we say that X is $\mathcal{G}$-measurable. In particular, if X is $\sigma(U)$-measurable, where U is another random element and if $\sigma(X)$ contains all sets of the form $X^{-1}(\{s\})$, $s \in S$, then X is a function of U.

The definition of a sigma-field does not provide any practical algorithm that can be used to decide whether or not a particular set belongs to a sigma field. For example, suppose that we have a Borel sigma-filed $\mathcal{B}$ on some topological space S, and we need to know whether or not $B \in \mathcal{B}$, for a given $B \subset S$. Then we need to either produce a formula that shows how to get B as a result of *countably many* unions, intersections and complements starting with open and closed sets, or to prove that such a formula does not exist. This is rarely obvious or straightforward, and sometimes it can require a considerable work. In cases when we want to show that a certain family of sets belongs to a given sigma-fields, the Dynkin's so-called "$\pi - \lambda$ theorem" is very useful. A collection $\mathcal{C}$ of subsets of a set S is called a π-system if $A \in \mathcal{C}, B \in \mathcal{C} \implies A \cap B \in \mathcal{C}$. It is called a λ-system if it has the following three properties: (1) $S \in \mathcal{C}$; (2) $A, B \in \mathcal{C}$ and $B \subset A \implies A \backslash B \in \mathcal{C}$; (3) For any sequence of sets $A_n \in \mathcal{C}$ with $A_n \subset A_{n+1}$ (increasing sets), it holds that $\sum_{i=1}^{+\infty} A_n \in \mathcal{C}$. Then we have the following.

Dynkin's $\pi - \lambda$ Theorem Let $\mathcal{A}$ be a π-system, $\mathcal{B}$ a λ-system and $\mathcal{A} \subset \mathcal{B}$. Then $\sigma(\mathcal{A}) \subset \mathcal{B}$.

Integration

Let X be a random variable that maps $(\Omega, \mathcal{F}, P)$ into $(\mathbb{R}, \mathcal{B}, \mu)$, where $\mathbb{R}$ is the set of reals, $\mathcal{B}$ is a Borel

sigma-algebra and μ is the distribution of X. The expectation of X is defined as

$$EX = \int_\Omega X(\omega)\mathrm{d}P(\omega) = \int_{\mathbb{R}} x\mathrm{d}\mu(x),$$

provided the integrals exist in the Lebesgue sense. By the construction of Lebesgue integral, E X exists if and only if E $|X|$ exists; in that case we say that X is integrable. To emphasize that the expectation is with respect to measure P, the notation $\mathrm{E}_P X$ can be used.

Let f be a measurable function $\mathbb{R} \to \mathbb{R}$ (in $\mathbb{R}$ we assume the Borel sigma-field if not specified otherwise). Then $f(X)$ is again a random variable, that is, the mapping $\omega \mapsto f(X(\omega))$ is $(\Omega, \mathcal{F}) - (\mathbb{R}, \mathcal{B})$ -measurable, and

$$Ef(X) = \int_\Omega f(X(\omega))\mathrm{d}P(\omega) = \int_{\mathbb{R}} f(x)\mathrm{d}\mu(x),$$

if the integral on the right hand side exists, and then we say that f is integrable. Expectations can be defined in the same way in more general spaces of values of f or X, for instance in $\mathbb{R}^d, d > 1$ or in any normed vector space.

Radon-Nikodym Theorem Suppose that P and Q are positive countably additive and sigma-finite measures (not necessarily probabilities) on the same space $(\Omega, \mathcal{F})$. We say that P is absolutely continuous with respect to Q (in notation $P \ll Q$) if $P(B) = 0$ for all $B \in \mathcal{F}$ with $Q(B) = 0$.

If $P \ll Q$, then there exists a non-negative measurable function f such that

$$P(A) = \int_\Omega I_A(\omega) f(\omega)\mathrm{d}Q(\omega), \quad \text{and}$$
$$\int_\Omega g(\omega)\mathrm{d}P(\omega) = \int_\Omega g(\omega) f(\omega)\mathrm{d}Q(\omega),$$

for any measurable g. The function f is called a *Radon-Nikodym derivative*, in notation $f = \frac{\mathrm{d}P}{\mathrm{d}Q}$, and it is Q-almost surely unique.

If Q is the Lebesgue measure and P a probability measure on $\mathbb{R}$, then the function f is called a *density* of P or of a corresponding random variable with the distribution P; distributions P on $\mathbb{R}$ that are absolutely continuous with respect to Lebesgue measure are called *continuous distributions*.

If both P and Q are probabilities and $P \ll Q$, then the ►Radon-Nikodym theorem yields that there exists a random variable $\Lambda \geq 0$ with $\mathrm{E}_Q \Lambda = 1$ such that

$$P(A) = \mathrm{E}_Q I_A \Lambda \qquad \text{and} \qquad \mathrm{E}_P X = \mathrm{E}_Q X\Lambda$$

for any random variable X.

Cross References

►Axioms of Probability
►Foundations of Probability
►Probability Theory: An Outline
►Radon–Nikodým Theorem
►Random Variable
►Stochastic Processes

References and Further Reading

Kolmogorov AN (1956) Foundations of the theory of probability, 2nd English edn. Chelsea, New York
Solovay RM (1970) A model of set-theory in which every set of reals is Lebesgue measurable. Ann Math Second Ser 92:1–56
Yosida K, Hewitt E (1952) Finitely additive measures. Trans Am Math Soc 72:46–66

Measurement Error Models

Alexander Kukush
Professor
National Taras Shevchenko University of Kyiv,
Kyiv, Ukraine

A (nonlinear) measurement error model (MEM) consists of three parts: (1) a *regression model* relating an observable regressor variable z and an unobservable regressor variable ξ (the variables are independent and generally vector valued) to a response variable y, which is considered here to be observable without measurement errors; (2) a *measurement model* relating the unobservable ξ to an observable surrogate variable x; and (3) a *distributional model* for ξ.

Parts of MEM

The *regression model* can be described by a conditional distribution of y given (z, ξ) and given an unknown parameter vector θ. As usual this distribution is represented by a probability density function $f(y|z, \xi; \theta)$ with respect to some underlying measure on the Borel σ-field of $\boldsymbol{R}$. We restrict our attention to distributions that belong to the exponential family, i.e., we assume f to be of the form

$$f(y|z, \xi; \beta, \varphi) = \exp\left(\frac{y\eta - c(\eta)}{\varphi} + a(y, \varphi)\right) \qquad (1)$$

with

$$\eta = \eta(z, \xi; \beta). \qquad (2)$$

Here β is the regression parameter vector, φ a scalar dispersion parameter such that $\theta = (\beta^T, \varphi)^T$, and a, c, and η are known functions. This class comprises the class of generalized linear models, where $\eta = \eta(\beta_0 + z^T\beta_z + \xi^T\beta_\xi)$, $\beta = (\beta_0, \beta_x^T, \beta_\xi^T)^T$.

The *classical measurement model* assumes that the observed variable x differs from the latent ξ by a measurement error variable δ that is independent of z, ξ, and y:

$$x = \xi + \delta \tag{3}$$

with $\mathbf{E}\delta = 0$. Here we assume that $\delta \sim N(0, \Sigma_\delta)$ with Σ_δ known. The observable data are independent realizations of the model (x_i, y_i), $i = 1, \ldots, n$.

Under the *Berkson measurement model*, the latent variable ξ differs from the observed x by a centered measurement error δ that is independent of z, x, and y:

$$\xi = x + \delta. \tag{4}$$

Thus, the values of x are fixed in advance, whereas the unknown true values, ξ, are fluctuating.

The *distributional model* for ξ either states that the ξ are unknown constants (*functional case*) or that ξ is a random variable (*structural case*) with a distribution given by a density $h(\xi; \gamma)$, where γ is a vector of nuisance parameters describing the distribution of ξ. In the structural case, we typically assume that

$$\xi \sim N(\mu_\xi, \Sigma_\xi), \tag{5}$$

although sometimes it is assumed that ξ follows a mixture of normal distributions. In the sequel, for the structural case we assume γ to be known. If not, it can often be estimated in advance (i.e., pre-estimated) without considering the regression model and the data y_i. For example, if ξ is normal, then μ_ξ and Σ_ξ can be estimated by $\overline{x}$ and $S_x - \Sigma_\delta$, respectively, where $\overline{x}$ and S_x are the empirical mean vector and the empirical covariance matrix of the data x_i, respectively.

The goal of measurement error modeling is to obtain nearly unbiased estimates of the regression parameter β by fitting a model for y in terms of (z, x). Attainment of this goal requires careful analysis. Substituting x for ξ in the model (1) – (2), but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously.

In the structural case, the *regression calibration* (RC) estimator can be constructed by substituting $\mathbf{E}(\xi|x)$ for unobservable ξ. In both functional and structural cases, another, the simulation-extrapolation (*SIMEX*) estimator, becomes very popular. These estimators are not consistent in general, although they often reduce the bias significantly; see Carroll et al. (2006).

Polynomial and Poisson Model

We mention two important examples of the classical MEM (1) – (3) where for simplicity the latent variable is scalar and the observable regressor z is absent. The *polynomial model* is given by

$$y = \beta_0 + \beta_1 \xi + \ldots + \beta_k \xi^k + \varepsilon,$$

where $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$ and ε is independent of ξ. Here

$$\eta = \sum_{r=0}^{k} \beta_r \xi^r, \; c(\eta) = \frac{1}{2}\eta^2,$$

and $\varphi = \sigma_\varepsilon^2$. Both cases are possible: (a) the measurement error variance σ_δ^2 is known and (b) the ratio $\sigma_\varepsilon^2/\sigma_\delta^2$ is known; for the latter case see Shklyar (2008). In the particular case of $k = 1$, we obtain the *linear model*; an overview of methods in this MEM is given in Cheng and Van Ness (1999).

In the *loglinear Poisson model* we have $y \sim Po(\lambda)$ with $\lambda = \exp(\beta_0 + \beta_1 \xi)$; then $\eta = \log \lambda$, $c(\eta) = e^\eta$, and $\varphi = 1$.

Methods of Consistent Estimation in Classical MEM

Now, we deal with the general model (1) – (3). We distinguish between two types of estimators, functional and structural. The latter makes use the distribution of ξ, which therefore must be given, at least up to the unknown parameter, vector γ. The former does not need the distribution of ξ and works even when ξ is not random (functional case).

Functional Method: Corrected Score

If the variable ξ were observable, one could estimate β (and also φ) by the method of maximum likelihood (ML). The corresponding likelihood score function for β is given by

$$\psi(y, z, \xi; \beta, \varphi) = \frac{\partial \log f(y|z, \xi; \beta, \varphi)}{\partial \beta} = \frac{y - c'(\eta)}{\varphi} \frac{\partial \eta}{\partial \beta}.$$

We want to construct an unbiased estimating function for β in the observed variables. For this purpose, we need to find functions g_1 and g_2 of z, x, and β such that

$$\mathbf{E}[g_1(z, x; \beta)|z, \xi] = \frac{\partial \eta}{\partial \beta}, \quad \mathbf{E}[g_2(z, x; \beta)|z, \xi] = c'(\eta) \frac{\partial \eta}{\partial \beta}.$$

Then

$$\psi_C(y, z, x; \beta) = y g_1(z, x; \beta) - g_2(z, x; \beta)$$

is termed the corrected score function. The *Corrected Score* (CS) estimator $\hat{\beta}_C$ of β is the solution to

$$\sum_{i=1}^{n} \psi_C(y_i, z_i, x_i; \hat{\beta}_C) = 0.$$

The functions g_1 and g_2 do not always exist. Stefanski (1989) gives the conditions for their existence and shows how to find them if they exist. The CS estimator is consistent in

both functional and structural cases. It was first proposed by Stefanski (1989) and Nakamura (1990).

An alternative functional method, particularly adapted to ▶generalized linear models, is the conditional score method; see Stefanski and Carroll (1987).

Structural Methods: Quasi-Likelihood and Maximum Likelihood

The conditional mean and conditional variance of y given (z, ξ) are, respectively,

$$\mathbf{E}(y|z,\xi) = m^*(z,\xi;\beta) = c'(\eta),\ \mathbf{V}(y|z,\xi) = v^*(z,\xi;\beta) = \varphi c''(\eta).$$

Then the conditional mean and conditional variance of y given the observable variables are

$$m(z,x;\beta) = \mathbf{E}(y|z,x) = E[m^*(z,\xi;\beta)|x],$$
$$v(z,x;\beta) = \mathbf{V}(y|z,x) = \mathbf{V}[m^*(z,\xi;\beta)|x] + \mathbf{E}[v^*(z,\xi;\beta)|x].$$

For the quasi-likelihood (QL) estimator, we construct the quasi-score function

$$\psi_Q(y,z,x;\beta) = [y - m(z,x;\beta)]v(z,x;\beta)^{-1}\frac{\partial m(z,x;\beta)}{\partial \beta}.$$

Here we drop the parameter φ considering it to be known. We also suppress the nuisance parameter γ in the argument of the functions m and v, although m and v depend on γ. Indeed, in order to compute m and v, we need the conditional distribution of ξ given x, which depends on the distribution of ξ with its parameter γ. For instance, assume (5) where the elements of μ_ξ and Σ_ξ make up the components of the parameter vector γ. Then $\xi|x \sim N(\mu(x), T)$ with

$$\mu(x) = \mu_\xi + \Sigma_\xi(\Sigma_\xi + \Sigma_\delta)^{-1}(x - \mu_\xi),$$
$$T = \Sigma_\delta - \Sigma_\delta(\Sigma_\xi + \Sigma_\delta)^{-1}\Sigma_\delta.$$

The QL estimator $\hat{\beta}_Q$ of β is the solution to

$$\sum_{i=1}^{n} \psi_Q(y_i, z_i, x_i; \hat{\beta}_C) = 0.$$

The equation has a unique solution for large n, but it may have multiple roots if n is not large. Heyde and Morton (1998) develop methods to deal with this case.

Maximum likelihood is based on the conditional joint density of x, y given z. Thus, while QL relies only on the error-free mean and variance functions, ML relies on the whole error-free model distribution. Therefore, ML is more sensitive than QL with respect to a potential model misspecification because QL is always consistent as long as at least the mean function (along with the density of ξ) has been correctly specified. In addition, the likelihood function is generally much more difficult to compute than the quasi-score function. This often justifies the use of the relatively less efficient QL instead of the more efficient ML method.

Efficiency Comparison

For CS and QL, $\hat{\beta}$ is asymptotically normal with asymptotic covariance matrix (ACM) Σ_C and Σ_Q, respectively. In the structural model, it is natural to compare the relative efficiencies of $\hat{\beta}_C$ and $\hat{\beta}_Q$ by comparing their ACMs. In case there are no nuisance parameters, it turns out that

$$\Sigma_C \geq \Sigma_Q \tag{6}$$

in the sense of the Loewner order for symmetric matrices. Moreover, under mild conditions the strict inequality holds.

These results hold true if the nuisance parameters γ are known. If, however, they have to be estimated in advance, (6) need not be true anymore. For the Poisson and polynomial structural models, Kukush et al. (2007) prove that (6) still holds even if the nuisance parameters are preestimated. Recently Kukush et al. (2009) have shown that QL can be modified so that, in general, $\Sigma_C \geq \Sigma_Q$; for this purpose the γ must be estimated together with β and not in advance.

Estimation in Berkson Model

Now, we deal with the model (1), (2), and (4). Substituting x for ξ in the regression model (1) – (2) is equivalent to RC. Therefore, it leads to estimates with a typically small bias.

A more precise method is ML. The conditional joint density of x and y given z has a simpler form compared with the classical MEM. That is why ML is more reliable in the Berkson model.

Nonparametric Estimation

We mention two nonparametric problems overviewed in Carroll et al. (2006), Ch. 12: the estimation of the density ρ of a random variable ξ, and the nonparametric estimation of a regression function f, both when ξ is measured with error. In these problems under normally distributed measurement error, the best mean squared error of an estimator of $\rho(x_0)$ or $f(x_0)$ converges to 0 at a rate no faster than the exceedingly slow rate of logarithmic order. However, under a more heavy-tailed measurement error, estimators can perform well for a reasonable sample size.

About the Author

Dr. Alexander Kukush is a Professor, Department of Mechanics and Mathematics, National Taras Shevchenko University of Kyiv, Ukraine. He is an Elected member of the International Statistical Institute (2004). He has authored and coauthored more than 100 papers on statistics and a book: *Theory of Stochastic Processes With Applications to Financial Mathematics and Risk Theory* (with D. Gusak, A. Kulik, Yu. Mishura, and A. Pilipenko, Problem Books in Mathematics, Springer, 2009). Professor Kukush has received the Taras Shevchenko award for a cycle of papers on regression (National Taras Shevchenko University of Kyiv, 2006).

Cross References

►Astrostatistics
►Bias Analysis
►Calibration
►Estimation
►Likelihood
►Linear Regression Models
►Nonparametric Estimation
►Normal Distribution, Univariate
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Probability Theory: An Outline

References and Further Reading

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall, London

Cheng CL, Van Ness JW (1999) Statistical regression with measurement error. Arnold, London

Heyde CC, Morton R (1998) Multiple roots in general estimating equations. Biometrika 85:967–972

Kukush A, Malenko A, Schneeweiss H (2007) Comparing the efficiency of estimates in concrete errors-in-variables models under unknown nuisance parameters. Theor Stoch Proc 13(29):4, 69–81

Kukush A, Malenko A, Schneeweiss H (2009) Optimality of the quasi score estimator in a mean-variance model with applications to measurement error models. J Stat Plann Infer 139:3461–3472

Nakamura T (1990) Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. Biometrika 77:127–137

Shklyar SV (2008) Consistency of an estimator of the parameters of a polynomial regression with a known variance relation for errors in the measurement of the regressor and the echo. Theor Probab Math Stat 76:181–197

Stefanski LA (1989) Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Commun Stat A – Theor 18:4335–4358

Stefanski LA, Carroll RJ (1987) Conditional scores and optimal scores in generalized linear measurement error models. Biometrika 74:703–716

Measurement of Economic Progress

Marat Ibragimov[1], Rustam Ibragimov[2]
[1]Associate Professor
Tashkent State University of Economics, Tashkent, Uzbekistan
[2]Associate Professor
Harvard University, Cambridge, MA, USA

Broadly defined, measurement of economic progress focuses on quantitative analysis of the standard of living or quality of life and their determinants. The analysis concerns many elements of the standard living such as its material components, human capital, including education and health, inequality and other factors [see, among others, Barro and Sala-i Martin (2004), Howitt and Weil (2008), Steckel (2008), and references therein].

Theoretical foundation for empirical analysis of determinants of economic growth is provided by the Solow growth model. The human capital-augmented version of the model with the Cobb-Douglas production function [see Mankiw et al. (1992)] assumes that, for country i at time t, the aggregate output $Y_i(t)$ satisfies $Y_i(t) = K_i(t)^{\alpha} H_i(t)^{\beta} (A_i(t)L_i(t))^{1-\alpha-\beta}$, where $K_i(t)$ is physical capital, $H_i(t)$ is human capital, $L_i(t)$ is labor supply and $A_i(t)$ is a productivity parameter (the efficiency level of each worker or the level of technology). The variables L and A are assumed to obey $L_i(t) = L_i(0)e^{n_i t}$ and $A(t) = A(0)e^{gt}$, where n_i and g are, respectively, the population growth rate and the rate of technological progress. Physical and human capital are assumed to follow continuous-time accumulation equations $dK_i(t)/dt = s_{K,i}Y_i(t) - \delta K_i(t)$ and $dH_i(t)/dt = s_{H,i}Y_i(t) - \delta H(t)$ with the depreciation rate δ and the savings rates $s_{K,i}$ and $s_{H,i}$. Under the above assumptions, the growth model leads to the regressions $\gamma_i = a_0 + a_1 \log y_i(0) + a_2 \log(n_i + g + \delta) + a_3 \log s_{K,i} + a_4 \log s_{H,i} + \epsilon_i$, where $\gamma_i = (\log y_i(t) - \log y_i(0))/t$ is the growth rate of output per worker $y_i(t) = Y_i(t)/L_i(t)$ between time 0 and t [see, among others, Barro and Sala-i Martin (2004), Durlauf et al. (2005)]. Cross-country growth regressions typically include additional regressors Z_i and focus on estimating models in the form $\gamma_i = \mathbf{a}\mathbf{X}_i + \mathbf{b}\mathbf{Z}_i + \epsilon_i$, where $\mathbf{a} = (a_0, a_1, ..., a_4) \in \mathbf{R}^5$, $\mathbf{b} = (b_1, b_2, ..., b_m) \in \mathbf{R}^m$, the components of $\mathbf{X}_i = (1, \log y_i(0), \log(n_i + g + \delta), \log s_{K,i}, \log s_{H,i})'$ are the growth determinants in the Solow model and $\mathbf{Z}_i \in \mathbf{R}^m$ is the vector of growth determinants outside the Solow growth theory.

The statistical analysis of economic progress and its determinants presents a number of challenges due to

the necessity of using proxy measures and corresponding weights for different components of the standard of living and factors affecting it. The material standard of living is typically measured as per capita Gross Domestic Product (GDP) adjusted for changes in price levels. Proxies for education and human capital used in growth economics include school-enrollment rates at the secondary and primary levels, literacy rates, average years of secondary and higher schooling and outcomes on internationally comparable examinations. Many works in the literature have also used student-teacher ratios as a measure of quality of education. The two most widely used measures of health are life expectancy at birth or age 1 and average height used as a proxy for nutritional conditions during the growing years.

Barro (1991) and Barro and Sala-i Martin (2004) find that the growth rate of real per capita GDP is positively related to initial human capital, including education and health, proxied by school-enrollment rates, upper-level schooling and life expectancy and negatively related to the initial level of real per capita GDP. The results in Barro (1991) also indicate statistically significant negative effects of political instability (measured using the number of revolutions and coups per year and the number of political assassinations per million population per year) on growth. Other factors used in the analysis in Barro (1991) and Barro and Sala-i Martin (2004) include fertility and the ratio of real government consumption to real GDP (with statistically significant negative effects on growth), investment ratio, inflation rate as well as proxies for market distortions, maintenance of the rule of law, measures for democracy, international openness, the terms of trade, indicators for economic systems and countries in sub-Saharian Africa and Latin America and other variables.

A number of works in theoretical and empirical growth economics have focused on the development and analysis of performance of models with endogenous technological progress. Many recent studies have also studied the factors that lead to the observed differences in the determinants of economic growth in different countries, including capital components, technology and efficiency. In particular, several works have emphasized the role of geographical differences, cultural factors, economic policies and institutions as fundamental causes of the differences in growth determinants (Howitt and Weil 2008).

Statistical study of economic growth determinants is complicated by relatively small samples of available observations, measurement errors in key variables, such as GDP, heterogeneity in observations and estimated parameters, dependence in data and large number of potential growth regressors under analysis. Related issues in the analysis of economic growth concern difficulty of causal interpretation of estimation results, robustness of the conclusions to alternative measures of variables in the analysis, and open-endedness of growth theories that imply that several key factors matter for growth at the same time. Levine and Renelt (1992) focus on the analysis of robustness of conclusions obtained using cross-country growth regressions. They propose assessing the robustness of the variable Z of interest using the variation of the coefficient b in cross-country regressions $\gamma_i = \mathbf{a}\mathbf{X}_i + bZ_i + \mathbf{c}\mathbf{V}_i + \epsilon_i$, where $\mathbf{X}_i$ is the vector of variables that always appear in the regressions (e.g., the investment share of GDP, initial level of income, a proxy for the initial level of human capital such as the school enrollment rate, and the rate of population growth in country i), and $\mathbf{V}_i$ is a vector of additional control variables taken from the pool of variables available. Departing from the extreme bounds approach in Levine and Renelt (1992) that requires the estimate of the coefficient of interest b to be statistically significant for any choice of control variables $\mathbf{V}$, several recent works [see Sala-i Martin et al. (2004), Ch. 12 in Barro and Sala-i Martin (2004), and references therein] propose alternative less stringent procedures to robustness analysis. Several recent works on the analysis of economic growth and related areas emphasize importance of models incorporating disasters and crises and probability distributions generating ►outliers and extreme observations, such as those with heavy-tailed and power-law densities [see Barro (1991), Gabaix (2009) and Ibragimov (2009)].

Acknowledgments

Marat Ibragimov gratefully acknowledges support by a grant R08-1123 from the Economics Education and Research Consortium (EERC), with funds provided by the Global Development Network and the Government of Sweden. Rustam Ibragimov gratefully acknowledges partial support by the National Science Foundation grant SES-0820124.

Cross References

►Composite Indicators
►Econometrics
►Economic Growth and Well-Being: Statistical Perspective
►Economic Statistics

References and Further Reading

Barro RJ (1991) Economic growth in a cross section of coutries. Q J Econ 106:407–443
Barro RJ, Sala-i Martin X (2004) Economic growth. MIT, Cambridge, MA

Durlauf S, Johnson P, Temple J (2005) Growth econometrics. In: Aghion P, Durlauf S (eds) Handbook of economic growth. North-Holland, Amsterdam

Gabaix X (2009) Power laws in economics and finance. Annu Rev Econ 1:255–293

Howitt P, Weil DN (2008) Economic growth. In: Durlauf SN, Blume LE (eds) New palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Washington, DC

Ibragimov, R (2009) Heavy tailed densities, In: *The New Palgrave Dictionary of Economics Online*, (Eds. S. N. Durlauf and L. E. Blume), Palgrave Macmillan. http://www.dictionaryofeconomics.com/article?id=pde2008_H000191

Levine R, Renelt D (1992) A sensitivity analysis of cross-country growth regressions. Am Econ Rev 82:942–963

Mankiw NG, Romer D, Wiel DN (1992) A contribution to the empirics of economic growth. Q J Econ 42:407–437

Sala-i Martin X, Doppelhofer G, Miller RI (2004) Determinants of long-term growth: A Bayesian averaging of classical estimates (bace) approach. Am Econ Rev 94:813–835

Steckel RH (2008) Standards of living (historical trends). In: Durlauf SN, Blume LE (eds) New palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Washington, DC

Measurement of Uncertainty

K. R. Muraleedharan Nair
Professor
Cochin University of Science and Technology, Cochin, India

The measurement and comparison of uncertainty associated with a random phenomenon have been a problem attracting a lot of researchers in Science and Engineering over the last few decades. Given a system whose exact description is unknown its ►entropy is the amount of information needed to exactly specify the state of the system. The Shannon's entropy, introduced by Shannon (1948), has been extensively used in literature as a quantitative measure of uncertainty. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, with respective probabilities $p_1, p_2, \ldots, p_n$, the Shannon's entropy is defined as

$$H_n(P) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

Earlier development in this area was centered on characterizing the Shannon's entropy using different sets of postulates. The classic monographs by Ash (1965), Aczel and Daroczy (1975) and Behra (1990) review most of the works on this aspect. Another important aspect of interest is that of identifying distributions for which the Shannon's entropy is maximum subject to certain restrictions on the underlying random variable. Depending on the conditions imposed, several maximum entropy distributions have been derived. For instance, if X is a random variable in the support of the set of non-negative real numbers, the maximum entropy distribution under the condition that the arithmetic men is fixed is the exponential distribution. The book by Kapur (1989) covers most of the results in this area.

For a continuous non-negative random variable X with probability density function $f(x)$ the continuous analogue of (1) takes the form

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{2}$$

Several modifications of the Shannon's entropy has been proposed and extensively studied. Renyi (1961) define the entropy of order α as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \frac{\sum_{i=1}^{n} p_i^\alpha}{\sum_{i=1}^{n} p_i}, \ \alpha \neq 1, \alpha > 0 \tag{3}$$

where $P = (P_{1,\ldots\ldots} P_n)$ is such that $p_i \geq 0$, *and* $\sum_{i-1}^{n} p_i = 1$.

As $\alpha \to 1$, (3) reduces to (1). Khinchin (1957) generalized the Shannon's entropy by choosing a convex function $\varphi(.)$, with $\varphi(1) = 0$ and defined the measure

$$H_\varphi(f) = -\int_{-\infty}^{\infty} f(x) \varphi[f(x)] dx. \tag{4}$$

Nanda and Paul (2006) studied (4) for two particular choices of φ in the form

$$H_1^\beta(f) = \frac{1}{\beta - 1}\left[1 - \int_0^\alpha f^\beta(x) dx\right] \tag{5}$$

and

$$H_2^\beta(f) = \frac{1}{1-\beta}\left[\log \int_0^\infty f^\beta(x) dx\right] \tag{6}$$

where the support of f is the set of non-negative reals and $\beta > 0$ with $\beta \neq 1$. As $\beta \to 1$, (5) and (6) reduces to the Shannon's entropy given in (2).

Recently Rao et al. (2004) introduced cumulative residual entropy defined by

$$E(X) = -\int_0^\infty \overline{F}(x) \log \overline{F}(x) dx$$

which is proposed as an alternative measure of uncertainty based on the cumulative survival function $\overline{F}(x) = P(X > x)$. For various properties and applications of this measure we refer to Rao (2005) and Asadi and Zohrevand (2007).

There are several other concepts closely related to the Shannon's entropy. Kullback and Leibler (1951) defines the directed divergence (also known as relative entropy or cross entropy) between two distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ with

$$p_i, q_i \geq 0 \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$$

as

$$D_n(P,Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. \tag{7}$$

Kannappan and Rathie (1973) and Mathai and Rathie (1975) have obtained characterization results based on certain postulates which naturally leads to (7). The continuous analogue of (7) turns out to be

$$D(f,g) = \int_{-\infty}^{\alpha} f(x) \log \frac{f(x)}{g(x)} dx \tag{8}$$

where $f(x)$ and $g(x)$ are probability density functions corresponding to two probability measures P and Q.

The concept of affinity between two distributions was introduced and studied in a series of works by Matusita [see Matusita (1961)]. This measure has been widely used as a useful tool for discrimination among distributions. Affinity is symmetric in distributions and has direct relationship with error probability when classification or discrimination is concerned. For two discrete distributions P and Q considered above the Matusita's affinity (Mathai and Rathie 1975) between P and Q is defined as

$$\delta(P,Q) = \sum_{i=1}^{n} (p_i q_i)^{1/2}. \tag{9}$$

If X and Y are non-negative random variables and if $f(x)$ and $g(x)$ are the corresponding probability density functions, the affinity between f and g takes the form

$$\delta(f,g) = \int_0^{\infty} \sqrt{f(x)g(x)} dx \tag{10}$$

$\delta(f,g)$ lies between 0 and 1.

Majernik (2004) has shown that

$$H(f,g) = 2[1 - \delta(f,g)]$$

where $H(f,g)$ is the Hellinger's distance defined by

$$H(f,g) = \int_0^{\infty} \left[\sqrt{f(x)} - \sqrt{g(x)}\right]^2 dx. \tag{11}$$

Affinity is a special case of the Chernoff distance considered in Akahira (1996) defined by

$$C(F,G) = -\log\left[\int f^{\alpha}(x) g^{1-\alpha} dx\right], 0 < \alpha < 1. \tag{12}$$

It may be noticed that when $\alpha = \frac{1}{2}$ (12) reduces to $-\log \delta(f,g)$, where $\delta(f,g)$ is the affinity defined in (10).

The concept of inaccuracy was introduced by Kerridge (1961). Suppose that an experimenter asserts that the probability for the i^{th} eventuality is q_i whereas the true probability is p_i, then the inaccuracy of the observer, as proposed by Kerridge, can be measured by

$$1(P,Q) = -\sum_{i=1}^{n} p_i \log q_i \tag{13}$$

where P and Q are two discrete probability distributions, considered earlier.

Nath (1968) extended the Kerridge's concept to the continuous situation. If $F(x)$ is the actual distribution function corresponding to the observations and $G(x)$ is the distribution assigned by the experimenter and $f(x)$ and $g(x)$ are the corresponding density functions the inaccuracy measure is defined as

$$1(F,G) = -\int_0^{\alpha} f(x) \log g(x) dx. \tag{14}$$

This measure has extensively been used as a useful tool for measurement of error in experimental results. In expressing statements about probabilities of various events in an experiment, two kinds of errors are possible: one resulting from the lack of enough information or vagueness in experimental results and the other from incorrect information. In fact, (14) can be written as

$$1(F,G) = -\int_0^{\infty} f(x) \log f(x) dx + \int_0^{\infty} f(x) \log \frac{f(x)}{f(x)} dx. \tag{15}$$

The first term on the right side of (15) represents the error due to uncertainty which is the Shannon's entropy while the second term is the Kullback–Leibler measure, defined in (8) representing the error due to wrongly specifying the distribution as $G(x)$. In this sense the measure of inaccuracy can accommodate the error due to lack of information as well as that due to incorrect information.

In many practical situations, complete data may not be observable due to various reasons. For instance, in lifetime studies the interest may be on the life time of a unit after a specified time, say t. If X is the random variable representing the life time of a component the random variable of interest is $X - t | X > t$. Ebrahimi (1996) defines the residual entropy function as the Shannon's entropy associated with the residual life distribution, namely

$$H(f,t) = -\int_t^{\infty} \frac{f(x)}{\overline{F}(t)} \log \frac{f(x)}{\overline{F}(x)}, \overline{F}(t) > 0. \tag{16}$$

In terms of the hazard rate $h(x) = \frac{f(x)}{F(x)}$, (16) can also be written as

$$H(f,t) = 1 - \frac{1}{\bar{F}(t)} \int_t^\infty f(x) \log h(x) dx. \qquad (17)$$

Ebrahimi points out that (16) can be used as a potential measure of stability of components in the reliability context. The problem of ordering life time distributions using this concept has been addressed in Ebrahimi and Kirmani (1996). Belzunce et al. (2004) has shown that the residual entropy function determines the distributions uniquely if $H(f,t)$ is increasing in t. Characterization of probability distributions using the functional form of the residual entropy function have been the theme addressed in Nair and Rajesh (1998), Sankaran and Gupta (1999), Asadi and Ebrahimi (2000) and Abraham and Sankaran (2005).

Recently Nanda and Paul (2006) has extended the definition of the Renyi entropy defined by (5) and (6) to the truncated situation. It is established that under certain conditions the Renyi's residual entropy function determines the distribution uniquely. They have also looked into the problem of characterization of probability distributions using the same.

Ebrahimi and Kirmani (1996) has modified the definition of the Kullback–Leibler measure to the truncated situation to accommodate the current age of a system. Recently Smitha et al. (2008) have extended the definition of affinity to the truncated situation and has obtained characterization results for probability distributions under the assumption of proportional hazard model. Nair and Gupta (2007) extended the definition of the measure of inaccuracy to the truncated situation and has characterized the generalized Pareto distributions using the functional form of the inaccuracy measure.

About the Author

Dr. K.R. Muraleedharan Nair is a senior Professor in the Department of Statistics of the Cochin University of Science and Technology, India. He had been teaching Statistics at the post graduate level for the past 39 years. He has served the University as the Head of the Department (2004–2007) and as the Controller of examinations (2000–2003). He is currently the Vice President of the Indian Society for Probability and Statistics, besides being reviewer for certain reputed journals. He has published 28 papers in international journals besides several conference papers. He is a member of the Board of Studies as well as Faculty of Science in some of the Indian Universities.

Cross References

►Diversity
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Kullback-Leibler Divergence
►Maximum Entropy Method for Estimation of Missing Data
►Probability Theory: An Outline
►Role of Statistics
►Statistical View of Information Theory

References and Further Reading

Abraham B, Sankaran PG (2005) Renyi's entropy for residual lifetime distributions, Stat Papers 46:17–30
Aczel J, Daroczy Z (1975) On measures of information and their characterization, Academic, New York. Ann Inst Stat Math 48:349–364
Akahira M (1996) Loss of information of a statistic for a family of non-regular distributions. Ann Inst Stat Math 48:349–364
Asadi M, Ebrahimi N (2000) Residual entropy and its characterizations in terms of hazard function and mean residual life function. Stat and Prob Letters 49:263–269
Asadi M, Zohrevand Y (2007) On the dynamic cumulative residual entropy. J Stat Plann Infer 137:1931–1941
Ash RB (1965) Information theory. Wiley, New York
Behra M (1990) Additive and non-additive measures of entropy. Wiley Eastern, New York
Belzunce F, Navarro J, Ruiz JM, del Aguila Y (2004) Some results on residual entropy function. Metrica 59:147–161
Ebrahimi N (1996) How to measure uncertainty in the residual life time distribution. Sankhya A 58:48–56
Ebrahimi N, Kirmani SUNA (1996) Some results on ordering survival function through uncertainty. Stat Prob Lett 29:167–176
Kannappan PI, Rathie PN (1973) On characterization of directed divergence. Inform Control 22:163–171
Kapur JN (1989) Maximum entropy models in science and engineering. Wiley Eastern, New Delhi
Kerridge DF (1961) Inaccuracy and inference. J R Stat Soc Series B, 23:184–194
Khinchin AJ (1957) Mathematical foundation of information theory. Dover, New York
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Majernik K (2004) A dissimilarity measure for an arbitrary number of probability distributions. Int J Gen Sys 33(6):673–678
Mathai AM, Rathie PN (1975) Basic concepts in information theory and statisticsaxiomatic foundations and applications. Wiley, New York
Matusita K (1961) Interval estimation based on the notion of affinity. Bull Int Stat Inst 38(4):241–244
Nanda AK, Paul P (2006) Some results on generalized residual entropy. Inform Sci 176:27–47
Nair KRM, Rajesh G (1998) Characterization of probability distribution using the residual entropy function. J Ind Stat Assoc 36:157–166
Nair NU, Gupta RP (2007) Characterization of proportional hazard models by properties of information measures. Int J Stat 6(Special Issue):223–231
Nath P (1968) Inaccuracy and coding theory. Metrika 13:123–135
Rajesh G, Nair KRM (1998) Residual entropy function in discrete time. Far East J Theor Stat 2(1):1–10

Rao M, Chen Y, Vemuri BC, Wang F (2004) Cumulative residual entropy: a new measure of information. IEE Trans Inform Theor 50(6):1220–1228
Rao M (2005) More on a concept of entropy and information. J Theor Probab 18:967–981
Renyi A (1961) On measures of entropy and information, Proceedings of Fourth Berkley Symposium on Mathematics, Statistics and Probability, 1960, University of California Press, vol 1, pp 547–561
Sankaran PG, Gupta RP (1999) Characterization of life distributions using measure of uncertainty. Cal Stat Assoc Bull 49:154–166
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 279–423:623–656
Smitha S, Nair KRM, Sankaran PG (2008) On measures of affinity for truncated distribution. Cal Stat Assoc Bull 59:151–162

Measures of Agreement

Elisabeth Svensson
Örebro University, Örebro, Sweden

Agreement in repeated assessments is a fundamental requirement for quality of data from assessments on ►rating scales. Scale assessments produce ordinal data, the ordered categories representing only a rank order of the intensity of a particular variable and not a numerical value in a mathematical sense, even when the assessments are numerically labeled.

The main quality concepts of scale assessments are *reliability* and *validity*. *Reliability* refers to the extent to which repeated measurements of the same object yield the same result, which means agreement. In *intra-rater reliability* studies the agreement in test-retest assessments is evaluated. *Inter-rater reliability* refers to the level of agreement between two raters judging the same object.

The *percentage agreement* (*PA*) in assessments is the basic agreement measure and is also called *overall agreement* or *raw agreement*. When $PA < 100\%$ the reasons for disagreement can be evaluated by a statistical approach by Svensson that takes account of the rank-invariant properties of ordinal data. The approach makes it possible to identify and measure systematic disagreement, when present, separately from disagreement caused by individual variability in assessments. Different frequency distributions of the two sets of ordinal assessments indicate that the two assessments disagree systematically regarding the use of the scale categories. When higher categories are more frequently used in one set of assessments, X, than in the other, Y, there is a systematic disagreement in position. The measure *Relative Position, RP,* estimates the parameter of a systematic disagreement in position defined by $\gamma = P(X < Y) - P(Y < X)$.

A systematic disagreement in how the two assessments are concentrated to the scale categories is measured by the *Relative Concentration, RC,* estimating the parameter of a systematic shift in concentration $\delta = P(X_{l_1} < Y_k < X_{l_2}) - P(Y_{l_1} < X_k < Y_{l_2})$.

The measure of individual variability, the relative rank variance, $0 \le RV \le 1$ is defined $RV = \frac{6}{n^3}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}\left[\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)}\right]^2$ where $\overline{R}_{ij}^{(X)}$ is the mean augmented rank of the observations in the ijth cell of an $m \times m$ square contingency table according to the assessments X. In the aug-rank approach $\overline{R}_{i,j-1}^{(X)} < \overline{R}_{i,j}^{(X)}$ and $\overline{R}_{i-1,j}^{(Y)} < \overline{R}_{i,j}^{(Y)}$. $RV = 0$ means that the observed disagreement is completely explained by the measures of systematic disagreement. In that case the two sets of aug-ranks are equal and the paired distribution is the *rank-transformable pattern of agreement* (see ►Ranks).

The advantage of separating the observed disagreement in the components of systematic and individual disagreements is that it is possible to improve the rating scales and/or the users of the scale. Systematic disagreement is population based and reveals a systematic change in conditions between test-.retest assessments or that raters interpret the scale categories differently. Large individual variability is a sign of poor quality of the rating scale as it allows for uncertainty in repeated assessments.

The Cohen's *coefficient kappa* (κ) is a commonly used measure of agreement adjusted for the chance expected agreement. There are limitations with kappa. The maximum level of kappa, $\kappa = 1$, requires equally skilled raters, in other words lack of systematic disagreement (bias). The value of weighted kappa depends on the choice of weights, and the weighting procedure ignores the rank-invariant properties of ordinal data. The kappa value increases when the number of categories decreases, and depends also on how the observations are distributed on the different categories, the prevalence. Therefore kappa values from different studies are not comparable.

The calculations of Cronbach's alfa and other so-called reliability coefficients are based on the assumption of quantitative, normally distributed data, which is not achievable in data from rating scales.

There is also a widespread misuse of correlation in reliability studies. The correlation coefficient measures the *degree of association* between two variables and does not measure the level of agreement, see Fig. 1. The PA is 12%, and the observed disagreement is mainly explained by a systematic disagreement in position. The negative RP value

A. The observed pattern							B. The rank-transformable pattern of agreement					
X / Y	C_1	C_2	C_3	C_4	total		X / Y	C_1	C_2	C_3	C_4	total
C_4			1	1	2		C_4				2	2
C_3		2	2	14	18		C_3			1	17	18
C_2	1	1	11	3	16		C_2			16		16
C_1	2	8	3	1	14		C_1	3	11			14
total	3	11	17	19	50			3	11	17	19	50

Measures of Agreement. Fig. 1 The frequency distribution of 50 pairs of assessments on a scale with four ordered categories, $C_1 < C_2 < C_3 < C_4$ and the corresponding rank-transformable pattern of agreement, defined by the marginal distributions

(-0.48) and the constructed RTPA shows that the assessments Y systematically used a lower category than did X. A slight additional individual variability, $RV = 0.08$ is observed. The Spearman rank-order correlation coefficient is 0.66 in A and 0.97 in B, ignoring the fact that the assessments are systematically biased and unreliable. The same holds for the coefficient kappa (-0.14).

About the Author

For biography *see* the entry ►Ranks.

Cross References

►Kappa Coefficient of Agreement
►Ranks
►Rating Scales

References and Further Reading

Svensson E (1997) A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical J 39:643–657
Svensson E (1998) Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. J Epidemiol Biostat 3(4):403–409

Measures of Dependence

Reza Modarres
Head and Professor of Statistics
The George Washington University, Washington, DC, USA

Let X and Y be continuous random variables with joint distribution function (DF) H and marginal DFs F and G. Three well-known measures of dependence are

1. Pearson's correlation:
$$\rho = \frac{1}{\sigma_X \sigma_Y} Cov(X, Y) = \frac{1}{\sigma_X \sigma_Y} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dxdy$$
where σ_x, σ_y and $Cov(X, Y)$ are the standard deviations and covariance of X and Y, respectively
2. Spearman's correlation: $s = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dF(x)dG(y)$,
3. Kendall's correlation: $\tau = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y)dH(x,y) - 1$

Pearson correlation measures the strength of linear relationship between X and Y and has well-studied theoretical properties. However, it can be unduly influenced by ►outliers, unequal variances, non-normality, and nonlinearity. Spearman's correlation reflects the monotone association between X and Y and measures the correlation between $F(X)$ and $G(Y)$. Kendall's correlation is the probability of concordance minus the probability of discordance. Spearman's and Kendall's correlations remain invariant under a monotone transformation. However, Pearson's correlation remains only invariant under a location and scale change.

Using the probability integral transformations $u = F(x)$ and $v = G(y)$, the copula (see also ►Copulas) of X and Y is defined as $C(u, v) = H(F^{-1}(u), G^{-1}(v))$. Hence,

$$\rho = \frac{1}{\sigma_X \sigma_Y} \iint_{I^2} [C(u,v) - uv]dF^{-1}(u)dG^{-1}(v),$$

$$s = 12 \iint_{I^2} [C(u,v) - uv]dudv,$$

$$\tau = 4 \iint_{I^2} C(u,v)dC(u,v) - 1$$

where I^2 is the unit square. Schweizer and Wolff (1981) note that $C(u, v) - uv$ is the signed volume between the surface $z = C(u, v)$ and $Z = uv$ (the independence copula).

Copula representation of ρ clearly shows its dependence on the marginal distributions. Therefore, it is not a measure of nonparametric dependence. Daniels (1950) shows that $-1 \le 3\tau - 2s \le 1$. Nelsen (1991) studies the relationship between s and τ for several families of copulas and Fredricks and Nelsen (2007) show that the ratio τ/s approaches 2/3 as H approaches independence.

Hoeffding (1940) and Frechét (1951) show that for all $(x,y) \in R^2$ the joint DF is bounded: $H_1(x,y) \le H(x,y) \le H_2(x,y)$ where $H_1(x,y) = max(0, F(x) + G(y) - 1)$ and $H_2(x,y) = min(F(x), G(y))$ are distribution functions. Perfect negative correlation is obtained when H_1 is concentrated on the line $F(x) + G(y) = 1$ whereas perfect positive correlation is obtained when H_2 is concentrated on the line $F(x) = G(y)$. In fact, $H_0(x,y) = F(x)G(y)$ for all $(x,y) \in R^2$ reflects independence of X and Y. Let $C_1(x,y) = max(0, u + v - 1)$, $C_2(x,y) = min(u,v)$ and $C_0(x,y)$ denote the Frechét lower, upper and independence copulas, respectively. Similarly, $C_1(u,v) \le C(u,v) \le C_2(u,v)$.

Using Hoeffding lemma (1948)

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[H(x,y) - F(x)G(y)]dxdy,$$

one can show $\rho_1 \le \rho \le \rho_2$ where ρ_1 and ρ_2 are the correlation coefficients associated with H_1 and H_2, respectively. Depending on the marginal distributions the range of ρ may be much smaller than $|\rho| \le 1$. For example, for the bivariate log-normal distribution with unit variances, one can show $\rho \in (-0.368, 1)$. Lancaster (1958) uses Chebyshev-Hermite polynomial to obtain the correlation coefficient of transformed bivariate random vectors. Freeman and Modarres (2005) obtain the form of the correlation after a ►Box-Cox transformation.

Moran (1967) states that the necessary and sufficient conditions for ρ to assume extreme values of +1 and −1 are

1. $X \overset{d}{=} aY + b$ for constants
2. $F(\mu + x) = 1 - F(\mu - x)$ where μ is the mean of X. Normal, uniform, double exponential and logistic distributions satisfy these conditions

Rényi (1959) considers a set of conditions that a symmetric nonparametric measure of dependence should satisfy. Schweizer and Wolff (1981) note that Rényi's conditions are too strong and suggest that any suitably normalized distance measure such as the L_p distance provides a symmetric measure of nonparametric dependence. They show that these distances, according to a modified set of Rényi conditions, enjoy many useful properties. Let $L_p = (K_p \int_{I^2} |C(u,v) - uv|^p dudv)^{1/p}$ where K_p is chosen such that L_p remains in $(0,1)$. We have

1. $L_1 = 12 \int_{I^2} |C(u,v) - uv| dudv$
2. $L_2 = \left(90 \int_{I^2} (C(u,v) - uv)^2 dudv\right)^{1/2}$
3. $L_\infty = 4\, Sup_{I^2} |C(u,v) - uv|$

In fact Hoeffding (1948) and Blum et al. (1961) base a nonparametric test of independence between X and Y on L_∞. Modarres (2007) studies several tests of independence, including a measure based on the likelihood of cut-points.

About the Author

Dr. Reza Modarres is a Professor and Head, Department of Statistics, George Washington University, Washington DC. He is an elected member of International Statistical Society. He has authored and co-authored more than 50 papers and is on the editorial board of several journals.

Cross References

►Bivariate Distributions
►Copulas: Distribution Functions and Simulation
►Correlation Coefficient
►Kendall's Tau
►Statistics on Ranked Lists
►Tests of Independence

References and Further Reading

Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free tests of independence based on the sample distribution function. Ann Math Stat 32:485–498

Daniels HE (1950) Rank correlation and population models. J R Stat Soc B 12:171–181

Frechét M (1951) Sur les tableaux de corrélation dont les marges sont données. Ann Univ Lyon Sec A 14:53–57

Fredricks GA, Nelsen RB (2007) On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. J Stat Plan Infer 137:2143–2150

Freeman J, Modarres R (2005) Efficiency of test for independence after Box-Cox transformation. J Multivariate Anal 95: 107–118

Hoeffding W (1940) Masstabinvariante korrelations-theorie. Schriften Math Inst Univ Berlin 5:181–233

Hoeffding W (1948) A nonparametric test of independence. Ann Math Stat 19:546–557

Lancaster HO (1958) The structure of bivariate distributions. Ann Math Stat 29:719–736

Modarres R (2007) A test of independence based on the likelihood of cut-points. Commun Stat Simulat Comput 36:817—825

Moran PAP (1967) Testing for correlation between non-negative variates. Biometrika 54:385–394

Nelsen RB (1991) Copulas and association. In: Dall'Aglio G, Kotz S, Salinetti G (eds) Advances in probability distributions with given marginals. beyond copulas. Kluwer Academic, London

Rényi A (1959) On measures of dependence. Acta Math Acad Sci Hunger 10:441–451

Schweizer B, Wolff EF (1981) On nonparametric measures of dependence for random variables. Ann Stat 9(4):879–885

Median Filters and Extensions

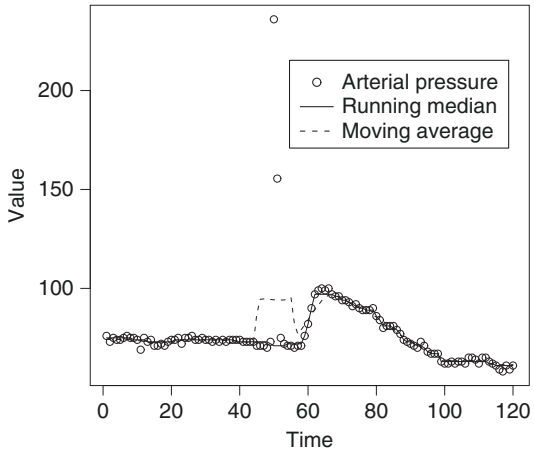
Roland Fried[1], Ann Cathrice George[2]
[1]Professor
TU Dortmund University, Dortmund, Germany
[2]TU Dortmund University, Dortmund, Germany

De-noising a time series, that is a sequence of observations of a variable measured at equidistant points in time, or an image, that is a rectangular array of pixels, is a common task nowadays. The objective is to extract a varying level (a "signal") representing the path followed by the time series or the true image which is overlaid by irrelevant noise.

Linear filters like moving averages are computationally simple and eliminate normal noise efficiently. However, their output is heavily affected by strongly deviating observations (called ▶outliers, spikes or impulses), which can be caused for instance by measurement artifacts. Moreover, linear filters do not preserve abrupt changes (also called step changes or jumps) in the signal or edges in an image. Tukey (1977) suggests median filters, also called running medians, for these purposes.

We focus on the time series setting in the following. Let $y_1, \ldots, y_N$ be observations of a variable at equidistant points in time. De-noising these data for extraction of the time-varying mean level underlying these data (the signal) can be accomplished by moving a time window $y_{t-k}, \ldots, y_t, \ldots, y_{t+k}$ of length $n = 2k+1$ through the series for estimation of the level μ_t in the center of the window. Whereas a moving average calculates the arithmetic average of the data in the time window for this, a running median uses the median of these values. If the window width is fixed throughout, we get estimates of the levels $\mu_{k+1}, \ldots, \mu_{N-k}$ at instances not very close to the start or the end of the time series. The levels at the start or the end of the time series can be estimated for instance by extrapolation of the results from the first and last window or by adding the first and the last observed value a sufficient number of times.

Figure 1 depicts observations of the arterial blood pressure of a patient in intensive care measured once a minute, as well as the outputs of a moving average and a running median, both with window width $n = 11$. The moving average is strongly affected by a few measurement artifacts, and it smooths the sudden increase at $t = 60$. The running median eliminates the spikes and preserves the shift.



Median Filters and Extensions. Fig. 1 Measurements of the arterial blood pressure of a patient and outputs of a running median and a moving average, both with window width $n = 11$

A possible disadvantage of running medians is that they implicitly rely on the assumption that the level is almost constant within each time window. While increasing the window width improves the reduction of noise if the signal is locally constant, this is no longer the case in trend periods. Davies et al. (2004) investigate application of robust regression to a moving time window to improve the approximation of trends in the presence of ▶outliers. Many further refinements of robust filters for signal extraction from time series or images and different rules for choosing a (possibly locally adaptive) window width from the data have been suggested in the literature. See Gather et al. (2006) for an overview on robust signal extraction from time series.

Cross References

▶Moving Averages
▶Outliers
▶Smoothing Techniques
▶Statistical Signal Processing
▶Time Series

References and Further Reading

Davies L, Fried R, Gather U (2004) Robust signal extraction for online monitoring data. J Stat Plan Infer 122:65–78

Gather U, Fried R, Lanius V (2006) Robust detail-preserving signal extraction. In: Schelter B, Winterhalder M, Timmer J (eds) Handbook of time series analysis. Wiley, New York, pp. 131–158

Tukey JW (1977) Exploratory data analysis (preliminary edition 1971). Addison-Wesley, Reading MA

Medical Research, Statistics in

B. S. Everitt
Professor Emeritus
Institute of Psychiatry, King's College, London, UK

Statistical science plays an important role in medical research. Indeed a major part of the key to the progress in medicine from the 17th century to the present day has been the collection and valid interpretation of empirical evidence provided by the application of statistical methods to medical studies. And during the last few decades, the use of statistical techniques in medical research has grown more rapidly than in any other field of application. Indeed, some branches of statistics have been especially stimulated by their applications in medical investigations, notably the analysis of ►survival data (see, for example, Collett 2003). But why has statistics (and statisticians) become so important in medicine? Some possible answers are:

- Medical practice and medical research generate large amounts of data. Such data can be full of uncertainty and variation and extracting the "signal," i.e. the substantive medical message in the data, form the 'noise' is usually anything but trivial.
- Medical research often involves asking questions that have strong statistical overtones, for example: 'How common is a particular disease?'; 'Which people have the greatest chance of contracting some condition or other?'; 'What is the probability that a patient diagnosed with breast cancer will survive more than five years?'
- The evaluation of competing treatments or preventative measures relies heavily on statistical concepts in both the design and analysis phase.

In a short article such as this it is impossible to cover all areas of medicine in which statistical methodology is of particular importance and so we shall concentrate on only three namely, clinical trials, imaging and molecular biology. (For a more comprehensive account of the use of statistics in medicine see Everitt and Palmer (2010)).

Clinical Trials

If a doctor claims that a certain type of psychotherapy will cure patients of their depression, or that taking large doses of vitamin C can prevent and even cure the common cold, how should these claims be assessed? What sort of evidence do we need to decide that claims made for the efficacy of clinical treatments are valid? One thing is certain: We should *not* rely either on the views of 'experts' unless they provide sound empirical evidence (measurements, observations, i.e., *data*) to support their views, nor should we credit the anecdotal evidence of people who have had the treatment and, in some cases, been 'miraculously' cured. (And it should be remembered that the plural of anecdote is not evidence.) Such 'wonder' treatments, which are often exposed as ineffectual when exposed to more rigorous examination, are particularly prevalent for those complaints for which conventional medicine has little to offer (see the discussion of alternative therapies in Chapter 13 of Everitt 2008).

There is clearly a need for some form of carefully controlled procedure for determining the relative effects of different treatments and this need has been met in the 20th and 21st centuries by the development of the clinical trial, a medical experiment designed to evaluate which (if any) of two or more treatments is the more effective. The quintessential components of a clinical trial, the use of a control group and, in particular the use of ►randomization as a way of allocating participants in the trial to treatment and control groups, were laid down in the first half of the 20th century. The randomization principle in clinical trials was indeed perhaps the greatest contribution made by arguably the greatest statistician of the 20th century, Sir Ronald Aylmer Fisher. Randomization achieves the following:

- It provides an impartial method, free of personal bias, for the assignment of participants to treatment and control groups. This means that treatment comparisons will not be invalidated by the way the clinician might chose to allocate the participants if left to his or her own judgment.
- It tends to balance treatment groups in terms of extraneous factors that might influence the outcome of treatment, even in terms of those factors the investigator may be unaware of.

Nowadays some 9,000–10,000 clinical trials are undertaken in all areas of medicine form the treatment of acne to the prevention of cancer and the randomized controlled clinical trial is perhaps the outstanding contribution of statistics to 20th century medical research. And in the 21st century statisticians have applied themselves to developing methods of analysis for such trials that can deal with the difficult problems of patient drop-out, the longitudinal aspects of most trials and the variety of measurement types used in such trials (see Everitt and Pickles 2004).

Imaging

Examples of medical imaging systems include conventional radiology (X-rays), positron-emission tomography (PET), magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI). A significant advantage often claimed for medical imaging is its ability to visualize structures or processes in the patient without the need for intrusive procedures, for example, surgery; but this may also be a disadvantage and the question that may need to be asked is how well do the conclusions from an imaging experiment correspond to the physical properties that might have been found from an intrusive procedure?

Imaging studies generate large amounts of data and a host of statistical techniques have been employed to analyze such data and to extract as much information as possible from what is in many cases very 'noisy' data. Autoregressive models, linear mixed effects models, finite mixture models and Gaussian random field theory have all been applied to mixture data with varying degrees of success. Some important references are Besag (1986), Silverman et al. (1990) and Lange (2003).

Molecular Biology

Molecular biology is the branch of biology that studies the structure and function of biological macromolecules of a cell and especially their genetic role. A central goal of molecular biology is to decipher the genetic information and understand the regulation of protein synthesis and interaction in the cellular process. Advances in biotechnology have allowed the cloning and sequencing of DNA and the massive amounts of data generated have given rise to the new field of ►bioinformatics which deals with the analysis of such data. A variety of statistical methods have been used in this area; for example, hidden Markov models have been used to model dependencies in DNA sequences and for gene finding (see Schliep et al. 2003) and data mining techniques (see ►Data Mining), in particular, cluster analysis (see, for example, Everitt et al. 2010) have been used to identify sets of genes according to their expression in a set of samples, and to cluster samples (see ►Cluster Sampling) into homogeneous groups (see Toh and Honimoto 2002).

Statistical methods are an essential part of all medical studies and increasingly sophisticated techniques now often get a mention in papers published in the medical literature. Some of these have been mentioned above but others which are equally important are Bayesian modeling (see Congdon 2001) and generalized estimating equations (see Everitt and Pickles 2004). In these days of evidence-based medicine (Sackett et al. 1996), collaboration between medical researchers and statisticians is essential to the success of almost all research in medicine.

About the Author

Brian Everitt retired from his post as Head of the Department of Computing and Statistics at the Institute of Psychiatry, King's College, London in 2005. He is the author (or joint author) of about 100 journal papers and 60 books. In retirement he continues to write and with colleagues has nearly completed the 5th edition of *Cluster Analysis*, first published in 1974. Apart from writing his interests are playing classical guitar (badly), playing tennis, walking and reading.

Cross References

►Biopharmaceutical Research, Statistics in
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Medical Statistics
►Research Designs
►Role of Statistics
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines
►Statistics: Nelder's view
►Survival Data
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Besag J (1986) On the statistical analysis of dirty pictures (with discussion). J Roy Stat Soc Ser B 48:259–302
Collett D (2003) Survival data in medical research. CRC/Chapman and Hall, London
Congdon P (2001) Bayesian statistical modelling. Wiley, Chichester
Everitt BS (2008) Chance rules, 2nd edn. Springer, New York
Everitt BS, Landau S, Leese M, Stahl D (2010) Cluster analysis, 5th edn. Wiley, Chickester, UK
Everitt BS, Palmer CR (2010) Encyclopaedic companion to medical statistics, 2nd edn. Wiley, Chichester, UK
Everitt BS, Pickles A (2004) Statistical aspects of the design and analysis of clinical trials. Imperial College Press, London
Lange N (2003) What can modern statistics offer imaging neuroscience? Stat Methods Med Res 12(5):447–469
Sackett DL, Rosenberg MC, Gray JA, Haynes RB, Richardson W (1996) Evidence-based medicine: what it is and what it isn't. Brit Med J 312:71–72
Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression data. Bioinformatics 19: 255–263
Silverman BW, Jones MC, Wilson JD, Nychka DW (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). J Roy Stat Soc Ser B 52:271–324
Toh H, Honimoto K (2002) Inference of a genetic network by a combined approach to cluster analysis and graphical Gaussian modelling. Bioinformatics 18:287–297

Medical Statistics

Vern T. Farewell[1], Daniel M. Farewell[2]
[1]Associate Director
Medical Research Council, Biostatistics Unit, Cambridge, UK
[2]School of Medicine, Cardiff University, Cardiff, UK

Historical Background

The term statistics has at least three, related, meanings. It may refer to data in raw form, or to summaries thereof, or to the analysis of uncertainty associated with data. The phrase medical statistics, therefore, may reasonably be applied to the specialization to medical science of any of these understandings of statistics.

Raw medical statistics date back at least to the London Bills of Mortality, collected weekly between 1603 and 1836 in order to provide an early warning of plague. The early demographic work of John Graunt (1620–1674) was based on these Bills. The summaries of vital statistics undertaken by William Farr (1807–1883), working at the General Registry Office of England and Wales, became the basis of many important health reforms. However, the founding editors of the journal *Statistics in Medicine* described modern medical statistics as "the deployment of the ideas, principles and methods of statistics to stimulate deeper understanding in medicine" (Colton et al. 1982), emphasizing the third understanding of the term.

The history of the link between statistics and medicine includes key figures in the development of statistics itself. For example, Arbuthnot (1667–1753) and Bernoulli (1700–1782), often cited in the early use of significance tests, were each qualified in both mathematics and in medicine. Many individuals have contributed to the emergence of medical statistics as a scientific discipline in its own right. The French writers, Pinel (1745–1826), Louis (1787–1872) and Gavarret (1809–1890) and the Danish physician, Heiberg (1868–1963) provided early impetus. Subsequently, Pearl (1879–1940) and Greenwood (1880–1949) established research programmes in medical statistics in the USA and the UK respectively. In 1937, Hill (1897–1991) published the highly influential book, *Principles of Medical Statistics*, Hill (1937), of which twelve editions were published over the next 55 years. Two other important contributions of Hill were arguably the first modern randomized clinical trial on the effect of streptomycin in tuberculosis, and his discussion of criteria for causality in epidemiological studies. A useful source for information on the history of medical statistics is the Lind Library [http://www.jameslindlibrary.org].

The Nature of Medical Statistics

Much activity in medical statistics is necessarily collaborative. Over the course of a career, statisticians engaged in medical research are likely to work closely with physicians, nurses, laboratory scientists and other specialists. Communication across disciplines can present challenges but, in addition to its scientific merit, also frequently stimulates worthwhile methodological and theoretical research. Further, since medical research often raises ethical issues, these too must be considered by medical statisticians. Hill (1936) stressed that the statistician "cannot sit in an armchair, remote and Olympian, comfortably divesting himself of all ethical responsibility."

A dominant characteristic of the statistical methods arising in medical statistics is that they must make allowance for known variability. Comparisons of groups should adjust for systematic discrepancies between groups, for instance in terms of demographics. This has been reflected for many years by the high profile given to regression methodology, which allows multiple explanatory variables to be incorporated. A more recent manifestation is in the monitoring of medical performance, where quality control procedures developed for industrial application have been modified to allow for predictable heterogeneity in medical outcomes (Grigg et al. 2003).

Illustrative Methodological Developments

In 1984, Cox identified three important periods in the development of modern statistical methodology. The first was linked to developments in agriculture, the second to industrial applications, and the third to medical research. Developments linked to medical research flourished in the 1970s; where earlier statistical methodology placed particular emphasis on normally distributed data, there was a need for methods more suited to survival (or time-to-event) and categorical data. A distinguished example of the former is Cox's own pioneering paper (Cox 1972), presenting a semiparametric regression model for ▶survival data that did not require full specification of an underlying survival distribution. In addition, and in contrast to virtually all other regression methods then available, this model allowed the incorporation of explanatory variables that varied over time. A wealth of subsequent extensions to this already very general methodology followed, many facilitated by Aalen's (1978) reformulation of the problem in a counting process framework [see also Andersen et al. (1993)].

An important application of statistical models for categorical data was to ▶case- control studies. These epidemiological investigations of the relationship between a disease

D and exposure E, a possible risk factor, involve separate sampling of diseased and disease-free groups, from which information on E and other disease risk factors is obtained. Binary ►logistic regression would seem to provide a natural tool for the analysis of these studies, but for the fact that it focuses on $\mathrm{pr}(D|E)$ whereas the sampling is from the distribution $\mathrm{pr}(E|D)$. Building on a series of earlier papers, Prentice and Pyke (1979) established how a prospective logistic regression model for $\mathrm{pr}(D|E)$ could be used with case-control data to provide valid estimates of the odds-ratio parameters. This rapidly became the standard methodology for the analysis of case-control studies (Breslow 1996).

Study Design

The design of medical studies is also a major area of activity for medical statisticians. The paradigmatic design is perhaps the Phase III clinical trial, of which a key aspect is often randomized treatment assignment. While ►randomization can provide a basis for statistical inference, its primary motivation in trials is to enable statements of causality, critical for Phase III trials where the aim is to establish treatment efficacy. Nevertheless, the need for, and methods of, randomization continue to generate discussion, since randomization can be seen to sacrifice potential individual advantage for collective gain. Other design questions arise in Phase I trials that establish the tolerability of treatments and basic pharmacokinetics, and Phase II trials aimed at finding potentially efficacious treatments or dosages.

For ethical reasons, ongoing monitoring of data during a clinical trial is often needed, and this has been an area of methodological investigation within medical statistics since the pioneering work of Armitage (1975) (a comprehensive discussion may be found in Jennison and Turnbull (2000)). There is also an increasing role for statisticians on formal committees that monitor trial data and safety, where their expertise is combined with that of physicians, ethicists, and community representatives to ensure the ethical conduct of trials more generally.

In the 1980s, two important variations on the standard case-control design emerged, namely case-cohort studies (Prentice 1986) and two stage case-control designs (Breslow and Cain 1988); both have proved very useful in epidemiology. Epidemiological cohorts where individuals are followed to observe disease incidence, or clinical cohorts for which information on patients with specified conditions is collected routinely – both usually implemented over long periods of time – also continue to present design and analysis challenges to the medical statistician.

More Recent Topics of Interest

Typically, medical studies are conducted not only to discover statistical associations, but also in the hopes of suggesting interventions that could benefit individuals or populations. This has led to a preference for investigations incorporating randomization or multiple waves of observation, based on the idea that cause should precede effect. Randomized or not, information gathered repeatedly on the same subjects is known as longitudinal data, and its analysis has become a major subdiscipline within medical statistics. Two distinct approaches to longitudinal data analysis have risen to prominence: likelihood-based models (incorporating both classical and Bayesian schools of thought) and estimating-equation techniques.

A consequence of this emphasis on studies monitoring subjects over several months (or even years) has been an increased awareness that data, as collected, are often quite different from what was intended at the design stage. This may be due to subjects refusing treatment, or choosing an alternate therapy, or dropping out of the investigations altogether. Likelihood approaches to longitudinal data may be extended to incorporate an explicit model for the observation process (Henderson et al. 2000), while estimating equations can be modified with subject- or observation-specific weights (Robins et al. 1995) to account for departures from the study design. Non-compliance, dynamic treatment regimes, and incomplete data are all areas of active methodological research within medical statistics.

Two other major areas of current interest are meta-analysis and genetic or genomic applications. Meta-analysis is often taken to refer to the technical aspects of combining information from different studies that address the same research question, although the term is sometimes used to describe the more general systematic review, which includes broader issues such as study selection. Study heterogeneity is an important aspect of ►meta-analysis that the statistician must address. The size and complexity of genetic and genomic data present major statistical and computational challenges, notably due to hypothesis test multiplicity.

Conclusion

Medicine remains a major area of application driving methodological research in statistics, and the demand for medical statisticians is considerable. A comprehensive introduction to the area can be found in Armitage et al. (2002) and a less technical introduction is Matthews and Farewell (2007).

About the Author

Prior to moving to the MRC Bistatistics Unit, Vern Farewell held professorial positions at the University of Washington, the University of Waterloo and University College London. He has published over 200 papers in the statistical and medical literature and is co-author of the four editions of the book *Using and Understanding Medical Statistics*. Since 2007, he has been Editor of *Statistics in Medicine*.

Cross References

►Biostatistics
►Case-Control Studies
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Hazard Regression Models
►Logistic Regression
►Medical Research, Statistics in
►Meta-Analysis
►Modeling Survival Data
►Psychiatry, Statistics in
►Statistical Analysis of Longitudinal and Correlated Data
►Statistical Genetics
►Statistical Methods in Epidemiology
►Statistics, History of
►Statistics: An Overview
►Survival Data

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Armitage P (1975) Sequential medical trials. Blackwell, Oxford

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research. Blackwell Science, Oxford

Breslow NE (1996) Statistics in epidemiology: the case control study. J Am Stat Assoc 91:14–28

Breslow NE, Cain KC (1988) Logistic regression for two-stage case-control data. Biometrika 75:11–20

Colton T, Freedman L, Johnson T (1982) Editorial. Stat Med 1:1–3

Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 34:187–220

Cox DR (1984) Present position and potential developments: some personal views: design of experiments and regression. J R Stat Soc A 147:306–315

Grigg OA, Farewell VT, Spiegelhalter DJ (2003) Use of risk adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Meth Med Res 12:147–170

Henderson R, Diggle P, Dobson A (2000) Joint modelling of repeated measurements and event time data. Biostatistics 1:465–480

Hill AB (1936) Medical ethics and controlled trials. Br Med J 3: 1043–1049

Hill AB (1937) Principles of medical statistics. Lancet, London

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman and Hall/CRC, New York

Matthews DE, Farewell VT (2007) Using and understanding medical statistics. Karger, Basel

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–12

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411

Robins JM, Rotnitsky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–121

Meta-Analysis

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, Bundoora, VIC, Australia

Introduction

Given several studies on the same topic, a *meta-analysis* synthesizes the information in them so as to obtain a more precise result. The proper procedure of conducting a *systematic review* of literature, the selection of which studies to include and the issues of *publication bias* and other possible biases are important aspects not covered here and we refer the interested reader to Cooper and Hedges (1994) and Higgins and Green (2008). We assume all studies estimate the same *effect*, which is often a comparison of outcomes for control and treatment groups via clinical trials. Examples for two binomial samples with parameters $(n_1, p_1), (n_2, p_2)$ are the *risk difference* $p_1 - p_2$, *relative risk* p_2/p_1 and *odds ratio* $\{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}$. Other examples comparing normal samples are the difference in means $\mu_1 - \mu_2$, or *effect sizes* such as the *standardized mean difference, or Cohen's-d* $d = (\mu_1 - \mu_2)/\sigma$ from Cohen (1988), where σ^2 is an assumed common variance, and Glass's $g = (\mu_1 - \mu_2)/\sigma_1$ from Glass (1976), where σ_1^2 is the variance of the control group.

Traditional Meta-Analysis Methodology

We are given K independent studies, in which the estimated effects $\hat{\theta}_k$ based on N_k observations are asymptotically normal such that $\hat{\theta}_k$ is for large enough N_k approximately normally distributed with mean θ_k and variance σ_k^2/N_k. This is denoted $\hat{\theta}_k \sim AN\left(\theta_k, \sigma_k^2/N_k\right)$ for each $k = 1, \ldots, K$. Examples satisfying the above assumptions are the risk difference, the log-relative risk, the log-odds ratio and the Cohen's-d. The goal is to combine the estimators $\hat{\theta}_k$ in some way so as to estimate a representative θ for all K studies, or even more ambitiously, for all potential studies of this type. Thus there is the conceptual question of how to define a representative θ, and the inferential problem of how to find a confidence interval for it.

Confidence Intervals for Effects

Note that for each individual study, one can already form large sample confidence intervals for individual θ_k, $k = 1, \ldots, K$. For *known* σ_k, a $100(1-\alpha)\%$ large-sample confidence interval for θ_k is $[L_k, U_k] = \left[\hat{\theta}_k - z_{1-\alpha/2}\sigma_k/N_k^{1/2},\ \hat{\theta}_k + z_{1-\alpha/2}\sigma_k/N_k^{1/2}\right]$, where $z_\beta = \Phi^{-1}(\beta)$ is the β quantile of the standard normal distribution. If σ_k is *unknown*, and there exists estimators $\hat{\sigma}_k$ with $\hat{\sigma}_k/\sigma_k \to 1$ in probability as $N_k \to \infty$, then the same can be said for $[L_k, U_k] = \left[\hat{\theta}_k - z_{1-\alpha/2}\hat{\sigma}_k/N_k^{1/2}, \hat{\theta}_k + z_{1-\alpha/2}\hat{\sigma}_k/N_k^{1/2}\right]$.

Unequal Fixed Effects Model (UFEM)

Standard meta-analysis proceeds by choosing a weight w_k for each study and combines the estimated $\hat{\theta}_k$ through weighted means. If we interpret θ_k as the true effect for the study k and if this effect is of interest in its own right, then the following definition can be adopted. Consider a representative effect for the K studies defined by $\theta_w = \sum_k w_k\theta_k/W$ with $W = \sum_j w_j$. This *weighted effect* is the quantity that we want to estimate by meta-analysis. There is a good dose of arbitraryness in this procedure, because the weighted effect does not necessarily have a readily interpreted meaning. An exception occurs if the weights are all equal to one, in which case θ_w is simply the average of the study effects.

The weights are, however, often chosen to be proportional to the reciprocals of the variances in order to give more weight to θ_k that are estimated more accurately. If this is the choice, it follows that $w_k = N_k/\sigma_k^2$ and $\hat{\theta}_w = \sum_k w_k\hat{\theta}_k/W$ satisfies $\hat{\theta}_w \sim AN(\theta_w, W^{-1})$. Therefore a $100(1-\alpha)\%$ large-sample confidence interval for θ_w is given by $[L, U] = [\hat{\theta}_w - z_{1-\alpha/2}W^{-1/2},\ \hat{\theta}_w + z_{1-\alpha/2}W^{-1/2}]$.

In practice the weights usually need to be estimated, (w_k by $\hat{w}_k$ and W by $\hat{W} = \sum_k \hat{w}_k$), but a large sample confidence interval for θ_w can be obtained by substituting $\hat{\theta}_{\hat{w}}$ for $\hat{\theta}_w$ and $\hat{W}$ for W in the above interval.

Fixed Effects Model (FEM)

When statisticians speak of the fixed effects model they usually mean *equal* fixed effects which makes the very strong assumption that all $\theta_k = \theta$. This has the appeal of simplicity. The UFEM just described includes the FEM as a special case. In particular the target parameter θ_w reduces to $\theta_w = \theta$ and thus becomes a meaningful quantity no matter what weights are chosen.

However, one of the preferred choices still uses the weights inversely proportional to the variance, because in this case $\sum_k w_k\hat{\theta}_k/W$ has the smallest asymptotic variance amongst all unbiased (for θ) linear combinations of the individual study estimators of θ. The same confidence interval given above for θ_w is used for θ. The methodology for the UFEM and FEM models is the same, but the target parameter θ_w of the UFEM has a different interpretation.

Random Effects Model (REM)

The REM assumes that the true effects θ_k, $k = 1, \ldots, K$ are the realized values of sampling from a normal population with mean θ and variance γ^2 for some unknown inter-study variance γ^2, and further that the above results for the UFEM are all *conditional* on the given θ_k, $k = 1, \ldots, K$. The justification for this assumption is that the K studies are a 'random sample' of all possible studies on this topic. Inference for θ can now be interpreted as saying something about the larger population of possible studies.

Formally, the REM assumes $\theta_1, \ldots, \theta_K$ are a sample from $N(\theta, \gamma^2)$, with both parameters unknown; and $\hat{\theta}_k|\theta_k \sim AN\left(\theta_k, \sigma_k^2/N_k\right)$ for each k. If the *conditional* distribution of $\hat{\theta}_k$, given θ_k, were exactly normal, then the *unconditional* distribution of $\hat{\theta}_k$ would be exactly $\hat{\theta}_k \sim N\left(\theta, \gamma^2 + \sigma_k^2/N_k\right)$. However, in general the unconditional distributions are only asymptotically normal $\hat{\theta}_k \sim AN\left(\theta, \gamma^2 + \sigma_k^2/N_k\right)$. It is evident that one needs an estimate $\hat{\gamma}^2$ of γ^2 in order to use the inverse variance weights approach described earlier, and this methodology will be described below.

Choosing between Fixed and Random Effects Models

Qualitative Grounds

If one assumes the K studies are a random sample from a larger population of potential studies and that the true effects θ_k are each $N(\theta, \gamma^2)$ then θ is the target effect, and γ^2 is a measure of inter-study variability of the effect. In

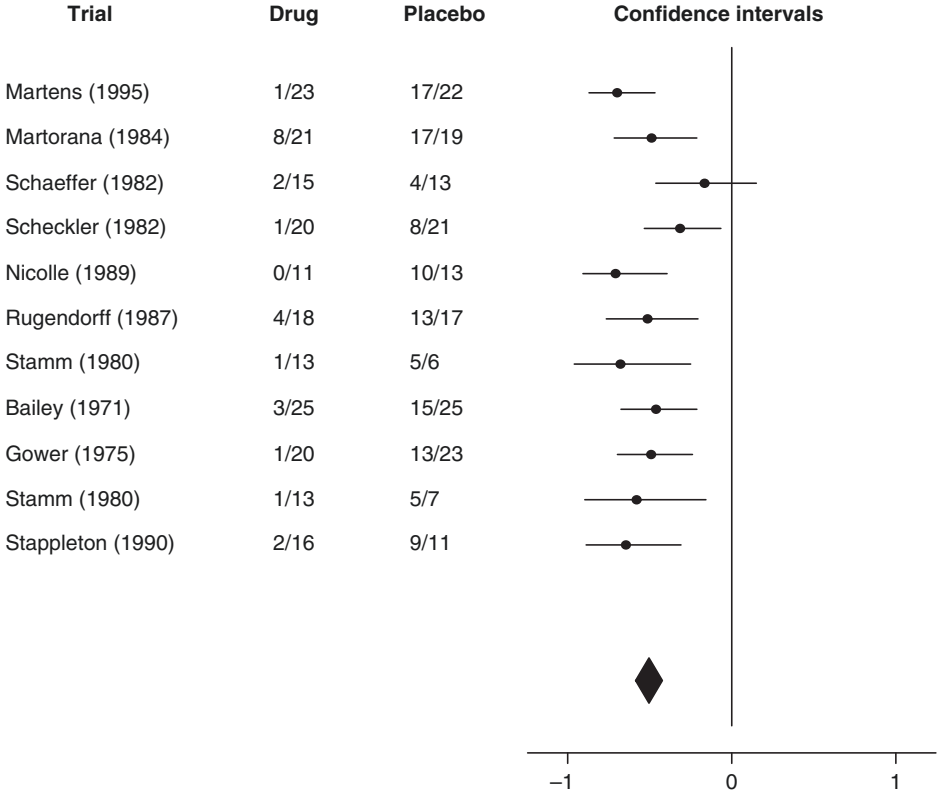
this case choose the REM. If there is reason to believe that the θ_k are different, but not the result of random sampling, then use the UFEM. In this case, it may be possible to explain a good part of the variation in the effects θ_k by *meta-regression*. The differences between the studies can sometimes be captured by variables that describe the circumstances of each study and by regressing the $\hat{\theta}_k$ on such variables, these differences can be explained and corrected. Meta-regression may thus turn a UFEM into a FEM. In both models, the target is $\theta_w = \sum_k w_k \theta_k / W$. If there is reason to believe all $\theta_k = \theta$, (the *homogeneous case*), use the FEM with target θ. For the FEM and UFEM inferential conclusions only apply to the K studies.

Quantitative Grounds

It is clear that if $\gamma^2 = 0$ in the REM, or all $\theta_k = \theta$ in the UFEM, one obtains the FEM. It is a special case of both. One way to test the null hypothesis of homogeneity (all $\theta_k = \theta$) is to use Cochran's Q, defined by $Q = \sum_k w_k (\hat{\theta}_k - \hat{\theta}_w)^2$, where w_k are the inverse variance weights and $\hat{\theta}_w = \sum_k w_k \hat{\theta}_k / W$. One can show that under the null hypothesis of homogeneity, and when each $\hat{\theta}_k$ is normally distributed, $Q \sim \chi^2_{K-1}$, so a level α test of homogeneity rejects when $Q \geq \chi^2_{K-1,1-\alpha}$. Further, under the UFEM model, the statistic Q has a non-central chisquared distribution $Q \sim \chi^2_{K-1}(\lambda)$, where $\lambda = \sum_k w_k (\theta_k - \theta_w)^2$. This result and others allowing for the weaker assumption $\theta_k \sim AN\left(\theta_k, \sigma_k^2 / N_k\right)$ and estimated weights are derived in Sect. 24.1, Kulinskaya et al. (2008). In the asymptotic case, the χ^2 distributions are only approximate. Testing for heterogeneity is strongly discouraged in Higgins and Green (2008) in favor of the quantification of inherently present heterogeneity.

Inference for the REM

Let $M_r = \sum_k w_k^r$ for inverse variance weights w_k, and $a = M_1 - M_2/M_1$. It can be shown that for this model $\mathrm{E}[Q] = K - 1 + a\gamma^2$. This "justifies" the DerSimonian and Laird (1986) estimator $\hat{\gamma}^2_{DL} = \{Q - (K-1)\}^+ / a$, where $\{\ldots\}^+$ means set the quantity in brackets equal to 0 if it is negative and otherwise leave it. Using this estimator and $\hat{\theta}_k \sim AN\left(\theta, \gamma^2 + w_k^{-1}\right)$, we have new weights $w_k^* = \left(\gamma^2 + w_k^{-1}\right)^{-1}$



Meta-Analysis. Fig. 1 The data of eleven independent studies of antibiotic treatment to prevent recurrent urinary tract infection are presented in this forest plot. The confidence intervals for the individual studies are shown on the right-hand side. The lozenge at the bottom shows the combined confidence interval, the result of the meta-analysis

and estimator $\hat{\theta}^* = \sum_k w_k^* \hat{\theta}_k / W^* \sim AN(\theta, \{W^*\}^{-1})$, where $W^* = \sum_k w_k^*$. In practice w_k^* is usually estimated by $\hat{w}_k^* = 1/\left(\hat{\gamma}_{DL}^2 + \hat{w}_k^{-1}\right)$. Anotherestimator of γ^2 is proposed in Biggerstaff and Tweedie (1997).

Meta-Regression

In some cases there is information regarding the K studies which may explain the inter-study variance. In this case the estimated effects $\hat{\theta}_k$ can be considered as responses to be regressed on explanatory variables $x_1, \ldots, x_p$, also called *moderators*. Thus one has $y_k = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_p x_{kp} + \epsilon_k$, where y_k is the estimated effect $\hat{\theta}_k$ (or a transformed effect), and ϵ_k is the random error in the kth study, $k = 1, \ldots, K$. Weighted least squares (with known or estimated weights) can be used to estimate the coefficients. When the variance stabilizing transformation is applied to estimated effects, generalized linear models techniques (see ►Generalized Linear Models) with Gaussian family of distributions can be used, see Chap. 14 of Kulinskaya et al. (2008).

Example

As illustration, consider a series of 11 studies of antibiotic treatment to prevent recurrent urinary tract infection. The sources of the data, the data themselves, and the confidence intervals are shown in Fig. 1. These studies are part of those reviewed by Albert et al. (2004) and have been discussed in Chap. 19 (p. 158) of Kulinskaya et al. (2008). The total sample sizes range from $N = 19$ to $N = 50$. The parameter of interest is the risk difference $p_1 - p_2$ between the placebo group and the treated groups. The studies show a more or less strong benefit of the treatment, while the meta-analysis gives a fairly convincing result. This depiction of results is known as a *forest plot*.

Additional Literature

The traditional approach is general, only requiring asymptotically normal effects and estimates for the weights. However the methodology is overly simple, because it assumes known weights, when in fact they usually need to be estimated. Recent studies indicate that typical sample sizes are woefully inadequate in order for the approximations that assume known weights to be reliable (Malzahn et al. 2000; Viechtbauer 2007). One way of overcoming this problem is to employ variance stabilization of the estimated effects before applying the traditional approach, see Kulinskaya et al. (2008). For further reading we recommend the classical work Hedges and Olkin (1985), as well as the recent books Böhning et al. (2008), Borenstern et al. (2009), Hartung et al. (2008) and Whitehead (2002).

About the Authors

Prof. Elena Kulinskaya is a recently appointed Aviva Chair in Statistics, University of East Anglia. Previously she has been Director of the Statistical Advisory Service at Imperial College London (2004–2010). She is also a Visiting Professor at The Center for Lifespan and Chronic Illness Research (CLiCIR), University of Hertfordshire. She has a long standing interest in statistical evidence and its applications in meta-analysis. She has authored and co-authored 78 papers, including numerous theoretical and applied papers on meta-analysis, and a recent book on meta analysis (*Meta-analysis: A Guide to Calibrating and Combining Statistical Evidence*, Wiley, 2008) co-authored with Stephan Morgenthaler and R.G. Staudte and dedicated to a new approach based on variance stabilization.

Dr. Stephan Morgenthaler is Professor of Applied Statistics in the Institute of Mathematics Ecole Polytechnique Fédérale de Lausanne in Switzerland. He has authored, co-authored and edited more than 80 papers and eight books. He is a member of the ISI and a Fellow of the American Statistical Association. He served as a vice-president of ISI from 2004 to 2008.

Dr. Robert G. Staudte is Professor and Head, Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia. He has authored and co-authored more than 50 papers and four books, including *Robust Estimation and Testing*, Wiley 1990, co-authored with Professor Simon J. Sheather; and *Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence*, Wiley 2008, co-authored with Professors Elena Kulinskaya and Stephan Morgenthaler. He was Associate Editor of the *Journal of Statistical Planning and Inference* (1995–1998).

Cross References

►Clinical Trials: Some Aspects of Public Interest
►Effect Size
►Forecasting Principles
►Medical Statistics
►Psychology, Statistics in
►P-Values, Combining of
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Albert X, Huertas I, Pereiró I, Sanfelix J, Gosalbes V, Perrota C (2004) Antibiotics for preventing recurrent urinary tract infection in non-pregnant women (Cochran Review). In: The Cochran Library, Issue 3. Wiley, Chichester, UK

Biggerstaff BJ, Tweedie RL (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Statistics in Medicine 16:753–768

Böhning D, Kuhnert R, Rattanasiri S (2008) Meta-analysis of Binary data using profile likelihood. Chapman and Hall/CRC Statistics. CRC, Boca Raton, FL
Borenstern M, Hedges LV, Higgins JPT, Rothstein H (2009) Introduction to meta analysis. Wiley, London
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Earlbaum Associates, Hillsdale, NJ
Cooper H, Hedges LV (eds) (1994) The handbook of research synthesis. Russell Sage Foundation, New York
DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. control Clin Trials 7:177–188
Glass GV (1976) Primary, secondary and meta-analysis of research. Educ Res 5:3–8
Hartung J, Knapp G, Sinha BK (2008) Statistical meta analysis with applications. Wiley, Chichester
Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic, Orlando
Higgins JPT, Green S (eds) (2008) Cochrane handbook for systematic review of interventions version 5.0.1. The Cochrane Collaboration: available on www.cochrane-handbook.org
Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis: a guide to calibrating and combining statistical evidence. Wiley, Chichester
Malzahn U, Bohning D, Holling H (2000) Nonparametric estimation of heterogeneity variance for the standardized difference used in meta-analysis. Biometrika 87(3):619–632
Viechtbauer W (2007) Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol 60:29–60
Whitehead A (2002) Meta-analysis of controlled clinical trials. Applied statistics. Wiley, Chichester

Method Comparison Studies

Graham Dunn
Professor of Biomedical Statistics and Head of the Health Methodology Research Group
University of Manchester, Manchester, UK

We are here concerned with the comparison of the performance to two or more measurement devices or procedures. At its simplest, a method comparison study involves the measurement of a given characteristic on a sample of subjects or specimens by two different methods. One possible question is then whether measurements taken by the two different methods are interchangeable. Another is whether one of the two methods is more or less precise than the other. A third, more difficult task, is to calibrate one set of fallible measurements (using Device A, for example) against another set of fallible measurements produced by device B. A potentially-serious problem in all of these situations is the possibility that the measurement errors arising from the use of these two devices may be correlated. A slightly more complicated study involves replication of each of the sets of measurements taken using the two different procedures or devices, usually carried out on the naïve assumption that the measurement errors of the within-device replicates will be uncorrelated and that replication will enable the investigator to obtain an unbiased estimate of the instruments' precisions (based on the standard deviations of the replicates).

Let's return to the simplest situation – measurement of a given characteristic on a sample of subjects by two different methods that are assumed to provide independent measurement errors. Are the two methods interchangeable? How closely do the measurements agree with each other? Is this agreement good enough for all our practical purposes? A method suggested by Bland and Altman (1986) is to determine *limits of agreement*. One simply subtracts the measurement arising from one method from the corresponding measurement using the other. The average of these differences tells us about the possibility of relative bias (and the so-called Bland-Altman plot – a graph of the difference against the average of the two measurements – may tell us that the bias is changing with the amount of the characteristic being measured, but it is not 100% fool-proof since a relationship between the difference between and the average of the two measures may arise from differences in the instruments' precisions). The standard deviation of the differences tells us about the variability of the difference of the two measurement errors. The 95% limits of agreement are simply defined as the range of differences between the 2.5th and 97.5th percentiles or, assuming normality, approximately two standard deviations either side of the mean. If the measurement errors for the two methods are positively correlated then the variability of the differences will be less than one would expect if they were uncorrelated and the limits of agreement will be too small. If the measurement methods use different scales (comparison of temperatures in °C and °F, for example) then this simple procedure will break down and the limits of agreement will fail to tell the investigator that the two methods are interchangeable (after suitable rescaling).

One might be tempted to plot results using one of the methods (in °F, for example) against the other (in °C) and carry out a simple regression to calibrate one against the other. But the hitch is that both methods are subject to error (the classical errors-in-variables problem) and the estimate of the regression coefficient would be biased (attenuated towards zero). If one knows the ratio of the variances of the measurement errors for the two methods then it is possible to use orthogonal regression, widely-known as Deming's regression, to solve the problem. The

catch is that one does not normally have an unbiased estimate of the ratio of these two variances – the problem again arising from the lack of independence (i.e., correlation) of any replicate measures used to determine these variances (Carroll and Ruppert 1996).

A third relatively simple approach is to look for and make use of an instrumental variable (IV) through IV or ►two-stage least squares (2SLS) regression methods. Here we need a variable (not necessarily a third measurement of the characteristic, but it may be) that is reasonably highly correlated with the characteristic being measured but can be justifiably assumed to be uncorrelated with the associated measurement errors. If we label the measurements using the two methods as X and Y, and the corresponding values of the instrumental variable as Z, then the instrumental variable estimator of the slope of Y on X is given by the ratio Cov(Y,Z)/Cov(X,Z) – see Dunn (2004, 2007). From here it's a relatively simple move into factor analysis models for data arising from the comparison of three or methods (Dunn 2004).

Statistical analyses for the data arising from more the informative designs, with more realistic measurement models (heteroscedasticity of measurement errors, for example), is beyond the scope of this article but the methods are described in considerable detail in Dunn (2004). The methods typically involve software developed for covariance structure modelling. Analogous methods for the comparison of binary measurements (diagnostic tests) can also be found in Dunn (2004).

About the Author

For biography *see* the entry ►Psychiatry, Statistics in.

Cross References

►Calibration
►Instrumental Variables
►Measurement Error Models
►Two-Stage Least Squares

References and Further Reading

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310

Carroll RJ, Ruppert D (1996) The use and misuse of orthogonal regression in linear errors-in-variables models. Am Stat 50:1–6

Dunn G (2004) Statistical evaluation of measurement errors. Arnold, London

Dunn G (2007) Regression models for method comparison data. J Biopharm Stat 17:739–756

Methods of Moments Estimation

Martin L. Hazelton
Chair of Statistics
Massey University, Palmerston North, New Zealand

The method of moments is a technique for estimating the parameters of a statistical model. It works by finding values of the parameters that result in a match between the sample moments and the population moments (as implied by the model). This methodology can be traced back to Pearson (1894) who used it to fit a simple mixture model. It is sometimes regarded as a poor cousin of maximum likelihood estimation since the latter has superior theoretical properties in many settings. Nonetheless, the method of moments and generalizations thereof continue to be of use in practice for certain (challenging) types of estimation problem because of their conceptual and computational simplicity.

Consider a statistical model defined in terms of a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathsf{T}}$. We denote by $\mu_k = \mathsf{E}[X^k]$ the kth moment about zero of a random variable X generated by our model. This moment will be a function of $\boldsymbol{\theta}$, and so we will write $\mu_k = \mu_k(\boldsymbol{\theta})$ to emphasize this dependence.

Suppose that we have a (univariate) random sample $X_1, \ldots, X_n$ from the model, which we want to use to estimate the components of $\boldsymbol{\theta}$. From this we can compute the kth sample moment, $\hat{\mu}_k = n^{-1} \sum_{i=1}^{n} X_i^k$. The rationale for the method of moments is that the sample moments are natural estimators of the corresponding model-based moments, and so a good estimate of $\boldsymbol{\theta}$ will reproduce these observed moments. In practice it is usual (although not essential) to use moments of the lowest possible orders in order to obtain parameter estimates. The method of moments estimator $\hat{\boldsymbol{\theta}}$ is hence defined to be the solution of the system of equations

$$\mu_k(\boldsymbol{\theta}) = \hat{\mu}_k \qquad k = 1, 2, \ldots, q$$

where q is the smallest integer for which this system has a unique solution.

As an example, suppose that $X_1, \ldots, X_n$ are drawn from a ►gamma distribution with shape parameter α and scale parameter β. Then $\mu_1 = \alpha\beta$ and $\mu_2 = \alpha(\alpha + 1)\beta^2$. The method of moments estimators $\hat{\alpha}$ and $\hat{\beta}$ therefore satisfy the pair of equations

$$\hat{\alpha}\hat{\beta} = \hat{\mu}_1$$
$$\hat{\alpha}(\hat{\alpha} + 1)\hat{\beta}^2 = \hat{\mu}_2.$$

Solving these we obtain

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} \quad \text{and} \quad \hat{\beta} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1}.$$

Method of moments estimators are, in general, consistent. To see this, note that the (weak) law of large numbers ensures that the sample moments converge in probability to their population counterparts. It then follows that if $\mu_k(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ for $k = 1, \ldots, q$ then the method of moments estimators will converge in probability to their true values. However, method of moments estimators are less efficient than maximum likelihood estimators, at least in cases where standard regularity conditions hold and the two estimators differ. Furthermore, unlike maximum likelihood estimation, the method of moments can produce infeasible parameter estimates in practice. For example, if $X_1, \ldots, X_n$ are drawn from a uniform distribution (see ►Uniform Distribution in Statistics) on $[0, \theta]$ then the method of moments estimator is $\hat{\theta} = 2\bar{X}$, but this estimate is infeasible if $\max\{X_i\} > 2\bar{X}$.

Despite the theoretical advantages of maximum likelihood estimation, the method of moments remains an important tool in many practical situations. One reason for this is that method of moments estimates are straightforward to compute, which is not always the case for maximum likelihood estimates. (For example, the maximum likelihood estimators for the gamma distribution parameters considered above are only available implicitly as the solution to the non-linear likelihood equations.) Furthermore, estimation by the method of moments does not require knowledge of the full data generating process. This has led to various extensions of the basic method of moments that can be applied in complex modeling situations.

One such extension is the generalized method of moments Hansen (1982) which is a type of generalized estimating equation methodology, widely used in econometrics. This technique works by utilizing sample and population moment conditions (or "orthogonality conditions") of the statistical model, and can provide estimates of parameters of interest in a model even when other model parameters remain unspecified. Another useful extension is the simulated method of moments (e.g., Gelman 1995). This technique can be employed when the model is so complex that neither the density function for the data nor the theoretical moments are available in closed form. It therefore provides a means of fitting micro-simulation and mechanistic stochastic models (Diggle and Gratton 1984).

About the Author

Professor Hazelton was appointed to the Chair of Statistics at Massey University in 2006. His current research interests include modeling and inference for transport networks, and multivariate smoothing problems. Professor Hazelton is an Associate Editor of the *Journal of the Korean Statistical Society* and a member of the Editorial Advisory Board for *Transportation Research Part B*.

Cross References

►Estimation
►Estimation: An Overview
►Social Network Analysis
►Statistical Inference for Stochastic Processes
►Statistics of Extremes
►Univariate Discrete Distributions: An Overview

References and Further Reading

Diggle P, Gratton J (1984) Monte Carlo methods of inference for implicit statistical models. J R Stat Soc B 46:193–227
Gelman A (1995) Method of moments using Monte Carlo simulation. J Comput Graph Stat 3:36–54
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054
Pearson K (1894) Contribution to the mathematical theory of evolution. Philos Tr R Soc S-A 185:71–110

M

Minimum Variance Unbiased

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

The term *minimum variance unbiased* refers to a property of statistical decision rules.

Idea. Any statistical experiment may be perceived as a random channel transforming a deterministic quantity θ (parameter) into a random quantity X (observation). *Point estimation* is a reverse process of regaining θ from X according to a rule $\widehat{\theta} = \delta(X)$ called *estimator*. Formally, estimator is a function from the set $\mathcal{X}$, of possible values of X, into the set Θ, of possible values of θ. As a measure of imprecision of such estimator one can use the function $R_\delta(\theta) = E_\theta(\delta(X) - \theta)^2$ called the Mean Squared Error. It may be rewritten in the form

$$var_\theta \delta(X) + [b(\theta)]^2, \text{ where } b(\theta) = E_\theta \delta(X) - \theta$$

is the bias of δ.

If $b(\theta) = 0$ for all θ then $\widehat{\theta} = \delta(X)$ is said to be *unbiased*. Minimizing the MSE among the unbiased estimators reduces to minimizing its variance. Any estimator δ_0 realizing this minimum (if such exists) is said to be a *minimum variance unbiased estimator* (MVUE). Searching for such estimator or verifying whether it is a MVUE needs some special statistical tools.

Example 1 (Urn problem). An urn contains N balls, where any ball is black or white, while the number θ of black balls is unknown. To search θ we draw without replacement n balls. Let k be the number of black balls in the sample. Estimate θ.

A potential number X of black balls in the sample has the hypergeometric distribution (see ►Hypergeometric Distribution and Its Application in Statistics) taking values k with probabilities

$$P_\theta(X = k) = p_{\theta,k} = \begin{cases} \frac{\binom{\theta}{k}\binom{N-\theta}{n-k}}{\binom{N}{n}} & \text{if } k \in [\max(0, n-N+\theta), \\ & \quad \min(n, \theta)] \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Since $EX = \frac{n\theta}{N}$, the rule $\widehat{\theta} = \frac{N}{n}X$ is an unbiased estimator of θ. This is, formally, not acceptable unless n is a divisor of N, because $\widehat{\theta}$ takes values outside the parameter set. Thus one can seek for an acceptable unbiased estimator. According to the formula (1) we get

$$p_{0,k} = \begin{cases} 1, & \text{if } k = 0 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$p_{1,k} = \begin{cases} \frac{N-n}{N}, & \text{if } k = 0 \\ \frac{n}{N}, & \text{if } k = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus any unbiased estimator $\widehat{\theta} = \widehat{\theta}(X)$ must satisfy the conditions $\widehat{\theta}(X) = 0$ if $X = 0$ and $\frac{N}{n}$ if $X = 1$. Therefore the desired estimator exists if and only if n is a divisor on N.

Basic Concepts. Let $X = (X_1, \ldots, X_n)$ be a random vector, interpreted as a potential observation in a statistical experiment. Assume that distribution P of the vector belongs to a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is an unknown parameter identifying P. Thereafter by distribution we shall mean density or probability mass function. Any potential estimator of θ is a function $T = t(X)$ called a statistic. If T involves the entire information on θ then one can reduce the problem by considering only these estimators which depends on X through T.

We say that a statistic T is *sufficient* for θ if the conditional probability $P_\theta(X/T)$ does not depend on θ. Determining a sufficient statistic directly from this definition may be a laborious task. It may be simplified by the well known Fisher-Neyman factorization criterion. A statistic $T = t(X)$ is sufficient for θ, if and only if, P_θ may be presented in the form $P_\theta(x) = g_\theta[t(x)]h(\theta)$. A sufficient statistic T is minimal if it is a function of any other sufficient statistic. In particular, the vector statistic $T = [t_1(X), \ldots, t_k(X)]$ in so called exponential family $P_\theta(x) = C(\theta)\exp\left[\sum_{j=1}^{k} Q_j(\theta)t_j(x)\right]h(x)$, for $\theta \in \Theta$, is sufficient.

We say that a statistic T is *complete* if for any (measurable) function f the condition $E_\theta f(T) = 0$ for all θ implies that $P[f(T) = 0] = 1$. It is known that any complete sufficient statistic (if exists) is minimal but a minimal sufficient statistic may not be complete. Moreover the above sufficient statistic in the exponential family distributions is complete providing Θ contains a k-dimensional rectangle.

Now let us consider a family of densities $\{p(x, \theta) : \theta \in \Theta\}$, where Θ is an open interval of a real line, satisfying some regularity conditions. Function $I = I(\theta)$ defined by the formula $I(\theta) = E\left[\frac{\partial \log p(X,\theta)}{\partial \theta}\right]^2$ is said to be Fisher information.

Advanced Tools. Let $X = (X_1, \ldots, X_n)$ be a random vector with a distribution P belonging to a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let $T = t(X)$ be a sufficient statistic for θ. In searching MVUE's one can use the following results.

►*Rao-Blackwell theorem*: If $U = u(X)$ is an unbiased estimator of a parametric function $g(\theta)$ then the conditional expectation $E[U/T]$ is also unbiased and its variance is not greater than $var(U)$.

Lehmann-Scheffé theorem: If T is, moreover, complete then any statistic $h(T)$ is a MVUE of its expectation. This MVUE is unique (with probability 1).

Rao-Cramer inequality: Let $\{p(x, \theta) : \theta \in \Theta\}$, where Θ is an open interval of a real line, be a family of densities satisfying some regularity conditions, such that $I(\theta) > 0$ for all θ. Then for any statistic $U = u(X)$ the inequality $var_\theta(U) \geq \frac{1}{I(\theta)}$ is met.

It is worth to add that the equality in the Rao-Cramer inequality is attained if and only if the family $\mathcal{P}$ of distributions is exponential. However this condition is not necessary for existing a MVUE; for instance, if $X_1, \ldots, X_n$ are i.i.d. according to the normal law $N\left(\alpha^{\frac{1}{3}}, 1\right)$. In this case the attainable minimum variance is $\frac{9\alpha^4}{n} + \frac{18\alpha^2}{n^2} + \frac{6}{n^3}$ while $\frac{1}{I(\theta)} = \frac{9\alpha^4}{n}$.

Example 2 (Bernoulli trials). Let $X_1, \ldots, X_n$ be independent and identically distributed zero-one distributions with probability $P(X_i = 1) = \theta$, where θ is unknown for $i = 1, \ldots, n$. In this case the family $\mathcal{P} = \{P_\theta : \theta \in (0,1)\}$ is exponential with complete sufficient statistic $\overline{X} = \frac{1}{n}\sum_i X_i$. Since $E\overline{X} = \theta$, the statistic $\overline{X}$ is the unique MVUE of θ. In this case the Fisher information takes the form $I(\theta) = \frac{n}{\theta(1-\theta)}$ while $var_\theta(\overline{X}) = \frac{\theta(1-\theta)}{n}$. Thus the lower bound $\frac{1}{I(\theta)}$ in the Rao-Cramer inequality is attained. It is worth to note that, similarly as in Example 1, this unique MVUE takes, with positive probability, the values 0 and 1, which lie outside the parameter set $(0,1)$.

Minimum Variance Invariant Unbiased Estimator. If distribution of the observation vector depends on several parameters, some of them may be out of our interest and play the role of nuisance parameters. Such a situation occurs, for instance, in linear models. In this case the class of all unbiased estimators is usually too large for handle. Then we may seek for an estimator which is invariant with respect to a class of transformations of observations or its variance does not depend on the nuisance parameters. An estimator minimizing variance in such a reduced class is called a minimum variance invariant unbiased estimator.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Cramér–Rao Inequality
►Estimation
►Properties of Estimators
►Rao–Blackwell Theorem
►Sufficient Statistics
►Unbiased Estimators and Their Applications

References and Further Reading

Cramér H (1946) Mathematical methods of statistics, Princeton University Press, Princeton, NJ
Kadec MN (1979) Sufficient statistic. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 2. Soviet Encyclopedia, Moscow, pp 375–377 (in Russian)
Nikulin MS (1984) Rao-Cramer inequality. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 4, Soviet Encyclopedia, Moscow, pp 867–868, (in Russian)
Nikulin MS (1993) Unbiased estimator. In: Hazewinkel M (ed) Encyclopaedia of mathematics. vol 9, pp 305–307
Lehmann EL (1983) Theory of point estimation. Wiley, New York
Rao CR (1973) Linear statistical inference, 2nd edn. Wiley, New York

Misuse and Misunderstandings of Statistics

Atsu S. S. Dorvlo
Professor
Sultan Qaboos University, Muscat, Sultanate of Oman

Introduction

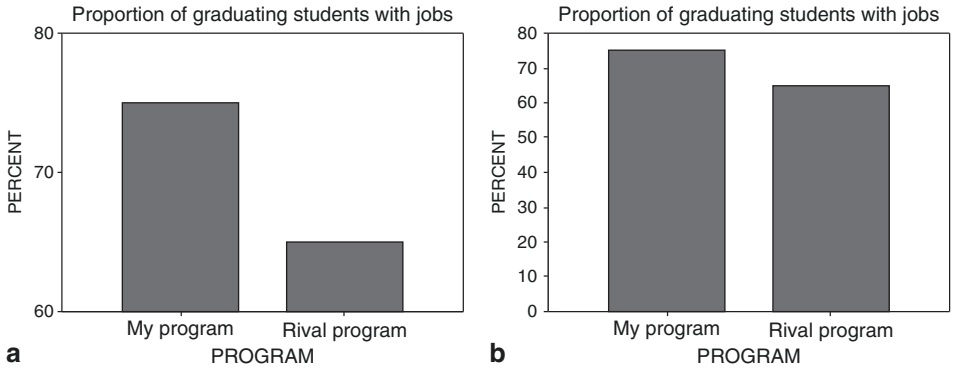
Because of the advent of high speed computers statistics has become more visible. Almost any discipline has an element of statistics in it. In fact one cannot publish in most journals when the statistics used or misused is not stated. Newspapers, magazines, etc are now awash with one form or other of "statistics". Now it is fashionable to take data, shove it into a computer and come out with nice tables, graphs and ►p-values. Clearly such practices are a gross ►misuse of statistics and do a disservice to the subject. There is no wonder we are in the company of "lies, damned lies and statistics."

So What Is Statistics?

There are several definitions of statistics, some not so flattering:

1. The American heritage dictionary says: Statistics is the mathematics of collection, organization and interpretation of numerical data.
2. Brase and Brase, in their beginning level statistics textbook define statistics as the science of how to collect, organize, analyze and interpret numerical information from data.
3. Evan Esar says statistics is the only science that enables different experts using the same figures to draw different conclusions.

The first two capture the essence of statistics. Ms. Esar captures the abuse that is possible. However, these definitions do not capture the true essence of statistics and that is: to make a deduction in the face of uncertainty. The true essence of statistics is captured when it is stated that statistics is the science that tells whether something we observe can be generalized or applied to a new or different but similar situation (the author of this statement is unknown). That is I observe a group of people in a community and found that 20% have cancer, can I generalized to say that the cancer rate in that community is 20%? Of course not without first saying how the sample was observed. The other definitions come into play then. I need to know how the data was collected/observed, how it was organized, analyzed, and then the interpretation.



In this author's opinion most of the problems, misunderstandings and misrepresentations in statistics originate from the observation – collection process. Invariably the data is observed/collected before thought is put in what to do with it. So therefore the inference which is finally made does not take account of how the data was observed in the first. Maybe in the everyday sense it is natural to observe first and then ask what to do with the data observed. However in complex tasks the research questions need to be asked first. Then thought put into how to collect the relevant data, organize and analyze it and make the inference supporting the research question or refuting it. Hence in large scale work, effort should be put in the "how to collect" the data stage. If this is done, only the relevant data will be collected, and there will be savings on resources, time and money.

In most instances the way data is collected, the data type collected determines the types of analysis that can be carried out. Data collection is an expensive, time consuming activity. It is unfortunate that lots of time and effort are wasted on collecting data only to find out that the data is not useful or the exercise could have been done in an easier and cheaper manner. Should 50 experiments be performed or can 10 be sufficient? Unfortunately more data does not necessarily equate to more valid or better results. In fact the opposite could be the case. Hence the design of the experiment or data collection, the estimation of the necessary sample sizes taking into consideration the error, precision and last but not least the use to which the results will be put, such as, will the results be generalized, should be well thought out at the very beginning of the study.

Another area where statistics has a bad name is the pictorial representation of results. The saying goes that "a picture is worth a thousand words." Simple clear graphs can help bring out the important aspects of the study. However there is room for abuse. More often than not attention is not paid to the scale of the graph. For example in comparing two teaching programs, what impression is graph (a) conveying? Are our students actually better? It is the duty of statisticians to point out at every opportunity the pitfalls that need to be avoided when reading graphs.

With the advent of fast computers computations that were near impossible or would take ages to accomplish a few years ago, now takes only seconds of computer time. Coupled with this is the fact that there are very good and easy to use software. Are computers taking the place of statisticians, especially applied statisticians? There is a lot more to data analysis than calculations. The computer is there to remove the drudgery out of number crunching. What calculations to perform, that is what analysis to do and foremost, the check of the validity of assumption under which the procedures are valid, is the domain of the statistician.

Conclusion

In my view statistics is simply whether one can generalize ones observation to a different or future situation. The difficulty is how the "observation" was obtained – data collection – and the generalization made – summarized, analyzed and interpreted. In all these the expert input of a statistician is invaluable.

Cross References

►Misuse of Statistics

References and Further Reading

Brase C, Brase C (2008) Understandable statistics, 9th edn. Brooks-Cole

Evan Esar (1899–1995) Quotations www.quotationspage.com/quotes or Esar's Comic Dictionary

Misuse of Statistics

Chamont Wang
Professor
The College of New Jersey, Ewing, NJ, USA

Statistics as an academic discipline is widely held as a science that is related to experiments and the quantification of uncertainty. This is true, but if used without caution, statistics can add more uncertainty to an already murky problem. A rich source on this topic would be "*How to Lie with Statistics* Turns Fifty," a 56-page Special Section of *Statistical Science* (2005, p. 205–260).

Misuses of statistics at a non-technical level can be roughly grouped in the following three categories, often with the three types of misuses feeding each other in a complicated, dynamic fashion.

1. Data Quality: A complete statistical project consists of the following components: (a) data collection, (b) data preprocessing, (c) data exploration, (d) data analysis and statistical modeling, and (e) summary report. The process is not entirely linear and often goes from one middle step back to another, and roughly 60–95% of the project effort is needed on data quality to ensure that the entire process will not go off the rails.

   In their 2005 article, "How to Lie with Bad Data," De Veaux and Hand pointed out that "Data can be bad in an infinite variety of ways." This is not an exaggeration. Fortunately, statistical design of experiments and survey methodology, if done right, are capable of producing data with high-quality. In the real world, the problem is that the majority of data are collected in non-controlled environments without much statistical guidance. Consequently, data might have been corrupted, distorted, wrong-headed, ill-defined, and with loads of missing values – the list goes on forever. De Veaux and Hand (2005) provided suggestions on how to detect data errors and how to improve data quality. The suggestions are very useful for practitioners.

   In journals and real-world applications, statistical reports often shine with tremendous amounts of energy on exotic models but with questionable effort (and insufficient details) on data quality. Statistics as a science is supposed to provide a guiding light for research workers and decision-makers. Without good data, exotic statistical models are unlikely to help. The situation is like a person who is nearly blinded by cataracts and tries to sharpen the lenses for better vision. The effort will be futile unless an operation is conducted to take away the clouding.

   A related note on data quality is the ▶outliers and unusual numbers in the data. Resistant and robust statistical procedures are often used to handle this kind of problem. But if the data was not collected in controlled experiments, then the efforts are mostly misguided. Furthermore, outliers often are the most interesting numbers that may reveal surprising features of the study. Blind applications of ▶robust statistics thus can be counterproductive if not altogether misleading.
2. Statistical tests and ▶*p*-values: A continuing source of mistake is the confusing of *statistical significance* with *practical significance*. Mathematically, if the sample size increases indefinitely, then the power of the statistical test will increase as well. Consequently, even a tiny difference between observed and the predicted values can be statistically highly significant. Certain large scale examples regarding the confusion of *practical significance* are discussed in Wang (1993, pp. 1–2, 117–119, 128). Other cautions on the misuse of statistical tests can be found in Freedman et al. (2007) and in the "What Can Go Wrong" sections of De Veaux et al. (2009, pp. 523, 549, 570, 604–605, 634–635, 662–663, 708) which discuss "no peeking at the data" and other caveats on the tests of significance.

   Freedman (2008a) further pointed out a potential problem in research journals when publications are "driven by the search for significance." The problem can be rather acute when research grants or academic careers hinge on publications. In short, researchers may conduct many tests, ignore contradictory results and only submit findings that meet the 5% cutoff. A possibility to deal with this problem, according to Freedman (2008a), is a journal requirement to document search efforts in the research process.
3. Statistical Inference of Cause-and-Effect: Causal inference is a foundation of science and is indeed a very tricky business. As an example, Aristotle maintained that cabbages produce caterpillars daily – a well-known assertion only to be refuted by controlled experiments carried out by Francesco Redi in 1668. For new comers to the field of statistics, it may be baffling that much of the practice of modern statistics is still Aristotelian in nature. For instance, a rough estimate indicates that in clinical research, "80% of observational studies fail to replicate or the initial effects are much smaller on retest" (Young et al. 2009; a la Ioannidis 2005).

Freedman (2008a) further discussed the related controversies and a diverse set of large-scale contradictory studies. The problem should be a concern to the statistical community as our trade is indeed widely used. For example, in the study of coronary heart disease, there are more than 3,600 statistical articles published each year (Ayres 2007, p. 92), and this is only the tip of the iceberg.

A potential problem with statistical causality is the use of regression models, directed graphs, path analysis, structural equations, and other law-like relationships. Take the example of regression; on a two-dimensional scatterplot, it is easy to see that *mathematically* it does not matter whether we put a variable on the left or the right of the equation. Any software package would produce the estimates of the slope and the intercept, plus a host of diagnostic statistics that often says the model is an excellent fit. Compounding the problem of causal inference, a third variable may be the reason behind the phenomenon as displayed by the scatterplot. For instance, a scatterplot can be drawn to show that the incidence of polio (Y-variable) increases when soft-drink sales (X-variable) increases, but in fact a lurking variable (warm weather) is the driving force behind the rise (Freedman et al. 1978, p. 137).

The problem quickly turns worse in higher-dimensional spaces. Try the following example in a regression class: draw 20 or 30 right triangles and then measure the values of (X_1, X_2, Y), with X_1, X_2 being the adjacent sides of the 90° angle. The Pythagorean Theorem says that $Y = \sqrt{X_1^2 + X_2^2}$. In an experiment (Wang 1993, p. 73–77), students of regression came up with all kinds of equations with R^2 of 96–99.93%. The equations all passed stringent tests of diagnostic statistics, but none of them comes close to the Pythagorean equation. A further twist makes the problem statistically intractable when the legs of the triangles are not orthogonal (Wang 1993, p. 77–78).

For causal inference, the misgivings of statistical models happen not only in the observational studies, but also in the analysis of experimental data. In an in-depth discussion, Freedman (2008b) examined the ▶Kaplan-Meier estimator and proportional-hazards models which are frequently used to analyze data from randomized controlled experiments. Specifically, Freedman investigated journal papers on the efficacy of screening for lung cancer (*New England Journal of Medicine*), the impact of negative religious feelings on survival (*Archives of Internal Medicine*), and the efficacy of hormone replacement therapy (*New England Journal of Medicine* and *Journal of the American Medical Association*). Freedman discussed reverse causation plus a host of other issues such as measurements, omitted variables, and the justification of the models. Freedman concluded that "the models are rarely informative," that "as far as the model is concerned, the ▶randomization is irrelevant," that "randomization does not justify the model," and that it "is a mistake" to apply the models in the first place.

In yet another example, Freedman (2008c) investigated ▶logistic regression in the experimental setting for drawing conclusions on cause-and-effect. Again, Freedman noted that the model is not justified by randomization. He further questioned "Why would the logit specification be correct rather than the probit – or anything else? What justifies the choice of covariates? Why are they exogenous? If the model is wrong, what is $\widehat{\beta}_2$ supposed to be estimating?" Furthermore, in a summary of a vast variety of investigations, Freedman (2008a) concluded that "Experimental data are frequently analyzed through the prism of models. This is a mistake."

Taken together, Freedman et al. (1978, 1991, 1998, 2007), Freedman (2005, 2008a, b, c), Wang (1993, p. 72–79), and a very long list of references all indicate that sophisticated statistical models are often detached from the underlying mechanism that generated the data. In other words, many law-like equations produced by statistical models are as structure-less as *Amoeba Regression* (Wang 1993) and need to be viewed with caution. This is indeed a big disappointment to countless researchers who spend their lives on statistical models (see, e.g., Pearl 2009, p. 100), but this is a truth that we have to face.

Nevertheless, the models should be treasured for a number of reasons. To begin with, recall Newton's theory on celestial mechanics. The story is well-known and is relevant to statistical modeling in the following ways: (1) The Newtonian theory relies on observational studies, yet its prediction accuracy rivals most of the tightly controlled experiments. In other words, there is nothing wrong with observational studies, as long as they are accurate and they are consistent in subsequent studies. (2) Statistical models represent the intellectual accomplishment of the statistical community that may one day produce useful results on both experimental data and observational studies. History is the witness that ivory tower research often produces surprising results decades or hundreds of years later. And when the model is correct, the consequences can be enormous. Take the example of proportional-hazards model,

even Freedman (2008b, p. 116) acknowledged that "Precise measures of the covariates are not essential" and that if the model "is right or close to right, it works pretty well." (3) If used for descriptive or exploratory purposes, fancy statistical models may indeed reveal unexpected features in the data. For certain examples on non-parametric structural equations and counterfactual analysis, see references in Pearl (2009). For another example on hot spot detection, see Wang et al. (2008).

As a matter of fact, in the past 15 years or so, statistical models have taken a new life in the realm of ►data mining, predictive modeling, and statistical learning (see, e.g., Wang et al. 2008). In these applications, the concerns are not cause-and-effect or the specific mechanism that generates the data. Instead, the focus is the prediction accuracy that can be measured by profit, false positive, false negative, and by other criteria to assess the model utility. This is a sharp departure from causation to prediction. The great news is that the new applications have been ranked by the 2001 *MIT Technology Review* as one of the ten emerging technologies that will change the world – and it is arguable that the successes of this new technology will eventually feedback to traditional statistics for other breakthroughs. In fact, countless examples with ingenious twists have already happened (see, e.g., Ayres 2007). It is a triumph of statistical models.

A cautionary note is that statistical learning and the new breed of predictive modeling can easily go wrong and misinformation can propagate with unprecedented speed in the modern age of internet blogging and social networks. Newcomers to the field should consult, for examples, "Top 10 Data Mining Mistakes" (Elder 2009) and "Myths and Pitfalls of Data Mining" (Khabaza 2009). For unsupervised learning, one may want to read "The Practice of Cluster Analysis" (Kettenring, 2006) and "A Perspective on Cluster Analysis" (Kettenring 2008). For supervised learning, given a dozen or thousands of predictors, statistical tools are frequently used to generate predictor importance scores, but these scores are often wildly different from one algorithm to the next (see e.g., Wang et al. 2008, Sect. 4).

For yet another example, a model such as a Neural Network may produce higher profit and higher prediction accuracy than other tools, yet the model may also be more volatile in repeated uses and hence pose considerable hazards in the long run. ►Sensitivity analysis and similar techniques are thus needed to prevent misleading conclusions (see, e.g., Wang et al. 2009).

The hallmark of empirical science is its replicability. Much of the current statistical practice, unfortunately, does not really meet this criterion. Just look at how many authors are unwilling to disclose their data and how many journals are unwilling to archive the datasets and the code (see also Freedman, 2008a, c). Exceptions include *American Economic Review, American Economic Journals* and *Science*.

Data disclosure reduces the cost of research and cost of replicating results. It also deters unprofessional conduct and improves collective findings of the research community. Certain online journals (see e.g., http://www.bentley.edu/csbigs/csbigs-v1-n1.cfm) post both the research article and the data side-by-side. If more journals are willing to make available the datasets used in their publications, the situation of misuse and misconduct of statistics will be greatly improved.

About the Author

Dr. Chamont Wang received the Ph.D. degree in Statistics from Michigan State University, East Lansing (1983). He is Full Professor at the Department of Mathematics and Statistics, the College of New Jersey, serving as an Associate Editor of a research journal, CSBIGS (*Case Studies in Business, Industry and Government Statistics*), serving as an expert witness of a premier expert witness referral firm. He is author of the book, *Sense and nonsense of statistical inference: controversy, misuse, and subtlety* (Taylor and Francis, 1993), and also of journal papers in the field of Chaos and Dynamical Systems. He is a member of American Statistical Association, the Mathematical Association of America, and the Institute of Mathematical Statistics.

Cross References

►Discriminant Analysis: Issues and Problems
►Economic Growth and Well-Being: Statistical Perspective
►Fraud in Statistics
►Misuse and Misunderstandings of Statistics
►Role of Statistics
►Significance Tests: A Critique
►Statistical Fallacies
►Statistical Fallacies: Misconceptions, and Myths
►Statistics and the Law
►Statistics: Controversies in Practice

References and Further Reading

Ayres I (2007) Super crunchers: why thinking-by-numbers is the new way to be smart. Bantom, New York
De Veaux R, Hand D (2005) How to lie with bad data. Stat Sci 20(3):231–238
De Veaux R, Velleman P, Bock D (2009) Intro Stats, 3rd edn. Pearson
Elder JF IV (2009) Top 10 data mining mistakes. Handbook of statistical analysis and data mining applications, Elsevier, pp 733–754

Freedman D (2005) Statistical models: theory and practice. Cambridge University Press, Cambridge
Freedman DA (2008a) Oasis or mirage? Chance 21(1):59–61
Freedman DA (2008b) Survival analysis: a primer. Am Stat 62(2):110–119
Freedman DA (2008c) Randomization does not justify logistic regression. Stat Sci 23(2):237–249
Freedman DA, Pisani R, Purves R (1978, 1991, 1998, 2007) Statistics. W.W. Norton, USA
Ioannidis J (2005) Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. J Am Med Assoc 294:218–228
Kettenring JR (2006) The Practice of Cluster Analysis. Journal of Classif 23(1):3–30
Kettenring JR (2008) A Perspective on Cluster Analysis. Stat Anal Data Mining 1(1):52–53
Khabaza T (2009) Hard hat area: myths and pitfalls of data mining. An *SPSS* Executive Brief, http://viewer.bitpipe.com/viewer/viewDocument.do?accessId=10318929
Pearl J (2009) Causal inference in statistics: an overview. Stat Surv 3:96–146, http://www.i-journals.org/ss/
Wang C (1993) Sense and nonsense of statistical inference: controversy, misuse, and subtlety. Marcel Dekker, Inc., New York
Wang C, Liu B (2008) Data mining for large datasets and hotspot detection in an urban development project. J Data Sci 6(3): 389–414. http://proj1.sinica.edu.tw/~jds/JDS-501.pdf
Wang C, Zhuravlev M (2009) An analysis of profit and customer satisfaction in consumer finance. Case Stud Bus Indus Govern Stat 2(2):147–156, http://www.bentley.edu/csbigs/documents/Wang.pdf
Young SS, Bang H, Oktay K (2009) Cereal-induced gender selection? Most likely a multiple testing false positive. Proc R Soc B 276:1211–1212

Mixed Membership Models

Elena A. Erosheva[1], Stephen E. Fienberg[2]
[1]Associate Professor
University of Washington, Seattle, WA, USA
[2]Maurice Falk University Professor
Carnegie Mellon University, Pittsburgh, PA, USA

The notion of mixed membership arises naturally in the context of multivariate data analysis (see ▶Multivariate Data Analysis: An Overview) when attributes collected on individuals or objects originate from a mixture of different categories or components. Consider, for example, an individual with both European and Asian ancestry whose mixed origins correspond to a statement of mixed membership: "1/4 European and 3/4 Asian ancestry." This description is conceptually very different from a probability statement of "25% chance of being European and 75% chance of being Asian". The assumption that individuals or objects may combine attributes from several basis categories in a stochastic manner, according to their proportions of membership in each category, is a distinctive feature of mixed membership models. In most applications, the number and the nature of the basis categories, as well as individual membership frequencies, are typically considered latent or unknown. Mixed membership models are closely related to latent class and finite ▶mixture models in general. Variants of these models have recently gained popularity in many fields, from genetics to computer science.

Early Developments

Mixed membership models arose independently in at least three different substantive areas: medical diagnosis and health, genetics, and computer science. Woodbury et al. (1978) proposed one of the earliest mixed membership models in the context of disease classification, known as the *Grade of Membership* or GoM model. The work of Woodbury and colleagues on the GoM model is summarized in the volume *Statistical Applications Using Fuzzy Sets* (Manton et al. 1994).

Pritchard et al. (2000) introduced a variant of the mixed membership model which became known in genetics as the *admixture model* for multilocus genotype data and produced remarkable results in a number of applications. For example, in a study of human population structure, Rosenberg et al. (2002) used admixture models to analyze genotypes from 377 autosomal microsatellite loci in 1,056 individuals from 52 populations. Findings from this analysis indicated a typology structure that was very close to the "traditional" five main racial groups.

Among the first mixed membership models developed in computer science and machine learning for analyzing words in text documents were a multivariate analysis method named Probabilistic Latent Semantic Analysis (Hofmann 2001) and its random effects extension by Blei et al. (2003a, b). The latter model became known as *Latent Dirichlet Allocation* (LDA) due to the imposed Dirichlet distribution assumption for the mixture proportions. Variants of LDA model in computer science are often referred to as *unsupervised generative topic models*. Blei et al. (2003a, b) and Barnard et al. (2003) used LDA to combine different sources of information in the context of analyzing complex documents that included words in main text, photographic images, and image annotations. Erosheva et al. (2004) analyzed words in abstracts and references in bibliographies from a set of research reports published in the *Proceeding of the National Academy of Sciences* (PNAS), exploring

an internal mixed membership structure of articles and comparing it with the formal PNAS disciplinary classifications. Blei and Lafferty (2007) developed another mixed membership model replacing the Dirichlet assumption with a more flexible logistic normal distribution for the mixture proportions. Mixed membership developments in machine learning have spurred a number of applications and further developments of this class of models in psychology and cognitive sciences where they became known as *topic models* for semantic representations (Griffiths et al. 2007).

Basic Structure

The basic structure of a mixed membership model follows from the specification of assumptions at the population, individual, and latent variable levels, and the choice of a sampling scheme for generating individual attributes (Erosheva et al. 2004). Variations in these assumptions can provide us with different mixed membership models, including the GoM, admixture, and generative topic models referred to above.

Assume K basis subpopulations. For each subpopulation $k = 1, \ldots, K$, specify $f(x_j|\theta_{kj})$, a probability distribution for attribute x_j, conditional on a vector of parameters θ_{kj}. Denote individual-level membership score vector by $\lambda = (\lambda_1, \ldots, \lambda_K)$, representing the mixture proportions in each subpopulation. Given λ, the subject-specific conditional distribution for *jth* attribute is

$$Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj}).$$

In addition, assume that attributes x_j are independent, conditional on membership scores. Assume membership scores, the latent variables, are random realizations from some underlying distribution D_α, parameterized by α. Finally, specify a sampling scheme by picking the number of observed distinct attributes, J, and the number of independent replications for each attribute, R.

Combining these assumptions, the marginal probability of observed responses $\left\{x_1^{(r)}, \ldots, x_J^{(r)}\right\}_{r=1}^{R}$, given model parameters α and $\boldsymbol{\theta}$, is

$$Pr\left(\left\{x_1^{(r)}, \ldots, x_J^{(r)}\right\}_{r=1}^{R} | \alpha, \boldsymbol{\theta}\right) = \int \left(\prod_{j=1}^{J}\prod_{r=1}^{R}\sum_{k=1}^{K}\lambda_k f\left(x_j^{(r)}|\theta_{kj}\right)\right) dD_\alpha(\lambda). \quad (1)$$

In general, the number of observed attributes need not be the same across subjects, and the number of replications need not be the same across attributes. In addition, instead of placing a probability distribution on membership scores, some mixed membership model variants may treat latent variables as fixed but unknown constants. Finally, other extensions can be developed by specifying further dependence structures among sampled individuals or attributes that may be driven by particular data forms as, e.g., in relational or network data (Airoldi et al. 2008b; Chang and Blei 2010; Xing et al. 2010).

Estimation

A number of estimation methods have been developed for mixed membership models that are, broadly speaking, of two types: those that treat membership scores as fixed and those that treat them as random. The first group includes the numerical methods introduced by Hofmann (2001), and joint maximum likelihood type methods described in Manton et al. (1994) and Cooil and Varki (2003), and related likelihood approaches in Potthoff et al. (2000) and Varki et al. (2000). The statistical properties of the estimators in these approaches, such as consistency, identifiability, and uniqueness of solutions, are yet to be fully understood (Haberman 1995) – empirical evidence suggests that the likelihood function is often multi-modal and can have bothersome ridges. The second group uses Bayesian hierarchical structure for direct computation of the posterior distribution, e.g., with Gibbs sampling based on simplified assumptions (Pritchard et al. 2000; Griffiths and Steyvers 2004) or with fully Bayesian MCMC sampling (Erosheva 2003). Variational methods used by Blei et al. (2003a, b), or expectation-propagation methods developed by Minka and Lafferty (2002), can be used to approximate the posterior distribution. The Bayesian hierarchical methods solve some of the statistical and computational problems, and variational methods in particular scale well for higher dimensions. Many other aspects of working with mixed membership models remain as open challenges, e.g., dimensionality selection (Airoldi et al. 2008a).

Relationship to Other Methods of Multivariate Analysis

It is natural to compare mixed membership models with other latent variable methods, and, in particular, with factor analysis and latent class models (Bartholomew and Knott 1999). For example, the GoM model for binary outcomes can be thought of as a constrained factor analysis model: $E(x|\lambda) = A\lambda$, where x is a column-vector of observed attributes $x = (x_1, \ldots, x_J)'$, $\lambda = (\lambda_1, \ldots, \lambda_K)'$ is a column-vector of factor (i.e., membership) scores, and A is

a $J \times K$ matrix of factor loadings. The respective constraints in this factor model are $\lambda' I_K = 1$ and $AI_K = I_K$, where I_K is a K-dimensional vector of 1s.

Mixed membership models can also address objectives similar to those in ►Correspondence Analysis and Multidimensional Scaling methods for contingency tables. Thus, one could create a low-dimensional map from a contingency table data and graphically examine membership scores (representing table rows or individuals) in the convex space defined by basis or extreme profiles (representing columns or attributes) to address questions such as whether some table rows have similar distribution over the table columns categories.

Finally, there is a special relationship between the sets of mixed membership and latent class models, where each set of models can be thought of as a special case of the other. Manton et al. (1994) and Potthoff et al. (2000) described how GoM model can be thought of as an extension of latent class models. On the other hand, Haberman (1995) first pointed out that GoM model can be viewed as a special case of latent class models. The fundamental representation theorem of equivalence between mixed membership and population-level mixture models clarifies this nonintuitive relationship (Erosheva et al. 2007).

About the Authors

Elena Erosheva is a Core member of the Center for Statistics and the Social Sciences, University of Washington.

For biography of Professor Fienberg *see* the entry ►Data Privacy and Confidentiality.

Acknowledgments

Supported in part by National Institutes of Health grant No. R03 AG030605-01 and by National Science Foundation grant DMS-0631589.

Cross References

►Correspondence Analysis
►Factor Analysis and Latent Variable Modelling
►Multidimensional Scaling
►Multivariate Data Analysis: An Overview

References and Further Reading

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008a) Mixed-membership stochastic blockmodels. J Mach Learn Res 9:1981–2014

Airoldi EM, Fienberg SE, Joutard C, Love TM (2008b) Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. In: Poncelet P, Masseglia F, Teisseire M (eds) Data mining patterns: new methods and applications. pp 240–275

Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3: 1107–1135

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis, 2nd edn. Arnold, London

Blei DM, Lafferty JD (2007) A correlated topic model of Science. Ann Appl Stat 1:17–35

Blei DM, Ng AY, Jordan MI (2003a) Latent Dirichlet allocation. J Mach LearnRes 3:993–1002

Blei DM, Ng AY, Jordan MI (2003b) Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 127–134

Chang J, Blei DM (2010) Hierarchical relational models for document networks. Ann Appl Stat 4, pp 124–150

Cooil B, Varki S (2003) Using the conditional Grade-of-Membership model toassess judgement accuracy. Psychometrika 68:453–471

Erosheva EA (2003) Bayesian estimation of the Grade of Membership Model. In: Bernardo J et al (eds) Bayesian statistics 7. Oxford University Press, Oxford, pp 501–510

Erosheva EA, Fienberg SE (2004) Partial membership models with application to disability survey data. In: Weihs C, Caul W (eds) Classification – the ubiquitous challenge. Springer, Heidelberg, pp 11–26

Erosheva EA, Fienberg SE, Lafferty J (2004) Mixed membership models of scientific publications. Proc Natl Acad Sci 101 (suppl 1):5220–5227

Erosheva EA, Fienberg SE, Joutard C (2007) Describing disability through individual-level mixture models for multivariate binary data. Ann Appl Stat 1:502–537

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101 (suppl 1):5228–5235

Griffiths TL, Steyvers M, Tenenbaum JB (2007) Topics in Semantic Representation. Psychol Rev 114(2):211–244

Haberman SJ (1995) Book review of "Statistical applications using fuzzy sets," by K.G. Manton, M.A. Woodbury and H.D. Tolley. J Am Stat Assoc 90:1131–1133

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42:177–196

Manton KG, Woodbury MA, Tolley HD (1994) Statistical applications using fuzzy sets. Wiley, New York

Minka TP, Lafferty JD (2002) Expectation-propagation for the generative aspect model. In: Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI–2002), Morgan Kaufmann, San Francisco, pp 352–359

Potthoff RF, Manton KG, Woodburry MA (2000) Dirichlet generalizations of latent-class models. J Classif 17:315–353

Pritchard P, Stephens JK, Donnely P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Varki S, Cooil B, Rust RT (2000) Modeling fuzzy data in qualitative marketing research. J Market Res 37:480–489

Woodbury MA, Clive J, Garson A (1978) Mathematical typology: a grade of membership technique for obtaining disease definition. Comput Biomed Res 11:277–298

Xing E, Fu W, Song L (2010) A state-space mixed membership blockmodel for dynamic network tomography. Ann Appl Stat 4, in press

Mixture Models

Wilfried Seidel
Professor, President of the German Statistical Society
Helmut-Schmidt-Universität, Hamburg, Germany

Introduction

Mixture distributions are convex combinations of "component" distributions. In statistics, these are standard tools for modeling heterogeneity in the sense that different elements of a sample may belong to different components. However, they may also be used simply as flexible instruments for achieving a good fit to data when standard distributions fail. As good software for fitting mixtures is available, these play an increasingly important role in nearly every field of statistics.

It is convenient to explain finite mixtures (i.e., finite convex combinations) as theoretical models for cluster analysis (see ▶Cluster Analysis: An Introduction), but of course the range of applicability is not at all restricted to the clustering context. Suppose that a feature vector X is observed in a heterogeneous population, which consists of k homogeneous subpopulations, the "components." It is assumed that for $i = 1, \ldots, k$, X is distributed in the i-th component according to a (discrete or continuous) density $f(x, \theta_i)$ (the "component density"), and all component densities belong to a common parametric family $\{f(x, \theta),\ \theta \in \Theta\}$, the "component model." The relative proportion of the i-th component in the whole population is p_i, $p_1 + \cdots + p_k = 1$. Now suppose that an item is drawn randomly from the population. Then it belongs to the i-th component with probability p_i, and the conditional probability that X falls in some set A is $\Pr(X \in A \mid \theta_i)$, calculated from the density $f(x, \theta_i)$. Consequently, the marginal probability is

$$\Pr(X \in A \mid P) = p_1 \Pr(X \in A \mid \theta_1) + \cdots + p_k \Pr(X \in A \mid \theta_k)$$

with density

$$f(x, P) = p_1 f(x, \theta_1) + \cdots + p_k f(x, \theta_k), \tag{1}$$

a "simple finite mixture" with parameter $P = ((p_1, \ldots, p_k), (\theta_1, \ldots, \theta_k))$. The components p_i of P are called "mixing weights," the θ_i "component parameters." For fixed k, let $\mathcal{P}_k$ be the set of all vectors P of this type, with $\theta_i \in \Theta$ and nonnegative mixing weights summing up to one. Then $\mathcal{P}_k$ parameterizes all mixtures with not more than k components. If all mixing weights are positive and component densities are different, then k is the exact number of components. The set of all simple finite mixtures is parameterized by $\mathcal{P}_{\text{fin}}$, the union of all $\mathcal{P}_k$.

This model can be extended in various ways. For example, all component densities may contain additional common parameters (variance parameters, say), they may depend on covariables (mixtures of regression models), and also the mixing weights may depend on covariables. Mixtures of time series models are also considered. Here I shall concentrate on simple mixtures, as all relevant concepts can be explained very easily in this setting. These need not be finite convex combinations; there is an alternative and more general definition of simple mixtures: Observe that the parameter P can be considered as a discrete probability distribution on Θ which assigns probability mass p_i to the parameter θ_i. Then Eq. 1 is an integral with respect to this distribution, and if ξ is an arbitrary probability distribution on Θ, a mixture can be defined by

$$f(x, \xi) = \int_\Theta f(x, \theta)\, d\xi(\theta)\,. \tag{2}$$

It can be considered as the distribution of a two-stage experiment: First, choose a parameter θ according to the distribution ξ, then choose x according to $f(x, \theta)$. Here, ξ is called a "mixing distribution," and mixture models of this type can be parameterized over every set Ξ of probability distributions on Θ.

In statistical applications of mixture models, a nontrivial key issue is identifiability, meaning that different parameters describe different mixtures. In a trivial sense, models parameterized over vectors P are never identifiable: All vectors that correspond to the same probability distribution on Θ describe the same mixture model. For example, any permutation of the sequence of components leaves the mixing distribution unchanged, or components may be added with zero mixing weights. Therefore identifiability can only mean that parameters that correspond to different mixing distributions describe different mixture models. However, also in this sense identifiability is often violated. For example, the mixture of two uniform distributions with supports $[0, 0.5]$ and $[0.5, 1]$ and equal mixing weights is the uniform distribution with support $[0, 1]$. On the other hand, finite mixtures of many standard families (normal, Poisson, ...) are identifiable, see for example Titterington et al. (1985). Identifiability of mixtures of regression models has been treated among others by Hennig (2000). A standard general reference for finite mixture models is McLachlan and Peel (2000).

Statistical Problems

Consider a mixture model with parameter η (vector or probability measure). In the simplest case, one has i.i.d.

data $x_1, \ldots, x_n$ from $f(x, \eta)$, from which one wants to gain information about η. Typical questions are estimation of (parameters of) η, or mixture diagnostics: Is there strong evidence for a mixture (in contrast to homogeneity in the sense that η is concentrated at some single parameter θ)? What is the (minimum) number of mixture components?

A variety of techniques has been developed. The data provide at least implicitly an estimate of the mixture, and Eqs. 1 and 2 show that mixture and mixing distribution are related by a linear (integral) equation. Approximate solution techniques have been applied for obtaining estimators, and moment estimators have been developed on basis of this structure. Distance estimators exhibit nice properties. Traditionally, mixture diagnostics has been handled by graphical methods. More recent approaches for estimation and diagnostics are based on Bayesian or likelihood techniques; likelihood methods will be addressed below. Although Bayesian methods have some advantages over likelihood methods, they are not straightforward (for example, usually no "natural" conjugate priors are available, therefore posteriors are simulated using MCMC. Choice of "noninformative" priors is not obvious, as improper priors usually lead to improper posteriors. Nonidentifiability of $\mathcal{P}_k$ causes the problem of "label switching"). A nice reference for Bayesian methods is Frühwirth-Schnatter (2006).

Let me close this section with a short discussion of robustness. Robustness with respect to ▶outliers is treated by Hennig (2004). Another problem is that mixture models are extremely nonrobust with respect to misspecification of the component model. Estimating the component model in a fully nonparametric way is of course not possible, but manageable alternatives are for example mixtures of log-concave distributions. Let me point out, however, that issues like nonrobustness and nonidentifiability only cause problems if the task is to interpret the model parameters somehow. If the aim is only to obtain a better data fit, one need not worry about them.

Likelihood Methods

In the above setting, $l(\eta) = \log(f(x_1, \eta)) + \cdots + \log(f(x_n, \eta))$ is the log likelihood function. It may have some undesirable properties: First, the log likelihood is often unbounded. For example, consider mixtures of normals. If the expectation of one component is fixed at some data point and the variance goes to zero, the likelihood goes to infinity. Singularities usually occur at the boundary of the parameter space. Second, the likelihood function is usually not unimodal, although this depends on the parameterization. For example, if the parameter is a probability distribution as in Eq. 2 and if the parameter space Ξ is a convex set (with respect to the usual linear combination of measures), the log likelihood function is concave. If it is bounded, there is a nice theory of "nonparametric likelihood estimation" (Lindsay 1995), and "the" "nonparametric maximum likelihood estimator" is in some sense uniquely defined and can be calculated numerically (Böhning 2000; Schlattmann 2009).

Nonparametric methods, however, work in low dimensional component models, whereas "parametric" estimation techniques like the Expectation-Maximization (EM) method work in nearly any dimensional. The EM is a local maximizer for mixture likelihoods in $\mathcal{P}_k$. Here the mixture likelihood is usually multimodal; moreover, it can be very flat. Analytic expressions for likelihood maxima usually do not exist, they have to be calculated numerically. On the other hand, even for unbounded likelihoods, it is known from asymptotic theory, that the simple heuristics of searching for a large local maximum in the interior of the parameter space may lead to reasonable estimators. However, one must be aware that there exist "spurious" large local maxima that are statistically meaningless. Moreover, except from simple cases, there is no manageable asymptotics for likelihood ratio.

Some of the problems of pure likelihood approaches can be overcome by considering penalized likelihoods. However, here one has the problem of choosing a penalization parameter. Moreover, the EM algorithm is a basic tool for a number of estimation problems, and it has a very simple structure for simple finite mixtures. Therefore it will be outlined in the next section.

EM Algorithm

The EM algorithm is a local maximization technique for the log likelihood in $\mathcal{P}_k$. It starts from the complete-data log-likelihood. Suppose that for observation x_i the (fictive) component membership is known. It is defined by a vector $z_i \in \Re^k$ with $z_{ij} = 1$, if x_i belongs to j-th component, and zero elsewhere. As a random variable Z_i, it has a ▶multinomial distribution with parameters $k, p_1, \ldots, p_k$. Then the complete data likelihood and log likelihood of P, respectively, are $L_c(P) = \prod_{i=1}^n \prod_{j=1}^k (p_j f(x_i, \theta_j))^{z_{ij}}$ and $l_c(P) = \log(L_c(P)) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log p_j + \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log f(x_i, \theta_j)$.

The EM needs a starting value P_0, and then proceeds as an iteration between an "E-step" and an "M-step" until "convergence." The first E-step consists in calculating the conditional expectation $E_{P_0}(l_c(P) \mid x_1, \ldots, x_n)$ of $l_c(P)$ for

arbitrary P, given the data, under P_0. As the only randomness is in the z_{ij}, we obtain

$$E_{P_0}(l_c(P)\,|\,x_1,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j(x_i|P_0)\log p_j + \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j(x_i|P_0)\log f(x_i,\,\theta_j),$$

where

$$\tau_j(x_i|P_0) = \Pr_{P_0}(Z_{ij}=1\,|\,x_i) = \frac{p_j f(x_i,\,\theta_j)}{f(x_i,\,P_0)}$$

is the conditional probability that the i-th observation belongs to component j, given the data, with respect to P_0.

In the following M-step, $E_{P_0}(l_c(P)\,|\,x_1,\ldots,x_n)$ is maximized with respect to P. As it is the sum of terms depending on the mixing weights and on the parameters only, respectively, both parts can be maximized separately. It is easily shown that the maximum in the p_j is achieved for $p_j^{(1)} = (1/n)\sum_{i=1}^{n}\tau_j(x_i|P_0), j = 1,\ldots,n$. For component densities from exponential families, similar simple solutions exist for the θ_j, therefore both the E-step and the M-step can be carried out here analytically. It can be shown that (1) the log-likelihood is not decreasing during the iteration of the EM, and (2) that under some regularity conditions it converges to a stationary point of the likelihood function. However, this may also be a saddle point.

It remains to define the stopping rule and the starting point(s). Both are crucial, and the reader is referred to the literature. There are also techniques that prevent from convergence to singularities or spurious maxima. A final nice issue of the EM is that it yields a simple tool for classification of data points: If $\hat{P}$ is an estimator, then $\tau_j(x_i|\hat{P})$ is the posterior probability that x_i belongs to class j with respect to the "prior" $\hat{P}$. The Bayesian classification rule assigns observation i to the class j that maximizes $\tau_j(x_i|\hat{P})$, and the $\tau_j(x_i|\hat{P})$ measure the plausibility of such a clustering.

Number of Components, Testing and Asymptotics

Even if one has an estimator in each $\mathcal{P}_k$ from the EM, the question is how to assess the number of components (i.e., how to choose k). Usually information criteria like AIC and BIC are recommended. An alternative is to perform a sequence of tests of k against $k+1$ components, for $k = 1, 2\ldots$.

There are several tests for homogeneity, i.e., for the "component model", as for example goodness of fit or dispersion score tests. For testing k_0 against k_1 components, a likelihood ratio test may be performed. However, the usual χ^2-asymptotics fails, so critical values have to be simulated. Moreover, the distribution of the test statistic usually depends on the specific parameter under the null hypothesis. Therefore some sort of bootstrap (see ►Bootstrap Methods) is needed, and as estimators have to be calculated numerically, likelihood ratio tests are computationally intensive.

Let me close with some remarks on asymptotics. Whereas ►asymptotic normality of estimators is guaranteed under some conditions, the usual asymptotics for the likelihood ratio test fails. The reason is that under the null hypothesis, the parameter P_0 is on the boundary of the parameter space, it is not identifiable and the Fisher information matrix in P_0 is singular. There is an asymptotic theory under certain restrictive assumptions, but it is usually hard to calculate critical values from it.

About the Author

Professor Seidel was the Editor of "*AStA – Advances of Statistical Analysis*" (Journal of the German Statistical Society) (2004–2008). He is Dean of the Faculty of Economics and Social Sciences of Helmut-Schmidt-Universität (since January 2009), and has been elected next President of Helmut-Schmidt-University, starting in October 2010.

Cross References

►Bayesian Statistics
►Contagious Distributions
►Identifiability
►Likelihood
►Modeling Count Data
►Multivariate Statistical Distributions
►Nonparametric Estimation
►Optimum Experimental Design

References and Further Reading

Böhning D (2000) Finite mixture models. Chapman and Hall, Boca Raton
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York
Hennig C (2000) Identifiability of models for clusterwise linear regression. J Classif 17:273–296
Hennig C (2004) Breakdown points for ML estimators of location-scale mixtures. Ann Stat 32:1313–1340
Lindsay BG (1995) Mixture models: theory, geometry and applications. NSC-CBMS Regional Conference Series in Probability and Statistics, 5
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Schlattmann P (2009) Medical applications of finite mixture models. Springer, Berlin
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions, Wiley, New York

Model Selection

Walter Zucchini[1], Gerda Claeskens[2], Georges Nguefack-Tsague[3]
[1]Professor
Georg-August-Universität, Göttingen, Germany
[2]Professor
Leuven, Belgium
[3]University of Yaoundé I, Yaoundé, Cameroon

Introduction

In applications there are usually several models for describing a population from a given sample of observations and one is thus confronted with the problem of model selection. For example, different distributions can be fitted to a given sample of univariate observations; in polynomial regression one has to decide which degree of the polynomial to use; in multivariate regression one has to select which covariates to include in the model; in fitting an autoregressive model to a stationary time series one must choose which order to use.

When the set of models under consideration is nested, as is the case in polynomial regression, the fit of the model to the sample improves as the complexity of the model (e.g., the number of parameters) increases but, at some stage, its fit to the population deteriorates. That is because the model increasingly moulds itself to the features of the sample rather than to the "true model," namely the one that characterizes the population. The same tendency occurs even if the models are not nested; increasing the complexity eventually leads to deterioration. Thus model selection needs to take both goodness of the fit and the complexity of the competing models into account.

Reference books on model selection include Linhart and Zucchini (1986), Burnham and Anderson (2002), Miller (2002), Claeskens and Hjort (2008). An introductory article is Zucchini (2000).

Information Criteria – Frequentist Approach

The set of models considered for selection can be thought of as approximating models which, in general, will differ from the true model. The answer to the question "Which approximation is best?" depends, of course, on how we decide to measure the quality of the fit. Using the Kullback-Leibler distance for this leads to the popular ▶Akaike Information Criterion (AIC, Akaike 1973):

$$AIC(M) = 2\log(L(\hat{\theta})) - 2p,$$

where M is the model, L the likelihood, and $\hat{\theta}$ the maximum likelihood estimator of the vector of the model's p parameters. The first term of the AIC measures the fit of the model to the *observed sample*; the fit improves as the number of parameters in the model is increased. But improving the fit of the model to the sample does not necessarily improve its fit to the population. The second term is a penalty term that compensates for the complexity of the model. One selects the model that maximizes the AIC. Note, however, that in much of the literature the AIC is defined as minus the above expression, in which case one selects the model that minimizes it.

A *model selection criterion* is a formula that allows one to compare models. As is the case with the AIC, such criteria generally comprise two components: one that quantifies the fit to the data, and one that penalizes complexity. Examples include Mallows' C_p criterion for use in ▶linear regression models, Takeuchi's model-robust information criterion TIC, and refinements of the AIC such as the 'corrected AIC' for selection in linear regression and autoregressive time series models, the network information criterion NIC, which is a version of AIC that can be applied to model selection in ▶neural networks, and the generalized information criterion GIC for use with influence functions. Several of these criteria have versions that are applicable in situations where there are outlying observations, leading to robust model selection criteria; other extensions can deal with missing observations.

Alternative related approaches to model selection that do not take the form of an information criterion are *bootstrap* (see, e.g., Zucchini 2000) and *cross-validation*. For the latter the idea is to partition the sample in two parts: the calibration set, that is used to fit the model, and the validation sample, that is used to assess the fit of the model, or the accuracy of its predictions. The popular "leave-one-out cross-validation" uses only one observation in the validation set, but each observation has a turn at comprising the validation set. In a model selection context, we select the model that gives the best results (smallest estimation or prediction error) averaged over the validation sets. As this approach can be computationally demanding, suggestions have been made to reduce the computational load. In "five-fold cross-validation" the sample is randomly split in five parts of about equal size. One of the five parts is used as validation set and the other four parts as the calibration set. The process is repeated until each of the five sets is used as validation set.

Bayesian Approach

The Bayesian regards the models available for selection as candidate models rather than approximating models; each of them has the potential of being the true model. One begins by assigning to each of them a prior probability, $P(M)$, that it is the true model and then, using ▶Bayes'

theorem, computes the posterior probability of it being so:

$$P(M|\,\text{Data}) = \frac{P(\,\text{Data}\,|M)P(M)}{P(\,\text{Data}\,)}.$$

The model with the highest posterior probability is selected. The computation of $P(\,\text{Data}\,|M)$ and $P(\,\text{Data}\,)$ can be very demanding and usually involves the use of Markov chain Monte Carlo (MCMC) methods (see ►Markov Chain Monte Carlo) because, among other things, one needs to 'integrate out' the distribution of the parameters of M (see e.g., Wasserman 2000).

Under certain assumptions and approximations (in particular the Laplace approximation), and taking all candidate models as a priori equally likely to be true, this leads to the Bayesian Information Criterion (BIC), also known as the Schwarz criterion (Schwarz 1978):

$$\text{BIC}(M) = 2\log(L(\hat{\theta})) - p\log(n),$$

where n is the sample size and p the number of unknown parameters in the model. Note that although the BIC is based on an entirely different approach it differs from the AIC only in the penalty term.

The difference between the frequentist and Bayesian approaches can be summarized as follows. The former addresses the question "Which model is best, in the sense of least wrong?" and the latter the question "Which model is most likely to be true?".

The Deviance Information Criterion (Spiegelhalter et al. 2002) is an alternative Bayesian method for model selection. While explicit formulae are often difficult to obtain, its computation is simple for situations where MCMC simulations are used to generate samples from a posterior distribution.

The principle of minimum description length (MDL) is also related to the BIC. This method tries to measure the complexity of the models and selects the model that is the least complex. The MDL tries to minimize the sum of the description length of the model, plus the description length of the data when fitted to the model. Minimizing the description length of the data corresponds to maximizing the log likelihood of the model. The description length of the model is not uniquely defined but, under certain assumptions, MDL reduces to BIC, though this does not hold in general (Rissanen 1996). Other versions of MDL come closer to approximating the full Bayesian posterior $P(M|)$ Data. See Grünwald (2007) for more details.

Selecting a Selection Criterion

Different selection criteria often lead to different selections. There is no clear-cut answer to the question of which criterion should be used. Some practitioners stick to a single criterion; others take account of the orderings indicated by two or three different criteria (e.g., AIC and BIC) and then select the one that leads to the model which seems most plausible, interpretable or simply convenient in the context of the application.

An alternative approach is to tailor the criterion to the particular objectives of the study, i.e., to construct it in such a way that selection favors the model that best estimates the quantity of interest. The Focussed Information Criterion (FIC, Claeskens and Hjort 2003) is designed to do this; it is based on the premise that a good estimator has a small mean squared error (MSE). The FIC is constructed as an estimator of the MSE of the estimator of the quantity of interest. The model with the smallest value of the FIC is the best.

Issues such as consistency and efficiency can also play a role in the decision regarding which criterion to use. An information criterion is called *consistent* if it is able to select the true model from the candidate models, as the sample size tends to infinity. In a weak version, this holds with probability tending to one; for strong consistency, the selection of the true model is almost surely. It is important to realize that the notion of consistency only makes sense in situations where one can assume that the true model belongs to the set of models available for selection. Thus will not be the case in situations in which researchers "believe that the system they study is infinitely complicated, or there is no way to measure all the important variables" (McQuarrie and Tsai 1998). The BIC is a consistent criterion, as is the Hannan-Quinn criterion that uses $\log\log(n)$ instead of $\log(n)$ in the penalty term.

An information criterion is called *efficient* if the ratio of the expected mean squared error (or expected prediction error) under the selected model and the expected mean squared error (or expected prediction error) under its theoretical minimizer converges to one in probability. For a study of the efficiency of a model selection criterion, we do not need to make the assumption that the true model is one of the models in the search list. The AIC, corrected AIC, and Mallows's C_p are examples of efficient criteria. It can be shown that the BIC and the Hannan-Quinn criterion are not efficient. This is an observation that holds in general: consistency and efficiency cannot occur together.

Model Selection in High Dimensional Models

In some applications, e.g., in radiology and biomedical imaging, the number of unknown parameters in the model is larger than the sample size, and so classical model selection procedures (e.g., AIC, BIC) fail because the parameters cannot be estimated using the method of maximum likelihood. For these so-called high-dimensional models regularized or penalized methods have been suggested in

the literature. The popular Lasso estimator, introduced by Tibshirani (1996), adds an l_1 penalty for the coefficients in the estimation process. This has as a particular advantage that it not only can shrink the coefficients towards zero, but also sets some parameters equal to zero, which corresponds to variable selection. Several extensions to the basic Lasso exist, and theoretical properties include consistency under certain conditions. The Dantzig selector (Candes and Tao 2008) is another type of method for use with high-dimensional models.

Post-model Selection Inference

Estimators that are obtained in a model that has been selected by means of a model selection procedure, are referred to as *estimators-post-selection* or *post-model-selection estimators*. Since the data are used to select the model, the selected model that one works with, is random. This is the main cause of inferences to be wrong when ignoring model selection and pretending that the selected model had been given beforehand. For example, by ignoring the fact that model selection has taken place, the estimated variance of an estimator is likely to be too small, and confidence and prediction intervals are likely to be too narrow. Literature on this topic includes Pötscher (1991), Hjort and Claeskens (2003), Shen et al. (2004), Leeb and Pötscher (2005).

Model selection can be regarded as the special case of model averaging in which the selected model takes on the weight one and all other models have weight zero. However, regarding it as such does not solve the problem because selection depends on the data, and so the weights in the estimator-post-selection are random. This results in non-normal limiting distributions of estimators-post-selection, and requires adjusted inference techniques to take the randomness of the model selection process into account. The problem of correct post-model selection inference has yet to be solved.

About the Authors

Walter Zucchini previously held the Chair of Statistics at the University of Cape Town. He is a Fellow of the Royal Statistical Society and the Royal Society of South Africa. He is Past President of the South African Statistical Association (1992) and Editor of the *South African Statistical Journal* (1986–1989). He was awarded the "Herbert Sichel Medaille" of the South African Statistical Association (2008), and the Shayle Searle Visiting Fellowship in Statistics, Victoria University, New Zealand (2008). Walter Zucchini is the co-author of the text *Model Selection* (with H. Linhart, Wiley 1986).

Gerda Claeskens is Professor at the Faculty of Business and Economics of the K.U. Leuven (Belgium). She is Elected member of the International Statistical Institute and recipient of the Noether Young Scholar Award (2004) "for outstanding achievements and contributions in nonparametric statistics." She is the author of more than 40 papers and of the book *Model selection and model averaging* (with N.L. Hjort, Cambridge University Press, 2008). Currently she is Associate editor of the *Journal of the American Statistical Association*, of *Biometrika*, and of the *Journal of Nonparametric Statistics*.

Georges Nguefack-Tsague is lecturer of Biostatistics in the Department of Public Health at the University of Yaounde I, Cameroon. He is head of the Biostatistics Unit and deputy speaker of the Master Program in Public Health. He was awarded a Lichtenberg Scholarship for his PhD studies, which he completed at the University of Goettingen (Germany). The title of his PhD thesis was *Estimating and Correcting the Effects of Model Selection Uncertainty*. He was teaching assistant (2001–2003) in the Department of Statistics and Econometrics at the University Carlos III of Madrid (Spain). Other awards included a Belgium Ministry of External Affairs (MSc) Scholarship and a Cameroon Ministry of Economy and Finance (MA) Scholarship.

Cross References

►Akaike's Information Criterion
►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bootstrap Methods
►C_p Statistic
►Exponential and Holt-Winters Smoothing
►Kullback-Leibler Divergence
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Markov Chain Monte Carlo
►Sensitivity Analysis
►Statistical Evidence
►Structural Time Series Models
►Time Series

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds) Second international symposium on information theory, Akadémiai Kiadó, Budapest, pp 267–281

Burnham PK, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Candes E, Tao T (2008) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35:2313–2351
Claeskens G, Hjort NL (2003) The focussed information criterion (with discussion). J Am Stat Assoc 98:900–916
Claeskens G, Hjort NL (2008) Model selection and model averaging. Cambridge University Press, Cambridge
Grünwald P (2007) The minimum description length principle. MIT Press, Boston
Hjort NL, Claeskens G (2003) Frequentist model average estimators (with discussion). J Am Stat Assoc 98:879–899
Leeb H, Pötscher BM (2005) Model selection and inference: fact and fiction. Econometr Theor 21:21–59
Linhart H, Zucchini W (1986) Model selection. Wiley, New York
McQuarrie ADR, Tsai CL (1998) Regression and time series model selection. World Scientific, River Edge
Miller AJ (2002) Subset selection in regression, 2nd edn. Chapman and Hall/CRC, Boca Raton
Pötscher BM (1991) Effects of model selection on inference. Econometr Theor 7:163–185
Rissanen JJ (1996) Fisher information and stochastic complexity. IEEE Trans Inform Theory 42:40–47
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shen X, Huang HC, Ye J (2004) Inference after model selection. J Am Stat Assoc 99:751–762
Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J Roy Stat Soc B 64:583–639
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58(1):267–288
Wasserman L (2000) Bayesian model selection and model averaging. J Math Psychol 44:92–107
Zucchini W (2000) An introduction to model selection. J Math Psychol 44:41–61

Model-Based Geostatistics

Hannes Kazianka[1], Jürgen Pilz[2]
[1]University of Technology, Vienna, Austria
[2]Professor, Head
University of Klagenfurt, Klagenfurt, Austria

Stochastic Models for Spatial Data

Diggle and Ribeiro (2007) and Mase (2010) describe geostatistics as a branch of spatial statistics that deals with statistical methods for the analysis of spatially referenced data with the following properties. Firstly, values Y_i, $i = 1, \ldots, n$, are observed at a discrete set of sampling locations $\boldsymbol{x}_i$ within some spatial region $\mathcal{S} \subset \mathbb{R}^d$, $d \geq 2$. Secondly, each observed value Y_i is either a measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon, $Z(\boldsymbol{x})$, at the corresponding sampling location $\boldsymbol{x}_i$. The term model-based geostatistics refers to geostatistical methods that rely on a stochastic model. The observed phenomenon is viewed as a realization of a continuous stochastic process in space, a so-called random field.

Such a random field $Z(\boldsymbol{x})$ is fully determined by specifying all multivariate distributions, i.e., $P(Z(\boldsymbol{x}_1) \leq z_1, \ldots, Z(\boldsymbol{x}_n) \leq z_n)$ for arbitrary $n \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{S}$. Since a full characterization of a random field is usually hopeless, the mean function $m(\boldsymbol{x}) = E(Z(\boldsymbol{x}))$ and the covariance function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ play a prominent role. Thereby, $m(\boldsymbol{x})$ represents the trend while $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ defines the dependence structure of the random field. It is typical that the assumption of weak (second-order) isotropy is made about the random field, i.e., its mean function is constant and its covariance function $K(\boldsymbol{x}_1, \boldsymbol{x}_2)$ depends on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ only through $h = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$, where $\|.\|_2$ denotes the Euclidean distance. In this case K is called an isotropic autocovariance function. The covariance function is directly related to smoothness properties of the random field such as mean square continuity and differentiability. A widely used parametric family of isotropic autocovariance functions is the Matern family

$$K_{\sigma^2,\boldsymbol{\theta}}(h) = \sigma^2 \left((1 - \vartheta_2) + \frac{\vartheta_2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{2\kappa^{\frac{1}{2}} h}{\vartheta_1} \right)^{\kappa} \mathcal{K}_{\kappa} \left(\frac{2\kappa^{\frac{1}{2}} h}{\vartheta_1} \right) \right),$$

where $\mathcal{K}_{\kappa}$ denotes the modified Bessel function of order $\kappa > 0$, $\vartheta_1 > 0$ is a called the "range parameter" controlling how fast the covariance decays as the distance h gets large, $\vartheta_2 \in [0,1]$ is called the "nugget parameter" and describes a measurement error, σ^2 controls the variance and $\boldsymbol{\theta} = (\vartheta_1, \vartheta_2, \kappa)$ denotes the vector of correlation parameters. The parameter κ controls the smoothness of the corresponding process. A thorough mathematical introduction to the theory of random fields is given in Stein (1999) and Yaglom (1987).

The most important geostatistical model is the linear Gaussian model

$$Y_i = \boldsymbol{f}(\boldsymbol{x}_k)^T \boldsymbol{\beta} + Z(\boldsymbol{x}_i), \quad i = 1, \ldots, n, \tag{1}$$

where $Z(\boldsymbol{x})$ is a weakly isotropic zero-mean Gaussian random field with autocovariance function $K_{\sigma^2,\boldsymbol{\theta}}$, $\boldsymbol{f}$ is a vector of location-dependent explanatory variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression parameters. The

likelihood function for the linear Gaussian model is

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\right) = (2\pi)^{-\frac{n}{2}} \left|\sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right)\right\},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ denotes the correlation matrix, $\boldsymbol{F}$ is the design matrix and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is the vector of observations. The maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 in the linear Gaussian model are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{Y}, \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{Z} - \mathbf{F}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\mathbf{Z} - \mathbf{F}\hat{\boldsymbol{\beta}}\right). \tag{3}$$

Plugging these estimates into the log-likelihood, we arrive at the so-called profiled log-likelihood, which just contains the parameters $\boldsymbol{\theta}$

$$\log p\left(\boldsymbol{Y} \mid \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\theta}\right) = -\frac{n}{2}\left(\log(2\pi) + 1\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right| - \frac{n}{2}\log\left(\hat{\sigma}^2\right).$$

To obtain $\hat{\boldsymbol{\theta}}$ we have to maximize the latter equation for $\boldsymbol{\theta}$ numerically. Note that this maximization problem is a lot simpler than the maximization of the complete likelihood where $\boldsymbol{\beta}$ and σ^2 are additional unknowns, especially when p is large. Spatial prediction, which is often the goal in geostatistics, is performed based on the estimated parameters. The plug-in predictive distribution for the value of the random field at an unobserved location $\boldsymbol{x}_0$ is Gaussian

$$Y_0 \mid \boldsymbol{Y}, \hat{\sigma}^2, \hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{Y} + \boldsymbol{s}^T \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 - \boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{k} + \hat{\sigma}^2 \boldsymbol{s}^T \left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{s}\right), \tag{4}$$

where $\boldsymbol{K} = \hat{\sigma}^2 \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{s} = \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{k}, \boldsymbol{k} = Cov(\boldsymbol{Z}, Z(\boldsymbol{x}_0))$, $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))^T$.

Weak isotropy is a rather strong assumption and environmental processes are typically not direction independent but show an anisotropic behavior. A popular extension to isotropic random fields is to consider random fields that become isotropic after a linear transformation of the coordinates (Schabenberger and Gotway 2005). This special variant of anisotropy is called geometric anisotropy. Let $Z_1(\boldsymbol{x})$ be an isotropic random field on $\mathbb{R}^d$ with autocovariance function K_1 and mean μ. For the random field $Z(\boldsymbol{x}) = Z_1(\boldsymbol{T}\boldsymbol{x})$, where $\boldsymbol{T} \in \mathbb{R}^{d \times d}$, we get that $E(Z(\boldsymbol{x})) = \mu$ and the corresponding autocovariance function is $Cov(Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)) = K_1\left(\|\boldsymbol{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\|_2\right)$. When correcting for geometric anisotropy we need to revert the coordinate transformation. $Z\left(\boldsymbol{T}^{-1}\boldsymbol{x}\right)$ has the same mean as $Z(\boldsymbol{x})$ but isotropic autocovariance function K_1. When correcting for stretching and rotation of the coordinates we have

$$\boldsymbol{T}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

Here, λ and φ are called the anisotropy ratio and anisotropy angle, respectively. All the models that we consider in this chapter can be extended to account for geometric anisotropy by introducing these two parameters.

Bayesian Kriging

The first steps towards Bayesian modeling and prediction in geostatistics were made by Kitanidis (1986) and Omre (1987) who developed a Bayesian version of universal kriging. One of the advantages of the Bayesian approach, besides its ability to deal with the uncertainty about the model parameters, is the possibility to work with only a few measurements. Assume a Gaussian random field model in the form of the form Eq. 1 with known covariance matrix $\boldsymbol{K}$ but unknown parameter vector $\boldsymbol{\beta}$. From Bayesian analysis we know that it is natural to assume a prior of the form $\boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{m}_b, \sigma^2 \boldsymbol{V}_b\right)$ for $\boldsymbol{\beta}$, where $\boldsymbol{V}_b$ is a positive semidefinite matrix. It can be shown that the posterior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \boldsymbol{Z} \sim \mathrm{N}\left(\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{V}_{\tilde{\beta}}\right),$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{V}_{\tilde{\beta}}\left(\sigma^2 \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{Z} + \boldsymbol{V}_b^{-1} \boldsymbol{m}_b\right)$ and $\boldsymbol{V}_{\tilde{\beta}} = \left(\sigma^2 \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F} + \boldsymbol{V}_b^{-1}\right)^{-1}$. The predictive distribution of $Z(\boldsymbol{x}_0)$ is also Gaussian and given by

$$Z(\boldsymbol{x}_0) \mid \boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{Z} + \boldsymbol{s}^T \tilde{\boldsymbol{\beta}}, \sigma^2 - \boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{k} + \sigma^2 \boldsymbol{s}^T \boldsymbol{V}_{\tilde{\beta}} \boldsymbol{s}\right),$$

where $\boldsymbol{F}$, $\boldsymbol{s}$ and $\boldsymbol{k}$ are defined as in Section "►Stochastic Models for Spatial Data". From the above representation of the Bayesian kriging predictor it becomes clear that Bayesian kriging bridges the gap between simple and universal kriging. We get simple kriging in case of complete knowledge of the trend, which corresponds to $\mathbf{V}_b = \mathbf{0}$, whereas we get the universal kriging predictor if we have no knowledge of $\boldsymbol{\beta}$ ($\mathbf{V}_b^{-1} = \mathbf{0}$ in the sense that the smallest eigenvalue of $\mathbf{V}_b$ converges to infinity). Interestingly, the Bayesian universal kriging predictor has a smaller or equal variance than the classical universal kriging predictor (see Eq. 4) since $\left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F} + \sigma^{-2} \boldsymbol{V}_b^{-1}\right)^{-1} \preceq \left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F}\right)^{-1}$, where $\preceq$ denotes the Loewner partial ordering.

Bayesian universal kriging is not fully Bayesian because $\boldsymbol{K}$ is assumed known. Diggle and Ribeiro (2007) summarize the results for a fully Bayesian analysis of Gaussian random field models of the form Eq. 1, where $K_{\sigma^2,\theta} = \sigma^2 \Sigma_{\vartheta_1}$ and ϑ_1 is the range parameter of an isotropic autocorrelation function model.

Transformed Gaussian Kriging

Probably the most simple way to extend the Gaussian random field model is to assume that a differentiable transformation of the original random field, $Z_1(\boldsymbol{x}) = g(Z(\boldsymbol{x}))$, is Gaussian. The mean of the transformed field is unknown and parameterized by $\boldsymbol{\beta}$, $E(Z_1(\boldsymbol{x})) = \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{\beta}$. If we assume that the transformation function g and the covariance function K of $Y(\boldsymbol{x})$ are known, the optimal predictor for $Z(\boldsymbol{x}_0)$ can be derived using the results from Section "►Stochastic Models for Spatial Data". However, in practice neither K nor g is known and we have to estimate them from the data.

A family of one-parameter transformation functions g_λ that is widely used in statistics is the so-called Box-Cox family

$$g_\lambda(z) = \begin{cases} \frac{z^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(z), & \lambda = 0. \end{cases}$$

The ►Box-Cox transformation is valid for positive-valued random fields and is able to model moderately skewed, unimodal data.

The likelihood of the data $\boldsymbol{Y}$ in the transformed Gaussian model can be written as

$$p(\mathbf{Y} \mid \boldsymbol{\Theta}) = J_\lambda(\mathbf{Y}) (2\pi)^{-\frac{n}{2}} \left|\sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (g_\lambda(\mathbf{Y}) - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (g_\lambda(\mathbf{Y}) - \boldsymbol{F}\boldsymbol{\beta})\right],$$

where, $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \lambda)$, $J_\lambda(\mathbf{Y})$ is the determinant of the Jacobian of the transformation, $g_\lambda(\boldsymbol{Y}) = (g_\lambda(Y_1), \ldots, g_\lambda(Y_n))$ and λ is the transformation parameter. De Oliveira et al. (1997) point out that the interpretation of $\boldsymbol{\beta}$ changes with the value of λ, and the same is true for the covariance parameters σ^2 and $\boldsymbol{\theta}$, to a lesser extent though. To estimate the parameters λ and $\boldsymbol{\theta}$, we make use of the profile likelihood approach that we have already encountered in Section "►Stochastic Models for Spatial Data". For fixed values of λ and $\boldsymbol{\theta}$, the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 are given by Eqs. 2 and 3 with $\boldsymbol{Y}$ replaced by $g_\lambda(\boldsymbol{Y})$. Again, the estimates for λ and $\boldsymbol{\theta}$ cannot be written in closed form and must be found numerically by plugging $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ in the likelihood for numerical maximization.

The estimated parameters $\hat{\boldsymbol{\Theta}}$ are subsequently used for spatial prediction. To perform a plug-in prediction we make use of the conditional distribution of the Gaussian variable $Y_0 \mid \boldsymbol{Y}, \hat{\boldsymbol{\Theta}}$ and back-transform it to the original scale by g_λ^{-1}. A Bayesian approach to spatial prediction in the transformed Gaussian model is proposed in De Oliveira et al. (1997).

The copula-based geostatistical model (Kazianka and Pilz 2009) also works with transformations of the marginal distributions of the random field and is a generalization of transformed Gaussian kriging. In this approach all multivariate distributions of the random field are described by a copula (Sempi 2010) and a family of univariate marginal distributions. Due to the additional flexibility introduced by the choice of the copula and of the marginal distribution, these models are able to deal with extreme observations and multi-modal data.

Generalized Linear Geostatistical Models

►Generalized linear models (McCullagh and Nelder 1989) provide a unifying framework for regression modeling of both continuous and discrete data. Diggle and Ribeiro (2007) extend the classical generalized linear model to what they call the generalized linear geostatistical model (GLGM). The responses Y_i, $i = 1, \ldots, n$, corresponding to location $\boldsymbol{x}_i$ are assumed to follow a family of univariate distributions indexed by their expectation, μ_i, and to be conditionally independent given $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))$. The μ_i are specified through

$$h(\mu_i) = \boldsymbol{f}(\boldsymbol{x}_i)^T \boldsymbol{\beta} + Z(\boldsymbol{x}_i),$$

where $Z(\boldsymbol{x})$ is a Gaussian random field with autocovariance function $K_{\boldsymbol{\theta}}$ and h is a pre-defined link function. The two most frequently applied GLGMs are the Poisson log-linear model, where Y_i is assumed to follow a Poisson distribution and the link function is the logarithm, and the binomial logistic-linear model, where Y_i is assumed to follow a Bernoulli distribution with probability $\mu_i = p(\boldsymbol{x}_i)$ and $h(\mu_i) = \log(p(\boldsymbol{x}_i) / (1 - p(\boldsymbol{x}_i)))$. These models are suitable for representing spatially referenced count data and binary data, respectively.

Since maximum likelihood estimation of the parameters is difficult, a Markov chain Monte Carlo (Robert and Casella 2004) approach (see ►Markov Chain Monte Carlo) is proposed to sample from the posteriors of the model parameters as well as from the predictive distributions at unobserved locations $\boldsymbol{x}_0$. The algorithm proceeds by sampling from $\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\theta}$, from $\boldsymbol{\theta} \mid \boldsymbol{Z}$ and from $\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y}$ with the help of Metropolis-Hastings updates. At iteration $t+1$ and

actual sample $\left(\boldsymbol{Z}^t, \boldsymbol{\theta}^t, \boldsymbol{\beta}^t, Z^t(\boldsymbol{x}_0)\right)$, perform the following steps:

- Update $\boldsymbol{Z}$. For $i = 1, \ldots, n$, sample a new proposal $Z'(\boldsymbol{x}_i)$ from the conditional Gaussian distribution $p\left(Z(\boldsymbol{x}_i) \mid \boldsymbol{\theta}^t, \boldsymbol{Z}^t_{-i}\right)$, where $\boldsymbol{Z}^t_{-i}$ denotes $\boldsymbol{Z}^t = (Z^t(\boldsymbol{x}_1), \ldots, Z^t(\boldsymbol{x}_n))$ with its ith element removed. Accept $Z'(\boldsymbol{x}_i)$ with probability $r = \min\left\{1, \frac{p(Y_i \mid \boldsymbol{\beta}^t, Z'(\boldsymbol{x}_i))}{p(Y_i \mid \boldsymbol{\beta}^t, Z^t(\boldsymbol{x}_i))}\right\}$.
- Update $\boldsymbol{\theta}$. Sample a new proposal $\boldsymbol{\theta}'$ from a proposal distribution $J\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t\right)$. Accept the new proposal with probability $r = \min\left\{1, \frac{p(\boldsymbol{Z}^{t+1} \mid \boldsymbol{\theta}')J(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}')}{p(\boldsymbol{Z}^{t+1} \mid \boldsymbol{\theta}^t)J(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^t)}\right\}$.
- Update $\boldsymbol{\beta}$. Sample a new proposal $\boldsymbol{\beta}'$ from a proposal distribution $J\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}^t\right)$. Accept the new proposal with probability $r = \min\left\{1, \frac{\prod_{i=1}^n p(Y_i \mid Z^{t+1}(\boldsymbol{x}_i), \boldsymbol{\beta}')J(\boldsymbol{\beta}^t \mid \boldsymbol{\beta}')}{\prod_{i=1}^n p(Y_i \mid Z^{t+1}(\boldsymbol{x}_i), \boldsymbol{\beta}^t)J(\boldsymbol{\beta}' \mid \boldsymbol{\beta}^t)}\right\}$
- Draw a sample $Z^{t+1}(\boldsymbol{x}_0)$ from the conditional Gaussian distribution $Z(\boldsymbol{x}_0) \mid \boldsymbol{Z}^{t+1}, \boldsymbol{\theta}^{t+1}$.

If point predictions for $Z(\boldsymbol{x}_0)$ are needed, the Monte Carlo approximation to the expected value of $Z(\boldsymbol{x}_0) \mid \boldsymbol{Y}$ can be used, i.e., $E(\boldsymbol{Z}(\boldsymbol{x}_0) \mid \boldsymbol{Y}) \approx \frac{1}{M}\sum_{t=1}^M Z^t(\boldsymbol{x}_0)$, where M is the number of simulations.

About the Author

For the biography *see* the entry ►Statistical Design of Experiments

Cross References

►Analysis of Areal and Spatial Interaction Data
►Box–Cox Transformation
►Gaussian Processes
►Generalized Linear Models
►Geostatistics and Kriging Predictors
►Markov Chain Monte Carlo
►Random Field
►Spatial Statistics

References and Further Reading

De Oliveira V, Kedem B, Short D (1997) Bayesian prediction of transformed Gaussian fields. J Am Stat Assoc 92:1422–1433
Diggle P, Ribeiro P (2007) Model-based geostatistics. Springer, New York
Sempi C (2010) Copulas. (this volume)
Kazianka H, Pilz J (2009) Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch Env Res Risk Assess, doi: 10.1007/s00477-009-0353-8
Kitanidis P (1986) Parameter uncertainty in estimation of spatial function: Bayesian analysis. Water Resour Res 22: 499–507
Mase S (2010) Geostatistics and kriging predictors. (this volume)
McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall/CRC, Boca Raton
Omre H (1987) Bayesian kriging – merging observations and qualified guesses in kriging. Math Geol 19:25–39
Robert C, Casella G (2004) Monte Carlo statistical methods. Springer, New York
Schabenberger O, Gotway C (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC, Boca Raton
Stein M (1999) Interpolation of spatial data. Springer, New York
Yaglom A (1987) Correlation theory of stationary and related random functions. Springer, New York

Modeling Count Data

Joseph M. Hilbe
Emeritus Professor
University of Hawaii, Honolulu, HI, USA
Adjunct Professor of Statistics
Arizona State University, Tempe, AZ, USA
Solar System Ambassador
California Institute of Technology, Pasadena, CA, USA

Count models are a subset of discrete response regression models. Count data are distributed as non-negative integers, are intrinsically heteroskedastic, right skewed, and have a variance that increases with the mean. Example count data include such situations as length of hospital stay, the number of a certain species of fish per defined area in the ocean, the number of lights displayed by fireflies over specified time periods, or the classic case of the number of deaths among Prussian soldiers resulting from being kicked by a horse during the Crimean War.

►Poisson regression is the basic model from which a variety of count models are based. It is derived from the Poisson probability mass function, which can be expressed as

$$f(y_i;\lambda_i) = \frac{e^{-t_i\lambda_i}(t_i\lambda_i)^{y_i}}{y_i!}, \qquad y = 0,1,2,\ldots;\ \mu > 0 \quad (1)$$

with y_i as the count response, λ_i as the predicted count or rate parameter, and t_i the area or time in which counts enter the model. When λ_i is understood as applying to individual counts without consideration of size or time, $t_i = 1$. When $t_i > 1$, it is commonly referred to as an exposure, and is modeled as an offset.

Estimation of the Poisson model is based on the log-likelihood parameterization of the Poisson probability distribution, which is aimed at determining parameter values

making the data most likely. In exponential family form it is given as:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\}, \quad (2)$$

where μ_i is typically used to symbolize the predicted counts in place of λ_i. Equation 2, or the deviance function based on it, is used when the Poisson model is estimated as a generalized linear model (*GLM*) (*see* ►Generalized Linear Models). When estimation employs a full maximum likelihood algorithm, μ_i is expressed in terms of the linear predictor, $x'\beta$. As such it appears as

$$\mu_i = \exp(x_i\beta). \quad (3)$$

In this form, the Poisson log-likelihood function is expressed as

$$L(\beta; y_i) = \sum_{i=1}^{n} \{y_i(x_i\beta) - \exp(x_i\beta) - \ln(y_i!)\}. \quad (4)$$

A key feature of the Poisson model is the equality of the mean and variance functions. When the variance of a Poisson model exceeds its mean, the model is termed overdispersed. Simulation studies have demonstrated that overdispersion is indicated when the Pearson χ^2 dispersion is greater than 1.0 (Hilbe 2007). The dispersion statistic is defined as the Pearson χ^2 divided by the model residual degrees of freedom. Overdispersion, common to most Poisson models, biases the parameter estimates and fitted values. When Poisson overdispersion is real, and not merely apparent (Hilbe 2007), a count model other than Poisson is required.

Several methods have been used to accommodate Poisson overdispersion. Two common methods are quasi-Poisson and negative binomial regression. Quasi-Poisson models have generally been understood in two distinct manners. The traditional manner has the Poisson variance being multiplied by a constant term. The second, employed in the glm() function that is downloaded by default when installing R software, is to multiply the standard errors by the square root of the Pearson dispersion statistic. This method of adjustment to the variance has traditionally been referred to as scaling. Using R's quasipoisson() function is the same as what is known in standard GLM terminology as the scaling of standard errors.

The traditional negative binomial model is a Poisson-gamma mixture model with a second ancillary or heterogeneity parameter, α. The mixture nature of the variance is reflected in its form, $\mu_i + \alpha\mu_i^2$, or $\mu_i(1 + \alpha\mu_i)$. The Poisson variance is μ_i, and the two parameter gamma variance is μ_i^2/ν. ν is inverted so that $\alpha = 1/\nu$, which allows for a direct relationship between μ_i, and ν. As a Poisson-gamma mixture model, counts are Poisson distributed as they enter into the model. α is the shape (gamma) of the manner counts enter into the model as well as a measure of the amount of Poisson overdispersion in the data.

The negative binomial probability mass function (*see* ►Geometric and Negative Binomial Distributions) may be formulated as

$$f(y_i; \mu_i, \alpha) = \binom{y_i + 1/\alpha - 1}{1/\alpha - 1} (1/(1 + \alpha\mu_i))^{1/\alpha} (\alpha\mu_i/(1 + \alpha\mu_i))_i^y, \quad (5)$$

with a log-likelihood function specified as

$$L(\mu_i; y_i, \alpha) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\frac{1}{\alpha}\right) \ln(1 + \alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right\}. \quad (6)$$

In terms of $\mu = \exp(x'\beta)$, the parameterization employed for maximum likelihood estimation, the negative binomial log-likelihood appears as

$$L(\beta; y_i, \alpha) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\alpha \exp(x_i'\beta)}{1 + \alpha \exp(x_i'\beta)}\right) - \left(\frac{1}{\alpha}\right) \ln\left(1 + \alpha \exp(x_i'\beta)\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right\}. \quad (7)$$

This form of negative binomial has been termed *NB2*, due to the quadratic nature of its variance function. It should be noted that the *NB2* model reduces to the Poisson when $\alpha = 0$. When $\alpha = 1$, the model is geometric, taking the shape of the discrete correlate of the continuous negative exponential distribution. Several fit tests exist that evaluate whether data should be modeled as Poisson or *NB2* based on the degree to which α differs from 0.

When exponentiated, Poisson and *NB2* parameter estimates may be interpreted as incidence rate ratios. For example, given a random sample of 1,000 patient observations from the German Health Survey for the year 1984, the following Poisson model output explains the years expected number of doctor visits on the basis of gender and marital status, both recorded as binary (1/0) variables, and the continuous predictor, age.

Docvis	IRR	OIM std. err.	z	P > |z|	[95% Conf. interval]	
Female	1.516855	0.054906	11.51	0.000	1.41297	1.628378
Married	0.8418408	0.0341971	−4.24	0.000	0.7774145	0.9116063
Age	1.018807	0.0016104	11.79	0.000	1.015656	1.021968

The estimates may be interpreted as

- Females are expected to visit the doctor some 50% more times during the year than males, holding marital status and age constant.

  Married patients are expected to visit the doctor some 16% fewer times during the year than unmarried patients, holding gender and age constant.

  For a one year increase in age, the rate of visits to the doctor increases by some 2%, with marital status and gender held constant.

It is important to understand that the canonical form of the negative binomial, when considered as a *GLM*, is not *NB2*. Nor is the canonical negative binomial model, *NB-C*, appropriate to evaluate the amount of Poisson overdispersion in a data situation. The NB-C parameterization of the negative binomial is directly derived from the negative binomial log-likelihood as expressed in Eq. 6. As such, the link function is calculated as $\ln(\alpha\mu/(1+\alpha\mu))$. The inverse link function, or mean, expressed in terms of $x'\beta$, is $1/(\alpha(\exp(-x'\beta)-1))$.

When estimated as a *GLM*, *NB-C* can be amended to *NB2* form by substituting $\ln(\mu)$ and $\exp(x'\beta)$ respectively for the two above expressions. Additional amendments need to be made to have the *GLM*-estimated *NB2* display the same parameter standard errors as are calculated using full maximum likelihood estimation. The *NB-C* log-likelihood, expressed in terms of μ, is identical to that of the *NB2* function. However, when parameterized as $x'\beta$, the two differ, with the *NB-C* appearing as

$$L(\beta;y_i,\alpha)=\sum_{i=1}^{n}\{y_i(x_i\beta)+(1/\alpha)\ln(1-\exp(x_i\beta)) + \ln\Gamma(y_i+1/\alpha)-\ln\Gamma(y_i+1)-\ln\Gamma(1/\alpha)\} \quad (8)$$

The *NB-C* model better fits certain types of count data than *NB2*, or any other variety of count model. However, since its fitted values are not on the log scale, comparisons cannot be made to Poisson or *NB2*.

The *NB2* model, in a similar manner to the Poisson, can also be overdispersed if the model variance exceeds its nominal variance. In such a case one must attempt to determine the source of the extra correlation and model it accordingly.

The extra correlation that can exist in count data, but which cannot be accommodated by simple adjustments to the Poisson and negative binomial algorithms, has stimulated the creation of a number of enhancements to the two base count models. The differences in these enhanced models relates to the attempt of identifying the various sources of overdispersion.

For instance, both the Poisson and negative binomial models assume that there exists the possibility of having zero counts. If a given set of count data excludes that possibility, the resultant Poisson or negative binomial model will likely be overdispersed. Modifying the loglikelihood function of these two models in order to adjust for the non-zero distribution of counts will eliminate the overdispersion, if there are no other sources of extra correlation. Such models are called, respectively, zero-truncated Poisson and zero-truncated negative binomial models.

Likewise, if the data consists of far more zero counts that allowed by the distributional assumptions of the Poisson or negative binomial models, a zero-inflated set of models may need to be designed. Zero-inflated models are ▶mixture models, with one part consisting of a 1/0 binary response model, usually a ▶logistic regression, where the probability of a zero count is estimated in difference to a non-zero-count. A second component is generally comprised of a Poisson or negative binomial model that estimates the full range of count data, adjusting for the overlap in estimated zero counts. The point is to (1) determine the estimates that account for zero counts, and (2) to estimate the adjusted count model data.

Hurdle models are another type mixture model designed for excessive zero counts. However, unlike the zero-inflated models, the hurdle-binary model estimates the probability of being a non-zero count in comparison to a zero count; the hurdle-count component is estimated on the basis of a zero-truncated count model. Zero-truncated, zero-inflated, and hurdle models all address abnormal

Modeling Count Data. Table 1 Models to adjust for violations of Poisson/NB distributional assumptions

Response	Example models
1: no zeros	Zero-truncated models (*ZTP*; *ZTNB*)
2: excessive zeros	Zero-inflated (*ZIP*; *ZINB*; *ZAP*; *ZANB*); hurdle models
3: truncated	Truncated count models
4: censored	Econometric and survival censored count models
5: panel	*GEE*; fixed, random, and mixed effects count models
6: separable	Sample selection, finite mixture models
7: two-responses	Bivariate count models
8: other	Quantile, exact, and Bayesian count models

Modeling Count Data. Table 2 Methods to directly adjust the variance (from Hilbe 2007)

Variance function	Example models
0: μ	Poisson
1: $\mu(\Phi)$	Quasi-Poisson; scaled SE; robust SE
2: $\mu(1+\alpha)$	Linear *NB* (*NB1*)
3: $\mu(1+\mu)$	Geometric
4: $\mu(1+\alpha\mu)$	Standard *NB* (*NB2*); quadratic *NB*
5: $\mu(1+(\alpha v)\mu)$	Heterogeneous *NB* (*NH-H*)
6: $\mu(1+\alpha\mu^{\rho})$	Generalized *NB* (*NB-P*)
7: $V[R]V'$	Generalized estimating equations

zero-count situations, which violate essential Poisson and negative binomial assumptions.

Other violations of the distributional assumptions of Poisson and negative binomial probability distributions exist. Table 1 below summarizes major types of violations that have resulted in the creation of specialized count models.

Alternative count models have also been constructed based on an adjustment to the Poisson variance function, μ. We have previously addressed two of these. Table 2 provides a summary of major types of adjustments.

Three texts specifically devoted to describing the theory and variety of count models are regarded as the standard resources on the subject. Other texts dealing with discrete response models in general, as well as texts on generalized linear models (*see* Generalized Linear Models), also have descriptions of many of the models mentioned in this article.

About the Author

For biography *see* the entry ►Logistic Regression.

Cross References

►Dispersion Models
►Generalized Linear Models
►Geometric and Negative Binomial Distributions
►Poisson Distribution and Its Application in Statistics
►Poisson Regression
►Robust Regression Estimation in Generalized Linear Models
►Statistical Methods in Epidemiology

References and Further Reading

Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, New York
Hilbe JM (2007) Negative binomial regression. Cambridge University Press, Cambridge, UK
Hilbe JM (2011) Negative binomial regression, 2nd edn. Cambridge University Press, Cambridge, UK
Winkelmann R (2003) Econometric analysis of count data, 4th edn. Springer, Heidelberg

Modeling Randomness Using System Dynamics Concepts

Mahender Singh[1], Frank M. Guess[2], Timothy M. Young[2], Lefei Liu[3]
[1]Research Director of Supply Chain 2020
Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Professor
University of Tennessee, Knoxville, TN, USA
[3]University of South Carolina, Columbia, SC, USA

L. J. Savage (1980) and others understood the importance of better computational tools for utilizing Bayesian insights data in real life applications long ago. Such computational tools and software are now available that use subjective (or soft) data as well as quantitative (or hard) data. But

despite the availability of new tools and buildup of massive databases, the increased complexity and integration of economic and other systems involving people poses a significant challenge to a solely statistical driven view of the system. More importantly, evidence suggests that relying solely on standard statistical models is inadequate to represent real life systems effectively for management insights and decisions.

Unpredictability characterizes most real life systems due to non-linear relationships and multiple time-delayed feedback loops between interconnected elements. Senge (1990) describes it as *dynamic complexity* – *"situations* where the cause and effect are subtle, and the effects over time of interventions are not obvious." As a result, such systems are unsuitable for quantitative "only" representations without some subjective expert views. System Dynamics models offer a helpful alternative to modeling randomness that is based on hard data and soft data that models a real world system; see for example Sterman (2000) and his references.

According to , Forrester (1980) three types of data are required to develop the foundation of an effective model: numerical, written and mental data; compare, also, Sterman (2000) discussion on these points. In most cases, however, only a small fraction of the data needed to model a real world system may be available in the form of numerical data. Perhaps, the most important data to build a model, namely the mental data, is difficult to represent only numerically. But due to heavy influence of quantitative bias in model development, some modelers disregard key qualitative information in favor of information that can be estimated statistically. Sterman (2000) considers this reasoning counterintuitive and counterproductive in practice with realistic systems. He states that "omitting structures and variables known to be important because numerical data are unavailable is actually less scientific and less accurate than using your best judgment to estimate their values." This is in line with Forrester's views (1961) asserting that, "to omit such variables is equivalent to saying they have zero effect - probably the only value that is known to be wrong!" A suitable approach in such cases is to iteratively improve the accuracy and reliability of data by leveraging deeper insights into the system and interaction between various variables over time, along with sensitivity analysis of various contingencies.

A key to understanding a dynamic real world system is to identify and study the causal loops (or sub-systems) of the system. An analysis of the structure-behavior relationship in a model can uncover causal loops that are primarily responsible for the observed behavior of the model, i.e., identify the "dominant" loop. The dominant loop is the most influential structure in determining the overall behavior of a system depending on the specific conditions of a system. It is possible for any loop to be the dominant loop at a point in time but then as conditions change the same loop can be displaced by another loop as the dominant loop in a different time frame. Due to the shifting dominance of the loops in determining system performance over time, it is necessary that a system is explored to isolate the interactions between the variables that form various causal loops. Clearly, collecting such information is challenging on many fronts. First, the sheer volume of data required to map a real world system is a challenge; secondly, this kind of information is often qualitative in nature (mental, experiential or judgment) and hence not easy to capture; and thirdly, the information keeps changing over time.

Viewing system performance as a series of connected dominant loop behaviors is a fundamentally different way to study a system. In effect, this point of view suggests that it may not be possible or necessary to find the "one best" single representation to describe the system's performance over time. Instead, we can now treat the system as a composite structure that may be formed by the amalgamation of a number of different sub representations that collectively describe the system performance. This perspective alleviates the unnecessary difficulty that is imposed on a single representation to capture the logic of possibly disconnected patterns. Indeed, this approach has its own challenges in terms of how to superimpose the various patterns to model reality.

Note both Bayesian and System Dynamics have very helpful roles to play in the analysis of real life systems that do not yield easily to purely hard data or classical models. Accordingly, one can consider an integrated approach where a Bayesian model provides specific input to a System Dynamics model to complement the capabilities of the two approaches. A System Dynamics model enhanced by Bayesian inference will allow modelers to iteratively incorporate various data types into a comprehensive model and study the behavior of a system over time. This approach allows for the inclusion of both hard data and soft data into the model. Since the modeling process is iterative, the subjective views can be augmented or replaced with hard data as such information is acquired and improved over time. When appropriate data are available, it can be used as input to the System Dynamics model of various contingencies, such as "fear" curves, "hope" curves, or mixtures of them from a Bayesian perspective. When such data are not available, varied contingencies can still be incorporated as subjective expert views, but with the advantage that sensitivity analyses can be done to measure the impact on the system

performance over time under different assumptions. One can test better which subjective views might lead to more realistic insights using a system dynamic model. Software that helps in such modeling includes Vensim, Powersim, and ithink; compare Sterman (2000).

Cross References

►Bayesian Statistics
►Stochastic Processes

References and Further Reading

Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge, MA
Forrester JW (1980) Information sources for modeling the national economy. J Am Stat Assoc 75(371):555–574
Savage LJ (1980) The writing of Leonard Jimmie savage – a memorial collection. The American Statistical Association and the Institute of Mathematical Statistics
Senge P (1990) The fifth discipline: the art and practice of the learning organization. Doubleday, Boston
Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, New York

Modeling Survival Data

Edward L. Melnick
Professor of Statistics
New York University, New York, NY, USA

►Survival Data are measurements in time from a well defined origin until a particular event occurs. The event is usually death (e.g., lifetime from birth to death), but it could also be a change of state (e.g., occurrence of a disease or time to failure of an electrical component).

Of central importance to the study of risk is the probability that a system will perform and maintain its function (remain in a state) during a specified time interval $(0, t)$. Let $F(t) = P(T \le t)$ be the cumulative distribution function for the probability that a system fails before time t and conversely $R(t) = 1 - F(t)$ be the survival function for the system. Data from survival studies are often censored (the system has not failed during the study) so that survival times are larger than censored survival times. For example, if the response variable is the lifetime of an individual (or component), then the censored data are represented as (y_i, δ_i) where the indicator variable δ is equal to 1 if the event occurred during the study, and 0 if the event occurred after the study; i.e., $t_i = y_i$ if $\delta_i = 1$ and $t_i > y_i$ if $\delta_i = 0$. Further, if $f(t)dt$ is the probability of failure in the infinitesimal interval $(t, t + dt)$, then rate of a failure among items that have survived to time t is

$$h(t) = \frac{f(t)}{R(t)} = \frac{-d \ln R(t)}{dt}. \quad (1)$$

The function $h(t)$ is called the hazard function and is the conditional probability of failure, conditioned upon survival up to time t. The log likelihood function of (y_i, δ_i) is

$$\ln L = \delta_i \ln f(y_i) + (1 - \delta_i) \ln R(y_i), \quad (2)$$

and the cumulative hazard rate is

$$H(t) = \int_0^t h(x)dx. \quad (3)$$

The survival rate, $R(t)$, is equivalent to $R(t) = \exp(-H(t))$.

Examining the hazard function, it follows that

1. If $h(t)$ increases with age, $H(t)$ is an increasing failure rate. This would be the case for an object that wears out over time.
2. If $h(t)$ decreases with age, $H(t)$ is a decreasing failure rate. Examples of these phenomena include infant mortality and burn-in periods for engines.
3. If $h(t)$ is constant with age, $H(t)$ is a constant failure rate. In this situation failure time does not depend on age.

Note that $h(t)$ is a conditional probability density function since it is the proportion of items in *service* that fail per unit time. This differs from the probability density function $f(t)$, which is the proportion of the *initial* number of items that fail per unit time.

Distributions for failure times are often determined in terms of their hazard function. The exponential distribution function has a constant hazard function. The lognormal distribution function with standard deviation greater than 1 has a hazard function that increases for small t, and then decreases. The lognormal hazard function for standard deviation less than 1 has maximum at $t = 0$ and is often used to describe length of time for repairs (rather than modeling times to failure).

The ►Weibull distribution is often used to describe failure times. Its hazard function depends on the shape parameter m. The hazard function decreases when $m < 1$, increases when $m > 1$ and is constant when $m = 1$. Applications for this model include structured components in a system that fails when the weakest components fail, and for failure experiences that follow a bathtub curve. A bathtub failure time curve (convex function) has three stages: decreasing (e.g., infant mortality), constant (e.g., useful region), and increasing (e.g., wear out region). This curve is formed by changing m over the three regions. The basic

Modeling Survival Data. Table 1 Basic probability functions used to model survival data

Parametric		
Name	**Cumulative distribution function**	**Hazard function**
Exponential	$F(t) = 1 - \exp(-\lambda t) \quad \lambda > 0$	λ
Weibull	$F(t) = 1 - \exp(-\lambda t^m) \quad \lambda > 0$	$m\lambda$
Gumbel	$F(t) = 1 - \exp(-m(\exp(\lambda t) - 1)) \quad \lambda, m > 0$	$m\lambda \exp(\lambda t)$
Gompertz	$F(t) = 1 - \exp\left(\frac{m}{\lambda}(1 - \exp(\lambda t))\right) \quad \lambda, m > 0$	$m \exp(\lambda t)$
Nonparametric		
[a]Piecewise constant rates of change		$\sum_{i=1}^{n} \lambda_i I\{t_{i-1} < t < t_i\}$
[b]Kaplan–Meier	$\hat{F}(t) = 1 - \prod_{t_i \le t} \left(1 - \frac{d_i}{r_i}\right)$	$\frac{d_i}{r_i(t_{i+1} - t_i)}$
[c]Nelson–Aalen		$\hat{H}(t) = \sum_{t_i \le t} \left(1 - \frac{d_i}{r_i}\right)$

[a]The time axis is split into intervals such that $t_1 < t_2 < \ldots < t_n$ resulting in a non-continuous hazard function with jumps at the interval end points. The notation $I\{A\}$ is 1 if an event occurs in interval A, and is zero otherwise.
[b]The set $t_i \le \ldots \le t_n$ are the ordered event times where r_i are the number of individuals at risk at time t_i and d_i are the total number of individuals either experiencing the event or were censored at time t_i.
[c]The Nelson-Aalen statistic is an estimate of the cumulative hazard rate. It is based on the Poisson distribution.

probability functions used to model ▶survival data are in Table 1. These distributions are left skewed with support on $(0, \infty)$ for continuous distributions and support on the counting numbers $(0, n]$ for discrete distributions.

Nonparametric approaches have also been developed for estimating the survival function. A first approach might be the development of an empirical function such as:

$$\widehat{R}(t) = \frac{\text{Number of individuals with event times} \ge t}{\text{Number of individuals in the data set}}. \quad (4)$$

Unfortunately, this estimate requires that there are no censored observations. For example, an individual whose survival time is censored before time t cannot be used when computing the empirical function at t. This issue is addressed by introducing the ▶Kaplan–Meier estimator [see Kaplan and Meier (1958)]. Further, the variance of the Kaplan–Meier statistic can be estimated and confidence intervals can be constructed based on the normal distribution. Closely related to the Kaplan-Meier estimator is the Nelson–Aalen estimator (Nelson 1972; Aalen 1978) of the cumulative hazard rate function. The estimated variance and confidence interval can also be computed for this function.

Although the models already discussed assume that the occurrences of hazards are independent and identically distributed, often there are known risk factors such as environmental conditions and operating characteristics that affect the quality of a system.

In many problems a researcher is not only interested in the probability of survival, but how a set of explanatory variables affect the survival rate. Cox (1972) proposed the proportional hazard model that allows for the presence of covariates and the partial likelihood estimation procedure for estimating the parameters in the model. The proportional hazard model is of the form:

$$\lambda(t|\underline{Z}) = \lambda_0(t) \exp(\underline{Z}^T \underline{\beta}) \quad (5)$$

where

$\lambda_0(t)$ is the hazard function of unspecified shape (the subscript 0 implies all covariates are zero at time t).

$\underline{Z}$ is a vector of risk factors measured on each individual.

$\underline{\beta}$ is a vector of parameters describing the relative risk associated with the factors.

$\lambda(t|\underline{Z})$ is the hazard function at time t conditioned on the covariates.

The proportional hazard model is semi-parametric because no assumptions are made about the base hazard function but the effect of the risk factors is assumed to be linear on the log of the hazard function; i.e., $\lambda_0(t)$ is an infinite dimensional parameter and $\underline{\beta}$ is finite dimensional.

The proportionality assumption implies that if an individual has a risk of an event twice that of another individual, then the level of risk will remain twice as high for all time. The usual application of the model is to study the effect of the covariates on risk when absolute risk is less important. For example, consider a system where two types of actions can be taken, let

$$Z = \begin{cases} 1 & \text{if the high risk action is taken} \\ 0 & \text{if the low risk action is taken} \end{cases}$$

and let β be the relative risk associated with Z. The relative risk of the two types of actions is computed from the hazard ratio:

$$\frac{\lambda(t|Z=1)}{\lambda(t|Z=0)} = \exp\beta, \tag{6}$$

the instantaneous risk conditioned on survival at time t. In this problem the model describes relative risks and removes the effect of time. In a more general context, the ratio of hazards is the difference of covariates assuming the intercept is independent of time.

In many applications $\lambda_0(t)$ is unknown and cannot be estimated from the data. For example, the proportional hazard model is often used in credit risk modeling for corporate bonds based on interest rates and market conditions. A nonparametric estimation procedure for the conditional proportional hazard function is based on the exponential regression model:

$$\frac{\lambda(t|Z)}{\lambda_0(t)} = \exp(\underline{Z}^T\underline{\beta})$$

where the underlying survival function is estimated with a Kaplan–Meier estimator, a measure of time until failure.

If, however, the absolute risk is also important (usually in prediction problems), then the Nelson–Aalen estimate is preferred over the Kaplan–Meier estimator. The state space time series model [see Commandeur and Koopman (2007)] is useful for predicting risk over time and by using the Kalman Filter, can also include time varying covariates.

The proportional hazard model assumes event times are independent, conditioned on the covariates. The ▶frailty model relaxes this assumption by allowing for the presence of unknown covariates (random effects model). In this model event times are conditionally independent when values are given for the frailty variable. A frailty model that describes unexplained heterogeneity resulting from unobserved risk factors has a hazard function of the form

$$\lambda_{T_{ji}}(t) = w_{ji}\lambda_0(t)\exp\left(\underline{Z}_i^T\underline{\beta}_i\right) \tag{7}$$

where

T_{ji} is the time to failure (event) j for individual i, and

w_{ji} is the frailty variable.

In this model the frailty variable is constant over time, is shared by subjects within a subgroup, and acts multiplicatively on the hazard rates of all members of the subgroup. The two sources of variation for this model are:

1. Individual random variation described by the hazard function.
2. Group variation described by the frailty variable.

The log likelihood function, Eq. 2, for this model can be expressed in simple form if the hazard function has a Gompertz distribution and the frailty variable has a ▶gamma distribution. Other commonly used distributions for the frailty variable are the gamma, compound Poisson, and the lognormal. Estimators for situations where the likelihood function does not have an explicit representation are derived from the penalized partial likelihood function or from algorithms such as EM or Gibbs sampling.

Survival models have also been extended to multivariate conditional frailty survival functions. In the univariate setting, frailty varies from individual to individual whereas in the multivariate setting, frailty is shared with individuals in a subgroup. Consider, for example, the multivariate survival function conditioned on the frailty variable w:

$$s(t_1,\ldots,t_k|w) = \exp\left[-w(\Lambda_1(t_1),\ldots,\Lambda_k(t_k))\right], \tag{8}$$

where $\Lambda_i(t_i)$ is the cumulative hazard rate for group i. By integrating over w, the survival function is:

$$s(t_1,\ldots,t_k) = E\exp\left[-w(\Lambda_1(t_1),\ldots,\Lambda_k(t_k))\right], \tag{9}$$

the Laplace transform of w. Because of the simplicity of computing derivatives from the Laplace transform, this method is often used to derive frailty distributions. The most often assumed distributions are those from the gamma family. See Hougaard (2008) for a complete discussion on modeling multivariate survival data.

Conclusion

This paper presents a discussion for analyzing and modeling time series survival data. The models are then extended to include covariates primarily based upon regression modeling, and finally generalized to include multivariate models. Current research is focused on the development of multivariate time series models for survival data.

About the Author

Edward Melnick is Professor of Statistics and former Chair of the Department of Statistics and Operations Research at

Leonard N. Stern School of Business, New York University. He is an editor (with Brian Everitt) of the four volume *Encyclopedia of Quantitative Risk Analysis and Assessment* (Wiley Blackwell 2008), "valuable reference work ... and a rather beautiful work" (David Hand, *International Statistical Review*, Volume 77, Issue 2, p. 314). The number and impact of his publications were recognized by the American Statistical Association (ASA) when he became Fellow of the ASA. He is also Fellow of the Royal Statistical Society, and Elected Member of the International Statistical Institute. He was Chairman of the Risk Analysis section of the American Statistical Association (2004). Professor Melnick has won 16 teaching awards at NYU including the NYU Distinguished Teaching Award. Currently, he is an Associate Editor of the *Journal of Forecasting*.

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
►Population Projections
►Statistical Inference in Ecology
►Survival Data
►Time Series Models to Determine the Death Rate of a Given Disease
►Weibull Distribution

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes, Ann Stat 6:701–726
Commandeur JJ F, Koopman SJ (2007) An introduction to state space time series analysis. Oxford University Press, Oxford
Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 74:187–220
Hougaard P (2000) Analysis of multivariate survival data. Springer, New York
Jia J, Dyer JS, Butler JC (1999) Measures of perceived risk. Manage Sci 45:519–532
Johnson N, Kotz S, Kemp A (1993) Univariate discrete distributions, 2nd edn. Wiley, New York
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
Nelson W (1972) Theory and applications of hazard plotting for censored failure data, Technometrics 14:945–965
Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton

Models for Z_+-Valued Time Series Based on Thinning

Emad-Eldin A. A. Aly
Professor
Kuwait University, Safat, Kuwait

Introduction

Developing models for integer-valued time series has received increasing attention in the past two decades. Integer-valued time series are useful in modeling dependent count data. They are also useful in the simulation of dependent discrete random variables with specified distribution and correlation structure.

Lawrance and Lewis (1977) and Gaver and Lewis (1980) were the first authors to construct autoregressive processes with non-Gaussian marginals. This has essentially motivated all the research on integer-valued time series. The present review is far from being exhaustive. Our focus is on models for Z_+-valued first-order autoregressive processes $INAR(1)$. We will consider five approaches which are based on "thinning" for developing these models.

First construction

To introduce integer-valued autoregressive moving average processes, McKenzie (1986, 1988) and Al-Osh and Alzaid (1987) used the binomial thinning operator $\odot$ of Steutel and van Harn (1979). The operation $\odot$ is defined as follows: if X is a Z_+–valued random variable (rv) and $\alpha \in (0,1)$, then

$$\alpha \odot X = \sum_{i=1}^{X} Y_i,$$

where $\{Y_i\}$ is a sequence of *i.i.d.* Bernoulli(α) $rv's$ independent of X. A sequence $\{X_n\}$ is said to be an $INAR(1)$ process if for any $n \in Z$,

$$X_n = \alpha \odot X_{n-1} + \varepsilon_n, \tag{1}$$

where $\odot$ is as in (1) and $\{\varepsilon_n\}$ is a sequence of *i.i.d.* Z_+–valued $rv's$ such that ε_n is independent of $\eta \odot X_{n-1}$ and the thinning $\eta \odot X_{n-1}$ is performed independently for each n. McKenzie (1986) constructed stationary Geometric

and Negative Binomial $INAR(1)$ processes and Al-Osh and Alzaid (1987) and independently McKenzie (1988) studied the Poisson $INAR(1)$ process.

Second Construction

Du and Li (1991) generalized the model (1) by introducing the $INAR(p)$ process

$$X_n = \sum_{i=1}^{p} \alpha_i \odot X_{n-i} + \varepsilon_n, \tag{2}$$

where all the thinning processes are independent and for $j < n$,

$$\operatorname{cov}(X_j, \varepsilon_n) = 0.$$

They proved that (2) has a unique stationary Z_+−valued solution $\{X_n\}_{n\in Z}$ if the roots of

$$\lambda^p - \sum_{i=1}^{p} \alpha_i \lambda^{p-i} = 0$$

are inside the unit circle. The main feature of the work of Du and Li (1991) is that it allows for models whose autocorrelation function (ACF) mimics that of the Normal *ARIMA* models.

Latour (1998) generalized Du and Li (1991) model by introducing the general $INAR(p)$ process ($GINAR(p)$),

$$X_n = \sum_{i=1}^{p} \alpha_i \circ X_{n-i} + \varepsilon_n,$$

where

$$\alpha_i \circ X_{n-i} = \sum_{i=1}^{X_{n-i}} Y_i^{(n,i)}$$

$\left\{Y_j^{(n,j)}\right\}$ is a sequence of nonnegative *i.i.d.rv's* independent of the X's with finite mean $\alpha_j > 0, j = 1, 2, \ldots, p$ and finite variance β_j and the innovation, ε_n, is assumed to have a finite mean μ_ε and finite variance σ_ε^2. Latour (1998) proved the existence of a stationary $GINAR(p)$ process if $\sum_{j=1}^{p} \alpha_j < 1$. He also showed that a stationary $GINAR(p)$ process, centered around its mean μ_X, admits a standard $AR(p)$ representation with the spectral density

$$f(\lambda) = \frac{\mu_X \sum_{j=1}^{p} \beta_j + \sigma_\varepsilon^2}{2\pi \left|\alpha\left(\exp(-i\lambda)\right)\right|^2}, \lambda \in [-\pi, \pi],$$

where

$$\alpha(t) = 1 - \sum_{j=1}^{p} \alpha_j t^j.$$

Third Construction

In the third approach the $INAR(1)$ stationary time series model takes the form

$$X_n = A_n(X_{n-1}, \eta) + \varepsilon_n, \tag{3}$$

where $\{\varepsilon_n\}$ are *i.i.d.r.v.'s* from the same family as the marginal distribution of $\{X_n\}$ and $A_n(X_{n-1}, \eta)$ is a random contraction operation performed on X_{n-1} which reduces it by the "amount η." Let $G_\theta(\cdot; \lambda_i)$ be the distribution of $Z_i, i = 1, 2$ and assume that Z_1 and Z_2 are independent and $G_\theta(\cdot; \lambda_1) * G_\theta(\cdot; \lambda_2) = G_\theta(\cdot; \lambda_1 + \lambda_2)$, where $*$ is the convolution operator. Let $G(\cdot; x, \lambda_1, \lambda_2)$ be the conditional distribution of Z_1 given $Z_1 + Z_2 = x$. The distribution of the random operator $A(X, \eta)$ given $X = x$, is defined as $G(\cdot; x, \eta\lambda, (1-\eta)\lambda)$. The distribution of $A(X, \eta)$ is $G_\theta(\cdot; \eta\lambda)$ when the distribution of X is $G_\theta(\cdot; \lambda)$. Now, if the distributions of X_0 and ε_1 are respectively $G_\theta(\cdot; \lambda)$ and $G_\theta(\cdot; (1-\eta)\lambda)$, then $\{X_n\}$ of (3) is stationary with marginal distribution $G_\theta(\cdot; \lambda)$. This construction was employed by Al-Osh and Alzaid (1991) for the Binomial marginal and Alzaid and Al-Osh (1993) for the Generalized Poisson marginal. This construction was generalized to the case when X_0 is infinitely divisible by Joe (1996) and to the case when X_0 is in the class of Exponential Dispersion Models by Jørgensen and Song (1998).

Fourth Construction

This construction is based on the expectation thinning operator $K(\eta)\circledast$ of Zhu and Joe (2003). The expectation thinning operator $K(\eta)\circledast$ is defined as follows: if X is a Z_+−valued *rv* and $\eta \in (0,1)$, then

$$K(\eta) \circledast X = \sum_{i=1}^{X} K_i(\eta),$$

where $K_i(\eta)$ are *i.i.d.r.v.'s* and the family $\{K(\alpha) : 0 \le \alpha \le 1\}$ is self-generalized, i.e., $E\{K(\eta) \circledast X | X = x\} = \eta x$ and $K(\eta') \circledast K(\eta) = K(\eta\eta')$. The corresponding $INAR(1)$ stationary time series model takes the form

$$X_n \overset{d}{=} K(\eta) \circledast X_{n-1} + \epsilon(\eta) = \sum_{i=1}^{X_{n-1}} K_i(\eta) + \epsilon(\eta).$$

The marginal distribution of X_n must be generalized discrete self-decomposable with respect to K, that is, $P_{X_n}(z)/P_{X_n}(P_{K(\alpha)}(z))$ must be a proper probability generating function (PGF) for every $\alpha \in [0,1]$. The ACF at lag$k$ is $\rho(k) = \eta^k$. The expectation thinning $K(\eta)\circledast$ governs the serial dependence. Several families of self-generalized *r.v.'s* $\{K(\eta)\}$ are known and the corresponding stationary distributions of $\{X_n\}$ are overdispersed with respect to Poisson (e.g., Generalized Poisson, Negative Binomial, Poisson-Inverse Gaussian). When a marginal distribution is possible for more than one self-generalized family then different $\{K(\eta)\}$ lead to differing amounts of conditional heteroscedasticity.

Fifth Construction

The fifth approach makes use of the thinning operator $\odot_{\mathcal{F}}$ of van Harn et al. (1982) and van Harn and Steutel (1993) which is defined as follows. Let $\mathcal{F} := (F_t, t \geq 0)$ be a continuous composition semigroup of PGF's such that $F_t(0) \neq 1, \delta = \delta(\mathcal{F}) = -\ln F_1'(1) > 0, F_{0+}(z) = z.$ and $F_{\infty-}(z) = 1.$ The infinitesimal generator U of $\mathcal{F}$ is given for $|z| \leq 1$ by

$$U(z) = \lim_{t \to 0+} \frac{F_t(z) - z}{t} = a\{H(z) - z\},$$

where a is a constant and $H(z) = \sum_{n=0}^{\infty} h_n z^n$ is a PGF of a Z_+ valued rv with $h_1 = 0$ and $H'(1) \leq 1$. For a Z_+ valued rv X and $\eta \in (0,1)$

$$\eta \odot_{\mathcal{F}} X = \sum_{i=1}^{X} Y_i,$$

where $\{Y_i\}$ is a sequence of *i.i.d.r.v.'s* independent of X with common PGF $F_{-\ln \eta} \in \mathcal{F}$. The corresponding $\mathcal{F}-$first order integer-valued autoregressive ($\mathcal{F}-INAR(1)$) model takes the form

$$X_n = \eta \odot_{\mathcal{F}} X_{n-1} + \varepsilon_n, \tag{4}$$

where $\{\varepsilon_n\}$ is a sequence of *i.i.d.* Z_+ valued *rv's* such that ε_n is independent of $\eta \odot_{\mathcal{F}} X_{n-1}$ and the thinning $\eta \odot_{\mathcal{F}} X_{n-1}$ is performed independently for each n. Note that $\{X_n\}$ is a Markov chain (see ▶Markov Chains). In terms of PGF's (4) reads

$$P_{X_n}(z) = P_{X_{n-1}}(F_{-\ln \eta}(z)) P_\varepsilon(z). \tag{5}$$

A distribution on Z_+ with PGF $P(z)$ is $\mathcal{F}$-self-decomposable (van Harn et al. (1982)) if for any t there exists a PGF $P_t(z)$ such

$$P(z) = P(F_t(z)) P_t(z).$$

Aly and Bouzar (2005) proved that any $\mathcal{F}$-self-decomposable distribution can arise as the marginal distribution of a stationary $\mathcal{F}-INAR(1)$ model. On assuming that the second moments of each of $H(\cdot), \varepsilon$ and X_n are finite for any $n \geq 0$, Aly and Bouzar (2005) proved that (1) the regression of X_n on X_{n-1} is linear, (2) the variance of X_n given X_{n-1} is linear, (3) the ACF at lag k, $\rho(X_{n-k}, X_n) = \eta^{\delta k} \sqrt{V(X_{n-k})/V(X_n)}$. Moreover, if $\{X_n\}$ is stationary, then $\rho(k) = \rho(X_{n-k}, X_n) = \eta^{\delta k}$.

We consider some important stationary time series models based on the composition semigroup

$$F_t^{(\theta)}(z) = 1 - \frac{\overline{\theta} e^{-\overline{\theta} t}(1-z)}{\overline{\theta} + \theta(1 - e^{-\overline{\theta} t})(1-z)}, t \geq 0, |z| \leq 1,$$
$$\overline{\theta} = 1 - \theta, 0 \leq \theta < 1$$

of van Harn et al. (1982). Note that when $\theta = 0$, $F_t^{(0)}(z) = 1 - e^{-t} + e^{-t} z$ and the corresponding thinning is the Binomial thinning of Steutel and van Harn (1979). In this case (4) becomes

$$P_X(z) = P_X(1 - \eta + \eta z) P_\varepsilon(z). \tag{6}$$

Particular $INAR(1)$ of (6) are the Poisson (Al-Osh and Alzaid 1987; McKenzie 1988), the Geometric and the Negative Binomial (McKenzie 1986), the Mittag-Leffler (Pillai and Jayakumar 1995) and the discrete Linnik (Aly and Bouzar 2000). Particular $INAR(1)$ time series models when $0 < \theta < 1$ are the Geometric, the Negative Binomial and the Poisson Geometric (Aly and Bouzar 1994) and the Negative Binomial (Al-Osh and Aly 1992).

Remarks

We mention some methods of parameter estimation. The most direct approach is using moment estimation based on the Yule-Walker equations. The conditional least squares method with some modifications, e.g., a two-stage procedure, in order to be able to estimate all the parameters (see, for example, Brännäs and Quoreshi 2004) may be used. Joe and Zhu (2006) used the method of maximum likelihood after using a recursive method to calculate the probability mass function of the innovation. Neal and Subba Rao (2007) used the MCMC approach for parameter estimation. For additional references on parameter estimation we refer to Brännäs (1994), Jung and Tremayne (2006), Silva and Silva (2009) and the references contained therein. Finally, we note that Hall and Scotto (2006) studied the extremes of integer-valued time series.

About the Author

Dr Emad-Eldin A. A. Aly is a Professor since 1994 at the Department of Statistics and Operations Research, Kuwait University, Kuwait. He was the Chair of the Department (2002–2006), and the Vice Dean for Academic Affairs of the Faculty of Graduate Studies, Kuwait University (1996–2002). He was a Faculty member at The University of Alberta, Edmonton, Alberta, Canada (1984–1995) and the Chair of the Department of Statistics and Applied Probability, The University of Alberta (1991–1994). He has authored and co-authored more than 75 papers. He was an Associate Editor of the *Journal of Nonparametric Statistics*. He was awarded (jointly with Professor A. Alzaid of King Saud University) the 1995 Kuwait Prize of the Kuwait Foundation for the Advancement of Sciences for his research in Mathematical Statistics.

Cross References

►Box–Jenkins Time Series Models
►Generalized Quasi-Likelihood (GQL) Inferences
►Time Series

References and Further Reading

Al-Osh MA, Aly E-EAA (1992) First order autoregressive time series with negative binomial and geometric marginals. Commun Statist Theory Meth 21:2483–2492

Al-Osh MA, Alzaid A (1987) First order integer-valued autoregressive (INAR(1)) process. J Time Ser Anal 8:261–275

Al-Osh MA, Alzaid A (1991) Binomial autoregressive moving average models. Commun Statist Stochastic Models 7:261–282

Aly E-EAA, Bouzar N (1994) Explicit stationary distributions for some Galton Watson processes with immigration. Commun Statist Stochastic Models 10:499–517

Aly E-EAA, Bouzar N (2000) On geometric infinite divisibility and stability. Ann Inst Statist Math 52:790–799

Aly E-EAA, Bouzar N (2005) Stationary solutions for integer-valued autoregressive processes. Int J Math Math Sci 1:1–18

Alzaid AA, Al-Osh MA (1993) Some autoregressive moving average processes with generalized Poisson marginal distributions. Ann Inst Statist Math 45:223–232

Brännäs K (1994) Estimation and testing in integer-valued AR(1) models. Umeå Economic Studies No. 335

Brännäs K, Quoreshi AMMS (2004) Integer-valued moving average modeling of the number of transactions in stocks. Umeå Economic Studies No. 637

Du JG, Li Y (1991) The integer-valued autoregressive INAR(p) model. J Time Ser Anal 12:129–142

Gaver DP, Lewis PAW (1980) First-order autoregressive gamma sequences and point processes. Adv Appl Probab 12:724–745

Hall A, Scotto MG (2006) Extremes of periodic integer-valued sequences with exponential type tails Revstat 4:249–273

Joe H (1996) Time series models with univariate margins in the convolution-closed infinitely divisible class. J Appl Probab 33:664–677

Jørgensen B, Song PX-K (1998) Stationary time series models with exponential dispersion model margins. J Appl Probab 35:78–92

Jung RC, Tremayne AR (2006) Binomial thinning models for integer time series. Statist Model 6:81–96

Latour A (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive process. J Time Ser Anal 19:439–455

Lawrance AJ, Lewis PAW (1977) An exponential moving average sequence and point process, EMA(1). J Appl Probab 14:98–113

McKenzie E (1986) Autoregressive-moving average processes with negative binomial and geometric marginal distributions. Adv Appl Probab 18:679–705

McKenzie E (1988) Some ARMA models for dependent sequences of Poisson counts. Adv Appl Probab 20:822–835

Neal P, Subba Rao T (2007) MCMC for integer valued ARMA Models. J Time Ser Anal 28:92–110

Pillai RN, Jayakumar K (1995) Discrete Mittag-Leffler distributions. Statist Probab Lett 23:271–274

Silva I, Silva ME (2009) Parameter estimation for INAR processes based on high-order statistics. Revstat 7:105–117

Steutel FW, van Harn K (1979) Discrete analogues of self-decomposability and stability. Ann Probab 7:893–899

van Harn K, Steutel FW (1993) Stability equations for processes with stationary independent increments using branching processes and Poisson mixtures. Stochastic Process Appl 45:209–230

van Harn K, Steutel FW, Vervaat W (1982) Self-decomposable discrete distributions and branching processes. Z Wahrsch Verw Gebiete 61:97–118

Zhu R, Joe H (2003) A new type of discrete self-decomposability and its application to continuous-time Markov processes for modelling count data time series. Stochastic Models 19:235–254

Zhu R, Joe H (2006) Modelling count data time series with Markov processes based on binomial thinning. J Time Ser Anal 27: 725–738

Moderate Deviations

Jayaram Sethuraman
Robert O. Lawton Distinguished Professor, Professor Emeritus
Florida State University, Tallahassee, FL, USA

M

Moderate Deviations

Consider the familiar simple set up for the central limit theorem (CLT, see ►Central Limit Theorems). Let $X_1, X_2, \ldots$ be independently and identically distributed real random variables with common distribution function $F(x)$. Let $Y_n = \frac{1}{n}(X_1 + \cdots + X_n), n = 1, 2, \ldots$. Suppose that

$$\int xF(dx) = 0, \int x^2 F(dx) = l \tag{1}$$

Then the central limit theorem states that

$$P\left(|Y_n| > \frac{a}{\sqrt{n}}\right) \to 2[1 - \Phi(a)] \tag{2}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt$ and $a > 0$.

In other words, the CLT gives an approximation to the two-sided deviation of size $\frac{a}{\sqrt{n}}$ of Y_n and the approximation is a number in $(1/2, 1)$. Deviations of the this type are called *ordinary deviations*.

However, one needs to study deviations larger than ordinary deviations to understand finer properties of the distributions of Y_n and to approximate expectations of other functions of Y_n. Thus a deviation of magnitude λ_n will be called a *excessive deviation* if $n\lambda_n^2 \to \infty$. In the particular case of $\lambda_n = \lambda$ where λ is a constant, it is called a *large deviation* (see also ►Large Deviations and Applications).

The following, due to Cramér (1938), Chernoff (1952), Bahadur and Rao (1960), etc., is a classical result on large deviations. Let

$$\int \exp(tx)F(dx) < \infty \text{ for } t \text{ in some neighborhood of } 0. \tag{3}$$

Then

$$\frac{1}{n}\log P(|Y_n| > \lambda) \to -I(\lambda) \tag{4}$$

where

$$I(\lambda) = \sup_t(t\lambda - \log\phi(t)) \tag{5}$$

and $0 < I(\lambda) \leq \infty$. This result is usually read as "the probability of large deviations tends to zero exponentially." For sequences of random variables $\{Y_n\}$ distributed in more general spaces like $R^k, C([0,1]), D([0,1])$, etc. (i.e., ▶stochastic processes), there is no preferred direction for deviations. The appropriate generalization of the large deviation result (4) is the *large deviation principle*, which states that for all Borel sets A

$$-I(A^0) \leq \underline{\lim}_n \frac{1}{n}\log P(Y_n \in A) \leq -I(\overline{A}) \tag{6}$$

where $A^0, \overline{A}$ denote the interior and closure of A, and

$$I(A) = \inf_{\lambda \in A} I(\lambda) \tag{7}$$

for some function $I(\lambda)$ whose level sets $\{\lambda : I(\lambda) \leq K\}$ are compact for $K < \infty$. The function $I(x)$ is called the *large deviation rate function*.

When the moment generating function condition (3) holds, Cramér (1938) has further shown that

$$P(|Y_n| > \lambda_n) \sim \frac{2}{\sqrt{2\pi n\lambda_n^2}}\exp\left(\frac{-n\lambda_n^2}{2}\right) \tag{8}$$

when $n\lambda_n^3 \to 0$ and $n\lambda_n^2 \to \infty$. This excludes large deviations ($\lambda_n = \lambda$), but it gives a rate for the probability (and not just the logarithm of the probability) of a class of excessive deviations and is therefore called a *strong excessive deviation result*.

Rubin and Sethuraman (1965a) called deviations λ_n with $\lambda_n = c\sqrt{\frac{\log n}{n}}$ where c is a constant as *moderate deviations*. Moderate deviations found their first applications in Bayes risk efficiency which was introduced in Rubin and Sethuraman (1965b). Cramér's result in (8) reduces to

$$P(|Y_n| > c\sqrt{\frac{\log n}{n}}) \sim \frac{2}{c\sqrt{2\pi\log n}}n^{-c^2/2} \tag{9}$$

and holds under the moment generating function condition (3). Rubin and Sethuraman (1965a) showed that the moderate deviation result (9) holds under the weaker condition

$$E(|X_1|^{c^2+2+\delta}) < \infty \text{ for some } \delta > 0. \tag{10}$$

They also showed that when (9) holds we have

$$E(|X_1|^{c^2+2-\delta}) < \infty \text{ for all } \delta > 0. \tag{11}$$

Slastnikov (1978) showed that the strong moderate deviation result (9) if and only if

$$\lim_{t\to\infty} t^{2+c}(\log(t))^{-(1+c)/2}P(|X_1| > t) = 0. \tag{12}$$

Since (8) was called a strong excessive deviation result, we should call (9) as a *strong moderate deviation result*. Analogous to the logarithmic large deviation result (4) is the *logarithmic moderate deviation result* which states that

$$\frac{1}{\log(n)}\log P(|Y_n| \geq c\sqrt{\frac{\log(n)}{n}}) \sim n^{-c^2/2} \tag{13}$$

which may be the only possible result for more complicated random variables $\{Y_n\}$ than are not means of i.i.d. random variables,

For random variables $\{Y_n\}$ which take values in $R^k, C([0,1]), D([0,1])$ etc., we can, under some conditions, establish the *moderate deviation principle* which states

$$-J(A^0) \leq \underline{\lim}_n \frac{1}{\log(n)}P(\sqrt{\frac{n}{\log(n)}}Y_n \in A) \leq -J(\overline{A}) \tag{14}$$

where $J(A) = \inf_{x\in A} J(x)$ for some function $J(x)$ whose level sets are compact. The function $J(x)$ is then called the *moderate deviation rate function*. This is analogous to the large deviation principle (6).

Following the paper of Rubin and Sethuraman (1965a), there is a vast literature on moderate deviations for a large class of random variables $\{Y_n\}$ that arise in a multitude of contexts. The asymptotic distribution of $\{Y_n\}$ can be more general than Gaussian. We will give just a brief summary below.

We stated the definition of two-sided moderate deviations and quoted Slastnikov's necessary and sufficient condition. One can also consider one-sided moderate deviations results and the necessary and sufficient conditions are slightly different and these are given in Slastnikov (1978). Without assuming á priori that the mean and variance of the i.i.d. random variables $X_1, X_2 \ldots$ are 0 and 1 respectively, one can ask for necessary and sufficient conditions for moderate deviations. This problem has been completely addressed in Amosova (1979). Another variant of moderate deviations has been studied in Davis (1968).

The case where $\{Y_n\}$ is the sum of triangular arrays of independent random variables or a U-statistic were begun in Rubin and Sethuraman (1965). Ghosh (1974) studied moderate deviations for sums of m-dependent random variables. Michel (1974) gave results on rates of convergence in the strong moderate deviation result (9). Gut (1980) considered moderate deviations for random variables with multiple indices. Dembo (1996) considered moderate deviations for ►martingales.

Moderate deviations in general topological spaces with applications in Statistical Physics and other areas can be found in Borovkov and Mogulskii (1978), (1980), Deo and Babu (1981), De Acosta (1992), Liming (1995), Djellout and Guillin (2001).

About the Author

Professor Jayaram Sethuraman earned a Ph.D. in statistics from the Indian Statistical Institute in 1962. Professor Sethuraman has received many recognitions for his contributions to the discipline of statistics: the U.S. Army S. S. Wilks Award (1994), the Teaching Incentive Program Award, FSU (1995), the Professorial Excellence Award, FSU (1996), an ASA Service Award (2001), the President's Continuing Education Award, FSU (2002), and the Bhargavi and C. R. Rao Prize, Pennsylvania State University (2005).

"Sethuraman has been a superior researcher throughout his career, making important contributions in many areas including asymptotic distribution theory, large deviations theory, moderate deviations theory for which he was the pioneer, limit theory, nonparametric statistics, Dirichlet processes and Bayesian nonparametrics, stopping times for sequential estimation and testing, order statistics, stochastic majorization, Bahadur and Pitman efficiency, Markov chain Monte Carlo, reliability theory, survival analysis and image analysis." (Myles Hollander (2008). A Conversation with Jayaram Sethuraman, *Statistical Science* 23, 2, 272–285).

Cross References

►Central Limit Theorems
►Estimation: An Overview
►Large Deviations and Applications
►Prior Bayes: Rubin's View of Statistics
►Statistics on Ranked Lists

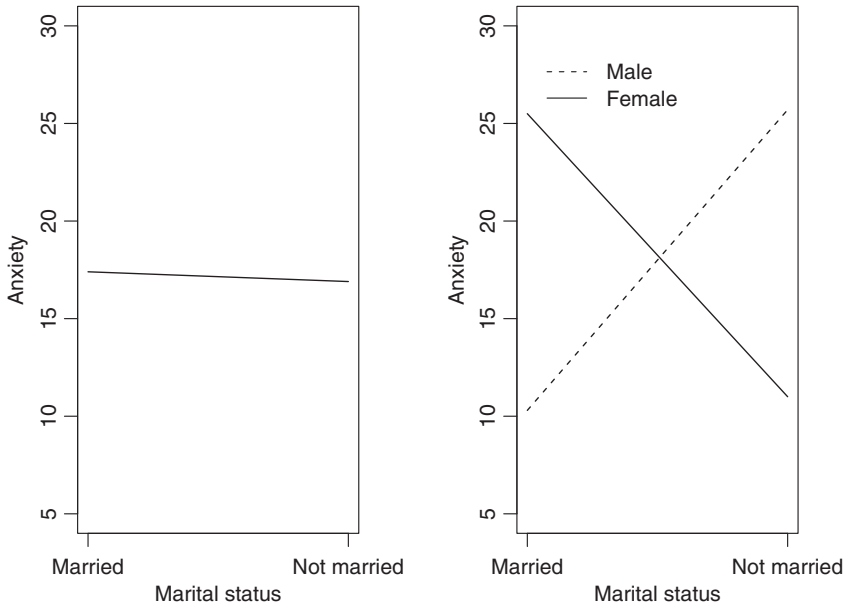
References and Further Reading

Borovkov AA, Mogulskii AA (1978) Probabilities of large deviations in topological vector space I. Siberian Math J 19:697–709
Borovkov AA, Mogulskii AA (1980) Probabilities of large deviations in topological vector space II. Siberian Math J 21:12–26
Cramér H (1938) Sur un nouveau théorème limite de la probabilités. Actualites Sci Indust 736:5–23
Davis AD (1968) Convergence rates for probabilities of moderate deviations. Ann Math Statist 39:2016–2028
De Acosta A (1992) Moderate deviations and associated Laplace approximations for sums of independent random vectors. Trans Am Math Soc 329:357–375
Dembo A (1996) Moderate deviations for martingales with bounded jumps. Elec Comm Probab 1:11–17
Deo CM, Babu JG (1981) Probabilities of moderate deviations in a Banach space. Proc Am Math Soc 24:392–397
Djellout H, Guillin A (2001) Moderate deviations for Markov chains with atom. Stoch Proc Appl 95:203–217
Gao FQ (2003) Moderate deviations and large deviations for kernel density estimators. J Theo Probab 16:401–418
Ghosh M (1974) Probabilities of moderate deviations under m-dependence. Canad J Statist 2:157–168
Gut A (1980) Convergence rates for probabilities of moderate deviations for sums of random variables with multidimensional indices. Ann Probab 8:298–313
Liming W (1995) Moderate deviations of dependent random variables related to CLT. Ann Probab 23:420–445
Michel R (1974) Results on probabilities of moderate deviations. Ann Probab 2:349–353
Rubin H, Sethuraman J (1965a) Probabilities of moderate deviations. Sankhya Ser A 27:325–346
Rubin H, Sethuraman J (1965b) Bayes risk efficiency. Sankhya Ser A 27:347–356
Slastnikov AD (1978) Limit theorems for moderate deviation probabilities. Theory Probab Appl 23:322–340

Moderating and Mediating Variables in Psychological Research

Petar Milin[1], Olga Hadžić[2]
[1]Associate Professor
University of Novi Sad, Novi Sad, Serbia
[2]Professor
University of Novi Sad, Novi Sad, Serbia

Moderating and mediating variables, or simply *moderators* and *mediators*, are related but distinct concepts in both general statistics and its application in psychology. A moderating variable is a variable that affects the relationship between two other variables. This effect is usually referred to as an *interaction*. The simplest case of an interaction can occur in ►analysis of variance (ANOVA).



Moderating and Mediating Variables in Psychological Research. Fig. 1 The main effect of one categorical variable on a continuous dependent variable (*left-hand panel*), and how it is moderated by the third categorical variable (*right-hand panel*)

For example, we tested whether there is a significant difference in the *level of anxiety* (as measured with an appropriate standardized psychological test) between married and unmarried participants (i.e., variable *marital status*). The effect was not statistically significant. However, when we enter the third variable – *gender* (female/male) – it appears that, on average, unmarried males are significantly more anxious than married males, while for females the effect is the reverse. Figure 1 represents the results from two models described above. In the left-hand panel, we can see that, on average, there are no differences between married and unmarried participants in the level of anxiety. From the right-hand panel, we can conclude that gender moderates the effect of marital status on the level of anxiety: married males and unmarried females are significantly less anxious than the other two groups (unmarried males and married females).

We can generalize the previous example to more complex models, with two independent variables having more than just two levels for comparison, or even with more than two independent variables. If all variables in the model are continuous variables, we would apply multiple regression analysis, but the phenomenon of a moderating effect would remain the same, in essence. For example, we confirmed a positive relationship between the *hours of learning* and the *result in an assessment test*. Yet, *music loudness* during learning can moderate test results. We can imagine this as if a hand on the volume knob of an amplifier rotates clockwise and turns the volume up, students get all the worse results the longer they learn. Depending on the music volume level, the relationship between the hours of learning and the knowledge assessment changes continuously. This outcome is presented in Fig. 2. On the left-hand side, we can observe a positive influence of the hours of learning on the results in the assessment test, while on the right-hand side, we can see how music loudness moderates this relationship.

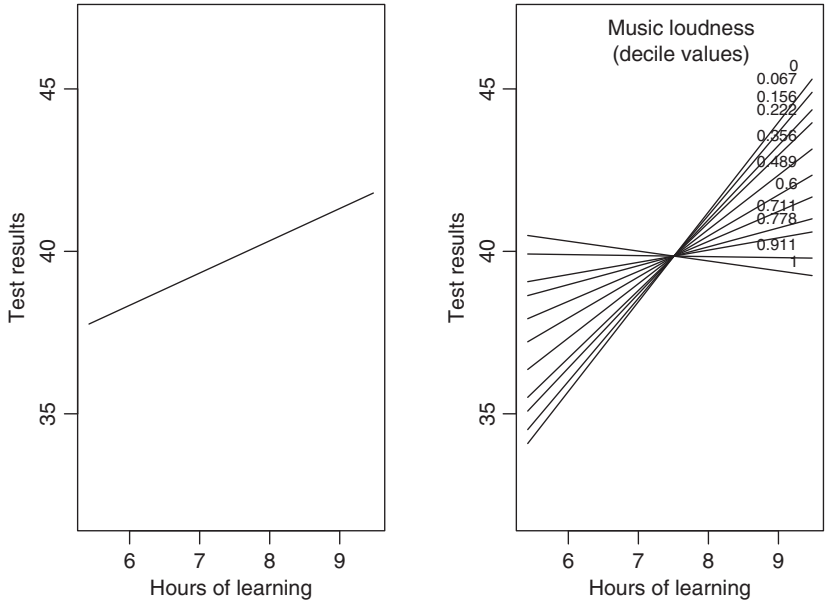
The general linear form with one dependent, one independent, and one moderating variable is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon,$$

where β_3 evaluates the interaction between X_1 and X_2.

Mediating variables typically emerge in multiple regression analysis, where the influence of some independent variable (*predictor*) on the dependent variable (*criterion*) is not direct, but mediated through the third variable. For example, the correlation between *ageing* and the *number of work accidents* in the car industry appears to be strong and negative. Nevertheless, the missing link in this picture is *work experience*: it affects injury rate, and is itself affected by the age of worker.

In regression modeling, one can distinguish between *complete mediation* and *incomplete mediation*. In practice, if the effects of ageing on the number of work injuries



Moderating and Mediating Variables in Psychological Research. Fig. 2 The main effect of one continuous variable on another (*left-hand panel*), and how it is moderated by a third continuous variable (*right-hand panel*). Lines on the right panel represent decile values for the moderator variable

would not differ statistically from zero when work experience is included in the model, then mediation is complete. Otherwise, if this effect still exists (in the statistical sense), then mediation is incomplete. Complete and incomplete mediation are presented in Fig. 3.

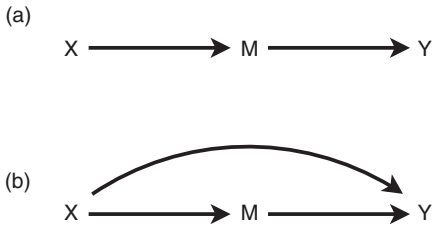
In principle, a mediating variable flattens the effect of an independent variable on the dependent variable. The opposite phenomenon would occur if the mediator variable would increase the effect. This is called *suppression*. It is a controversial concept in statistical theory and practice, but contemporary applied approaches take a more neutral position, and consider that suppression may provide better insights into the relationships between relevant variables.

The simplest case of linear regression with one dependent, one independent, and one mediating variable is defined by the following equations:

$$Y = \beta_0 + \beta_1 X + \varepsilon_1$$
$$M = \gamma_0 + \gamma_1 X + \varepsilon_2$$
$$Y = \beta_0' + \beta_1' X + \beta_2 M + \varepsilon_3,$$

where of particular interest are β_1, which is called the *total effect*, and β_1', named the *direct effect*. If suppression does not take place, which would occur if $\beta_1' > \beta_1$, then we can continue the analysis with a standard regression model. First, we ascertain whether mediation is complete or incomplete, depending on whether the direct effect drops to zero ($\beta_1' \approx 0$). The most important step in the analysis is the inference about the *indirect effect*, or the *amount of mediation*. It is defined as the reduction in the effect of the initial variable on the model outcome ($\beta_1 - \beta_1'$). In simple hierarchical regression models, the difference of the coefficients is exactly the same as the product of the effect of the independent variable on the mediating variable multiplied by the effect of the mediating variable on the dependent variable. In the general case, this equality only approximately holds.

Mediation and moderation can co-occur in statistical models. This is often the case in psychology. *Mediated moderation* takes place when the independent variable is actually an interaction ($X = X_A \times X_B$). Thus, the mediator acts between interacting variables (X_A and X_B) and the dependent variable (Y). For example, the effect of interacting variable *hours of learning* and *music loudness* on the dependent variable *result in an assessment test* can be mediated by the *importance of the test*, as rated by the participants. Conversely, *moderated mediation* is realized in two forms: (a) the effect of the independent variable on the mediator is affected by a moderator (γ_1 varies; as if the effect of *ageing* on *work experience* is moderated by a particular personality trait, like H. J. Eysenck's *Neuroticism*), or (b) a moderator may interact with the mediating variable (β_2 varies; as if the *work experience* and the *level of anxiety* would interact and mediate between *ageing* and *number of*



Moderating and Mediating Variables in Psychological Research. Fig. 3 Schematic representation of a complete mediation effect (panel a, *upper*), and an incomplete mediation effect (panel b, *lower*)

work accidents). If moderated mediation exists, inference about its type must be given.

Finally, special attention is required in moderation and mediation analyses since both can be influenced by ►*multicollinearity*, which makes estimates of regression coefficients unstable. In addition, in an analysis with a moderating term – i.e., an interaction effect – the product of the variables can be strongly related to either the independent or the moderating variable, or both of them. If two variables are collinear, one of them can be centred to its mean. In this way, half of its value will become negative, and consequently, collinearity will decrease. Another possibility is to regress the independent variable with a moderator or mediator, and then to use the *residuals* or unexplained values, of the independent variable in the main analysis. Thus, the independent variable will be orthogonal to the moderating or mediating variable, with zero correlation, which will bring collinearity under control. However, in applying the previous two remedies, and others that are available, one must choose a conservative approach. The risk of emphasizing, or even inventing, what is not present in the data ought to be as little as possible. In any circumstances, the ultimate way of securing more reliable estimates is simply to obtain enough data.

Acknowledgment

We would like to thank Professor David Kenny for reading a draft of this article, and providing us with comments and suggestions which resulted in many improvements.

About the Author

Dr. Olga Hadzic is Professor, Department of Mathematics and Informatics, University of Novi Sad, Serbia. She is an Elected Member of the Serbian Academy of Sciences and Arts (since 1984). Her research interests are in fixed point theory, functional analysis, probability theory, and organizational psychology. She has (co-)authored about 180 scientific papers, 5 monographs, and 4 textbooks, including, *Fixed Point Theory in Probabilistic Metric Spaces* (with Endre Pap, Kluwer Academic Publishers, Dordrecht 2001). Professor Hadzic was Rector (Chancellor) of the University of Novi Sad (1996–1998). She was an external adviser for two Ph.D. theses defended abroad.

Cross References

►Analysis of Variance
►Interaction
►Linear Regression Models
►Multilevel Analysis
►Psychology, Statistics in
►Variables

References and Further Reading

Baron R, Kenny D (1986) The moderator-mediator variable distinction in social psychological research – conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173–1182

Eysenck H (2006) The biological basis of personality. Transaction Publishers, London

Friedman L, Wall M (2005) Graphical views of suppression and multicollinearity in multiple linear regression. Am Stat 59(2): 127–136

Hayes A, Matthes J (2009) Computational procedures for probing interactions in ols and logistic regression: SPSS and SAS implementations. Behav Res Meth 41(3):924–936

Judd C, Kenny D, McClelland G (2001) Estimating and testing mediation and moderation in within-participant designs. Psychol Meth 6(2):115–134

Muller D, Judd C, Yzerbyt V (2005) When moderation is mediated and mediation is moderated. J Pers Soc Psychol 89(6):852–863

Shrout P, Bolger N (2002) Mediation in experimental and nonexperimental studies: new procedures and recommendations. Psychol Meth 7(4):422–445

Moment Generating Function

Jan Beran[1], Sucharita Ghosh[2]
[1]Professor
University of Konstanz, Konstanz, Germany
[2]Scientific Staff Member
Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

The moment generating function (mgf) of a real valued random variable X with distribution $F(x) = P(X \leq x)$ is defined by

$$M_X(t) = E\left[e^{tX}\right] = \int e^{tx} dF(x). \quad (1)$$

For distributions with a density function $f = F'$, M_X can also be interpreted as a (two-sided) Laplace transform of f. In order that M_X exists and is finite for $t \in (-a, a)$ and some $a > 0$, all moments $\mu_j = E\left[X^j\right]$ must be finite and such that $\sum \mu_j t^j/j!$ is a convergent series. We then have

$$M_X(t) = \sum_{j=0}^{\infty} \frac{\mu_j}{j!} t^j \tag{2}$$

so that

$$\mu_j = M_X^{(j)}(0) = \frac{d^j}{dt^j} M_X(t) \mid_{t=0} \tag{3}$$

which explains the name moment generating function. A counter example where M_X does not exist in any open neighborhood of the origin is the Cauchy distribution, since there even μ_1 is not defined. The lognormal distribution is an example where all μ_j are finite but the series in (2) does not converge. In cases where $X > 0$ and $M_X(t) = \infty$ for $t \neq 0$, the mgf of $-X$ may be used (see e.g., Severini (2005) for further results). Related to M_X are the characteristic function $\phi_X(t) = M_X(it)$ and the probability generating function $H_X(z) = E(z^X)$ for which $M_X(t) = H_X(e^t)$. Note however that, in contrast to M_X, $\phi_X(t) = E\left[\exp(itX)\right]$ always exists. A furhter important function is the cumulant generating function $K_X(t) = \log M_X(t)$ which can be written as power series

$$K_X(t) = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j \tag{4}$$

where κ_j are cumulants. The first two cumulants are $\kappa_1 = \mu = E(X)$ and $\kappa_2 = \sigma^2 = var(X)$. In contrast to the raw moments μ_j, higher order cumulants κ_j ($j \geq 3$) do not depend on the location μ and scale σ^2. For vector valued random variables $X = (X_1, ..., X_k)' \in \mathbb{R}^k$, M_X is defined in an analogous manner by $M_X(t) = E\left[\exp\left(t'X\right)\right] = E\left[\exp\left(\sum_{j=1}^k t_j X_j\right)\right]$. This implies

$$\frac{\partial^{j_1+j_2+\cdots+j_k}}{\partial t_1^{j_1} \partial t_1^{j_2} \cdots \partial t_1^{j_k}} M_X(0) = E\left[X_1^{j_1} X_2^{j_2} \cdots X_k^{j_k}\right] \tag{5}$$

and corresponding expressions for joint cumulants as derivatives of K_X. In particular,

$$\frac{\partial^2}{\partial t_i \partial t_j} K_X(0) = cov(X_i, X_j). \tag{6}$$

An important property is uniqueness: if $M_X(t)$ exists and is finite in an open interval around the origin, then there is exactly one distribution function with this moment generating function. For instance, if $\kappa_j = 0$ for $j \geq 3$, then $X \in \mathbb{R}$ is normally distributed with expected value $\mu = \kappa_1$ and variance $\sigma^2 = \kappa_2$. The moment generating function is very practical when handling sums of independent random variables. If X and Y are independent with existing moment generating function, then $M_{X+Y}(t) = M_X(t)M_Y(t)$ (and vice versa). For the cumulant generating function this means $K_{X+Y}(t) = K_X(t) + K_Y(t)$. For limit theorems, the following result is useful: Let X_n be a sequence of random variables with moment generating functions $M_{X_n}(t)$ which converge to the moment generating function $M_X(t)$ of a random variable X. Then X_n converges to X in distribution. This together with the additivity property of the cumulant generating function can be used for a simple proof of the central limit theorem (see ►Central Limit Theorems).

Moment Generating Function. Table 1 $M_X(t)$ for some important distributions

Distribution	$M_X(t)$
Binomial with n trials, success probability $p = 1 - q$	$[q + pe^t]^n$
Geometric distribution with success probability $p = 1 - q$	$pe^t\,(1 - qe^t)^{-1}$
Poisson with expected value λ	$\exp\left[\lambda\,(e^t - 1)\right]$
Uniform on $[a, b]$	$t^{-1}(b-a)^{-1}\left(e^{tb} - e^{ta}\right)$
Normal $N(\mu, \sigma^2)$	$\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$
Multivariate Normal $N(\mu, \Sigma)$	$\exp\left(\mu' t + \frac{1}{2} t' \Sigma t\right)$
Chi-square χ_k^2	$(1 - 2t)^{-\frac{k}{2}}$
Exponential with expected value λ^{-1}	$\left(1 - t\lambda^{-1}\right)^{-1}$
Cauchy distribution	not defined

The empirical counterparts of M_X, K_X and ϕ_X, defined by

$$m_n(t) = n^{-1} \sum_{i=1}^{n} \exp(tX_i), \tag{7}$$

$k_n(t) = \log m_n(t)$ and $\varphi_n(t) = \log m_n(it)$, are often useful for statistical inference. For instance, testing the null hypothesis that X and Y are independent can be done by testing $M_{X+Y} \equiv M_X M_Y$ or $\varphi_{X+Y} \equiv \varphi_X \varphi_Y$ (see e.g., Csörgő 1985; Feuerverger 1987). Testing normality of a random sample $X_1, ..., X_n$ is the same as testing $H_o : \partial^3/\partial t^3 K_X(t) \equiv 0$ (see Ghosh 1996; Fang et al. 1998). For further applications of empirical moment and cumulant generating functions see e.g., Csörgő (1982, 1986), Epps et al. (1982),

Feuerverger (1989), Feuerverger and McDunnough (1984), Knight and Satchell (1997), Ghosh and Beran (2000, 2006).

Cross References

►Bivariate Distributions
►Financial Return Distributions
►Random Variable
►Statistical Distributions: An Overview
►Univariate Discrete Distributions: An Overview

References and Further Reading

Csörgő S (1982) The empirical moment generating function. In: Gnedenko BV, Puri ML, Vincze I (eds) Nonparametric statistical inference: Coll Math Soc J Bolyai, 32, Amsterdam, North-Holland, pp 139–150

Csörgő S (1985) Testing for independence by the empirical characteristic function. J Multivariate Anal 16(3):290–299

Csörgő S (1986) Testing for normality in arbitrary dimension. Ann Stat 14:708–723

Epps TW, Singleton KJ, Pulley LB (1982) A test of separate families of distributions based on the empirical moment generating function. Biometrika 69:391–399

Fang K-T, Li R-Z, Liang J-J (1998) A multivariate version of Ghosh's T3-plot to detect non-multinormality. Comput Stat Data Anal 28:371–386

Feuerverger A (1987) On some ECF procedures for testing independence. In: MacNeill IB, Umphrey GJ, Festschrift J (eds) Time series and econometric modeling, Reidel, New York, pp 189–206

Feuerverger A (1989) On the empirical saddlepoint approximation. Biometrika 76(3):457–464

Feuerverger A, McDunnough P (1984) On statistical transform methods and their efficiency. Can J Stat 12:303–317

Ghosh S (1996) A new graphical tool to detect non-normality. J Roy Stat Soc B 58:691–702

Ghosh S, Beran J (2000) The two-sample T3 test – a graphical method for comparing two distributions. J Comput Graph Stat 9(1):167–179

Ghosh S, Beran J (2006) On estimating the cumulant generating function of linear processes. Ann Inst Stat Math 58:53–71

Knight JL, Satchell SE (1997) The cumulant generating function estimation method: implementation and asymptotic efficiency. Economet Theor 13(2):170–184

Severini TA (2005) Elements of distribution theory. Cambridge University Press, Cambridge

Monte Carlo Methods in Statistics

Christian Robert
Professor of Statistics
Université Paris-Dauphine, CEREMADE, Paris, France

Monte Carlo methods are now an essential part of the statistician's toolbox, to the point of being more familiar to graduate students than the measure theoretic notions upon which they are based! We recall in this note some of the advances made in the design of Monte Carlo techniques towards their use in Statistics, referring to Robert and Casella (2004, 2010) for an in-depth coverage.

The Basic Monte Carlo Principle and Its Extensions

The most appealing feature of Monte Carlo methods [for a statistician] is that they rely on sampling and on probability notions, which are the bread and butter of our profession. Indeed, the foundation of Monte Carlo approximations is identical to the validation of empirical moment estimators in that the average

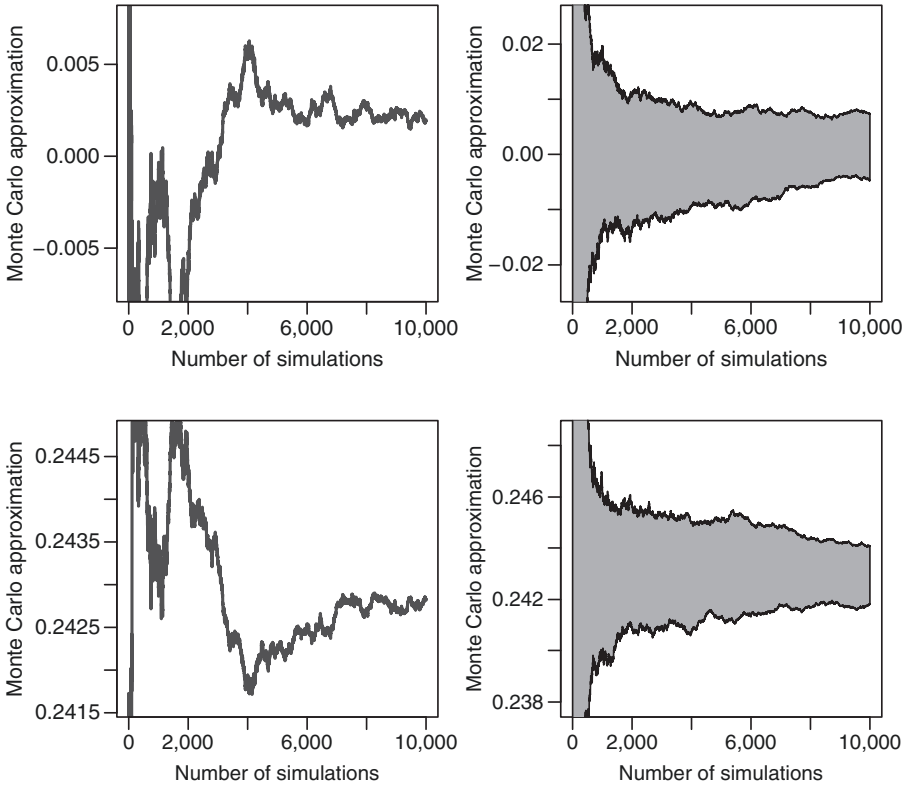
$$\frac{1}{T}\sum_{t=1}^{T} h(x_t), \quad x_t \sim f(x), \tag{1}$$

is converging to the expectation $\mathbb{E}_f[h(X)]$ when T goes to infinity. Furthermore, the precision of this approximation is exactly of the same kind as the precision of a statistical estimate, in that it usually evolves as $O(\sqrt{T})$. Therefore, once a sample $x_1, \ldots, x_T$ is produced according to a distribution density f, all standard statistical tools, including bootstrap (see ►Bootstrap Methods), apply to this sample (with the further appeal that more data points can be produced if deemed necessary). As illustrated by Fig. 1, the variability due to a single Monte Carlo experiment must be accounted for, when drawing conclusions about its output and evaluations of the overall variability of the sequence of approximations are provided in Kendall et al. (2007). But the ease with which such methods are analyzed and the systematic resort to statistical intuition explain in part why Monte Carlo methods are privileged over numerical methods.

The representation of integrals as expectations $\mathbb{E}_f[h(X)]$ is far from unique and there exist therefore many possible approaches to the above approximation. This range of choices corresponds to the importance sampling strategies (Rubinstein 1981) in Monte Carlo, based on the obvious identity

$$\mathbb{E}_f[h(X)] = \mathbb{E}_g[h(X)f(X)/g(X)]$$

provided the support of the density g includes the support of f. Some choices of g may however lead to appallingly poor performances of the resulting Monte Carlo estimates, in that the variance of the resulting empirical average may be infinite, a danger worth highlighting since often neglected while having a major impact on the quality of the approximations. From a statistical perspective, there exist some natural choices for the importance function



Monte Carlo Methods in Statistics. Fig. 1 Monte Carlo evaluation (1) of the expectation $\mathbb{E}[X^3/(1+X^2+X^4)]$ as a function of the number of simulation when $X \sim \mathcal{N}(\mu,1)$ using (*left*) one simulation run and (*right*) 100 independent runs for (*top*) $\mu = 0$ and (*bottom*) $\mu = 2.5$

g, based on Fisher information and analytical approximations to the likelihood function like the Laplace approximation (Rue et al. 2008), even though it is more robust to replace the normal distribution in the Laplace approximation with a t distribution. The special case of Bayes factors (Andrieu et al. 2005) (Andrieu et al. 2005)

$$B_{01}(x) = \int_\Theta f(x|\theta)\pi_0(\theta)\mathrm{d}\theta \Big/ \int_\Theta f(x|\theta)\pi_1(\theta)\mathrm{d}\theta,$$

which drive Bayesian testing and model choice, and of their approximation has led to a specific class of importance sampling techniques known as *bridge sampling* (Chen et al. 2000) where the optimal importance function is made of a mixture of the posterior distributions corresponding to both models (assuming both parameter spaces can be mapped into the same Θ). We want to stress here that an alternative approximation of marginal likelihoods relying on the use of *harmonic means* (Gelfand and Dey 1994; Newton and Raftery 1994) and of direct simulations from a posterior density has repeatedly been used in the literature, despite often suffering from infinite variance (and thus numerical instability). Another potentially very efficient approximation of Bayes factors is provided by Chib's (1995) representation, based on parametric estimates to the posterior distribution.

MCMC Methods

Markov chain Monte Carlo (MCMC) methods (see ►Markov Chain Monte Carlo) have been proposed many years (Metropolis et al. 1953) before their impact in Statistics was truly felt. However, once Gelfand and Smith (1990) stressed the ultimate feasibility of producing a Markov chain (see ►Markov Chains) with a given stationary distribution f, either via a Gibbs sampler that simulates each conditional distribution of f in its turn, or via a Metropolis–Hastings algorithm based on a proposal $q(y|x)$ with acceptance probability [for a move from x to y]

$$\min\left\{1, f(y)q(x|y)\big/f(x)q(y|x)\right\},$$

then the spectrum of manageable models grew immensely and almost instantaneously.

Due to parallel developments at the time on graphical and hierarchical Bayesian models, like generalized linear mixed models (Zeger and Karim 1991), the wealth of multivariate models with available conditional distributions (and hence the potential of implementing the Gibbs sampler) was far from negligible, especially when the availability of latent variables became quasi universal due to the slice sampling representations (Damien et al. 1999; Neal 2003). (Although the adoption of Gibbs samplers has primarily taken place within ►Bayesian statistics, there is nothing that prevents an artificial augmentation of the data through such techniques.)

For instance, if the density $f(x) \propto \exp(-x^2/2)/(1 + x^2 + x^4)$ is known up to a normalizing constant, f is the marginal (in x) of the joint distribution $g(x,u) \propto \exp(-x^2/2)\mathbb{I}(u(1 + x^2 + x^4) \leq 1)$, when u is restricted to $(0,1)$. The corresponding slice sampler then consists in simulating

$$U|X = x \sim \mathcal{U}(0, 1/(1 + x^2 + x^4))$$

and

$$X|U = u \sim \mathcal{N}(0,1)\mathbb{I}(1 + x^2 + x^4 \leq 1/u),$$

the later being a truncated normal distribution. As shown by Fig. 2, the outcome of the resulting Gibbs sampler perfectly fits the target density, while the convergence of the expectation of X^3 under f has a behavior quite comparable with the iid setting.

While the Gibbs sampler first appears as *the* natural solution to solve a simulation problem in complex models if only because it stems from the true target f, as exhibited by the widespread use of BUGS (Lunn et al. 2000), which mostly focus on this approach, the infinite variations offered by the Metropolis–Hastings schemes offer much more efficient solutions when the proposal $q(y|x)$ is appropriately chosen. The basic choice of a random walk proposal (see ►Random Walk) $q(y|x)$ being then a normal density centered in x) can be improved by exploiting some features of the target as in Langevin algorithms (see Andrieu et al. 2005 Sect. 7.8.5) and Hamiltonian or hybrid alternatives (Duane et al. 1987; Neal 1999) that build upon gradients. More recent proposals include particle learning about the target and sequential improvement of the proposal (Douc et al. 2007; Rosenthal 2007; Andrieu et al. 2010). Fig. 3 reproduces Fig. 2 for a random walk Metropolis–Hastings algorithm whose scale is calibrated towards an acceptance rate of 0.5. The range of the convergence paths is clearly wider than for the Gibbs sampler, but the fact that this is a generic algorithm applying to any target (instead of a specialized version as for the Gibbs sampler) must be borne in mind.

Another major improvement generated by a statistical imperative is the development of variable dimension generators that stemmed from Bayesian model choice requirements, the most important example being the reversible jump algorithm in Green (1995) which had a significant impact on the study of graphical models (Brooks et al. 2003).

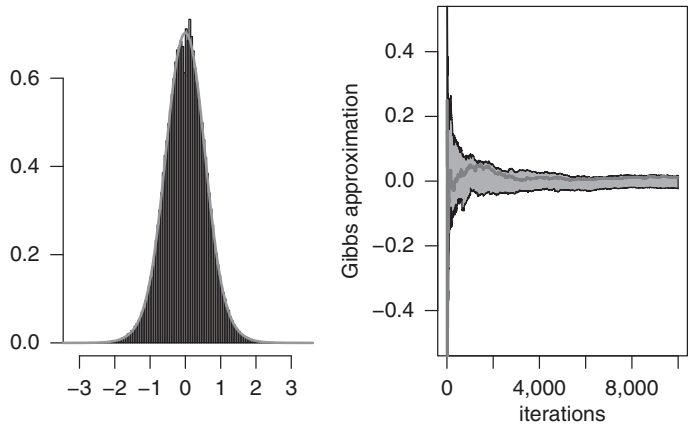
Some Uses of Monte Carlo in Statistics

The impact of Monte Carlo methods on Statistics has not been truly felt until the early 1980s, with the publication of Rubinstein (1981) and Ripley (1987), but Monte Carlo methods have now become invaluable in Statistics because they allow to address optimization, integration and exploration problems that would otherwise be unreachable. For instance, the calibration of many tests and the derivation of their acceptance regions can only be achieved by simulation techniques. While integration issues are often linked with the Bayesian approach – since Bayes estimates are posterior expectations like

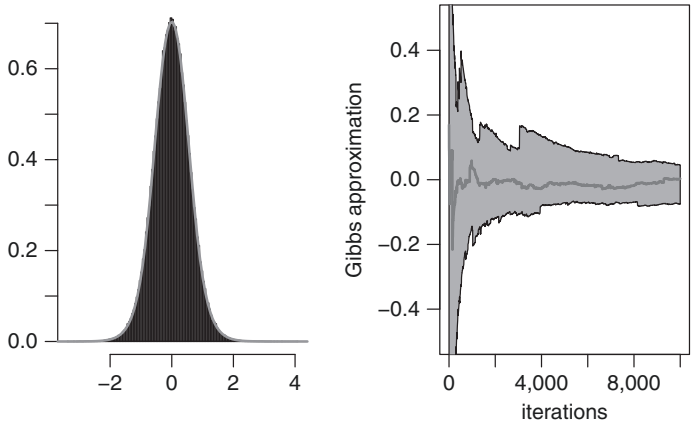
$$\int h(\theta)\pi(\theta|x)\,\mathrm{d}\theta$$

and Bayes tests also involve integration, as mentioned earlier with the Bayes factors, and optimization difficulties with the likelihood perspective, this classification is by no way tight – as for instance when likelihoods involve unmanageable integrals – and all fields of Statistics, from design to econometrics, from genomics to psychometry and environmics, have now to rely on Monte Carlo approximations. A whole new range of statistical methodologies have entirely integrated the simulation aspects. Examples include the bootstrap methodology (Efron 1982), where multilevel resampling is not conceivable without a computer, indirect inference (Gouriéroux et al. 1993), which construct a pseudo-likelihood from simulations, MCEM (Cappé and Moulines 2009), where the E-step of the EM algorithm is replaced with a Monte Carlo approximation, or the more recent approximated Bayesian computation (ABC) used in population genetics (Beaumont et al. 2002), where the likelihood is not manageable but the underlying model can be simulated from.

In the past fifteen years, the collection of real problems that Statistics can [afford to] handle has truly undergone a quantum leap. Monte Carlo methods and in particular MCMC techniques have forever changed the emphasis from "closed form" solutions to algorithmic ones, expanded our impact to solving "real" applied problems while convincing scientists from other fields that statistical solutions were indeed available, and led us into a world



Monte Carlo Methods in Statistics. Fig. 2 (*left*) Gibbs sampling approximation to the distribution $f(x) \propto \exp(-x^2/2)/(1+x^2+x^4)$ against the true density; (*right*) range of convergence of the approximation to $\mathbb{E}_f[X^3] = 0$ against the number of iterations using 100 independent runs of the Gibbs sampler, along with a single Gibbs run



Monte Carlo Methods in Statistics. Fig. 3 (*left*) Random walk Metropolis–Hastings sampling approximation to the distribution $f(x) \propto \exp(-x^2/2)/(1+x^2+x^4)$ against the true density for a scale of 1.2 corresponding to an acceptance rate of 0.5; (*right*) range of convergence of the approximation to $\mathbb{E}_f[X^3] = 0$ against the number of iterations using 100 independent runs of the Metropolis–Hastings sampler, along with a single Metropolis–Hastings run

where "exact" may mean "simulated." The size of the data sets and of the models currently handled thanks to those tools, for example in genomics or in climatology, is something that could not have been conceived 60 years ago, when Ulam and von Neumann invented the Monte Carlo method.

Acknowledgments

Supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75775 Paris) through the 2009–2012 project ANR-08-BLAN-0218 Big'MC. The author is grateful to Jean-Michel Marin for helpful comments.

About the Author

Dr. Christian P. Robert is Professor of Statistics in the Department of Mathematics, Université Paris-Dauphine, and Head of the Statistics Laboratory, Centre de Recherche en Economie et Statistique, Institut National de la Statistique et des Études Économiques (INSEE), Paris, France. He has authored and co-authored more than 130 papers and 9 books, including *The Bayesian Choice* (Springer Verlag, 2001), which received the DeGroot Prize in 2004, *Monte Carlo Statistical Methods* with George Casella (Springer Verlag, 2004), *Bayesian Core* with Jean-Michel Marin (Springer Verlag, 2007), and *Introducing Monte Carlo Methods with R* with George Casella (Springer Verlag, 2009). He was President of the International

Society for Bayesian Analysis (ISBA) in 2008. He is an IMS Fellow (1996) and an Elected member of the Royal Statistical Society (1998). Professor Robert has been the Editor of the *Journal of the Royal Statistical Society Series* (2005–2009) and an Associate Editor for *Annals of Statistics* (1998–2006), the *Journal of the American Statistical Society* (1996–1999 and 2005–2008), *Annals of the Institute of Statistical Mathematics* (2003–2005), *Statistical Science* (2000–2004), *Bayesian Analysis* (2003–2005), *TEST* (1994–1997 and 2000–2003), and *Sankhya* (1999–2002 and 2010).

Cross References

►Bootstrap Methods
►Computational Statistics
►Copulas: Distribution Functions and Simulation
►Entropy and Cross Entropy as Diversity and Distance Measures
►Frequentist Hypothesis Testing: A Defense
►Markov Chain Monte Carlo
►Multivariate Statistical Simulation
►Non-Uniform Random Variate Generations
►Numerical Integration
►Sensitivity Analysis
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Modeling of Financial Markets
►Uniform Distribution in Statistics
►Uniform Random Number Generators

References and Further Reading

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo (with discussion). J Roy Stat Soc B 72:269–342
Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035
Brooks S, Giudici P, Roberts G (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). J Roy Stat Soc B 65:3–55
Cappé O, Moulines E (2009) On-line expectation-maximization algorithm for latent data models. J Roy Stat Soc B, 71(3):593–613
Chen M, Shao Q, Ibrahim J (2000) Monte Carlo methods in Bayesian computation. Springer, New York
Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321
Damien P, Wakefield J, Walker S (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. J Roy Stat Soc B 61:331–344
Douc R, Guillin A, Marin J-M, Robert C (2007) Convergence of adaptive mixtures of importance sampling schemes. Ann Stat 35(1):420–448
Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett B 195:216–222
Efron B (1982) The Jacknife, the Bootstrap and other resampling plans, vol 38. SIAM, Philadelphia
Gelfand A, Dey D (1994) Bayesian model choice: asymptotics and exact calculations. J Roy Stat Soc B 56:501–514
Gelfand A, Smith A (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
Gouriéroux C, Monfort A, Renault E (1993) Indirect inference. J Appl Econom 8:85–118
Green P (1995) Reversible jump MCMC computation and Bayesian model determination. Biometrika 82:711–732
Kendall W, Marin J-M, Robert C (2007) Confidence bands for Brownian motion and applications to Monte Carlo simulations. Stat Comput 17:1–10
Lunn D, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Neal R (1999) Bayesian learning for neural networks, vol 118. Springer, New York
Neal R (2003) Slice sampling (with discussion). Ann Statist 31: 705–767
Newton M, Raftery A (1994) Approximate Bayesian inference by the weighted likelihood boostrap (with discussion). J Roy Stat Soc B 56:1–48
Ripley B (1987) Stochastic simulation. Wiley, New York
Rober C, Casella G (2004) Monte Carlo statistical methods. 2nd ed. Springer-Verlag, New York
Robert C, Casella G (2010) Introducing Monte Carlo methods with *R*. Springer, New York
Rosenthal J (2007) AMCM: an *R* interface for adaptive MCMC. Comput Stat Data Anal 51:5467–5470
Rubinstein R (1981) Simulation and the Monte Carlo method. Wiley, New York
Rue H, Martino S, Chopin N (2008) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). J Roy Stat Soc B 71(2):319–392
Zeger S, Karim R (1991) Generalized linear models with random effects; a Gibbs sampling approach. J Am Stat Assoc 86:79–86

Monty Hall Problem : Solution

Richard D. Gill
Professor, Faculty of Science, President of the Dutch Society for Statistics and Operations Research
Leiden University, Leiden, Netherlands

Introduction

The *Three Doors Problem*, or *Monty Hall Problem*, is familiar to statisticians as a paradox in elementary probability theory often found in elementary probability texts (especially in their exercises sections). In that context it is usually meant to be solved by careful (and elementary) application of ►Bayes' theorem. However, in different forms, it is much discussed and argued about and written

about by psychologists, game-theorists and mathematical economists, educationalists, journalists, lay persons, blog-writers, wikipedia editors.

In this article I will briefly survey the history of the problem and some of the approaches to it which have been proposed. My take-home message to you, dear reader, is that one should distinguish two levels to the problem.

There is an informally stated problem which you could pose to a friend at a party; and there are many concrete *versions* or *realizations* of the problem, which are actually the result of mathematical or probabilistic or statistical *modeling*. This modeling often involves adding supplementary assumptions chosen to make the problem well posed in the terms of the modeler. The modeler finds those assumptions perfectly natural. His or her students are supposed to guess those assumptions from various key words (like: "indistinguishable," "unknown") strategically placed in the problem re-statement. Teaching statistics is often about teaching the students to read the teacher's mind. Mathematical (probabilistic, statistical) modeling is, unfortunately, often solution driven rather than problem driven.

The very same criticism can, and should, be leveled at this very article! By cunningly presenting the history of *The Three Doors Problem* from my rather special point of view, I have engineered complex reality so as to convert the *Three Doors Problem* into an illustration of my personal Philosophy of Science, my Philosophy of Statistics.

This means that I have re-engineered the *Three Doors Problem* into an example of the point of view that Applied Statisticians should always be wary of the lure of *Solution-driven Science*. Applied Statisticians are trained to know Applied Statistics, and are trained to know how to convert real world problems into statistics problems. That is fine. But the best Applied Statisticians know that Applied Statistics is not the only game in town. Applied Statisticians are merely some particular kind of Scientists. They know lots about modeling uncertainty, and about learning from more or less random data, but probably not much about anything else. The Real Scientist knows that there is not a universal *disciplinary* approach to every problem. The *Real Statistical Scientist* modestly and persuasively and realistically offers what his or her discipline has to offer in synergy with others.

To summarize, we must distinguish between:

(0) the *Three-Doors-Problem Problem* [sic], which is to make sense of some real world question of a real person.
(1) a large number of solutions to this *meta*-problem, i.e., the many Three-Doors-Problem *Problems*, which are competing mathematizations of the meta-problem (0).

Each of the solutions at level (1) can well have a number of different solutions: nice ones and ugly ones; correct ones and incorrect ones. In this article, I will discuss three level (1) solutions, i.e., three different Monty Hall problems; and try to give three short correct and attractive solutions.

Now read on. Be critical, use your intellect, don't believe anything on authority, and certainly not on mine. Especially, don't forget the problem at meta-level (–1), not listed above.

C'est la vie.

Starting Point

I shall start not with the historical roots of the problem, but with the question which made the Three Doors Problem famous, even reaching the front page of the *New York Times*.

Marilyn vos Savant (a woman allegedly with the highest IQ in the world) posed the *Three Door Problem* or *Monty Hall Problem* in her "Ask Marilyn" column in *Parade* magazine (September 1990:16), as posed to her by a correspondent, a Mr. Craig Whitaker. It was, quoting vos Savant literally, the following:

> *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?*

Apparently, the problem refers to a real American TV quiz-show, with a real presenter, called Monty Hall.

The literature on the Monty Hall Problem is enormous. At the end of this article I shall simply list two references which for me have been especially valuable: a paper by Jeff Rosenthal (2008) and a book by Jason Rosenhouse (2009). The latter has a huge reference list and discusses the pre- and post-history of vos Savant's problem.

Briefly regarding the pre-history, one may trace the problem back through a 1975 letter to the editor in the journal *The American Statistician* by biostatistician Steve Selkin, to a problem called *The Three Prisoners Problem* posed by Stephen Gardner in his Mathematical Games column in *Scientific American* in 1959, and from there back to *Bertrand's Box Problem* in his 1889 text on Probability Theory. The internet encyclopedia `wikipedia.org` discussion pages (in many languages) are a fabulous though every-changing resource. Almost everything that I write here was learnt from those pages.

Despite making homage here to the two cited authors Rosenthal (2008) and Rosenhouse (2009) for their wonderful work, I emphasize that I strongly disagree with

both Rosenhouse ("the canonical problem") and Rosenthal ("the original problem") on what the essential Monty Hall problem is. I am more angry with certain other authors, who will remain nameless but for the sake of argument I'll just call Morgan et al. for unilaterally declaring in *The American Statistician* in 1981 *their* Monty Hall problem to be the only possible sensible problem, for calling everyone who solved different problems stupid, and for getting an incorrect theorem (I refer to their result about the situation when we do not know the quiz-master's probability of opening a particular door when he has a choice, and put a uniform prior on this probability.) published in the peer-reviewed literature.

Deciding unilaterally (Rosenhouse 2009) that a certain formulation is *canonical* is asking for a schism and for excommunication. Calling a particular version *original* (Rosenthal 2008) is asking for a historical contradiction. In view of the pre-history of the problem, the notion is not well defined. Monty Hall is part of folk-culture, culture is alive, the Monty Hall problem is not owned by a particular kind of mathematician who looks at such a problem from a particular point of view, and who adds for them "natural" extra assumptions which merely have the role of allowing their solution to work. Presenting any "canonical" or "original" Monty Hall problem together with a solution, is an example of *solution driven science* – you have learnt a clever trick and want to show that it solves lots of problems.

Three Monty Hall Problems

I will concentrate on three different particular Monty Hall problems. One of them (Q-0) is simply to answer the question literally posed by Marilyn vos Savant, "would you switch?". The other two (Q-1, Q-2) are popular mathematizations, particularly popular among experts or teachers of elementary probability theory: one asks for the unconditional probability that "always switching" would gets the car, the other asks for the conditional probability given the choices made so far. Here they are:

Q-0: Marilyn vos Savant's (or Craig Whitaker's) question "*Is it to your advantage to switch?*"

Q-1: A mathematician's question "*What is the unconditional probability that switching gives the car?*"

Q-2: A mathematician's question "*What is the conditional probability that switching gives the car, given everything so far?*"

The free, and freely editable, internet encyclopedia Wikipedia is the scene of a furious debate as to which mathematization Q-1 or Q-2 is the right starting point for answering the verbal question Q-0 (to be honest, many of the actors claim another "original" question as *the* original question). Alongside that, there is a furious debate as to which supplementary conditions are obviously implicitly being made. For each protagonist in the debate, those are the assumptions which ensure that his or her question has a unique and nice answer. My own humble opinion is "neither Q-1 nor Q-2, though the unconditional approach comes closer." I prefer Q-0, and I prefer to see it as a question of *game theory* for which, to my mind, [almost] no supplementary conditions need to be made.

Here I admit that I will suppose that the player knows game-theory and came to the quiz-show prepared. I will also suppose that the player wants to get the Cadillac while Monty Hall, the quizmaster, wants to keep it.

My analysis below of both problems Q-1 and Q-2 yields the good answer "2/3" under minimal assumptions, and almost without computation or algebraic manipulation. I will use Israeli (formerly Soviet Union) mathematician Boris Tsirelson's proposal on Wikipedia talk pages to use symmetry to deduce the conditional probability from the unconditional one. (Boris graciously gave me permission to cite him here, but this should not be interpreted to mean that anything written here also has his approval).

You, the reader, may well prefer a calculation using Bayes' theorem, or a calculation using the definition of conditional probability; I think this is a matter of taste.

I finally use a game-theoretic point of view, and von Neumann's minimax theorem, to answer the question Q-0 posed by Marilyn vos Savant, on the assumptions just stated.

Let the three doors be numbered in advance 1, 2, and 3. I add the universally agreed (and historically correct) additional assumptions: Monty Hall knows in advance where the car is hidden, Monty Hall always opens a door revealing a goat.

Introduce four random variables taking values in the set of door-numbers $\{1, 2, 3\}$:

C: the quiz-team hides the Car (a Cadillac) behind door C,
P: the Player chooses door P,
Q: the Quizmaster (Monty Hall) opens door Q,
S: Monty Hall asks the player if she'ld like to Switch to door S.

Because of the standard story of the Monty Hall show, we certainly have:

$Q \neq P$, the quizmaster *always* opens a door different to the player's first choice,
$Q \neq C$, opening that door *always* reveals a goat,
$S \neq P$, the player is *always* invited to switch to another door,
$S \neq Q$, *no* player wants to go home with a goat.

It does not matter for the subsequent mathematical analysis whether probabilities are subjective (Bayesian) or objective (frequentist); nor does it matter whose probabilities they are supposed to be, at what stage of the game. Some writers think of the player's initial choice as fixed. For them, P is degenerate.

I simply merely down some mathematical assumptions and deduce mathematical consequences of them.

Solution to Q-1: Unconditional Chance That Switching Wins

By the rules of the game and the definition of S, if $P \neq C$ then $S = C$, and vice-versa. A "switcher" would win the car if and only if a "stayer" would lose it. Therefore:

If $\Pr(P = C) = 1/3$ *then* $\Pr(S = C) = 2/3$, *since the two events are complementary.*

Solution to Q-2: Probability Car is Behind Door 2 Given You Chose Door 1, Monty Hall Opened 3

First of all, suppose that P and C are uniform and independent, and that given (P, C), suppose that Q is uniform on its possible values (unequal to those of P and of C). Let S be defined as before, as the third door-number different from P and Q. The joint law of C, P, Q, S is by this definition invariant under renumberings of the three doors. Hence $\Pr(S = C|P = x, Q = y)$ is the same for all $x \neq y$. By the law of total probability, $\Pr(S = C)$ (which is equal to 2/3 by our solution to Q-1) is equal to the weighted average of all $\Pr(S = C|P = x, Q = y), x \neq y \in \{1, 2, 3\}$. Since the latter are all equal, all these six conditional probabilities are equal to their average 2/3.

Conditioning on $P = x$, say, and letting y and y' denote the remaining two door numbers, we find the following corollary:

Now take the door chosen by the player as fixed, $P \equiv 1$, say. We are to compute $\Pr(S = C|Q = 3)$. *Assume that all doors are equally likely to hide the car and assume that the quizmaster chooses completely at random when he has a choice.* Without loss of generality we may as well pretend that P was chosen in advance completely at random. Now we have embedded our problem into the situation just solved, where P and C are uniform and independent.

▶ *If* $P \equiv 1$ *is fixed,* C *is uniform, and* Q *is symmetric, then "switching gives car" is independent of quizmaster's choice, hence*

$$\Pr(S = C|Q = 3) = \Pr(S = C|Q = 2') = \Pr(S = C) = 2/3.$$

Some readers may prefer a direct calculation. Using Bayes' theorem in the form "posterior odds equal prior odds times likelihoods" is a particularly efficient way to do this. The probabilities and conditional probabilities below are all conditional on $P = 1$, or if your prefer with $P \equiv 1$.

We have uniform prior odds

$$\Pr(C = 1) : \Pr(C = 2) : \Pr(C = 3) = 1 : 1 : 1.$$

The likelihood for C, the location of the car, given data $Q = 3$, is (proportional to) the discrete density function of Q given C (and P)

$$\Pr(Q = 3|C = 1) : \Pr(Q = 3|C = 2) : \Pr(Q = 3|C = 3) = \frac{1}{2} : 1 : 0.$$

The posterior odds are therefore proportional to the likelihood. It follows that the posterior probabilities are

$$\Pr(Q = 3|C = 1) = \frac{1}{3}, \quad \Pr(Q = 3|C = 2) = \frac{2}{3},$$
$$\Pr(Q = 3|C = 3) = 0.$$

Answer to Marilyn Vos Savant's Q-0: Should *You* Switch Doors?

Yes. Recall, *You only know that Monty Hall always opens a door revealing a goat.* You didn't know what strategy the quiz-team and quizmaster were going to use for their choices of the distribution of C and the distribution of Q given P and C, so naturally (since you know elementary Game Theory) you had picked your door uniformly at random. Your strategy of choosing C uniformly at random guarantees that $\Pr(C = P) = 1/3$ and hence that $\Pr(S = C) = 2/3$.

It was easy for you to find out that this combined strategy, which I'll call "symmetrize and switch," is your so-called minimax strategy.

On the one hand, "symmetrize and switch" guarantees you a 2/3 (unconditional) chance of winning the car, whatever strategy used by the quizmaster and his team.

On the other hand, if the quizmaster and his team use their "symmetric" strategy "hide the car uniformly at random and toss a fair coin to open a door if there is choice", then you cannot win the car with a *better* probability than 2/3.

The fact that your "symmetrize and switch" strategy gives you "at least" 2/3, while the quizmaster's "symmetry" strategy prevents you from doing better, proves that these are the respective minimax strategies, and 2/3 is the game-theoretic value of this two-party zero-sum game. (Minimax strategies and the accompanying "value" of the game exist by virtue of John von Neumann's (1929) minimax theorem for finite two-party zero-sum games).

There is not much point for you in worrying about your conditional probability of winning conditional on

your specific initial choice and the specific door opened by the quizmaster, say doors 1 and 3 respectively. You don't know this conditional probability anyway, since you don't know the strategy used by quiz-team and the quizmaster. (Even though you know probability theory and game theory, they maybe don't). However, it is maybe comforting to learn, by easy calculation, that if the car is hidden uniformly at random, then your conditional probability cannot be *smaller* than 1/2. So in that case at least, it certainly never *hurts* to switch door.

Discussion

Above I tried to give short clear mathematical solutions to three mathematical problems. Two of them were problems of elementary probability theory, the third is a problem of elementary game theory. As such, it involves not much more than elementary probability theory and the beautiful minimax theorem of John von Neumann (1928). That a finite two-party zero-sum game has a saddle-point, or in other words, that the two parties in such a game have matching minimax strategies (if ►randomization is allowed), is not obvious. It seems to me that probabilists ought to know more about game theory, since every ordinary non-mathematician who hears about the problem starts to wonder whether the quiz-master is trying to cheat the player, leading to an infinite regress: if I know that he knows that I know that....

I am told that the literature of mathematical economics and of game theory is full of Monty Hall examples, but no-one can give me a nice reference to a nice game-theoretic solution of the problem. Probably game-theorists like to keep their clever ideas to themselves, so as to make money from playing the game. Only losers write books explaining how the reader could make money from game theory.

It would certainly be interesting to investigate more complex game-theoretic versions of the problem. If we take Monty Hall as a separate player to the TV station, and note that TV ratings are probably helped if nice players win while annoying players lose, we leave elementary game theory and must learn the theory of Nash equilibria.

Then there is a sociological or historical question: who "owns" the Monty Hall problem? I think the answer is obvious: no-one. A beautiful mathematical paradox, once launched into the real world, lives it own life, it evolves, it is re-evaluated by generation after generation. This point of view actually makes me believe that Question 0: *would you switch* is the right question, and no further information should be given beyond the fact that you know that the quizmaster knows where the car is hidden, and always opens a door exhibiting a goat. Question 0 is a question you can ask a non-mathematician at a party, and if they have not heard of the problem before, they'll give the wrong answer (or rather, one of the two wrong answers: *no* because nothing is changed, or *it doesn't matter* because it's now 50–50). My mother, who was one of Turing's computers at Bletchley Park during the war, but who had had almost no schooling and in particular never learnt any mathematics, is the only person I know who immediately said: *switch*, by immediate intuitive consideration of the 100-door variant of the problem. The problem is a *paradox* since you can next immediately convince anyone (except lawyers, as was shown by an experiment in Nijmegen), that their initial answer is wrong.

The mathematizations Questions 1 and 2 are not (in my humble opinion!) *the* Monty Hall problem; they are questions which probabilists might ask, anxious to show off Bayes' theorem or whatever. Some people intuitively try to answer Question 0 via Questions 1 and 2; that is natural, I do admit. And sometimes people become very confused when they realize that the answer to Question 2 can only be given its pretty answer "2/3" under further conditions. It is interesting how in the pedagogical mathematical literature, the further conditions are as it were held under your nose, e.g., by saying "three *identical* doors," or replacing Marilyn's "say, door 1" by the more emphatic "door 1."

It seems to me that adding into the question explicitly the remarks that the three doors are equally likely to hide the car, and that when the quizmaster has a choice he secretly tosses a fair coin to decide, convert this beautiful paradox into a probability puzzle with little appeal any more to non experts.

It also converts the problem into one version of the three prisoner's paradox. The three prisoners problem is isomorphic to the conditional probabilistic three doors problem. I always found it a bit silly and not very interesting, but possibly that problem too should be approached from a sophisticated game theoretic point of view.

By the way, Marilyn vos Savant's original question is semantically ambiguous, though this might not be noticed by a non-native English speaker. Are the mentioned door numbers, huge painted numbers on the front of the doors *a priori*, or are we just for convenience *naming* the doors by the choices of the actors in our game *a posteriori*. Marilyn stated in a later column in *Parade* that she had originally been thinking of the latter. However, her own offered solutions are not consistent with a single unambiguous formulation. Probably she did not find the difference very interesting.

This little article contains nothing new, and only almost trivial mathematics. It is a plea for future generations to preserve the life of *The True Monty Hall paradox*, and not

let themselves be misled by probability purists who say "you *must* compute a conditional probability."

About the Author

Professor Gill has been selected as the 2010–2011 Distinguished Lorentz Fellow by the Netherlands Institute for Advanced Study in Humanities and Social Sciences. He is a member of the Royal Netherlands Academy of Arts and Sciences.

Cross References

►Bayes' Theorem
►Conditional Expectation and Probability
►Statistics and Gambling

References and Further Reading

Gill RD (2010) The one and only true Monty Hall problem. Submitted to *Statistica Neerlandica*. arXiv.org:1002.0651 [math.HO]

Rosenhouse J (2009) The Monty Hall problem. Oxford University Press, Oxford

Rosenthal JS (2008) Monty Hall, Monty Fall, Monty Crawl. Math Horizons September 2008:5–7. Reprint: http://probability.ca/jeff/writing/montyfall.pdf

Mood Test

Justice I. Odiase[1], Sunday M. Ogbonmwan[2]
[1]University of Benin, Benin City, Nigeria
[2]Professor and Dean, The Faculty of Physical Sciences University of Benin, Benin City, Nigeria

In 1954, A.M. Mood developed the square rank test for dispersion known as Mood test. It is based on the sum of squared deviations of the ranks of one sample from the mean rank of the combined samples. The null hypothesis is that there is no difference in spread against the alternative hypothesis that there is some difference. The Mood test assumes that location remains the same. It is assumed that differences in scale do not cause a difference in location. The samples are assumed to be drawn from continuous distributions.

In two-sample scale tests, the population distributions are usually assumed to have the same location with different spreads. However, Neave and Worthington (1988) cautioned that tests for difference in scale could be severely impaired if there is a difference in location as well.

In a two-sample problem composed of $X = \{x_1, x_2, \ldots, x_m\}$ with distribution $F(X)$ and $Y = \{y_1, y_2, \ldots, y_n\}$ with distribution $G(Y)$, arrange the combined samples in ascending order of magnitude and rank all the $N = m + n$ observations from 1 (smallest) to N (largest). Let W be the sum of squares of the deviations of one of the samples' (say X) ranks from the mean rank of the combined samples,

$$W = \sum_{i=1}^{m} \left(r_i - \frac{m+n+1}{2} \right)^2,$$

where r_i is the rank of the i^{th} X observation. The table of exact critical values can be found in Odiase and Ogbonmwan (2008).

Under the null hypothesis ($F = G$), the layout of the ranks of the combined samples is composed of N independent and identically distributed random variables, and hence conditioned on the observed data set, the mean and variance of W are $m(N^2-1)/12$ and $mn(N+1)(N^2-4)/180$, respectively. The large sample Normal approximation of W is

$$\frac{W - \frac{m(N^2-1)}{12}}{\sqrt{\frac{mn(N+1)(N^2-4)}{180}}}.$$

The efficiency of the two-sample Mood test against the normal alternative to the null hypothesis is $\frac{15}{2\pi^2} \cong 76\%$.

A Monte Carlo study of several nonparametric test statistics to obtain the minimum sample size requirement for large sample approximation was carried out by Fahoome (2002). Adopting Bradley's (1978) liberal criterion of robustness, Fahoome (2002) recommends the asymptotic approximation of the Mood test when $\min(m, n) = 5$ for the level of significance $\alpha = 0.05$ and $\min(m, n) = 23$ for $\alpha = 0.01$. However, Odiase and Ogbonmwan (2008) generated the exact distribution of the Mood test statistics by the permutation method and therefore provided the table of exact critical values at different levels of significance.

The idea of a general method of obtaining an exact test of significance originated with Fisher (1935). The essential feature of the method is that all the distinct permutations of the observations are considered, with the property that each permutation is equally likely under the hypothesis to be tested.

About the Authors

Dr. Justice Ighodaro Odiase is a Senior Lecturer, Department of Mathematics, University of Benin, Nigeria. He is the Scientific Secretary of the Statistics Research Group (SRG), Department of Mathematics, University of Benin. He is a member of the Nigerian Statistical Association (NSA), International Association for Statistical Computing (IASC), and The Society for Imprecise Probability:

Theories and Applications (SIPTA). He has authored and coauthored more than 30 papers.

Sunday Martins Ogbonmwan is a Professor of Statistics, Department of Mathematics, University of Benin, Benin City, Nigeria. He is the President of the Statistics Research Group (SRG), Department of Mathematics, University of Benin. He was the Head of Department of Mathematics, University of Benin (2006–2009). He is currently the Dean of the Faculty of Physical Sciences, University of Benin. He is a member of the Institute of Mathematical Statistics (IMS). He is also a member of the Nigerian Statistical Association (NSA). He has authored and coauthored more than 50 papers. He was the Editor-in-Chief of the *Journal of the Nigerian Statistical Association* (JNSA (1990–1995)). Professor Ogbonmwan was an award winner in a competition organized by the International Statistical Institute for young statisticians in developing countries (Madrid, Spain, 1983).

Cross References

►Asymptotic Normality
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Tests for Homogeneity of Variance

References and Further Reading

Bradley JV (1978) Robustness? Br J Math Stat Psychol 31:144–152
Fahoome G (2002) Twenty nonparametric statistics and their large sample approximations. J Mod Appl Stat Meth 1:248–268
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Mood AM (1954) On the asymptotic efficiency of certain nonparametric two-sample tests. Ann Math Stat 25:514–522
Neave HR, Worthington PL (1988) Distribution-free tests. Unwin Hyman, London
Odiase JI, Ogbonmwan SM (2008) Critical values for the Mood test of equality of dispersion. Missouri J Math Sci 20(1):40–52

Most Powerful Test

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

This notion plays a key role in testing statistical hypotheses. Testing is a two-decision statistical problem.

Case Study

A producer of hydraulic pumps applies plastic gaskets purchased from a deliverer. The gaskets are supplied in batches of 5,000. Since the cost of repairing a pump found to be faulty is far higher than the cost of the gasket itself, each batch is subject to testing. Not only the testing is costly but also any gasket used in the process is practically damaged. Thus the producer decides to verify 50 gaskets taken randomly from each batch.

Assume the deliverer promised that the fraction of defective gaskets would not exceed 5%. Suppose 4 defective gaskets were disclosed in a sample of size 50. Is this enough to reject the batch? The situation is illustrated by the following table

Batch\decision	Accept	Reject
Good	+	Type I Error
Bad	Type II Error	+

Since the decision is taken on the basis of a random variable (the number of defective gaskets), the quality of test may be expressed in terms of the probabilities of these two errors. We would like to minimize these probabilities simultaneously. However, any decrease of one of these probabilities causes increase of the second one. Consequences of these two errors should also be taken into consideration. Similarly as in law, one presumes that the tested hypothesis is true. Thus the probability of the error of the first type should be under control. Theory of testing statistical hypotheses, regarding these postulates, was formalized in 1933 by Neyman and Pearson.

Neyman-Pearson Theory

Let X be a random variable (or: random vector) taking values in a sample space $(\mathcal{X}, \mathcal{A})$ with a distribution P belonging to a class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let Θ_0 be a proper subset of Θ. We are interested in deciding, on the basis of observation X, whether $\theta \in \Theta_0$ (decision d_0) or not (decision d_2).

Any statement of the form $H : \theta \in \Theta_0$ is called a statistical hypothesis. We consider also the alternative hypothesis $K : \theta \notin \Theta_0$, i.e., $\theta \in \Theta \setminus \Theta_0$. A criterion of rejecting H (called a test) may be assigned by a *critical region* $S \subseteq \mathcal{X}$, according to the rule: reject H if $X \in S$ and accept otherwise.

When performing a test one may arrive at the correct decision, or one may commit one of two errors: rejecting H when it is true or accepting when it is false. The upper bound of the probability $P_\theta(d_0(X))$ for all $\theta \in \Theta_0$ is called

the *size* while the function $\beta(\theta) = P_\theta(d_0)$ for $\theta \in \Theta \setminus \Theta_0$ is called the power function of the test.

The general principle in Neyman-Pearson theory is to find such a procedure that maximizes $\beta(\theta)$ for all $\theta \in \Theta \backslash \Theta_0$ under assumption that $P_\theta(d_0(X)) \le \alpha$ (significance level) for all $\theta \in \Theta_0$. Any such test (if exists) is called to be *uniformly most powerful* (UMP). The well known Neyman-Pearson fundamental lemma (see ▶Neyman-Pearson Lemma) states that for any two-element family of densities or mass probabilities $\{f_0, f_1\}$ such test always exists and it can be expressed by the likelihood ratio $r(x) = \frac{f_1(x)}{f_0(x)}$. In this case the power function β reduces to a scalar and the word *uniformly* is redundant.

It is worth to add that in the continuous case the size of the UMP test coincides with its significance level. However, it may not be true in the discrete case. The desired equality can be reached by considering the *randomized* decision rules represented by functions $\phi = \phi(x)$, taking values in the interval $[0,1]$ and interpreted as follows:

"If $X = x$ then reject H with probability $\phi(x)$
and accept it with probability $1 - \phi(x)$"

The size of the MP randomized test coincides with its significance level and its power may be greater than for the nonrandomized one. According to the Neyman-Pearson lemma, the randomized MP test has the form

$$\phi(x) = \begin{cases} 1, & \text{if} \quad p_1(x) > kp_0(x) \\ \gamma, & \text{if} \quad p_1(x) = kp_0(x \\ 0, & \text{if} \quad p_1(x) < kp_0(x \end{cases}$$

for some k induced by the significance level. If $\gamma = 0$ then it is non-randomized.

One-Side Hypothesis and Monotone Likelihood Ratio

In practical situations distribution of the observation vector depends on one or more parameters and we make use of composite hypotheses $\theta \in \Theta_0$ against $\theta \in \Theta \setminus \Theta_0$. Perhaps one of the simple situations of this type is testing one-side hypothesis $\theta \le \theta_0$ or $\theta \ge \theta_0$ in a scalar parameter family of distributions.

We say that a family of densities $\{f_\theta : \theta \in \Theta\}$ has *monotone likelihood ratio* if there exists a statistic $T = t(X)$ such that for any $\theta < \theta'$ the ratio $\frac{f_{\theta'}(x)}{f_\theta(x)}$ is a monotone function of T. It appears that for testing a hypothesis $H : \theta \le \theta_0$ against $K : \theta > \theta_0$ in such a family of densities there exists a UMP test of the form

$$\phi(x) = \begin{cases} 1 & \text{when } T(x) > C \\ \gamma & \text{when } T(x) = C \\ 0 & \text{when } T(x) < C. \end{cases}$$

An important class of families with monotone likelihood ratio are one-parameter exponential families with densities of type $f_\theta(x) = C(\theta)e^{Q(\theta)T(x)}h(x)$. In a discrete case with integer parameter instead the mononotonity condition it suffices to verify that the ratio $\frac{P_{k+1}(x)}{P_k(x)}$ is a monotone function of T for all k.

Example 1 (Testing expectation in a simple sample from normal distribution with known variance). Let $X_1, ..., X_n$ be independent and identically distributed. Random variables with distribution $N(\mu, \sigma^2)$, where σ^2 is known. Consider the hypothesis $H : \mu \le \mu_0$ under the alternative $K : \mu > \mu_0$. The family of distributions has a monotone likelihood ratio with respect to the statistic $T = \sum_{i=1}^n X_i$. Therefore there exists a UMP test which rejects H if $\sum_{i=1}^n X_i$. is too large.

Example 2 (Statistical control theory). From a great number (N) of elements with an unknown number D of defective ones we draw without replacement a sample of size n. Then the potential number X of defective elements in the sample has the hypergeometric distribution

$$P_D(X = x) = \begin{cases} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, & \text{if } \max(0, n+D-N) < \\ & x < \min(n, D) \\ 0, & \text{otherwise.} \end{cases}.$$

One can verify that

$$\frac{P_{D+1}(x)}{P_D(x)} = \begin{cases} 0, & \text{if } x = n + D - N \\ \frac{D+1}{N-D}\frac{N-D-n+x}{D+1-x}, & \text{if } n + D + 1 - N \le x \le D \\ \infty & \text{if } x = D + 1 \end{cases}$$

is a monotone function of x. Therefore there exists a UMP test for the hypothesis $H : D \le D_0$ against $K : D > D_0$, which rejects H if x is too large.

Invariant and Unbiased Tests

If distribution of the observation vector depends on several parameters, some of them may be out of our interest and play the role of nuisance parameters. Such a situation occurs, for instance, in testing linear hypotheses. In this case the class of all unbiased estimators is usually too large for handle. Then we may seek for a test with maximum power in a class of tests which are invariant with respect to some transformations of observations or their powers do not depend on the nuisance parameters. This is called the

most powerful invariant test. The class of tests under consideration may be also reduced by unbiasedness condition. A member of this class with maximum power is then called the most powerful unbiased test. The standard tests for linear hypotheses in a linear normal model are most powerful in each of these classes.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Asymptotic Relative Efficiency in Testing
►Frequentist Hypothesis Testing: A Defense
►Neyman-Pearson Lemma
►Power Analysis
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Statistical Evidence
►Statistical Inference
►Statistics: An Overview
►Testing Variance Components in Mixed Linear Models

References and Further Reading

Lehmann EL, Romano JP (2005) Testing statistical hypotheses 3rd edn. Springer, New York
Neyman J, Pearson E (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans Roy Stat Soc London 231:289–337
Pfanzagl J (1994) Parametric statistical theory. Gruyter, Berlin
Zacks S (1981) Parametric statistical inference. Pergamon, Oxford

Moving Averages

Rob J. Hyndman
Professor of Statistics
Monash University, Melbourne, VIC, Australia

A moving average is a time series constructed by taking averages of several sequential values of another time series. It is a type of mathematical convolution. If we represent the original time series by $y_1, \ldots, y_n$, then a *two-sided moving average* of the time series is given by

$$z_t = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j}, \qquad t = k+1, k+2, \ldots, n-k.$$

Thus $z_{k+1}, \ldots, z_{n-k}$ forms a new time series which is based on averages of the original time series, $\{y_t\}$. Similarly, a *one-sided moving average* of $\{y_t\}$ is given by

$$z_t = \frac{1}{k+1} \sum_{j=0}^{k} y_{t-j}, \qquad t = k+1, k+2, \ldots, n.$$

More generally, weighted averages may also be used. Moving averages are also called running means or rolling averages. They are a special case of "filtering", which is a general process that takes one time series and transforms it into another time series.

The term "moving average" is used to describe this procedure because each average is computed by dropping the oldest observation and including the next observation. The averaging "moves" through the time series until z_t is computed at each observation for which all elements of the average are available.

Note that in the above examples, the number of data points in each average remains constant. Variations on moving averages allow the number of points in each average to change. For example, in a cumulative average, each value of the new series is equal to the sum of all previous values.

Moving averages are used in two main ways: Two-sided (weighted) moving averages are used to "smooth" a time series in order to estimate or highlight the underlying trend; one-sided (weighted) moving averages are used as simple forecasting methods for time series. While moving averages are very simple methods, they are often building blocks for more complicated methods of time series smoothing, decomposition and forecasting.

Smoothing Using Two-Sided Moving Averages

It is common for a time series to consist of a smooth underlying trend observed with error:

$$y_t = f(t) + \varepsilon_t,$$

where $f(t)$ is a smooth and continuous function of t and $\{\varepsilon_t\}$ is a zero-mean error series. The estimation of $f(t)$ is known as smoothing, and a two-sided moving average is one way of doing so:

$$\hat{f}(t) = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j}, \qquad t = k+1, k+2, \ldots, n-k.$$

The idea behind using moving averages for smoothing is that observations which are nearby in time are also likely to be close in value. So taking an average of the points near an observation will provide a reasonable estimate of the trend at that observation. The average eliminates some of the randomness in the data, leaving a smooth trend component.

Moving averages do not allow estimates of $f(t)$ near the ends of the time series (in the first k and last k periods). This can cause difficulties when the trend estimate is used for forecasting or analyzing the most recent data.

Each average consists of $2k+1$ observations. Sometimes this is known as a $(2k+1)$ MA smoother. The larger the value of k, the flatter and smoother the estimate of $f(t)$ will be. A smooth estimate is usually desirable, but a flat estimate is biased, especially near the peaks and troughs in $f(t)$. When ε_t is a white noise series (i.e., independent and identically distributed with zero mean and variance σ^2), the bias is given by $\mathrm{E}[\hat{f}(x)] - f(x) \approx \frac{1}{6} f''(x)k(k+1)$ and the variance by $\mathrm{V}[\hat{f}(x)] \approx \sigma^2/(2k+1)$. So there is a trade-off between increasing bias (with large k) and increasing variance (with small k).

Centered Moving Averages

The simple moving average described above requires an odd number of observations to be included in each average. This ensures that the average is centered at the middle of the data values being averaged. But suppose we wish to calculate a moving average with an even number of observations. For example, to calculate a 4-term moving average, the trend at time t could be calculated as

$$\hat{f}(t-0.5) = (y_{t-2} + y_{t-1} + y_t + y_{t+1})/4$$

or

$$\hat{f}(t+0.5) = (y_{t-1} + y_t + y_{t+1} + y_{t+2})/4$$

That is, we could include two terms on the left and one on the right of the observation, or one term on the left and two terms on the right, and neither of these is centered on t. If we now take the average of these two moving averages, we obtain something centered at time t.

$$\begin{aligned}\hat{f}(t) &= \frac{1}{2}\left[(y_{t-2} + y_{t-1} + y_t + y_{t+1})/4\right] \\ &\quad + \frac{1}{2}\left[(y_{t-1} + y_t + y_{t+1} + y_{t+2})/4\right] \\ &= \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1}\frac{1}{8}y_{t+2}\end{aligned}$$

So a 4 MA followed by a 2 MA gives a *centered moving average*, sometimes written as 2×4 MA. This is also a weighted moving average of order 5, where the weights for each period are unequal. In general, a $2 \times m$ MA smoother is equivalent to a weighted MA of order $m+1$ with weights $1/m$ for all observations except for the first and last observations in the average, which have weights $1/(2m)$.

Centered moving averages are examples of how a moving average can itself be smoothed by another moving average. Together, the smoother is known as a *double moving average*. In fact, any combination of moving averages can be used together to form a double moving average. For example, a 3×3 moving average is a 3 MA of a 3 MA.

Moving Averages. Table 1 Weight functions a_j for some common weighted moving averages

Name	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
3 MA	.333	.333										
5 MA	.200	.200	.200									
2 × 12 MA	.083	.083	.083	.083	.083	.083	.042					
3 × 3 MA	.333	.222	.111									
3 × 5 MA	.200	.200	.133	.067								
S15 MA	.231	.209	.144	.066	.009	−.016	−.019	−.009				
S21 MA	.171	.163	.134	.037	.051	.017	−.006	−.014	−.014	−.009	−.003	
H5 MA	.558	.294	−.073									
H9 MA	.330	.267	.119	−.010	−.041							
H13 MA	.240	.214	.147	.066	.000	−.028	−.019					
H23 MA	.148	.138	.122	.097	.068	.039	.013	−.005	−.015	−.016	−.011	−.004

S, Spencer's weighted moving average.
H, Henderson's weighted moving average.

Moving Averages with Seasonal Data

If the centered 4 MA was used with quarterly data, each quarter would be given equal weight. The weight for the quarter at the ends of the moving average is split between the two years. It is this property that makes 2×4 MA very useful for estimating a trend in the presence of quarterly seasonality. The seasonal variation will be averaged out exactly when the moving average is computed. A slightly longer or a slightly shorter moving average will still retain some seasonal variation. An alternative to a 2×4 MA for quarterly data is a 2×8 or 2×12 which will also give equal weights to all quarters and produce a smoother fit than the 2×4 MA. Other moving averages tend to be contaminated by the seasonal variation.

More generally, a $2 \times (km)$ MA can be used with data with seasonality of length m where k is a small positive integer (usually 1 or 2). For example, a 2×24 MA may be used for estimating a trend in monthly seasonal data (where $m = 12$).

Weighted Moving Averages

A weighted k-point moving average can be written as

$$\hat{f}(t) = \sum_{j=-k}^{k} a_j y_{t+j}.$$

For the weighted moving average to work properly, it is important that the weights sum to one and that they are symmetric, that is $a_j = a_{-j}$. However, we do not require that the weights are between 0 and 1. The advantage of weighted averages is that the resulting trend estimate is much smoother. Instead of observations entering and leaving the average abruptly, they can be slowly downweighted. There are many schemes for selecting appropriate weights. Kendall et al. (1983, Chap. 46) give details.

Some sets of weights are widely used and have been named after their proposers. For example, Spencer (1904) proposed a $5 \times 4 \times 4$ MA followed by a weighted 5-term moving average with weights $a_0 = 1$, $a_1 = a_{-1} = 3/4$, and $a_2 = a_{-2} = -3/4$. These values are not chosen arbitrarily, but because the resulting combination of moving averages can be shown to have desirable mathematical properties. In this case, any cubic polynomial will be undistorted by the averaging process. It can be shown that Spencer's MA is equivalent to the 15-point weighted moving average whose weights are −.009, −.019, −.016, .009, .066, .144, .209, .231, .209, .144, .066, .009, −.016, −.019, and −.009. Another Spencer's MA that is commonly used is the 21-point weighted moving average. Henderson's weighted moving averages are also widely used, especially as part of seasonal adjustment methods (Ladiray and Quenneville 2001). The set of weights is known as the *weight function*. Table 1 shows some common weight functions. These are all symmetric, so $a_{-j} = a_j$.

Weighted moving averages are equivalent to kernel regression when the weights are obtained from a kernel function. For example, we may choose weights using the quartic function

$$Q(j,k) = \begin{cases} \left\{1 - [j/(k+1)]^2\right\}^2 & \text{for } -k \le j \le k; \\ 0 & \text{otherwise.} \end{cases}$$

Then a_j is set to $Q(j,k)$ and scaled so the weights sum to one. That is,

$$a_j = \frac{Q(j,k)}{\sum_{i=-k}^{k} Q(i,k)}. \quad (1)$$

Forecasting Using One-Sided Moving Averages

A simple forecasting method is to average the last few observed values of a time series. Thus

$$\hat{y}_{t+h|t} = \frac{1}{k+1} \sum_{j=0}^{k} y_{t-j}$$

provides a forecast of y_{t+h} given the data up to time t.

As with smoothing, the more observations included in the moving average, the greater the smoothing effect. A forecaster must choose the number of periods $(k + 1)$ in a moving average. When $k = 0$, the forecast is simply equal to the value of the last observation. This is sometimes known as a "naïve" forecast.

An extremely common variation on the one-sided moving average is the exponentially weighted moving average. This is a weighted average, where the weights decrease exponentially. It can be written as

$$\hat{y}_{t+h|t} = \sum_{j=0}^{t-1} a_j y_{t-j}$$

where $a_j = \lambda(1 - \lambda)^j$. Then, for large t, the weights will approximately sum to one. An exponentially weighted moving average is the basis of simple exponential smoothing. It is also used in some process control methods.

Moving Average Processes

A related idea is the moving average process, which is a time series model that can be written as

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

where $\{e_t\}$ is a white noise series. Thus, the observed series y_t, is a weighted moving average of the unobserved e_t

series. This is a special case of an Autoregressive Moving Average (or ARMA) model and is discussed in more detail in the entry ►Box-Jenkins Time Series Models. An important difference between this moving average and those considered previously is that here the moving average series is directly observed, and the coefficients $\theta_1, \ldots, \theta_q$ must be estimated from the data.

Cross References

►Box–Jenkins Time Series Models
►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Median Filters and Extensions
►Seasonality
►Smoothing Techniques
►Statistical Quality Control: Recent Advances
►Time Series
►Trend Estimation

References and Further Reading

Kendall MG, Stuart A, Ord JK (1983) Kendall's advanced theory of statistics. vol 3. Hodder Arnold, London

Ladiray D, Quenneville B (2001) Seasonal adjustment with the X-11 method, vol 158, of Lecture notes in statistics. Springer, Berlin

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications, 3rd edn. Wiley, New York

Spencer J (1904) On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893–1897. J Inst Actuaries 38:334–343

Multicollinearity

Vlasta Bahovec
Professor, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

One of the assumptions of the standard regression model $y = X\beta + \varepsilon$ is that there is no exact linear relationship among the explanatory variables, or equivalently, that the matrix X of explanatory variables has a full rank. The problem of multicollinearity occurs if two or more explanatory variables are linearly dependent, or near linearly dependent (including the variable $x_0' = [1, 1, \cdots, 1]$, which generates a constant term). There are two types of multicollinearity: perfect and near multicollinearity.

Perfect multicollinearity occurs if at least two explanatory variables are linearly dependent. In that case, the determinant of matrix $X'X$ equals zero (the $X'X$ matrix is singular), and therefore ordinary least squares (OLS) estimates of regression parameters $\beta' = (\beta_0, \beta_1, \cdots, \beta_k)$

$$\hat{\beta} = (X'X)^{-1}X'y = \frac{adj(X'X)}{\det(X'X)} \cdot X'y$$

are not unique. This type of multicollinearity is rare, but may occur if the regression model includes qualitative explanatory variables, whose effect is taken into account by ►dummy variables. Perfect multicollinearity occurs in a regression model with an intercept, if the number of dummy variables for each qualitative variable is not less than the number of groups of this variable. Perfect multicollinearity can easily be revealed. A more difficult problem is near or imperfect multicollinearity. This problem arises if at least two regressors are highly intercorrelated. In that case, $\det(X'X) \approx 0$, the matrix $X'X$ is ill conditioned, and therefore the estimated parameters are numerically imprecise. Furthermore, since the covariance matrix of estimated parameters is calculated by the formula $Cov(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$, the variances and covariances of the estimated parameters will be large. Large standard errors $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$ imply that empirical t-ratios $\left(t_j = \hat{\beta}_j / SE\left(\hat{\beta}_j\right)\right)$ could be insignificant, which may lead to an incorrect conclusion that some explanatory variables have to be omitted from the regression model. Also, large standard errors make interval parameter estimates imprecise.

Imperfect multicollinearity often arises in the time series regression model (see ►Time Series Regression), especially in data involving economic time series, while variables over time tend to move in the same direction.

The simplest way to detect serious multicollinearity problems is to analyze variances of estimated parameters, which are calculated with the following formula:

$$var(\hat{\beta}_j) = \sigma^2(X'X)^{-1}_{jj} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \cdot \left(1 - R_j^2\right)},$$

where R_j^2 is the coefficient of determination in the regression, variable x_j is the dependent, and the remaining x's are explanatory variables. If variable x_j is highly correlated with other regressors, R_j^2 will be large (near to 1), and therefore the variance of $\hat{\beta}_j$ will be large. There are some measures of multicollinearity included in standard statistical software: the variance inflation factor (VIF), tolerance (TOL), condition number (CN), and condition indices (CI). VIF and TOL are calculated with the following formulas:

$$VIF_j = \frac{1}{1 - R_j^2} \quad j = 1, 2, \cdots, k \quad TOL_j = \frac{1}{VIF_j} = 1 - R_j^2.$$

The multicollinearity problem is serious if $R_j^2 > 0.8$, consequently if $VIF_j > 5$, or equivalently if $TOL_j < 0.2$.

More sophisticated measures of multicollinearity are condition number, CN, and condition indices, CI_i, based on the use of eigenvalues of the $X'X$ matrix. CN is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue, and CI_i, $i = 1, 2, \ldots, k$, are square roots of the ratio of the largest eigenvalue to each individual eigenvalue. These measures, which are calculated with the formulas

$$CN = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \qquad CI_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}} \quad i = 1, 2, \cdots, k,$$

are measures of sensitivity of parameter estimates to small changes in data. Some authors, such as Belsley et al. (1980), suggested that a condition index of 30–100 indicates moderate to strong multicollinearity.

Several solutions have been suggested to rectify the multicollinearity problem. Some are the following: (1) increasing the sample size to reduce multicollinearity, as multicollinearity is a problem of the sample, and not the population; (2) dropping one or more variables suspected of causing multicollinearity; (3) transforming data as the first differences $\Delta X_t = X_t - X_{t-1}$ or ratios X_t/X_{t-1} $t = 2, 3, \ldots, n$ to eliminate linear or exponential trends; (4) ridge regression (see ►Ridge and Surrogate Ridge Regressions); and (5) principal component regression.

The problem of multicollinearity is approached differently by econometricians depending on their research goal. If the goal is to forecast future values of the dependent variable, based on the determined regression model, the problem of multicollinearity is neglected. In all other cases, this problem is approached more rigorously.

Cross References

►Dummy Variables
►Heteroscedasticity
►Linear Regression Models
►Multivariate Statistical Analysis
►Partial Least Squares Regression Versus Other Methods
►Ridge and Surrogate Ridge Regressions

References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: Identifying: Influential data and sources of collinearity. Wiley, New York

Green WH (2002) Econometric analysis, 5th edn. Prentice Hall, New Jersey

Gujarati DN (2002) Basic econometrics, 4th edn. McGraw-Hill/Irwin, New York

Maddala GS (2002) Introduction to econometrics, 3rd edn. Wiley, Chichester

Multicriteria Clustering

Anuška Ferligoj
Professor, Head of the Center of Informatics and Methodology, Faculty of Social Sciences
University of Ljubljana, Ljubljana, Slovenia

Some clustering problems cannot be appropriately solved with classical clustering algorithms because they require optimization over more than one criterion. In general, solutions optimal according to each particular criterion are not identical. Thus, the problem arises of how to find the best solution satisfying as much as possible all criteria considered. In this sense the set of Pareto efficient clusterings was defined: a clustering is Pareto efficient if it cannot be improved on any criterion without sacrificing some other criterion.

A multicriteria clustering problem can be approached in different ways:

- By reduction to a clustering problem with a single criterion obtained as a combination of the given criteria;
- By constrained clustering algorithms where a selected criterion is considered as the clustering criterion and all others determine the constraints;
- By direct algorithms: Hanani (1979) proposed an algorithm based on the dynamic clusters method using the concept of the kernel, as a representation of any given criterion. Ferligoj and Batagelj (1992) proposed modified relocation algorithms and modified agglomerative hierarchical algorithms.

Usual Clustering Problems

Cluster analysis (known also as classification and taxonomy) deals mainly with the following general problem: given a set of units, $\mathcal{U}$, determine subsets, called clusters, C, which are homogeneous and/or well separated according to the measured variables (e.g., Sneath and Sokal 1973; Hartigan 1975; Gordon 1981). The set of clusters forms a clustering. This problem can be formulated as an optimization problem:

Determine the clustering $\mathbf{C}^*$ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where $\mathbf{C}$ is a clustering of a given *set of units*, $\mathcal{U}$, Φ is the set of all feasible clusterings and $P : \Phi \to \mathbb{R}$ a *criterion function*.

As the set of feasible clusterings is finite a solution of the clustering problem always exists. Since this set is usually large it is not easy to find an optimal solution.

A Multicriteria Clustering Problem

In a *multicriteria clustering problem* $(\Phi, P_1, P_2, \ldots, P_k)$ we have several criterion functions $P_t, t = 1, \ldots, k$ over the same set of feasible clusterings Φ, and our aim is to determine the clustering $\mathbf{C} \in \Phi$ in such a way that

$$P_t(\mathbf{C}) \to \min, \quad t = 1, \ldots, k.$$

In the ideal case, we are searching for the dominant set of clusterings. The solution $\mathbf{C}_0$ is the *dominant* solution if for each solution $\mathbf{C} \in \Phi$ and for each criterion P_t, it holds that

$$P_t(\mathbf{C}_0) \le P_t(\mathbf{C}), \quad t = 1, \ldots, k.$$

Usually the set of dominant solutions is empty. Therefore, the problem arises of finding a solution to the problem that is as good as is possible according to each of the given criteria. Formally, the *Pareto-efficient* solution is defined as follows:

For $\mathbf{C}_1, \mathbf{C}_2 \in \Phi$, solution $\mathbf{C}_1$ *dominates* solution $\mathbf{C}_2$ if and only if

$$P_t(\mathbf{C}_1) \le P_t(\mathbf{C}_2), \quad t = 1, \ldots, k,$$

and for at least one $i \in 1, \ldots, k$ the strict inequality $P_i(\mathbf{C}_1) < P_i(\mathbf{C}_2)$ holds. We denote the dominance relation by $\prec$. $\prec$ is a strict partial order. The set of Pareto-efficient solutions, Π, is the set of minimal elements for the dominance relation:

$$\Pi = \{\mathbf{C} \in \Phi : \neg\exists \mathbf{C}' \in \Phi : \mathbf{C}' \prec \mathbf{C}\}$$

In other words, the solution $\mathbf{C}^* \in \Phi$ is *Pareto-efficient* if there exists no other solution $\mathbf{C} \in \Phi$ such that

$$P_t(\mathbf{C}) \le P_t(\mathbf{C}^*), \quad t = 1, \ldots, k,$$

with strict inequality for at least one criterion. A *Pareto-clustering* is a Pareto-efficient solution of the multicriteria clustering problem (Ferligoj and Batagelj 1992).

Since the optimal clusterings for each criterion are Pareto-efficient solutions the set Π is not empty. If the set of dominant solutions is not empty then it is equal to the set of Pareto-efficient solutions.

Solving Discrete Multicriteria Optimization Problems

Multicriteria clustering problems can be approached as a multicriteria optimization problem, that has been treated by several authors (e.g., Chankong and Haimes 1983; Ferligoj and Batagelj 1992). In the clustering case, we are dealing with discrete multicriteria optimization (the set of feasible solutions is finite), which means that many very useful theorems in the field of multicriteria optimization do not hold, especially those which require convexity. It was proven that if, for each of the given criteria, there is a unique solution, then the minimal number of Pareto-efficient solutions to the given multicriteria optimization problem equals the number of different minimal solutions of the single criterion problems.

Although several strategies haven been proposed for solving multicriteria optimization problems explicitly, the most common is the conversion of the multicriteria optimization problem to a single criterion problem.

Direct Multicriteria Clustering Algorithms

The multicriteria clustering problem can be approached efficiently by using direct algorithms. Two types of direct algorithms are known: a version of the relocation algorithm, and the modified agglomerative (hierarchical) algorithms (Ferligoj and Batagelj 1992).

Modified Relocation Algorithm

The idea of the *modified relocation* algorithm for solving the multicriteria clustering problem follows from the definition of a Pareto-efficient clustering. The solutions obtained by the proposed procedure can be only *local Pareto clusterings*. Therefore, the basic procedure should be repeated *many* times (at least hundreds of times) and the obtained solutions should be reviewed. An efficient review of the obtained solutions can be systematically done with an appropriate *metaprocedure* with which the true set of Pareto clusterings can be obtained.

Modified Agglomerative Hierarchical Approach

Agglomerative hierarchical clustering algorithms usually assume that all relevant information on the relationships between the n units from the set $\mathcal{U}$ is summarized by a symmetric pairwise dissimilarity matrix $D = [d_{ij}]$. In the case of multicriteria clustering we assume we have k dissimilarity matrices $D^t, t = 1, \ldots, k$, each summarizing all relevant information obtained, for example, in the k different situations. The problem is to find the best hierarchical solution which satisfies as much as is possible all k dissimilarity matrices.

One approach to solving the multicriteria clustering problem combines the given dissimilarity matrices (at each step) into a composed matrix. This matrix $D = [d_{ij}]$ can, for example, be defined as follows:

$$d_{ij} = \max\left(d_{ij}^t; t = 1, \ldots, k\right)$$
$$d_{ij} = \min\left(d_{ij}^t; t = 1, \ldots, k\right)$$
$$d_{ij} = \sum_{t=1}^{k} \alpha_t d_{ij}^t \ , \ \sum_{t=1}^{k} \alpha_t = 1$$

Following this approach, one of several *decision rules* (e.g., pessimistic, optimistic, Hurwicz, Laplace) for making decisions under uncertainty (Chankong and Haimes 1983) can be used at the composition and selection step of the agglomerative procedure.

Conclusion

The multicriteria clustering problem can be treated with the proposed approaches quite well if only a few hundreds units are analysed. New algorithms have to be proposed for large datasets.

About the Author

Anuška Ferligoj is Professor at the Faculty of Social Sciences at University of Ljubljana, head of the graduate program on Statistics at the University of Ljubljana and head of the Center of Methodology and Informatics at the Institute of Social Sciences. She is editor of the journal Advances in Methodology and Statistics (since 2004). She was awarded the title of Ambassador of Science of the Republic of Slovenia in 1997. Dr Ferligoj is a Fellow of the European Academy of Sociology. For the monograph *Generalized Blockmodeling* she was awarded the Harrison White Outstanding Book Award for 2007, the Mathematical Sociology Section of the American Sociological Association. In 2010 she received Doctor et Professor Honoris Causa at ELTE University in Budapest.

Cross References

►Cluster Analysis: An Introduction
►Data Analysis
►Distance Measures
►Fuzzy Logic in Statistical Data Analysis
►Hierarchical Clustering
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Random Permutations and Partition Models

References and Further Reading

Chankong V, Haimes YY (1983) Multiobjective decision making. North-Holland, New York
Ferligoj A, Batagelj V (1992) Direct multicriteria clustering algorithms. J Clas. 9:43–61
Gordon AD (1981) Classification. Chapman & Hall, London
Hanani U (1979) Multicriteria dynamic clustering, Rapport de Recherche No. 358, IRIA, Rocquencourt
Hartigan JA (1975) Clustering algorithms. Wiley, New York
Sneath PHA, Sokal RR (1973) Numerical taxonomy Freeman, San Francisco

Multicriteria Decision Analysis

Theodor J. Stewart
Emeritus Professor
University of Cape Town, Rondebosch, South Africa
University of Manchester, Manchester, UK

Basic Definitions

The field variously described as *multicriteria decision making (MCDM)* or *multicriteria decision analysis* or *aid (MCDA)* is that branch of operational research/management science (OR/MS) that deals with the explicit modeling of multiple conflicting goals or objectives in management decision making. Standard texts in OR/MS typically do include identification of objectives (often stated as plural) as a key step in the decision-making process, but the ensuing discussion appears to assume that such objectives are easily aggregated into a single measure of achievement which can formally be optimized. The field of MCDA, however, arose from a recognition that systematic and coherent treatment of multiple objectives requires structured decision support to ensure that all interests are kept in mind and that an informed balance is achieved. See, for example, the discussions and associated references in Chap. 2 of Belton and Stewart (2002) and Chap. 1 of Figueira et al. (2005).

The starting point of MCDA is the identification of the critical *criteria* according to which potential courses of action (choices, policies, strategies) may be compared and evaluated. In this sense, each *criterion* is a particular point of view or consideration according to which preference orders on action outcomes can (more-or-less) unambiguously be specified. Examples of such criteria may include issues such as investment costs, job creation, levels of river pollution etc., as well as more subjective criteria such as aesthetic appeal. With careful selection of the criteria, preference ordering according to each could be essentially self-evident apart from some fuzziness around the concept equality of performance.

Selection of criteria is a profound topic in its own right, but is perhaps beyond the scope of the present article. Some discussion may be found in Keeney and Raiffa (1976); Keeney (1992); Belton and Stewart (2010). In essence, the analyst needs to ensure that values and aspirations of the decision maker(s) have been fully captured by the chosen criteria, while still retaining a manageably small number of criteria (typically, one strives for not much more than 15 or 25 criteria in most applications). Care needs to be taken not

to double-count issues, and that preference orders can be understood on each criterion independently of the others.

Suppose then that say m criteria have been defined as above. For any specified course of action, say $a \in \mathscr{A}$ (the set of all possible actions), we define $z_i(a)$ to be a measure of performance of a according to the perspective of criterion i, for $i = 1, \ldots, m$. The scaling at this stage is not important, the only requirement being that action a is preferred to action b in terms of criterion i ($a \succ_i b$) if and only if $z_i(a) > z_i(b) + \epsilon_i$ for some tolerance parameter ϵ_i. Apart from the brief comments in the final section, we assume that these measures of performance are non-stochastic.

The primary aim of MCDA is to support the decision maker in aggregating the single-criterion preferences into an overall preference structure, in order to make a final selection which best satisfies all criteria, or to select a reduced subset of $\mathscr{A}$ for further discussion and evaluation. It is important to recognize that this aggregation phase contains fundamentally subjective elements, namely the value judgments and tradeoffs provided by the decision maker. We shall briefly review some of the support processes which are used. A comprehensive overview of these approaches may be found in Figueira et al. (2005).

Methods of Multicriteria Analysis

It is important to recognize that two distinct situations may arise in the context described above, and that these may lead to broadly different forms of analysis:

- *Discrete choice problems:* In this case, $\mathscr{A}$ consists of a discrete set of options, e.g., alternative locations for a power station. The discrete case arises typically at the level of high level strategic choices, within which many of the criteria may require subjective evaluation of alternatives.
- *Multiobjective optimization problems:* These problems are often defined in mathematical programming terms, i.e., an option will be defined in terms of a vector of *decision variables*, say $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$. The measures of performance for each criterion typically need to be defined quantitatively in terms of functions $f_i(\mathbf{x})$ mapping $\mathbb{R}^n \to \mathbb{R}$ for each i.

The methods adopted can be characterized in two ways:

- By the underlying paradigm for modeling human preferences (*preference modeling*);
- By the stage of the analysis at which the decision makers' judgments are brought into play (*timing of preference statements*).

We deal with each of these in turn.

Preference Modeling

At least four different paradigms can be identified.

1. **Value scoring or utility methods:** The approach is first to re-scale the performance measures $z_i(a)$ so as to be commensurate in some way, typically by means of transformation through a *partial value function*, say $v_i(z_i)$. This rescaling needs to ensure that equal-sized intervals in the transformed scale represent the same importance to the decision maker (in terms of trade-offs with other criteria) irrespective of where they occur along the scale. Relatively mild assumptions (under conditions of deterministic performance measures) imply that an overall value of a can be modeled additively, i.e., as $V(a) = \sum_{i=1}^{m} w_i v_i(z_i(a))$. The assessment of the partial values and weights (w_i) may be carried out by direct assessment (e.g., Dyer 2005), indirectly such as by the analytic hierarchy process approach (Saaty 2005), or by learning from previous choices (Siskos et al. 2005).
2. **Metric methods:** In this approach, some form of goal or aspiration is specified (by the decision maker) for each criterion, say G_i for each i. A search (discrete or by mathematical optimization) is then conducted to find the option for which the performance levels $z_1(a), z_2(a), \ldots, z_m(a)$ approach the goal levels $G_1, G_2, \ldots, G_m$ as closely as possible. Typically, L_1, L_2, or L_∞ metrics are used to define closeness, with provision for differential weighting of criteria. Differences do also arise in terms of whether over-achievement of goals adds additional benefits or not. Such approaches are termed (generalized) goal programming, and are reviewed in Lee and Olson; Wierzbicki (1999; 1999). Goal programming is primarily applied in the context of the multiobjective optimization class of model.
3. **Outranking methods:** These methods consider action alternatives pairwise in terms of their performance levels on all criteria, in order to extract the level of evidence in the data provided by the performance measures which either support (are concordant with) or oppose (are discordant with) a conclusion that the one action is better than the other. These considerations generate partial rankings of the actions, or at least a classification of the actions into ordered preference classes. Descriptions of different outranking approaches may be found in Part III of Figueira et al. (2005).
4. **Artificial intelligence:** Greco et al. (2005) describe how observed choices by the decision maker(s) can

be used to extract decision rules for future multicriteria decisions, without explicit or formal preference modeling along the lines described above.

Timing of Preference Statements

Three possible stages of elicitation of values and preferences from the decision maker may be recognized as described below (although in practice no one of these is used completely in isolation).

1. **Elicitation prior to analysis of options:** In this approach, a complete model of the decision maker preferences is constructed from a sequence of responses to questions about values, trade-offs, relative importance, etc. The resulting model is then applied to the elements of $\mathscr{A}$ in order to select the best alternative or a shortlist of alternatives. This approach is perhaps most often used with value scoring methods, in which a simple and transparent preference model (e.g., the additive value function) is easily constructed and applied.
2. **Interactive methods:** Here a tentative preference model, incomplete in many ways, is used to generate a small number of possible choices which are presented to the decision maker, who may either express strong preferences for some or dislike of others. On the basis of these stated preferences, models are refined and a new set of choices generated. Even in the prior elicitation approach, some degree of interaction of this nature will occur, where in the application of value scoring or outranking approaches to discrete choice problems, results will inevitably be fed back to decision makers for reflection on the value judgements previously specified. However, it is especially with continuous multiobjective optimization problems that the interaction becomes firmly designed and structured into the process. See Chap. 5 of Miettinen (1999) for a comprehensive coverage of such structured interaction.
3. **Posterior value judgements:** If each performance measure is to be maximized, then an action a is said to *dominate* action b if $z_i(a) \geq z_i(b)$ for all criteria, with strict inequality for at least one criterion. With discrete choice problems, the removal of dominated actions from $\mathscr{A}$ may at times reduce the set of options to such a small number that no more analysis is necessary – decision makers can make a holistic choice. In some approaches to multiobjective optimization (see also Miettinen 1999), a similar attempt is made to compute the "efficient frontier," i.e., the image in criterion space of all non-dominated options, which can be displayed to the decision maker for a holistic choice. In practice, however, this approach is restricted to problems with two or three criteria only which can be displayed graphically (although there have been attempts at graphical displays for slightly higher dimensionality problems).

Stochastic MCDA

As indicated at the start, we have focused on deterministic problems, i.e., in which a fixed (even if slightly "fuzzy") performance measure $z_i(a)$ can be associated with each action-criterion combination. However, there do of course exist situations in which each $z_i(a)$ will be a *random variable*.

The introduction of stochastic elements into the multicriteria decision making problem introduces further complications. Attempts have been made to adapt value scoring methods to be consistent with the von Neumann/Morgenstern axioms of expected utility theory, to link multicriteria decision analysis with scenario planning, and to treat probabilities of achieving various critical outcomes as separate "criteria." Discussion of these extensions is beyond the scope of space available for this short article, but a review is available in Stewart (2005).

About the Author

Professor Stewart is Past-President of both the Operations Research Society of South Africa (1978) and the South African Statistical Association (1989). He was Vice President of IFORS (International Federation of Operational Research Societies) for the period 2004–2006, and President of the International Society on Multiple Criteria Decision Making for the period 2004–2008. He is currently Editor-in-Chief of the *Journal of Multi-Criteria Decision Analysis*, and African Editor of *International Transactions in Operations*. He is a Member of the Academy of Science of South Africa. In 2008 Professor Stewart was awarded the Gold medal of the International Society on Multiple Criteria Decision Making (for marked contributions to theory, methodology and practice in the field), and has been awarded ORSSA's Tom Roszwadowski Medal (for written contributions to OR) on five occasions.

Cross References

►Decision Theory: An Introduction
►Decision Theory: An Overview

References and Further Reading

Belton V, Stewart TJ (2002) Multiple criteria decision analysis: an integrated approach. Kluwer, Boston

Belton V Stewart TJ (2010) Problem structuring and MCDA. In: Ehrgott M, Figueira JR, Greco S (eds) Trends in multiple criteria decision analysis, chapter 8. Springer, Berlin, pp 237–271

Dyer JS (2005) MAUT – multiattribute utility theory. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 7. Springer, New York, pp 265–295

Figueira J, Greco S, Ehrgott M (eds) (2005) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76. Springer, New York

Gal T, Stewart TJ, Hanne T (eds) (1999) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications. Kluwer, Boston

Greco S, Matarazzo B, Słowiński R (2005) Decision rule approach. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 13. Springer, New York, pp 507–561

Keeney RL (1992) Value-focused thinking: a path to creative decision making. Harvard University Press, Cambridge

Keeney RL, Raiffa H (1976) Decisions with multiple objectives. Wiley, New York

Lee SM, Olson DL (1999) Goal programming. In: Gal T, Stewart TJ, Hanne T (eds) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications, chapter 8. Kluwer, Boston

Miettinen K (1999) Nonlinear multiobjective optimization, International series in operations research and management science, vol 12. Kluwer, Dordrecht

Saaty TL (2005) The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 9. Springer, New York, pp 345–407

Siskos Y, Grigoroudis E, Matsatsinis N (2005) MAUT – multiattribute utility theory. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 8. Springer, New York, pp 299–343

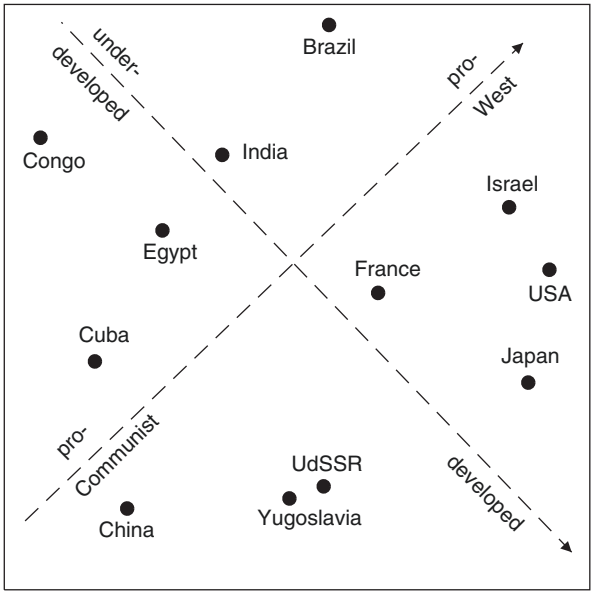
Stewart TJ (2005) Dealing with uncertainties in MCDA. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 11. Springer, New York, pp 445–470

Wierzbicki AP (1999) Reference point approaches. In: Gal T, Stewart TJ, Hanne T (eds) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications, chapter 9, Kluwer, Boston

Multidimensional Scaling

Ingwer Borg
Professor of Applied Psychological Methods
University of Giessen, Giessen, Germany
Scientific Director
GESIS, Mannheim, Germany

►Multidimensional scaling (MDS) is a family of methods that optimally map *proximity indices* of objects into distances between points of a multidimensional space with a given dimensionality (usually two or three dimensions). The main purpose for doing this is to visualize the data so that the user can test structural hypotheses or discover patterns "hidden" in the data.



Multidimensional Scaling. Fig. 1 MDS configuration for country similarity data

Historically, MDS began as a psychological model for judgments of (dis)similarity. A typical example of this early era is the following. Wish (1971) was interested to find out how persons generate overall judgments on the similarity of countries. He asked a sample of subjects to assess each pair of twelve countries with respect to their global similarity. For example, he asked "How similar are Japan and China?", offering a 9-point answer scale from "very dissimilar" to "very similar" for the answer. On purpose, "there were no instructions concerning the characteristics on which these similarity judgments were to be made; this was information to discover rather than to impose" (Kruskal and Wish 1978:30). The resulting numerical ratings were averaged over subjects, and then mapped via MDS into the distances among 12 points of a Euclidean plane. The resulting MDS configuration (Fig. 1) was interpreted to show that the ratings were essentially generated from two underlying dimensions.

As an MDS model, Wish (1971) used *ordinal MDS*, the most popular MDS model. It maps the proximities of the n objects (δ_{ij}) into distances d_{ij} of the $n \times m$ configuration $\mathbf{X}$ such that their ranks are optimally preserved. Hence, assuming that the δ_{ij}'s are dissimilarities, the function $f : \delta_{ij} \rightarrow d_{ij}(\mathbf{X})$ is monotone so that $f : \delta_{ij} < \delta_{kl} \rightarrow d_{ij}(\mathbf{X}) \leq d_{kl}(\mathbf{X})$, for all pairs (i,j) and (k,l) for which

data are given. Missing data impose no constraints onto the distances.

Another popular MDS model is *interval MDS*, where $f : \delta_{ij} \rightarrow a + b \cdot \delta_{ij} = d_{ij}(\mathbf{X})$. This model assumes that the data are given on an interval scale. Hence, both a and $b(\neq 0)$ can be chosen arbitrarily. In particular, they can be chosen such that the re-scaled proximities are equal to the distances of a given MDS configuration $\mathbf{X}$.

A second facet of an MDS model is the distance function that it uses. In psychology, the family of *Minkowski distances* has been studied extensively as a model of judgment. Minkowski distances can be expressed by the formula

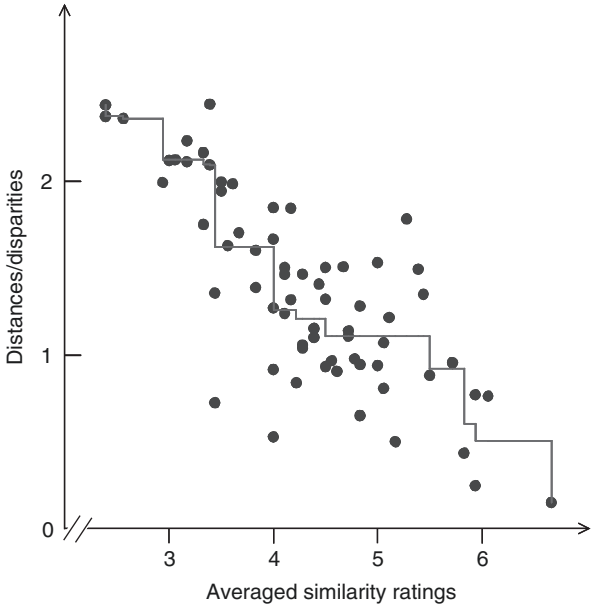
$$d_{ij}^{(p)}(\mathbf{X}) = \left(\sum_{a=1}^{m} |x_{ia} - x_{ja}|^p \right)^{1/p}, p \geq 1. \quad (1)$$

Setting $p = 1$ results in the *city-block metric*, setting $p = 2$ in the *Euclidean distance*. If p grows, d_{ij} is quickly dominated by its largest intra-dimensional difference (out of the $a = 1, \ldots, m$ dimensions). Such metrics supposedly explain fast and frugal (dis)similarity judgments. The city-block metric, in contrast, models careful judgments with important consequences for the individual. When MDS is used for exploratory purposes, however, only $p = 2$ should be used, because all other choices imply geometries with non-intuitive properties.

The fit of the MDS representation to the data can be seen from its *Shepard diagram*. For our country-similarity example, this is shown in Fig. 2. The plot exhibits how the data are related to the distances. It also shows the monotone regression line. The vertical scatter of the points about this regression line corresponds to the model's loss or misfit. It is measured as $\sum_{i<j} e_{ij}^2 = \sum_{i<j} (d_{ij}(\mathbf{X}) - f(\delta_{ij}))^2$, for all points i und j. The $f(\delta_{ij})$'s here are *disparities*, i.e., proximities that are re-scaled using all admissible transformations of the chosen scale level to optimally approximate the corresponding distances of the MDS configuration $\mathbf{X}$. The optimization is done by ordinal or linear regression (or, generally, by regression of type f) so that $f(\delta_{ij}) = \widehat{d}_{ij}(\mathbf{X})$. In order to obtain an interpretable measure of model misfit, the error sum is normed to yield the standard MDS loss function

$$Stress = \sqrt{\sum_{i<j} (d_{ij}(\mathbf{X}) - \widehat{d}_{ij})^2 / \sum_{i<j} d_{ij}^2(\mathbf{X})}. \quad (2)$$

A perfect MDS solution has a Stress of zero. In this case, the distances of the MDS solution correspond perfectly to the disparities. For the above example, we get Stress = 0.19. Evaluating if this is an acceptably low value is complex. A minimum criterion is that the observed Stress value should be clearly smaller than the Stress that results



Multidimensional Scaling. Fig. 2 Shepard diagram of MDS solution in Fig. 1

for random data. Other criteria (such as the number of points (n), the number of missing data, the restrictiveness of the MDS model, or the dimensionality of the MDS space (m)), but also the interpretability of the solution have to be taken into account. Indeed, it may be true that Stress is high but the configuration is nevertheless stable over replications of the data. This case can result if the data have a large random error component. MDS, then, acts as a *data smoother* that irons out the error in the distance representation.

MDS methods allow one to utilize many different proximity measures. One example is direct judgments of similarity or dissimilarity as in the example given above. Another example are intercorrelations of test items over a sample of persons. A third example are co-occurrence coefficients that assess how often an event X is observed together with another event Y.

MDS is also robust against randomly distributed missing data. Computer simulations show that some 80% of the proximities may be missing, provided the data contain little error and the number of points (n) is high relative to the dimensionality of the MDS space (m). The data can also be quite coarse and even dichotomous.

A popular variety of MDS is *Individual Differences Scaling* or INDSCAL (Carroll and Chang 1970). Here, we have N different proximity matrices, one for each of N persons. The idea of the model is that these proximities can

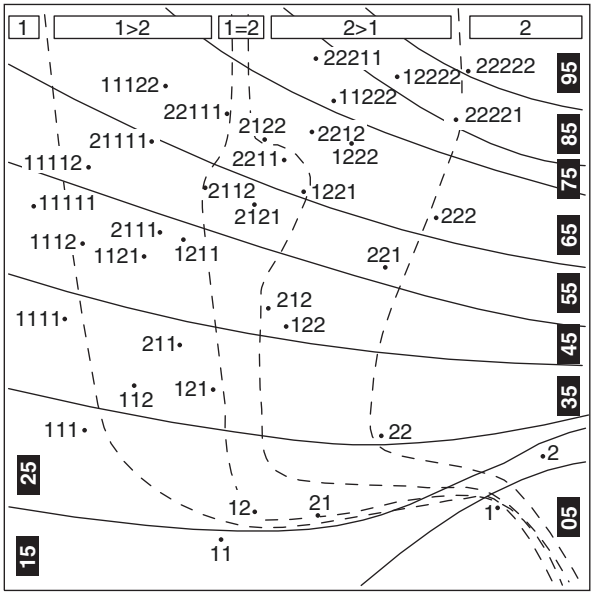
be explained by individually stretching or compressing a common MDS space along a fixed set of dimensions. That is,

$$d_{ij}^{(k)}(\mathbf{X}) = \sqrt{\sum_{a=1}^{m} w_a^{(k)} (x_{ia} - x_{ja})^2}, w_a^{(k)} \geq 0, \qquad (3)$$

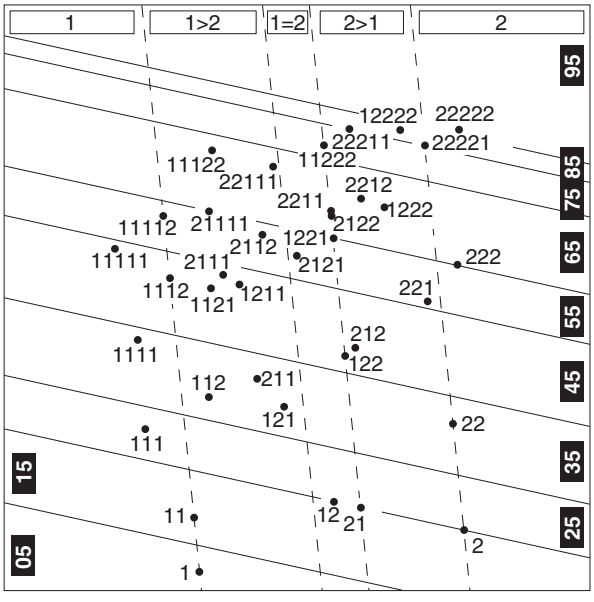
where $k = 1, \ldots, N$. The weight $w_a^{(k)}$ is interpreted as the salience of dimension a for individual k. Carroll and Wish (1974) used INDSCAL on the overall similarity ratings of different individuals for a set of countries, similar to the data discussed above. What they find is that one group of persons ("doves") pays much attention to economic development, while the other group ("falcons") emphasizes almost only political alignment of the countries with the West. Note, though, that these interpretations depend on the norming of $\mathbf{X}$. A more transparent way to analyze such data is to scale each individual's data matrix by itself, and then proceed by Procrustean fittings of the various solutions to each other, followed by finding optimal dimensions for an INDSCAL-type weighting model (Lingoes and Borg 1978).

A second popular variety of MDS is *Unfolding*. The prototypical data for this model are preference ratings of a set of persons for a set of objects. These data are mapped into distances between person-points and object-points in a "joint" space. The person-points are interpreted as "ideal" points that express the persons' points of maximal preference in the object space.

MDS solutions can be interpreted in different ways. The most popular approach is interpreting dimensions, but this is just a special case of interpreting regions. Regions are partitions of the MDS space which sort its points into subgroups that are equivalent in terms of substance. A systematic method for that purpose is *facet theory* (Borg and Shye 1995), an approach that offers methods to cross-classify the objects into substantively meaningful cells of a Cartesian product. The facets used for these classifications induce, one by one, partitions into the MDS space if they are empirically valid. The facets themselves are often based on theoretical considerations, but they can also be attributes that the objects possess by construction. Figure 3 shows an example. Here, (symmetrized) confusion probabilities of 36 Morse signals are represented as distances of a 2-dimensional MDS configuration. The space is partitioned by dashed lines into five regions that contain signals with only short beeps (coded as 1's); signals with more short than long (coded as 2's) beeps; etc. The solid lines cut the space into ten regions that each contain signals with equal duration (0.15 seconds to 0.95 seconds).



Multidimensional Scaling. Fig. 3 Exploratory MDS for confusion probabilities of 36 Morse signals



Multidimensional Scaling. Fig. 4 Confirmatory MDS for the Morse signals, enforcing linearized regions

The solution in Fig. 3 is found by *exploratory* ordinal MDS. There also exist various methods for *confirmatory* MDS that impose additional external constraints onto the MDS model. Figure 4 shows an example of an ordinal MDS with the additional constraint $\mathbf{X}$=$\mathbf{YC}$,

where **Y** is a 36 × 2 matrix of composition and duration codes, respectively, assigned to the 36 Morse signals; **C** is an unknown matrix of weights that re-scales **Y**'s columns monotonically. The confirmatory MDS procedure optimally represents the proximities in the sense of ordinal MDS while satisfying **X**=**YC**. The resulting configuration linearizes the regions of the MDS configuration which makes the solution easier to interpret. Provided its Stress is still acceptable, this is the preferred MDS representation, because it reflects a clear law of formation that is more likely to be replicable than an ad-hoc system of regions. Many alternative side constraints are conceivable. For example, an obvious modification is to require that **C** is diagonal. This enforces an orthogonal lattice of partitioning lines onto the solution in Fig. 4.

Many computer programs exist for doing MDS (for an overview, see Borg and Groenen (2005)). All large statistics packages offer MDS modules. One of the most flexible programs is Proxscal, one of the two MDS modules in Spss. The Spss package also offers Prefscal, a powerful program for unfolding. For R, De Leeuw and Mair (2009) have written a comprehensive MDS program called Smacof which can be freely downloaded from http://CRAN.R-project.org.

About the Author

Dr Ingwer Borg is Professor of Applied Psychological Methods at the University of Giessen (Giessen, Germany), and Scientific Director of the Department of Survey Design & Methodology at GESIS (Mannheim, Germany). He is Past President of the Facet Theory Association and of the International Society for the Study of Work and Organizational Values. He has published some 170 papers and 17 books, including *Modern Multidimensional Scaling* (with Patrick Groenen, Springer, 2005).

Cross References

►Data Analysis
►Distance Measures
►Multidimensional Scaling: An Introduction
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Sensometrics

References and Further Reading

Borg I, Groenen PJF (2005) Modern multidimensional scaling, 2nd edn. Springer, New York

Borg I, Shye S (1995) Facet theory: form and content. Sage, Newbury Park

Carroll JD, Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart-Young' decomposition. Psychometrika 35:283–320

Carroll JD, Wish M (1974) Multidimensional perceptual models and measurement methods. In: Carterette EC, Friedman MP (eds) Handbook of perception. Academic, New York, pp 391–447

Kruskal JB, Wish M (1978) Multidimensional scaling. Sage, Beverly Hills

Lingoes JC, Borg I (1978) A direct approach to individual differences scaling using increasingly complex transformations. Psychometrika, 43:491–519

Wish M (1971) Individual differences in perceptions and preferences among nations. In: King CW, Tigert D (eds) Attitude research reaches new heights. American Marketing Association, Chicago

Multidimensional Scaling: An Introduction

Nataša Kurnoga Živadinović
Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

►Multidimensional scaling (MDS), also called perceptual mapping, is based on the comparison of objects (persons, products, companies, services, ideas, etc.). The purpose of MDS is to identify the relationships between objects and to represent them in geometrical form. MDS is a set of procedures that allows the researcher to map distances between objects in a multidimensional space into a lower-dimensional space in order to show how the objects are related.

MDS was introduced by Torgerson (1952). It has its origins in psychology where it was used to understand respondents' opinions on similarities or dissimilarities between objects. MDS is also used in marketing, management, finance, sociology, information science, political science, physics, biology, ecology, etc. For example, it can be used to understand the perceptions of respondents, to identify unrecognized dimensions, for segmentation analysis, to position different brands, to position companies, and so on (for descriptions of various examples, see Borg and Groenen 2005 and Hair et al. 2010).

MDS starts from the proximities between the objects that express the similarity between them. There are different types of MDS: metric MDS (the similarities data are quantitative; input and output matrices are metric) and nonmetric MDS (the similarities data are qualitative; input matrix is nonmetric).

The steps involved in conducting MDS consist of problem formulation, selection of MDS procedure, determination of the number of dimensions, interpretation, and

validation. Problem formulation includes several tasks. First, the objectives of MDS should be identified. The nature of the variables to be included in MDS should be specified. Also, an appropriate number of variables should be chosen as the number of variables influences the resulting solution. The selection of MDS procedure depends on the nature of the input data (metric or nonmetric). Nonmetric MDS procedures assume that the input data is ordinal, but the resulting output is metric. Metric MDS procedures assume that both input and output data are metric. MDS procedures estimate the relative position of each object in a multidimensional space. The researcher must decide on a number of dimensions. The objective is to achieve an MDS solution that best fits the data in the smallest number of dimensions. Though the fit improves as the number of dimensions increases, the interpretation becomes more complicated. The interpretation of the dimensions and the configuration require subjective judgment, including some elements of judgment on the part of both the researcher and the respondent. The objectives of MDS are not achieved if an appropriate interpretation is lacking. Ultimately, the researcher must consider the quality of the MDS solution. (For detailed descriptions of MDS steps, see Cox and Cox 2001, Hair et al. 2010, and Kruskal and Wish 1978.)

To apply MDS, the distances between objects must first be calculated. The Euclidean distance is the most commonly used distance measure. The distance between objects A and B is given by $d_{AB} = \sqrt{\sum_{i=1}^{v}(x_{Ai} - x_{Bi})^2}$. MDS begins with a matrix $(n \times n)$ consisting of the distances between objects. From the calculated distances, a graph showing the relationship among objects is constructed.

The graphical representation used in MDS is a perceptual map, also called a spatial map. It represents the respondent's perceptions of objectives and shows the relative positioning of all analyzed objects. Let us suppose that there are five objects, A, B, C, D, and E. If objects A and B are judged by the respondents as most similar in comparison to all other pairs of objects (AC, AD, AE, BC, BD, etc.), the MDS procedures will position the objects A and B so that their distance is smaller than the distance of any other two objects. A perceptual map is constructed in two or more dimensions. In a two-dimensional map, objects are represented by points on a plane. In the case of a higher number of dimensions, graphical representation becomes more complicated.

MDS can be conducted at the individual or group level. At the individual level, perceptual maps should be constructed on a respondent-by-respondent base. At the group level, the average judgment of all respondents within a group should be established and the perceptual maps of one or more groups constructed.

Statistical packages such as statistical analysis system (SAS), statistical package for the social sciences (SPSS), Stata, and STATISTICA are suitable for MDS.

Methods closely related to MDS are factor analysis (see ►Factor Analysis and Latent Variable Modelling), ►correspondence analysis, and cluster analysis (see Borg and Groenen 2005, Hair et al. 2010; see also the entry ►Cluster Analysis: An Introduction).

Cross References

►Data Analysis
►Distance Measures
►Multidimensional Scaling
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis

References and Further Reading

Borg I, Groenen PJF (2005) Modern multidimensional scaling: theory and applications. Springer Series in Statistics. 2nd edn. Springer, New York

Cox TF, Cox AA (2001) Multidimensional scaling, 2nd edn. Chapman and Hall/CRC, Boca Raton

Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis: a global perspective, 7th edn. Pearson Education, Upper Saddle River

Kruskal JB, Wish M (1978) Multidimensional scaling. SAGE University Paper Series: Quantitative Applications in the Social Sciences. SAGE, Newbury Park

Torgerson WS (1952) Multidimensional scaling: I. Theory and method. Psychometrika, 17(4):401–419

Multilevel Analysis

Tom A. B. Snijders
Professor of Statistics
University of Oxford, Oxford, UK
Professor of Methodology and Statistics, Faculty of Behavioral and Social Sciences
University of Groningen, Groningen, Netherlands

Multilevel Analysis, Hierarchical Linear Models

The term "Multilevel Analysis" is mostly used interchangeably with "Hierarchical Linear Modeling," although strictly speaking these terms are distinct. Multilevel Analysis may be understood to refer broadly to the methodology of

research questions and data structures that involve more than one type of unit. This originated in studies involving several levels of aggregation, such as individuals and counties, or pupils, classrooms, and schools. Starting with Robinson's (1950) discussion of the *ecological fallacy*, where associations between variables at one level of aggregation are mistakenly regarded as evidence for associations at a different aggregation level (see Alker 1969, for an extensive review), this led to interest in how to analyze data including several aggregation levels. This situation arises as a matter of course in educational research, and studies of the contributions made by different sources of variation such as students, teachers, classroom composition, school organization, etc., were seminal in the development of statistical methodology in the 1980s (see the review in Chap. 1 of de Leeuw and Meijer 2008). The basic idea is that studying the simultaneous effects of variables at the levels of students, teachers, classrooms, etc., on student achievement requires the use of regression-type models that comprise error terms for each of those levels separately; this is similar to mixed effects models studied in the traditional linear models literature such as Scheffé (1959).

The prototypical statistical model that expresses this is the *Hierarchical Linear Model*, which is a mixed effects regression model for nested designs. In the two-level situation – applicable, e.g., to a study of students in classrooms – it can be expressed as follows. The more detailed level (students) is called the lower level, or level 1; the grouping level (classrooms) is called the higher level, or level 2. Highlighting the distinction with regular regression models, the terminology speaks of *units* rather than cases, and there are specific types of unit at each level. In our example, the level-1 units, students, are denoted by i and the level-2 units, classrooms, by j. Level-1 units are nested in level-2 units (each student is a member of exactly one classroom) and the data structure is allowed to be unbalanced, such that j runs from 1 to N while i runs, for a given j, from 1 to n_j. The basic two-level hierarchical linear model can be expressed as

$$Y_{ij} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj} z_{hij} + R_{ij}; \qquad (1a)$$

or, more succinctly, as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\ \mathbf{U} + \mathbf{R}. \qquad (1b)$$

Here Y_{ij} is the dependent variable, defined for level-1 unit i within level-2 unit j; the variables x_{hij} and z_{hij} are the explanatory variables. Variables R_{ij} are residual terms, or error terms, at level 1, while U_{hj} for $h = 0, \ldots, p$ are residual terms, or error terms, at level 2. In the case $p = 0$ this is called a *random intercept model*, for $p \geq 1$ it is called a *random slope model*. The usual assumption is that all R_{ij} and all vectors $U_j = (U_{0j}, \ldots, U_{pj})$ are independent, R_{ij} having a normal $\mathcal{N}(0, \sigma^2)$ and U_j having a multivariate normal $\mathcal{N}_{p+1}(\mathbf{0}, \mathbf{T})$ distribution. Parameters β_h are regression coefficients (fixed effects), while the U_{hj} are random effects. The presence of both of these makes (1) into a mixed linear model. In most practical cases, the variables with random effects are a subset of the variables with fixed effects ($x_{hij} = z_{hij}$ for $h \leq p; p \leq r$), but this is not necessary.

More Than Two Levels

This model can be extended to a three- or more-level model for data with three or more nested levels by including random effects at each of these levels. For example, for a three level structure where level-3 units are denoted by $k = 1, \ldots, M$, level-2 units by $j = 1, \ldots, N_k$, and level-1 units by $i = 1, \ldots, n_{ij}$, the model is

$$Y_{ijk} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hijk} + U_{0jk} + \sum_{h=1}^{p} U_{hjk} z_{hijk} + V_{0k} + \sum_{h=1}^{q} V_{hk} w_{hijk} + R_{ijk}, \qquad (2)$$

where the U_{hjk} are the random effects at level 2, while the V_{hk} are the random effects at level 3. An example is research into outcome variables Y_{ijk} of students (i) nested in classrooms (j) nested in schools (k), and the presence of error terms at all three levels provides a basis for testing effects of pupil variables, classroom or teacher variables, as well as school variables.

The development both of inferential methods and of applications was oriented first to this type of nested models, but much interest now is given also to the more general case where the restriction of nested random effects is dropped. In this sense, multilevel analysis refers to methodology of research questions and data structures that involve several sources of variation – each type of units then refers to a specific source of variation, with or without nesting. In social science applications this can be fruitfully applied to research questions in which different types of *actor* and *context* are involved; e.g., patients, doctors, hospitals, and insurance companies in health-related research; or students, teachers, schools, and neighborhoods in educational research. The word "level" then is used for such a type of units. Given the use of random effects, the most natural applications are those where each "level" is associated with some population of units.

Longitudinal Studies

A special area of application of multilevel models is longitudinal studies, in which the lowest level corresponds to repeated observations of the level-two units. Often the level-two units are individuals, but these may also be organizations, countries, etc. This application of mixed effects models was pioneered by Laird and Ware (1982). An important advantage of the hierarchical linear model over other statistical models for longitudinal data is the possibility to obtain parameter estimates and tests also under highly unbalanced situations, where the number of observations per individual, and the time points where they are measured, are different between individuals. Another advantage is the possibility of seamless integration with nesting if individuals within higher-level units.

Model Specification

The usual considerations for model specification in linear models apply here, too, but additional considerations arise from the presence in the model of the random effects and the data structure being nested or having multiple types of unit in some other way. An important practical issue is to avoid the ecological fallacy mentioned above; i.e., to attribute fixed effects to the correct level. In the original paper by Robinson (1950), one of the examples was about the correlation between literacy and ethnic background as measured in the USA in the 1930s, computed as a correlation at the individual level, or at the level of averages for large geographical regions. The correlation was .203 between individuals, and .946 between regions, illustrating how widely different correlations at different levels of aggregation may be.

Consider a two-level model (1) where variable X_1 with values x_{1ij} is defined as a level-1 variable – literacy in Robinson's example. For "level-2 units" we also use the term "groups." To avoid the ecological fallacy, one will have to include a relevant level-2 variable that reflects the composition of the level-2 units with respect to variable X_1. The mostly used composition variable is the group mean of X_1,

$$\bar{x}_{1.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{1ij}.$$

The usual procedure then is to include x_{1ij} as well as $\bar{x}_{1.j}$ among the explanatory variables with fixed effects. This allows separate estimation of the within-group regression (the coefficient of x_{1ij}) and the between-group regression (the sum of the coefficients of x_{1ij} and $\bar{x}_{1.j}$).

In some cases, notably in many economic studies (see Greene 2003), researchers are interested especially in the within-group regression coefficients, and wish to control for the possibility of unmeasured heterogeneity between the groups. If there is no interest in the between-group regression coefficients one may use a model with fixed effects for all the groups: in the simplest case this is

$$Y_{ij} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hij} + \gamma_j + R_{ij}. \qquad (3)$$

The parameters γ_j (which here have to be restricted, e.g., to have a mean 0 in order to achieve identifiability) then represent all differences between the level-two units, as far as these differences apply as a constant additive term to all level-1 units within the group. For example in the case of longitudinal studies where level-2 units are individuals and a linear model is used, this will represent all time-constant differences between individuals. Note that (3) is a linear model with only one error term.

Model (1) implies the distribution

$$\mathbf{y} \sim \mathcal{N}_p \left(\mathbf{X}\beta, \mathbf{Z}\,\mathbf{T}\,\mathbf{Z}' + \sigma^2 I \right).$$

Generalizations are possible where the level-1 residual terms R_{ij} are not i.i.d.; they can be heteroscedastic, have time-series dependence, etc. The specification of the variables Z having random effects is crucial to obtain a well-fitting model. See Chap. 9 of Snijders and Bosker (1999), Chap. 9 of Raudenbush and Bryk (2002), and Chap. 3 of de Leeuw and Meijer (2008).

Inference

A major reason for the take-off of multilevel analysis in the 1980s was the development of algorithms for maximum likelihood estimation for unbalanced nested designs. The EM algorithm (Dempster et al. 1981), Iteratively Reweighted Least Squares (Goldstein 1986), and Fisher Scoring (Longford 1987) were applied to obtain ML estimates for hierarchical linear models. The MCMC implementation of Bayesian procedures has proved very useful for a large variety of more complex multilevel models, both for non-nested random effects and for generalized linear mixed models; see Browne and Draper (2000) and Chap. 2 of de Leeuw and Meijer (2008).

Hypothesis tests for the fixed coefficients β_h can be carried out by Wald or Likelihood Ratio tests in the usual way. For testing parameters of the random effects, some care must be taken because the estimates of the random effect variances τ^2_{hh} (the diagonal elements of $\mathbf{T}$) are not approximately normally distributed if $\tau^2_{hh} = 0$. Tests for these parameters can be based on estimated fixed effects, using least squares estimates for U_{hj} in a specification where these are treated as fixed effects (Bryk and Raudenbush 2002, Chap. 3); based on appropriate distributions of the log likelihood ratio; or obtained as score tests (Berkhof and Snijders 2001).

About the Author

Professor Snijders is Elected Member of the European Academy of Sociology (2006) and Elected Correspondent of the Royal Netherlands Academy of Arts and Sciences (2007). He was awarded the Order of Knight of the Netherlands Lion (2008). Professor Snijders was Chairman of the Department of Statistics, Measurement Theory, and Information Technology, of the University of Groningen (1997–2000). He has supervised 52 Ph.D. students. He has been associate editor of various journals, and Editor of *Statistica Neerlandica* (1986–1990). Currently he is co-editor of *Social Networks*, Associate editor of *Annals of Applied Statistics*, and Associate editor of *Journal of Social Structure*. Professor Snijders has (co-)authored about 100 refereed papers and several books, including *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. (with Bosker, R.J., London etc.: Sage Publications, 1999). In 2005, he was awarded an honorary doctorate in the Social Sciences from the University of Stockholm.

Cross References

►Bayesian Statistics
►Cross Classified and Multiple Membership Multilevel Models
►Mixed Membership Models
►Moderating and Mediating Variables in Psychological Research
►Nonlinear Mixed Effects Models
►Research Designs
►Statistical Analysis of Longitudinal and Correlated Data
►Statistical Inference in Ecology

References and Further Reading

To explore current research activities and to obtain information training materials etc., visit the website www.cmm.bristol.ac.uk. There is also an on-line discussion group at www.jiscmail.ac.uk/lists/multilevel.html.

There is a variety of textbooks, such as Goldstein (2003), Longford (1993), Raudenbush and Bryk (2003), and Snijders and Bosker (1999). A wealth of material is contained in de Leeuw and Meijer (2008).

Alker HR (1969) A typology of ecological fallacies. In: Dogan M, Rokkan S (eds) Quantitative ecological analysis in the social sciences. MIT Press, Cambridge, pp 69–86

Berkhof J, Snijders TAB (2001) Variance component testing in multilevel models. J Educ Behav Stat 26:133–152

Browne WJ, Draper D (2000) Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Stat 15:391–420

de Leeuw J, Meijer E (2008) Handbook of multilevel analysis. Springer, New York

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance components models. J Am Stat Assoc 76:341–353

Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73:43–56

Goldstein H (2003) Multilevel statistical models, 3rd edn. Edward Arnold, London

Greene W (2003) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974

Longford NT (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika 74:812–827

Longford NT (1993) Random coefficient models. Oxford University Press, New York

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. Sage, Thousand Oaks

Robinson WS (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15:351–357

Scheffé H (1959) The analysis of variance. Wiley, New York

Snijders TAB, Bosker RJ (1999) Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage, London

Multinomial Distribution

George A. F. Seber
Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand

The Multinomial distribution arises as a model for the following experimental situation. An experiment or "trial" is carried out and the outcome occurs in one of k mutually exclusive categories with probabilities p_i, $i = 1, 2, \ldots, k$. For example, a person may be selected at random from a population of size N and their ABO blood phenotype recorded as A, B, AB, or O ($k = 4$). If the trial is repeated n times such that the trials are mutually independent, and if x_i is the frequency of occurrence in the ith category, then the joint probability function of the x_i is

$$P_1(x_1, x_1, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

where $\sum_{i=1}^{k} x_i = n$ and $\sum_{1=1}^{k} p_i = 1$. This would be the correct probability function for the genetics example if further people were chosen with replacement. In practice, sampling is without replacement and the correct distribution is the multivariate hypergeometric, a difficult distribution to deal with. Fortunately, all is not lost, as when the sampling fraction $f = n/N$ is small enough (say less than 0.1 or preferably less than 0.05), the Multinomial distribution

is a good approximation and is used extensively in genetics (e.g., Greenwood and Seber 1992). We note that when $k = 2$ we have the ▶Binomial distribution. Also the terms of P_1 can be obtained by expanding $(p_1 + p_2 + \cdots + p_k)^n$.

Various properties of the Multinomial distribution can be derived using extensive algebra. However, they are more readily obtained by noting that any subset of a multinomial distribution is also Multinomial. We simply group the categories relating to the remaining variables into a single category. For example x_i will have a Binomial distribution as there are just two categories, the ith and the rest combined. Hence the mean and variance of x_i are

$$\mathrm{E}(x_i) = np_i \text{ and } \mathrm{var}(x_i) = np_iq_i,$$

where $q_i = 1 - p_i$. Also, if we combine the ith and jth category and then combine the rest into single category, we see that $x_i + x_j$ is Binomial with probability parameter $p_i + p_j$ and variance $n(p_i + p_j)(1 - p_i - p_j)$. Hence the covariance of x_i and x_j is

$$\mathrm{cov}(x_i, x_j) = \frac{1}{2}[\mathrm{var}(x_i + x_j) - \mathrm{var}(x_i) - \mathrm{var}(x_j)] = -np_ip_j.$$

Another useful result that arises in comparing proportions p_i and p_j in a ▶questionnaire is

$$\begin{aligned}\mathrm{var}(x_i - x_j) &= \mathrm{var}(x_i) + \mathrm{var}(x_j) - 2\mathrm{cov}(x_i, x_j)\\ &= n[p_i + p_j - (p_i - p_j)^2]. \end{aligned} \quad (1)$$

It should be noted that the Multinomial distribution given above is a "singular" distribution as the random variables satisfy the linear constraint $\sum_{i=1}^{k} x_i = n$, which leads to a singular variance-covariance matrix. We can instead use the "nonsingular" version

$$P_2(x_1, x_2, \ldots, x_{k-1}) = \frac{n!}{x_1!x_2!\cdots\left(n - \sum_{i=1}^{k-1} x_i\right)!} \times p_1^{x_1}p_2^{x_2}\cdots p_k^{n-\sum_{i=1}^{k-1} x_i}.$$

We note that the joint ▶moment generating function of $\boldsymbol{x}$ is

$$M(\boldsymbol{t}) = (p_1e^{t_1} + p_2e^{t_2} + \cdots + p_{k-1}e^{t_{k-1}} + p_k)^n,$$

which can also be used to derive the above properties of the Multinomial distribution as well as the ▶asymptotic normality properties described next.

Let $\hat{p}_i = x_i/n$ be the usual estimate of p_i. Given the vectors $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k-1})'$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{k-1})'$, then the mean of $\hat{\boldsymbol{p}}$ is $\boldsymbol{p}$ and its variance-covariance matrix is $n^{-1}\boldsymbol{V}$, where $\boldsymbol{V} = (\mathrm{diag}\,\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}')$ and $\mathrm{diag}\,\boldsymbol{p}$ is a diagonal matrix with diagonal elements $p_1, p_2, \ldots p_{k-1}$. In the same way that a Binomial random variable is asymptotically normal for large n, $\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ is asymptotically multivariate Normal with mean vector 0 and variance-covariance matrix $\boldsymbol{V}$. If $\boldsymbol{V}^{-1}$ is the inverse of $\boldsymbol{V}$, then $\boldsymbol{V}^{-1} = n^{-1}\left((\mathrm{diag}\,\boldsymbol{p})^{-1} + p_k^{-1}1_{k-1}'1_{k-1}\right)$, where 1_{k-1} is a column $k-1$ ones (cf. Seber, 2008, 15.7). From the properties of the multivariate Normal distribution (cf. Seber 2008, 20.25),

$$n(\hat{\boldsymbol{p}} - \boldsymbol{p})'\boldsymbol{V}^{-1}(\hat{\boldsymbol{p}} - \boldsymbol{p}) = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} \quad (2)$$

will be asymptotically distributed as the ▶Chi-square distribution with $k-1$ degrees of freedom. If we use the singular version and include x_k to expand $\boldsymbol{V}$ to $\boldsymbol{V}_k$, we can obtain the result more quickly using a generalized inverse (cf. Seber, 2008, 20.29b using $\boldsymbol{A} = \boldsymbol{V}_k^{-} = (\mathrm{diag}\,(\boldsymbol{p}', p_k)')^{-1}$). This link with the Chi-square distribution forms the basis of a number of tests involving the Multinomial distribution mentioned below.

We see that $P_1(\cdot)$ above can be regarded conceptually as a nonsingular distribution for the x_i $(i = 1, 2, \ldots, k)$ with probabilities π_i, but conditional on $\sum_{i=1}^{k} x_i = n$ with $p_i = \pi_i / \sum_{i=1}^{k} \pi_i$. It therefore follows that the joint distribution of any subset of multinomial variables conditional on their sum is also multinomial. For example, the distribution of x_1 and x_2 given $x_1 + x_2 = n$ is Binomial with probability parameter $p_1/(p_1 + p_2)$. We get a similar result in ecology where we have a population of plants divided up into k areas with x_i in the ith area being distributed as the Poisson distribution with mean μ_i. If the x_i are mutually independent, then the joint distribution of the x_i conditional on the sum $\sum_{i=1}^{k} x_i$ is Multinomial with probabilities $p_i = \mu_i / \sum_{j=1}^{k} \mu_j$.

The last topic I want to consider briefly is inference for the multinomial distribution. Estimating p_i by $\hat{p}_i = x_i/n$, using the normal approximation, and applying (1), we can obtain a confidence interval for any particular p_i or any particular difference $p_i - p_j$. Simultaneous confidence interval procedures are also available for all the p_i or all differences using the Bonferroni method. We can also test $\boldsymbol{p} = \boldsymbol{p}_0$ using (2).

A common problem is testing the hypothesis H_0 : $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$, where $\boldsymbol{p}$ is a known function of some unknown t-dimensional parameter $\boldsymbol{\theta}$ (e.g., the genetics example above). This can be done using a derivation like the one that led to (2) above, giving the so-called "goodness of fit" statistic, but with $\boldsymbol{p}$ replaced by $\boldsymbol{p}(\hat{\boldsymbol{\theta}})$. Here $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate of $\boldsymbol{\theta}$, is asymptotically Normal so that $\boldsymbol{p}(\hat{\boldsymbol{\theta}})$ is also asymptotically Normal. Under H_0, it can be shown that the test statistic is approximately Chi-square with degrees of freedom now $k - 1 - t$.

One application of the above is to the theory of contingency tables. We have an $r \times c$ table of observations x_{ij}

$(i = 1, 2, \ldots, r; j = 1, 2, \ldots, c)$ and p_{ij} is the probability of falling in the (i,j)th category. Treating the whole array as a single Multinomial distribution, one hypothesis of interest is $H_0 : p_{ij} = \alpha_i \beta_j$, where $\sum_{i=1}^{r} \alpha_i = 1$ and $\sum_{j=1}^{c} \beta_j = 1$. In this hypothesis of row and column independence, we have $\boldsymbol{\theta}' = (\alpha_1, \ldots \alpha_{r-1}, \beta_1, \ldots, \beta_{c-1})$ with maximum likelihood estimates $\hat{\alpha}_i = R_i/n$ and $\hat{\beta}_j = c_j/n$, where r_i is the ith row sum of the table and c_j the jth column sum. The statistic for the test of independence is therefore

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_i - r_i c_j/n)^2}{r_i c_j/n}, \tag{3}$$

which, under H_0, is approximately Chi-square with $rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$ degrees of freedom. If the rows of the $r \times c$ table now represents r independent Multinomial distributions with $\sum_{j=1}^{c} p_{ij} = 1$ for $i = 1, 2, \ldots, r$, then the hypothesis that the distributions are identical is $H_0 : p_{ij} = \gamma_j$ for $i = 1, 2, \ldots, r$, where $\sum_{j=1}^{c} \gamma_j = 1$. Pooling the common distributions, the maximum likelihood estimate of γ_j is $\hat{\gamma}_j = c_j/n$ so that the term $np_{ij}(\hat{\theta})$ becomes $r_i \hat{\gamma}_j$ and the test statistic for testing homogeneity turns out to be the same as (3) with the same degrees of freedom.

The above chi-squared tests are not particularly powerful and need to be backed up with various confidence interval procedures. Other asymptotically equivalent tests are the likelihood ratio test and the so-called "score" (Lagrange multiplier) test. Log linear models can also be used. For further properties of the Multinomial distribution see Johnson et al. (1997, Chap. 35) and asymptotic background theory for the chi-squared tests is given by Bishop et al. (1975, Chap. 14). More recent developments are given by Agresti (2002).

About the Author

For biography *see* the entry ►Adaptive Sampling.

Cross References

►Binomial Distribution
►Categorical Data Analysis
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Divisible Statistics
►Entropy and Cross Entropy as Diversity and Distance Measures
►Geometric and Negative Binomial Distributions
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge
Greenwood SR, Seber GAF (1992) Estimating blood phenotypes probabilities and their products. Biometrics 48:143–154
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York
Seber GAF (2008) A matrix handbook for statisticians. Wiley, New York

Multi-Party Inference and Uncongeniality

Xiao-Li Meng
Professor, Chair
Harvard University, Cambridge, MA, USA

► *"Life is more complicated when you have three uncongenial models involved."*

The Multi-Party Inference Reality

Much of the statistical inference literature uses the familiar framework of "God's model versus my model." That is, an unknown model, "God's model," generates our data, and our job is to infer this model or at least some of its characteristics (e.g., moments, distributional shape) or implications (e.g., prediction). We first postulate one or several models, and then use an array of estimation, testing, selection, and refinement methods to settle on a model that we judge to be acceptable – according to some sensible criterion, hopefully pre-determined – for the inference goals at hand, even though we almost never can be sure that our chosen model resembles God's model in critical ways. Indeed, philosophically even the existence of God's model is not a universally accepted concept, just as theologically the existence of God is not an unchallenged notion.

Whether one does or does not adopt the notion of God's model, it is repeatedly emphasized in the literature that to select a reasonable model, an iterative process is necessary and hence multiple models are typically considered (e.g., see Box and Tiao 1973, Chap. 1; Gelman and Meng 1996). By *multiple models* we mean multiple sets of mathematically quantifiable assumptions (hence, not necessarily parametric models), which are compatible within each set but not across different sets. Indeed, if they are not incompatible across different sets then one is simply postulating a larger model; see McCullagh (2002). In this

sense we automatically take a "monotheistic" point of view that there is only one God's model; we assume God's model contains no self-contradiction (or at least none detectable by a human modeler). However, we do not go so far as to suggest that the modeler can always embed everything into one model, e.g., as in Bayesian model averaging, because contrasting models sometimes is as useful as, if not more so than, combining models.

Whereas many models may be entertained, the commonly accepted paradigm involves only two parties: the (hypothetical) God, and "me" – the modeler. Unfortunately, reality is far more complicated. To explain the complication, we must distinguish the *modeler's data* from *God's data*. The modeler's data are the data available to the modeler, whereas God's data are the realizations from God's model that the modeler's data were collected to *approximate*. Whereas any attempt to mathematically define such concepts is doomed to fail, it is useful to distinguish the two forms of data because the *approximation* process introduces an additional inference party (or parties).

For example, in the physical sciences, the modeler's data typically are results of a series of pre-processing steps to deal with limitations or irregularities in recording God's data (e.g., discarding "outliers" (see ▶Outliers); recalibration to account for instrument drift), and typically the modeler at best only has partial information about this process. For the social and behavioral sciences, some variables are not even what we normally think they are, such as responses to a questionnaire survey. Rather, they are so-called "constructed variables," typically from a deterministic algorithm converting a set of answers to an index that indicates, say, whether a subject is considered to suffer major depression. The algorithm is often a black box, and in some cases it is pitch black because the modeler is not even informed of what variables were used as inputs to produce the output. In the context of public-use data files, virtually all data sets contain imputations of some sort (see ▶Imputation) because of non-responses or other forms of missing data (e.g., missingness by design such as with matrix sampling), which means someone has "fixed the holes" in the data before they reach the modeler.

In all these examples, the key issue is not that there is data pre-processing step per se, but rather that during the journey from God's data to modeler's data, a set of assumptions has been introduced. There is no such thing as "assumption-free" pre-processing; any attempt to make the data "better" or "more usable" implies that a judgment has been made. Under the God-vs.-me paradigm, this intermediate "data cleaning" process has to be considered either as part of God's model, or of the modeler's model, or of both by somehow separating aspects of the process (e.g., one could argue that a refused answer to an opinion question is an opinion itself, whereas a refusal to an income question is a non-response). Regardless of how we conceptualize, we find ourselves in an extremely muddy – if not hopeless – situation. For example, if aspects of this intermediate process are considered to be part of God's model, then the modeler's inference is not just about God's model but also about someone else's assumptions about it. If we relegate the pre-processing to the modeler's model, then the modeler will need good information on the process. Whereas there has been an increasing emphasis on understanding the entire mechanism that leads to the modeler's data, the reality is that for the vast majority of real-life data sets, especially large-scale ones, it is simply impossible to trace back how the data were collected or pre-processed. Indeed, many such processes are nowhere documented, and some are even protected by confidentiality constraints (e.g., confidential information may be used for imputation by a governmental agency).

This intermediate "data cleaning" process motivates the *multi-party inference* paradigm. The term is self-explanatory: we acknowledge that there is more than one party involved in reaching the final inference. The key distinction between the multi-party paradigm and the God-vs.-me paradigm is not that the former involves more sets of assumptions, i.e., models – indeed under the latter we still almost always (should) consider multiple models. Rather, in the multi-party paradigm, we explicitly acknowledge the *sequential nature* of the parties' involvement, highlighted by how the intermediate party's assumptions impact the final inference, because typically they are necessarily incompatible with the modeler's assumptions, due both to the parties' having access to different amounts of information and to their having different objectives.

This situation is most vividly demonstrated by multiple imputation inference (Rubin 1987), where the intermediate party is the imputer. (There is often more than one intermediate party even in the imputation context, but the case of a single imputer suffices to reveal major issues.) In such a setting, the concept of *congeniality* (Meng 1994) is critical. In a nutshell, congeniality means that the imputation model and the analysis model are compatible for the purposes of predicting the missing data. In real life, this typically is not the case, even if the imputer and analyst are the same entity, because of the different aims of imputation (where one wants to use as many variables as possible even if causal directions are incorrectly specified) and of analysis (where one may be only interested in a subset of variables with specified causal directions). The next section demonstrates the importance

of recognizing *uncongeniality*, which directly affects the validity of the final inferences. The concept of uncongeniality was originally defined and has thus far been investigated in the context of multiple imputation inference, the most well-studied case of multi-party inference. However, its general implication is broad: to reach valid inference when more than one party is involved, we must consider the incompatibility/uncongeniality among their assumptions/models, even if each party has made assumptions that are consistent with God's model and has carried out its task in the best possible way given the information available at the time.

Uncongeniality in Multiple Imputation Inference

A common method for dealing with non-response in surveys and incomplete data in general is imputation (Little and Rubin 2002). Briefly, imputation is a prediction of the missing data from a posited (not necessarily parametric) model $p_I(Y_{mis}|Y_{obs})$, where Y_{mis} denotes the missing data and Y_{obs} the observed data. The trouble with single imputation, however sophisticated, is that the resulting data set cannot be analyzed in the same way as would an authentic complete data set, without sacrificing the validity of the inference. Multiple imputation (MI; Rubin 1987) attempts to circumvent this problem by providing multiple predictions from $p_I(Y_{mis}|Y_{obs})$, thereby permitting, via genuine replications, a direct assessment of uncertainties due to imputation.

Specifically, in the MI framework, we draw independently m times from $p_I(Y_{mis}|Y_{obs})$, resulting in m completed-data sets: $Y_{com}^{(\ell)} = \left\{Y_{obs}, Y_{mis}^{(\ell)}\right\}, \quad \ell = 1, \ldots, m.$ Suppose our complete-data analysis can be summarized by a point estimator $\hat{\theta}(Y_{com})$ and an associated variance estimator $U(Y_{com})$, where Y_{com} denotes $\{Y_{mis}, Y_{obs}\}$. The MI inference procedure consists of the following steps:

Step 1: Perform m complete-data analyses as if each $Y_{com}^{(\ell)}$ were real data:

$$\hat{\theta}_\ell \equiv \hat{\theta}\left(Y_{com}^{(\ell)}\right), \text{ and } U_\ell \equiv U\left(Y_{com}^{(\ell)}\right), \quad \ell = 1, \ldots, m.$$

Step 2: Use Rubin's Combining Rules:

$$\bar{\theta}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}_\ell, \text{ and } T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where

$$\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} U_\ell \text{ and } B_m = \frac{1}{m-1}\sum_{\ell=1}^{m} (\hat{\theta}_\ell - \bar{\theta}_m)(\hat{\theta}_\ell - \bar{\theta}_m)^\top$$

are respectively the within-imputation variance and the between-imputation variance, to reach the MI inference $\{\bar{\theta}_m, T_m\}$, with T_m the variance estimator of $\bar{\theta}_m$.

The justification of Rubin's combining rules is most straightforward under strict congeniality, which means that both the analyst and the imputer use (effectively) Bayesian models, and their Bayesian models are compatible. That is, we assume:

(I) The complete-data analysis procedure can be embedded into a Bayesian model, with

$$\hat{\theta}(Y_{com}) = E_A(\theta|Y_{com}) \text{ and } U(Y_{com}) = \mathrm{V}_A(\theta|Y_{com}),$$

where the subscript A indexes expectation with respect to the embedded analysis model;

(II) The imputer's model and the (embedded) analysis model are the same for the purposes of predicting missing data:

$$P_I(Y_{mis}|Y_{obs}) = P_A(Y_{mis}|Y_{obs}), \quad \textit{for all } Y_{mis} \textit{ (but the given } Y_{obs}).$$

Then for $\bar{\theta}_m$ as $m \to \infty$, we have

$$\begin{aligned}\bar{\theta}_\infty &= \mathrm{E}_I\left[\hat{\theta}(Y_{com})\middle|Y_{obs}\right] \\ &< \text{by (I)} > = \mathrm{E}_I\left[\mathrm{E}_A(\theta|Y_{com})\middle|Y_{obs}\right] \\ &< \text{by (II)} > = \mathrm{E}_A\left[\mathrm{E}_A(\theta|Y_{com})\middle|Y_{obs}\right] = \mathrm{E}_A(\theta|Y_{obs}).\end{aligned}$$

That is, the MI estimator $\bar{\theta}_m$ simply is a consistent (Monte Carlo) estimator of the posterior mean under the analyst's model based on the observed data Y_{obs}. The critical role of (II) is also vivid in establishing the validity of $T_m = \bar{U}_m + (1 + m^{-1})B_m$ as $m \to \infty$:

$$\begin{aligned}\bar{U}_\infty + B_\infty &= E_I[U(Y_{com})|Y_{obs}] + \mathrm{V}_I[\hat{\theta}(Y_{com})|Y_{obs}] \\ &< by\ (I) > = E_I[\mathrm{V}_A(\theta|Y_{com})|Y_{obs}] \\ &\qquad + \mathrm{V}_I[E_A(\theta|Y_{com})|Y_{obs}] \\ &< by\ (II) > = E_A[\mathrm{V}_A(\theta|Y_{com})|Y_{obs}] \\ &\qquad + \mathrm{V}_A[E_A(\theta|Y_{com})|Y_{obs}] = \mathrm{V}_A(\theta|Y_{obs}).\end{aligned}$$

Therefore, as $m \to \infty$, $\{\bar{\theta}_m, T_m\}$ reproduces the posterior mean and posterior variance under the analyst's model given Y_{obs}, because $\bar{\theta}_\infty = E_A(\theta|Y_{obs})$ and $T_\infty = \mathrm{V}_A(\theta|Y_{obs})$.

When congeniality fails, either because the analyst's procedure does not correspond to any Bayesian model or because the corresponding Bayesian model is incompatible with the imputer's model, the MI variance estimator T_m can overestimate or underestimate the variance of $\hat{\theta}_m$ even as $m \to \infty$. However, depending on the relationships

among God's model, the analyst's model and the imputer's model, we may still reach valid inference under uncongeniality. For example, under the assumption that the analyst's complete-data procedure is self-efficient (Meng 1994), if God's model is nested in the analyst's model, which in turn is nested in the imputer's model, then the MI confidence interval based on $\{\bar{\theta}_\infty, T_\infty\}$ is valid (asymptotically with respect to the size of the observed data). However, the MI estimator $\bar{\theta}_\infty$ may not be as efficient as the analyst's estimator (e.g., MLE) directly based on the observed data, because the additional assumptions built into the analysis model are not used by the imputer. But this comparison is immaterial when the analyst is unable to analyze the observed data directly, and therefore multiple imputation inference is needed (see ►Multiple Imputation).

However, the situation becomes more complicated if we assume God's model is nested in the imputer's model, which in turn is nested in the analyst's model. In such cases, it is possible to identify situations where the multiple imputation interval estimator is conservative in its own right, yet it is narrower than analyst's interval estimator (with the correct nominal coverage) directly based on the observed data (Xie and Meng 2010). This seemingly paradoxical phenomenon is due to the fact the imputer has introduced "secret" model assumptions into the MI inference, making it more efficient than the analyst's inference directly based on the observed data, which does not benefit from the imputer's assumptions. At the same time, since the analyst's complete-data procedure $\{\hat{\theta}(Y_{com}), U(Y_{com})\}$ is determined irrespective of the imputer's model, the imputer's secret assumption introduces uncongenality, which leads to the conservativeness of the MI interval. However, this is not to suggest that MI tends to be conservative, but rather to demonstrate the impact of imputation models on the MI inference and hence to provide practical guidelines on how to regulate the imputation models.

Even more complicated are situations where the analyst's and imputer's models do not nest, or where at least one of them does not contain God's model as a sub-model. Consequences of such are virtually undetermined at the present time, but one thing is clear. These complications remind us the importance of recognizing the multi-party inference paradigm, because the God-vs.-me paradigm sweeps all of them under the rug, or more precisely buries our heads in the sand, leaving our posteriors exposed without proper coverage.

Acknowledgment

The author thanks NSF for partial support, and Joseph Blitzstein, Yves Chretien and Xianchao Xie for very helpful comments and proofreading.

About the Author

Dr Xiao-Li Meng started his outstanding career in 1982 as Instructor of Mathematics in China Textile University and 22 years later has become Professor and Chair of Statistics at one of the most prestigious universities in the world, Harvard University (2004–Present), USA. In July 2007 he was appointed as Whipple V.N. Jones Professor of Statistics at his department. In 2001 he was awarded for "the outstanding statistician under the age of forty" by the Committee of Presidents of Statistical Societies. In 2002 he was ranked (by Science Watch) among the world top 25 most cited mathematicians for articles published and cited during 1991–2000. Professor Meng was Editor of *Bayesian Analysis* (2003–2005), and Co-Chair Editor, *Statistica Sinica* (2005–2008). He was an Associate editor for following journals: *Bernoulli* (2004–2005), *Biometrika* (2002–2005), The *Annals of Statistics* (1997–2003), *Journal of the American Statistical Association* (1996–2002) and *Statistica Sinica* (1992–1997). Currently, he is Editor of *Statistics Series, IMS Monograph and Textbook Series*. He is an Elected Fellow of the Institute of Mathematical Statistics (1997) and American Statistical Association (2004). Professor Meng is a recipient of the University of Chicago Faculty Award for Excellence in Graduate Teaching (1997–1998). He has published over 100 papers in leading statistical journals, and is widely known for his contributions in statistical analysis with missing data, Bayesian modeling, statistical computation, in particular Markov chain Monte Carlo and EM-type algorithms. (written by ML)

Cross References

►Data Analysis
►Data Privacy and Confidentiality
►Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)
►Imputation
►Model Selection
►Multiple Imputation
►Nonresponse in Surveys

References and Further Reading

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Wiley, New York
Gelman AE, Meng X-L (1996) Model checking and model improvement. In: Gilks W, Richardson S, Spiegelhalter D (eds) Practical Markov chain Monte Carlo, Chapman & Hall, London, pp 189–201
Little R, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
McCullagh P (2002) What is a statistical model? (with discussion). Ann Stat 30:1225–1310

Meng X-L (1994) Multiple-imputation inference with uncongenial sources of input (with discussion). Stat Sci 9: 538–573

Rubin DB (1987) Multiple imputaiton for nonresponse in surveys. Wiley, New York

Xie X, Meng X-L (2010) Multi-party inferences: what happens when there are three uncongenial models involved? Techincal Report, Department of Statistics, Harvard University

Multiple Comparison

Toshihiko Morikawa[1], Takeharu Yamanaka[2]
[1]Former Professor
Kurume University, Kurume, Japan
[2]Section Head
National Kyushu Cancer Center, Fukuoka, Japan

Multiplicity Issues

Statistical evidence is obtained by rejecting the null hypothesis at a "small" prespecified significance level α, say 0.05 or 0.01, which is an acceptable level of probability of the type I error (the error of rejecting the "true" null hypothesis). If we have a family of multiple hypotheses in a confirmatory experiment and test them simultaneously at each level α, the *overall* or *familywise type I error rate* (*FWER*), i.e., the probability of rejecting at least one "true" null hypothesis in the family, may inflate and exceed α, even if there exist no treatment differences. We call such inflation of the *FWER* a *multiplicity issue*.

Usually there may be some correlation structure between test statistics, and the inflation of the *FWER* might not be so remarkable. However, if we have multiple hypotheses to be tested for confirmatory purpose, we should adjust for multiplicity so as to control the *FWER* within α. This is called *multiplicity adjustment*. Testing procedures for multiplicity adjustment are called *multiple comparison procedures* (*MCPs*) or more generally *multiple testing procedures* (*MTPs*).

Multiplicity issues may arise in (1) multiple treatments (multiple comparisons), (2) multiple response variables (multiple endpoints), (3) multiple time points (longitudinal analysis), (4) multiple subgroups (subgroup analysis), and (5) multiple looks (interim analysis with group sequential methods or adaptive designs).

Hereafter we mainly concentrate on the multiple treatment comparisons, i.e., multiple comparisons in a traditional sense.

Multiple Comparisons

In a two group comparison of treatments A and B on their response means μ_A and μ_B, we have just one null hypothesis $H_0 : \mu_A = \mu_B$ to be tested and there is no need to adjust for multiplicity. However, when we compare three treatment groups, e.g., there are three treatments A, B and C, we may typically want to compare their means pairwisely, i.e., μ_A vs μ_B, μ_A vs μ_C and μ_B vs μ_C. Then there are three test hypotheses to be adjusted for multiplicity; namely, we need multiple comparison procedures.

All Pairwise Comparisons

The method to exactly control the *FWER* by adjusting the critical value in the above "all" pairwise comparisons is called *Tukey's method* (or *Tukey's multiple comparison test*). The method was developed for equal sample sizes, but even if the sample sizes are different between groups, the same critical value could be used conservatively, and such a method is known as the *Tukey-Kramer method*. The nonparametric version of Tukey's method is called the *Steel-Dwass test*.

Comparisons with a Control

The above three treatment example may have a structure that A and B are two (high and low) doses of a drug and C is a placebo (zero-dose). Then main interest in a formal analysis may be focused on the comparisons between each active dose and the placebo, i.e., μ_A vs μ_C and μ_B vs μ_C. This type of multiple comparison on treatment means can be performed by *Dunnett's method* (or *Dunnett's multiple comparison test*), and the common reference C is called a *control* or *control group*. The nonparametric version of Dunnett's method is called *Steel's test*.

If we assume the monotonicity of response means, such as $\mu_A \geq \mu_B \geq \mu_C$ or $\mu_A \leq \mu_B \leq \mu_C$, then in the comparison with a control, we can apply the *Williams test*, which is more powerful than Dunnett's test when the monotone dose-response relationship holds. The nonparametric version of the Williams test is known as the *Shirley-Williams test*.

Any Contrast Comparisons

More generally in a $k(\geq 3)$ treatment comparison, various hypotheses on any contrasts, such as, $\sum_{i=1}^{k} c_i \mu_i = 0$ where $\sum_{i=1}^{k} c_i = 0$, can be tested using *Scheffe's method* to control the *FWER*. For all pairwise comparisons or comparisons with a control, Scheffe's method is not recommended because it is "too" conservative in such cases. A nonparametric version of the Scheffe type multiple comparison method can be easily constructed.

Fixed Number of Comparisons

When the number of comparisons is fixed, the *Bonferroni method* (or *Dunn's method*) is simpler and easier to apply. The method only adjusts the significance level to α/m for each single test, where m is the number of interested comparisons. It is known that the method controls the *FWER* because the well-known *Bonferroni inequality*, $Pr\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} Pr(E_i)$ holds, where E_i is an event to reject hypothesis H_i. In the above three treatment example, the Bonferroni method could be applied with $m = 3$ for Tukey-type, and with $m = 2$ for Dunnett-type multiple comparisons, although it might be rather conservative.

Stepwise Procedures

All the methods described above (except the Williams test) are called "*simultaneous tests*" or "*single step tests*", because none of tests considered are affected by the results of others, and statistical testing for each hypothesis can be done simultaneously or in a single step manner. They control the *FWER* and can be used to easily construct the corresponding simultaneous confidence intervals, but there is some tradeoff in that they have a low statistical power in compensation for controlling the *FWER*.

Recently, more powerful test procedures than single step or simultaneous test procedures have been developed and become popular. Most of them are based on the *closed testing procedure* (*CTP*) proposed by Marcus, Peritz and Gabriel (1976) and they have a stepwise property in their nature. *CTPs* give a very general scheme of stepwise *MCPs* (or *MTPs*).

Closed Testing Procedures (CTPs)

Suppose that we have a family of m null hypotheses $F = \{H_1, H_2, \ldots, H_m\}$ to be tested and let $N = \{1, 2, \ldots, m\}$ be an *index set* that indicates the set of hypotheses considered. Then there are $2^m - 1$ possible intersections of null hypotheses H_i. We denote a set or family of such *intersection hypotheses* by $G = \{H_I = \bigcap_{i \in I} H_i : I \subseteq N,\ I \neq \emptyset\}$, where $\emptyset$ is an empty set and each intersection hypothesis H_I means that all hypotheses $H_i, i \in I$ hold simultaneously and thus H_I represents one possibility of the "true" null hypothesis. Because we do not know which H_I is true, a given *MCP* (or *MTP*) should control the *FWER* under any H_I. This is called a *strong control of the FWER*. If we control the *FWER* only under the *complete* or *global null hypothesis*, $H_N = \bigcap_{i \in N} H_i$, it is called a *weak control of the FWER*.

CTPs are testing procedures in which each *elementary hypothesis* $H_i, i = 1, \ldots, m$, is rejected only if all the intersection hypotheses including H_i, i.e., all $H_I = \bigcap_{j \in I} H_j, i \in I$, are rejected by the *size* α *test*. It is easily shown that any *CTP* controls the *FWER* in a strong sense. The procedure is equivalent to a test that starts with the test of *complete null hypothesis* H_N at level α and then proceeds in a stepwise manner that any *intersection hypothesis* H_I, $I \subset N$, is tested at level α only if all the intersection hypotheses $H_J = \bigcap_{i \in J} H_i$ which *imply* H_I, i.e., $J \supset I$, are rejected.

Some well known stepwise methods for the Tukey type multiple comparisons, e.g., *Fisher's protected LSD* (least significant difference) *test*, the *Newman-Keuls test*, and *Duncan's multiple range test*, control the *FWER* only in a weak sense, and should not be used. Instead, we can use the *Tukey-Welsch method* and *Peritz's method*. Also the *step-down Dunnett method* can be applied for the Dunnett type comparisons. They are *CTPs* and control the *FWER* in a strong sense. Note that the Williams test is also a *CTP*.

Modified Bonferroni Procedures (MBPs)

Modified Bonferroni procedures (*MBPs*) are extensions of the classical Bonferroni procedure, which use the Bonferroni's or similar criterion to test the intersection hypotheses H_I in *CTPs*. They use only *individual p-values* for multiplicity adjustment and are easy to apply. *Holm*, *Hochberg*, *Hommel* and *Rom procedures* are some of typical *MBPs*.

Gatekeeping Procedures (GKPs)

Most recently the new methods called the *gatekeeping procedures* (*GKPs*) have been rapidly developed. *GKPs* utilize the order and logical relationship between hypotheses or families of hypotheses and construct a *MTP* satisfying these relationships. They are usually based on *CTPs* and control the *FWER* in a strong sense. They include *serial GKP*, *parallel GKP*, *tree GKP*, and *truncated GKP*, etc. *GKPs* are especially useful for multiple endpoints and various combination structures of multiple comparisons, multiple endpoints and other multiplicities.

About the Authors

Dr. Toshihiko Morikawa is former professor of Kurume University, Japan. He is well-known as an author of the paper on a combined test of non-inferiority and superiority (Morikawa and Yoshida, *J. Biopharm. Statist.* 5, 297–306, 1995). He contributed to ICH as an expert working group (EWG) member of ICH E10 guideline. He is an elected member of ISI.

Dr. Takeharu Yamanaka is Chief Researcher in the Cancer Biostatistics Laboratory, National Kyushu Cancer Center, Japan. He has worked primarily on the design and analysis of clinical trials in areas including cancer. He has also served on the Data Safety Monitoring Boards for several international multi-center clinical trials.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►False Discovery Rate
►Multiple Comparisons Testing from a Bayesian Perspective
►Simes' Test in Multiple Testing

References and Further Reading

Dmitrienko A et al (2005) Analysis of clinical Trials Using SAS: A Practical Guide. SAS Press, Cary, NC
Dmitrienko A et al (2010) Multiple Testing Problems in Pharmaceutical Statistics Chapman & Hall/CRC, Boca Raton, FL
Hochberg Y, Tamhane AC (1987) Multiple Comparison Procedures John Wiley and Sons, New York
Hsu JC (1996) Multiple comparisons: Theory and Methods. Chapman & Hall, London
Miller RG (1981) Simultaneous Statistical Inference, 2nd edn. Springer-Verlag, New York
Morikawa T, Terao A, Iwasaki M (1996) Power evaluation of various modified Bonferroni procedures by a Monte Carlo study. J Biopharm Stat 6:343–359

Multiple Comparisons Testing from a Bayesian Perspective

Andrew A. Neath[1], Joseph E. Cavanaugh[2]
[1]Professor
Southern Illinois University Edwardsville, Edwardsville, IL, USA
[2]Professor
The University of Iowa, Iowa City, IA, USA

A General Multiple Comparisons Problem

In this note, we examine a general multiple comparisons testing problem from a Bayesian viewpoint. Suppose we observe independent random samples from I normally distributed populations with equal variances. The goal of our problem is to determine which pairs of groups have equal means.

Write

$$\{X_{ij}\} \,|\, \{\mu_i\}, \sigma^2 \sim \text{indep } N\left(\mu_i, \sigma^2\right). \tag{1}$$

We are interested in testing $H^{(a,b)} : \mu_a = \mu_b$ for each (a, b); a total of $I(I-1)/2$ distinct, but related hypotheses. A typical frequentist test is based on the decision rule of accept $H^{(a,b)}$ when

$$\left|\overline{X}_b - \overline{X}_a\right| \le Q_{a,b}. \tag{2}$$

The overall error rate is the probability of falsely rejecting any of the true hypotheses in the set $\left\{H^{(a,b)}\right\}$. The determination of $Q_{a,b}$ in (2) depends on how the overall error rate is to be controlled. A classical book featuring this multiple comparisons problem in detail is Scheffé (1959). For an applied review, see, for example, Kutner et al. (2004) or Montgomery (2008). A modern theoretical treatment is offered by Christensen (2002).

An overview to multiple comparisons under the Bayesian framework is given by Berry and Hochberg (1999). Westfall et al. (1997) consider the preceding problem of controlling the overall error rate from a Bayesian perspective. Here, our main focus is to show how a Bayesian approach can offer a logically pleasing interpretation of multiple comparisons testing.

A major point of difficulty to multiple comparisons procedures based on an accept / reject $H^{(a,b)}$ philosophy is illustrated by a case where one decides to accept $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, but reject $\mu_1 = \mu_3$. Such an outcome is possible under decision rule (2), but an interpretation is difficult to provide since the overall decision is not logically consistent. Employing a Bayesian philosophy, we may restate the goal of the problem as quantifying the evidence from the data in favor of each hypothesis $H^{(a,b)}$.

To implement this philosophy, we will require a measure of prior/posterior belief in $H^{(a,b)}$, represented by point mass probabilities. The construction of prior probabilities over the set of hypotheses $\left\{H^{(a,b)}\right\}$ must account for the fact that the collection does not consist of mutually exclusive events. For example, $H^{(1,2)}$ true $(\mu_1 = \mu_2)$ may occur with $H^{(2,3)}$ true $(\mu_2 = \mu_3)$ or with $H^{(2,3)}$ false $(\mu_2 \neq \mu_3)$. One cannot develop a prior by comparing relative beliefs in each of the pairwise hypotheses. Furthermore, certain combinations of hypotheses in the set $\left\{H^{(a,b)}\right\}$ represent impossibilities. For example, the event with $H^{(1,2)}$ true $(\mu_1 = \mu_2)$, $H^{(2,3)}$ true $(\mu_2 = \mu_3)$, $H^{(1,3)}$ false $(\mu_1 \neq \mu_3)$ should be assigned zero probability.

Allowable decisions can be reached through the formation of equal mean clusters among the I populations. For example, the clustering $\mu_1 = \mu_2$, $\mu_3 = \mu_4$ implies $H^{(1,2)}$ true, $H^{(3,4)}$ true, and all others false. Designating a clustering of equal means will define a model nested within (1). When two or more means are taken as equal, we merely combine all relevant samples into one. The smaller model is of the same form as (1), only for $I' < I$. The problem can now be stated in terms of Bayesian ►model selection, where each allowable combination of hypotheses will correspond to a candidate model.

We provide a short review of Bayesian model selection in the general setting using the notation of Neath

and Cavanaugh (1997). Let Y_n denote the observed data. Assume that Y_n is to be described using a model M_k selected from a set of candidate models $\{M_1, \ldots, M_L\}$. Assume that each M_k is uniquely parameterized by θ_k, an element of the parameter space $\Theta(k)$. In the multiple comparisons problem, the class of candidate models consists of all possible mean clusterings. Each candidate model is parameterized by the mean vector $\mu = (\mu_1, \ldots, \mu_I)$ and the common variance σ^2, with the individual means restricted by the model-defined clustering of equalities. That is, each model determines a corresponding parameter space where particular means are taken as equal.

Let $L(\theta_k|Y_n)$ denote the likelihood for Y_n based on M_k. Let $\pi(k)$, $k = 1, \ldots, L$, denote a discrete prior over the models $M_1, \ldots, M_L$. Let $g(\theta_k|k)$ denote a prior on θ_k given the model M_k. Applying Bayes' Theorem, the joint posterior of M_k and θ_k can be written as

$$f(k, \theta_k|Y_n) = \frac{\pi(k)g(\theta_k|k)L(\theta_k|Y_n)}{h(Y_n)},$$

where $h(Y_n)$ denotes the marginal distribution of Y_n.

The posterior probability on M_k is given by

$$\pi(k|Y_n) = h(Y_n)^{-1}\pi(k)\int_{\Theta(k)} g(\theta_k|k)L(\theta_k|Y_n)\,d\theta_k. \quad (3)$$

The integral in (3) requires numerical methods or approximation techniques for its computation. Kass and Raftery (1995) provide a discussion of the various alternatives. An attractive option is one based upon the popular Bayesian information criterion (Schwarz 1978). Define

$$B_k = -2\ln L(\hat{\theta}_k|Y_n) + dim(\theta_k)\ln(n),$$

where $\hat{\theta}_k$ denotes the maximum likelihood estimate obtained by maximizing $L(\theta_k|Y_n)$ over $\Theta(k)$. It can be shown under certain nonrestrictive regularity conditions (Cavanaugh and Neath 1999) that

$$\pi(k|Y_n) \approx \frac{\exp(-B_k/2)}{\sum_{l=1}^{L}\exp(-B_l/2)}. \quad (4)$$

The advantages to computing the posterior model probabilities as (4) include computational simplicity and a direct connection with a popular and well-studied criterion for Bayesian model selection. The justification of approximation (4) is asymptotic for the general case of prior $g(\theta_k|k)$, but Kass and Wasserman (1995) argue how the approximation holds under a noninformative prior on θ_k even for moderate and small sample sizes.

Regardless of which technique is used for computing $\pi(k|Y_n)$, we compute the probability on hypothesis $H^{(a,b)}$ by summing over the probabilities on those models for which $\mu_a = \mu_b$. This gives a nice approach to determining the evidence in favor of each of the pairwise equalities. The probability approach to presenting results for multiple comparisons testing provides more information than merely an accept / reject decision and is free of the potential contradictions alluded to earlier.

Example

We illustrate the Bayesian approach to multiple comparisons testing using data from Montgomery (2008). The $I = 5$ groups correspond to different cotton blends. Five fabric specimens are tested for each blend. The response measurements reflect tensile strength (in pounds per square inch). See Table 1 for the data and summary statistics. For ease of notation, treatments are identified in ascending order of the observed sample means.

A glance at the data suggests a potentially strong clustering of μ_1, μ_2 and a clustering to a lesser degree among μ_3, μ_4, μ_5. We shall see how these notions can be quantified by computing Bayesian posterior probabilities on the pairwise equalities. The top five most likely pairwise equalities are displayed in Table 2.

The hypothesis $\mu_1 = \mu_2$ is well-supported by the data ($P[H^{(1,2)}] \approx .8$), as was suspected. There is also some evidence in favor of $\mu_3 = \mu_4$ ($P[H^{(3,4)}] \approx .6$) and a non-negligible probability of $\mu_4 = \mu_5$ ($P[H^{(4,5)}] > .1$). Yet, there is good evidence against $\mu_3 = \mu_5$ ($P[H^{(3,5)}] < .02$).

Consider the clustering among μ_3, μ_4, μ_5. Tukey's multiple comparison procedure gives a critical range of $Q = 5.37$. A pair of means is deemed equal only if the corresponding sample difference is less than Q in magnitude. One reaches the decision of accept $\mu_3 = \mu_4$, accept $\mu_4 = \mu_5$, but reject $\mu_3 = \mu_5$. This decision is not logically consistent and is lacking any probabilistic detail. The proposed Bayesian approach bridges this probabilistic gap

Multiple Comparisons Testing from a Bayesian Perspective.

Table 1 Data for example

Group (cotton blend)	Response (tensile strength in lb/in^2)	Sample mean	Sample s.d.
1	7,7,9,11,15	9.8	3.35
2	7,10,11,11,15	10.8	2.86
3	12,12,17,18,18	15.4	3.13
4	14,18,18,19,19	17.6	2.07
5	19,19,22,23,25	21.6	2.61

Multiple Comparisons Testing from a Bayesian Perspective.

Table 2 Probabilities of pairwise equalities

Hypothesis	Posterior
$\mu_1 = \mu_2$	.7976
$\mu_3 = \mu_4$	.6015
$\mu_4 = \mu_5$	.1200
$\mu_2 = \mu_3$	.0242
$\mu_3 = \mu_5$	.0191

and provides a nice presentation for multiple comparisons. Bayesian inference has an advantage over traditional frequentist approaches to multiple comparisons in that degree of belief is quantified. One can avoid illogical conclusions which arise from an accept/reject decision process.

For computing details and continued analysis on this example, see Neath and Cavanaugh (2006).

About the Author

For the biographies *see* the entry ►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements.

Cross References

►Bayesian Statistics
►False Discovery Rate
►Multiple Comparison
►Simes' Test in Multiple Testing

References and Further Reading

Berry D, Hochberg Y (1999) Bayesian perspectives on multiple comparisons. J Stat Plan Infer 82:215–227
Cavanaugh J, Neath A (1999) Generalizing the derivation of the Schwarz information criterion. Commun Stat 28:49–66
Christensen R (2002) Plane answers to complex questions, 3rd edn. Springer, New York
Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc 90:773–795
Kass R, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934
Kutner M, Nachtsheim C, Neter J, Li W (2004) Applied linear statistical models, 5th edn. McGraw-Hill/Irwin, New York
Montgomery D (2008) Design and analysis of experiments, 7th edn. Wiley, New York
Neath A, Cavanaugh J (1997) Regression and time series model selection using variants of the Schwarz information criterion. Commun Stat 26:559–580
Neath A, Cavanaugh J (2006) A Bayesian approach to the multiple comparisons problem. J Data Sci 4:131–146
Scheffé H (1959) The analysis of variance. Wiley, New York
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Westfall P, Johnson W, Utts J (1997) A Bayesian perspective on the Bonferroni adjustment. Biometrika 84:419–427

Multiple Imputation

Christian Heumann
Ludwig-Maximilian University, Munich, Germany

Multiple Imputation and Combining Estimates

Missing data substantially complicates the statistical analysis of data. A common approach to circumvent the problem of analyzing a data set with missing data is to replace/impute the missing values by some estimates or auxiliary values. Subsequently, the data are then analyzed as if they would have been complete. While it is often straightforward to get a point estimate $\hat{\theta}$ for a quantity or parameter of interest, θ, an estimate for the variance of $\hat{\theta}$ is typically difficult to obtain, since the uncertainty due to the imputed values is not reflected correctly. This is exactly where multiple imputation (Rubin 1978, 1996) steps in: by creating several datasets by imputing several values for each missing position in the dataset, multiple imputation tries to reflect the uncertainty due to the imputed values. Note, that this uncertainty is additional to the usual uncertainty arising from the sampling process. Finally, the estimate $\hat{\theta}$ is computed for each of the completed datasets and these estimates are then combined into a single estimate for θ. In the following we give the algorithmic scheme for computing the combined point estimate and an estimated covariance matrix of it, that is, we directly address the case of a vector valued parameter θ. Strategies on how proper imputations can be created are discussed in the next paragraph.

Algorithm for inference under multiple imputation

1. Create m imputed datasets.
2. For each imputed dataset, $j = 1, \ldots, m$, compute the point estimate $Q^{(j)} = \hat{\theta}^{(j)}$ and its corresponding estimated (probably asymptotic) covariance matrix $U^{(j)} = \widehat{\mathrm{Cov}}(\hat{\theta}^{(j)})$. Usually, the "MI"-paradigm (Schafer 1999) assumes that $Q^{(j)}$ is asymptotically normal.
3. The multiple-imputation point estimate for θ is then

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} Q^{(j)} = \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}^{(j)}. \qquad (1)$$

4. The estimated covariance matrix of $\overline{Q}$ consists of two components, the within-imputation covariance and the between-imputation covariance. The within-imputation covariance $\overline{U}$ is given by

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U^{(j)} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\mathrm{Cov}}(\hat{\theta}^{(j)}). \qquad (2)$$

The between-imputation covariance B is given by

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(Q^{(j)} - \overline{Q}\right)\left(Q^{(j)} - \overline{Q}\right)^{T}, \qquad (3)$$

where T means the transposed vector, i.e. B is a quadratic matrix where the dimensions are equal to the length of the vector θ. Now we can combine the two estimates to the total variance T which is our estimated covariance matrix of $\overline{Q}$:

$$T = \widehat{\mathrm{Cov}}(\overline{Q}) = \overline{U} + (1 + m^{-1})B. \qquad (4)$$

5. A problem is that while the distribution of $T^{-\frac{1}{2}}(\theta - \overline{Q})$ can be approximated by a t-distribution with ν degrees of freedom,

$$\nu = (m-1)\left[1 + \frac{\overline{U}}{1 + m^{-1}B}\right]^2, \qquad (5)$$

in the *scalar* case, the same is not trivial for the vector valued case, see Schafer (1997).

Approaches to Create Multiple Imputations

So far we have discussed how MI works in principal and how the estimates for the completed datasets can be combined. Now we address how the imputations can be generated. We assume a missing data process that is ignorable. This relates essentially to a missing at random mechanism (MAR) plus the assumption that the parameters of the data model and the parameters of the missing data process are distinct (in likelihood inference this means that the combined parameter space is the product of the two parameter spaces, in a Bayesian analysis this means roughly that the prior distributions are independent). We note, that extensions to the case of nonignorable data situations are possible (although in general this is not easy), especially if one uses a Bayesian approach. The following subsections cannot reflect the whole research which has been done in the past. They only represent a small number of methods selected by the authors.

MI from Parametric Bayesian Models

Let D^{obs} be the observed data and D^{mis} the missing part of a dataset D, with $D = (D^{\mathrm{obs}}, D^{\mathrm{mis}})$. Then, m proper multiple imputations can be obtained via the predictive posteriori distribution of the missing data given the observed data

$$p(D^{\mathrm{mis}}|D^{\mathrm{obs}}) = \int p(D^{\mathrm{mis}}|D^{\mathrm{obs}};\theta)\, p(\theta|D^{\mathrm{obs}})\, d\theta \qquad (6)$$

or an approximation thereof. Note, that $p(\theta|D^{\mathrm{obs}})$ denotes the posteriori distribution of θ. Typically, two distinct approaches are considered to generate multiple imputations from (6): joint modeling and fully conditional modeling. The first approach assumes that the data follow a specific multivariate distribution, e.g. $D \sim N(\mu, \Sigma)$. Under a Bayesian framework draws from $p(D^{\mathrm{mis}}|D^{\mathrm{obs}})$ can be either generated directly (in some trivial cases) or simulated via suitable algorithms (in most cases) such as the IP-algorithm (see, e.g., Schafer [1997]). The second approach specifies an individual conditional distribution $p(D_j|D_{-j}, \theta_j)$ for each variable $D_j \in D$ and creates imputations as draws from these univariate distributions. It can be shown that the process of iteratively drawing and updating the imputed values from the conditional distributions can be viewed as a Gibbs sampler, that converges to draws from the (theoretical) joint distribution (if it exists). Further discussions and details on these issues can be found, e.g., in Drechsler and Rässler (2008) and the references therein.

An additional important remark refers to the fact that the imputations are called improper if we only draw imputations from

$$p(D^{\mathrm{mis}}|D^{\mathrm{obs}}, \tilde{\theta}),$$

where $\tilde{\theta}$ is a reasonable point estimate of θ (such as maximum likelihood, posterior mode or posterior mean), see also section "Other Pragmatic Approaches". That is why the above mentioned IP algorithm always includes the P-Step which samples also a new value of θ from $p(\theta|D^{\mathrm{obs}})$ before using this value to create a new imputed data set.

Nonparametric Methods

Another method to create proper multiple imputations is the so-called ABB (Approximate Bayesian Bootstrap). We refer the reader to Litte and Rubin (2002, Chap. 5.4).

Bootstrap EM

If the EM (Expectation-Maximization) algorithm is applied to an incomplete dataset, then a common problem is that only a point estimate (maximum likelihood estimate) is generated, but not an estimated (co-)variance matrix of this estimate. A typical approach to handle that issue corresponds to the use of the bootstrap (see ▶Bootstrap Methods) to create multiple imputations which then can be used to calculate such an estimate as shown in section "Multiple

Imputation and Combining Estimates". The following steps are repeated for $j = 1, \ldots, m$:

1 Draw a bootstrap sample $D^{(j)}$ from the data with replacement (including all data, complete and incomplete) with the same sample size as the original data. Obtain the maximum likelihood estimate $\hat{\theta}^{(j)}$ from the EM algorithm applied to $D^{(j)}$.
2 Use $\hat{\theta}^{(j)}$ to create an imputed dataset j from $p(D^{\text{mis}}|D^{\text{obs}}; \hat{\theta}^{(j)})$.

Other Pragmatic Approaches

Since Rubin introduced the MI paradigm in the late 1970s, there have been proposed several more or less ad-hoc methods to create multiple imputations that do not rely directly on random draws of the predictive posteriori distribution (6). A common approach refers to types of regression imputation (see, e.g., Little and Rubin [2002]), whereby missing values are replaced by predicted values from a regression of the missing item on the items observed based upon the subsample of the complete cases. This may be interpreted as an approximation to $p(D^{\text{mis}}|D^{\text{obs}}; \theta)$ from (6) with the simple constraint, that the uncertainty due to estimation of θ is not sufficiently reflected and hence $p(\theta|D^{\text{obs}})$ is apparently neglected. As an approach to consider this source of uncertainty anyhow and generate pragmatic multiple imputations (PMI), one might add an stochastic error to the imputation value and/or draw a random value from the conditional estimated distribution resulting from the prediction of the regression. Further extensions on regression imputation, e.g. the use of flexible nonparametric models and a recursive algorithm (GAMRI, Generalized Additive Model based Recursive Imputation), are discussed in Schomaker et al. (2010). Of course, the combination of values form different single imputation procedures might be seen as another type of PMI as well. Various strategies, such as nearest neighbor imputation (Chen and Shao 2000), Hot Deck imputations (Little and Rubin 2002) and others can be used for that approach.

Proper Versus Pragmatic Multiple Imputation

We recommend to create proper multiple imputations based on the predictive posteriori distribution of the missing data given the observed data. As mentioned in section "Software", a variety of statistical software packages nowadays provide fast and reliable tools to create proper multiple imputations even for users with less statistical expertise in missing-data-procedures. In situations where numerical algorithms fail to do so (sparse data, small datasets) pragmatic multiple imputations can be seen as a first approach to model imputation uncertainty.

Problems and Extensions

A number of problems arise along with multiple imputation procedures. Often they are not exclusively related to multiple imputation but to the general problem of misspecification in statistical models. If, e.g., the data model is misspecified because it assumes independent observations on the sampling units, but the observations are temporally or/and spatially correlated, also the results based on MI may become erroneous. An additional problem is ►model selection in general, especially if it is applied on high dimensional data. Also fully Bayesian inference, which often takes a lot of time for one specific model, is often too time consuming to be realistically applied to such problems. The same applies to model averaging (Frequentist or Bayesian) which may be thought of being an alternative to model selection.

Software

Recent years have seen the emergence on software that not only allows for valid inference with multiple imputation but also enables users with less statistical expertise to handle missing-data problems. We shortly introduce two packages that highlight the important progresses that lately have been made in easy-to-use Open-Source-Software. A broader description, discussion and comparison on MI-software can be found in Horton and Kleinman (2007).

- `Amelia II` (Honaker et al. 2008) is a package strongly related to the statistical Software *R* (R Development Core Team 2009) and performs proper multiple imputations by using an new, bootstrapping-based EM-algorithm that is both fast and reliable. All imputations are created via the `amelia()` function. For valid inference the quantities of the m imputed data sheets can be combined (i) in *R* using the `zelig()` command of `Zelig` (Imai et al. 2006), (ii) by hand using (1) and (4), respectively, or (iii) in separate software such as SAS, Stata etc. The Amelia II Software (named after the famous "missing" pilot Amelia Mary Earhart) is exceedingly attractive as it provides many useful options, such as the analysis of time-series data, the specification of priors on individual missing cell values, the handling of ordinal and nominal variables, the choice of suitable transformations and other useful tools. For further details see King et al. (2001) and Honaker and King (2010).

- `MICE` (Multiple Imputations by Chained Equations, van Buuren and Oudshoorn (2007)) is another package provided for *R* and *S-Plus*. It implements the chained equation approach proposed from van Buuren et al. (1999), where proper multiple imputations are generated via Fully Conditional Specification and Gibbs Sampling. The imputation step is carried out using the `mice()` function. As bugs of earlier versions seem to be removed, the MICE software can be attractive especially to the advanced user since he/she may specify his/her own imputation functions without much additional effort.

Cross References

►Imputation
►Incomplete Data in Clinical and Epidemiological Studies
►Multi-Party Inference and Uncongeniality
►Multivariate Statistical Distributions
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sampling From Finite Populations
►Statistical Software: An Overview

References and Further Reading

Chen JH, Shao J (2000) Nearest neighbor imputation for survey data. J Off Stat 16:113–131
R Development Core Team (2009) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org
Drechsler J, Rässler S (2008) Does convergence really matter? In: Shalabh, Heumann C (eds) Recent advances in linear models and related areas. Physica, pp 341–355
Honaker and King (2010) What to do about missing data in time series cross-secttion data. Am J Polit Sci 54(2):561–581
Honaker J, King G, Blackwell M (2008) Amelia II: a program for missing data. http://gking.harvard.edu/amelia
Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models. Am Stat 61:79–90
Imai K, King G, Lau O (2009) Zelig software website. http://gking.harvard.edu/zelig/
King G, Honaker J, Joseph A, Scheve K (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am Polit Sci Rev 95:49–69
Little R, Rubin D (2002) Statistical analysis with missing data. Wiley, New York
Rubin DB (1978) Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In: American Statistical Association Proceedings of the Section on Survey Research Methods, pp 20–40
Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489
Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall, London
Schafer J (1999) Multiple imputation: a primer. Stat Meth Med Res 8:3–15
Schomaker M, Wan ATK, Heumann C (2010) Frequentist model averaging with missing observations. Comput Stat Data Anal, in press
Van Buuren S, Oudshoorn CGM (2007) MICE: multivariate imputation by chained equations. R package version 1.16. http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm
van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of blood pressure covariates in survival analysis. Stat Med 18:681–694

Multiple Statistical Decision Theory

Deng-Yuan Huang
Professor
Fu Jen Catholic University, Taipei, Taiwan

In the theory and practice of statistical inference, multiple decision problems are encountered in many experimental situations. The classical methods for analyzing data customarily employ hypothesis testing in most situations. In such cases, when the hypothesis is rejected, one wants to know on which of a number of possible ways the actual situations fit our goal. If in the formulation of the problem, we consider only two decisions (reject or not reject the hypothesis), we will not only neglect to differentiate between certain alternative decisions but may also be using an inappropriate acceptance region for the hypothesis. Moreover, the traditional approach to hypothesis testing problems is not formulated in a way to answer the experimenter's question, namely, how to identify the hypothesis that satisfies the goal. Furthermore, when performing a test one may commit one of two errors: rejecting the hypothesis when it is true or accepting it when it is false. Unfortunately, when the number of observations is given, both probabilities cannot be controlled simultaneously by the classical approach (Lehmann 1959). Kiefer (1977) gave an example to show that for some sample values an appropriate test does not exhibit any detailed data-dependent measure of conclusiveness that conveys our strong feeling in favor of the alternative hypothesis. To enforce Kiefer's point, Schaafsma (1969) pointed out the Neyman–Pearson formulation is not always satisfactory and reasonable (Gupta and Huang 1981).

In the preceding paragraphs, we have discussed various difficulties associated with the hypothesis testing formulation. Thus, there arises the need for a modification of this theory and for alternative ways to attack such problems.

The approach in terms of Wald's decision theory (1950) provides an effective tool to overcome the above-mentioned difficulties in some reasonable ways. Actually, the problems of hypothesis testing can be formulated as general multiple decision problems. To this end, we first define that the space A of actions of the statistician consists of a finite number ($k \geq 2$) of elements, $A = \langle a_1, a_2, \ldots, a_k \rangle$. In practice, there are two distinct types of multiple decision problems. In one the parameter space Θ is partitioned into k subsets $\Theta_1, \Theta_2, \ldots, \Theta_k$, according to the increasing value of a real-valued function $r(\underline{\theta})$, $\underline{\theta} \in \Theta$. The action a_i is preferred if $\underline{\theta} \in \Theta_i$. This type of multiple decision problem is called monotone. This approach has been studied by Karlin and Rubin (1956) and Brown et al. (1976). For example, in comparing two treatments with means θ_1 and θ_2, an experimenter may have only a finite number of actions available, among these the experimenter might have preference based on the magnitudes of the differences of the means $\theta_2 - \theta_1$: A particular case occurs when one may choose from the three alternatives:

1. Prefer treatment 1 over treatment 2
2. Prefer treatment 2 over treatment 1
3. No preference (Ferguson 1967)

Another important class of multiple decision problems arises – selection problems where the treatments are classified into a superior category (the selected items) and an inferior one. In general, selection problems have been treated under several different formulations (Gupta and Panchapakesan 1979).

Recently, the modification of the classical hypothesis testing is considered the null hypothesis and several alternative hypotheses. Some multiple decision procedures are proposed to test the hypotheses. Under controlling the type I error, the type II error is the probability of incorrect decision. The type I and type II errors are given, the sample size can be determined. In general, one's interest is not just testing H_0 against the global alternative. Formulating the problem as one of choosing a subset of a set of alternatives has been studied (Lin and Huang 2007).

About the Author

Dr. Deng-Yuan Huang is Professor and Director, Institute of Applied Statistics, and Dean of the College of Management at Fu-Jen Catholic University in Taipei, Taiwan. He received his Ph.D. degree in Statistics from Purdue University in 1974. He is a renowned scholar in multiple decision theory, and has published numerous books and journal articles. Professor Huang has held positions of great honor in the research community of his country. He has also served as a member of the Committee on Statistics and the Committee on the Census of the Directorate General of Budget Accounting and Statistics of Taiwan. Before beginning his doctoral studies under Professor Shanti Gupta, he received the B.S. in mathematics from National Taiwan Normal University and the M.S. in Mathematics from National Taiwan University. Professor Huang is a member of the Institute of Mathematical Statistics, the Chinese Mathematical Association, and the Chinese Statistical Association. In 2002, he received the Distinguished Alumnus Award from Purdue University. In his honor, the International Conference on Multiple Decision Theory was held in Taiwan in 2007.

Cross References

►Decision Theory: An Introduction
►Decision Theory: An Overview

References and Further Reading

Brown LD, Cohen A, Strawderman WE (1976) A complete class theorem for strict monotone likelihood ratio with applications. Ann Stat 4:712–722
Ferguson TS (1967) Mathematical statistics: a decision theoretic approach. Academic, New York
Gupta SS, Huang DY (1981) Multiple decision theory: recent developments. Lecture notes in statistics, vol 6. Springer, New York
Gupta SS, Panchapakesan S (1979) Multiple decision procedures: theory and methodology of selecting and ranking populations. Wiley, New York, Republished by SIAM, Philadelphia, 2002
Karlin S, Rubin H (1956) The theory of decision procedures for distribution rules. Ann Math Stat 27:272–299
Kiefer J (1977) Conditional confidence statements and confidence estimators. JASA 72:789–827 (with comments)
Lehmann L (1959) Testing statistical hypotheses. Wiley, New York
Lin CC, Huang DY (2007) On some multiple decision procedures for normal variances Communication in statistics. Simulat Comput 36:265–275
Schaafsma W (1969) Minimal risk and unbiaseness for multiple decision procedures of type I. Ann Math Stat 40:1684–1720
Wald A (1950) Statistical decision function. Wiley, New York

Multistage Sampling

David Steel
Professor, Director of Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, NSW, Australia

Probability and Single Stage Sampling

In probability sampling each unit in the finite population of interest has a known, non-zero, chance of selection, π_i. In

single stage sampling the units in the sample, s, are selected directly from the population and information is obtained from them. For example, the finite population of interest may consist of businesses and a sample of businesses is selected. In these cases the population units and sampling units are the same. To obtain a single stage sample a sampling frame consisting of a list of the population units and means of contacting them are usually required. Simple random sampling (SRS) can be used, in which each possible sample of a given size has the same chance of selection. SRS leads to each unit in the population having the same chance of selection and is an equal probability selection method (EPSEM). Other EPSEMs are available. A probability sampling method does not need to be an EPSEM. As long as the selection probabilities are known it is possible to produce an estimator that is design unbiased, that is unbiased over repeated sampling. For example the ►Horvitz-Thompson estimator of the population total can be used, $\hat{T}_y = \sum_{i \in s} \pi_i^{-1} y_i$.

Stratification is often used, in which the population is divided into strata according to the values of auxiliary variables known for all population units. An independent sample is then selected from each stratum. The selection probabilities may be the same in each stratum, but often they are varied to give higher sampling rates in strata that are more heterogeneous and/or cheaper to enumerate. Common stratification variables are geography, size and type, for example industry of a business.

Cluster and Multistage Sampling

Instead of selecting a sample of population units directly it may be more convenient to select sampling units which are groups that contain several population units. The sampling unit and the population unit differ. The groups are called Primary Sampling Units (PSUs). If we select all population units from each selected PSU we have ►cluster sampling. If we select a sample of the units in the selected PSUs we have multistage sampling. Each population unit must be uniquely associated with only one PSU through coverage rules. These methods are often used when there is some geographical aspect to the sample selection and there are significant travel costs involved in collecting data and/or when there is no suitable population list of the population units available. A common example of a PSU is a household, which contains one or more people (Clark and Steel 2002). Another common example is area sampling (see Kish 1963, Chap. 9).

In a multistage sample the sample is selected in stages, the sample units at each stage being sampled from the larger units chosen at the previous stage. At each successive stage smaller sampling units are defined within those selected at the previous stage and further selections are made within each of them. At each stage a list of units from which the selections are to be made is required only within units selected at the previous stage.

For example, suppose we wish to select a sample of visitors staying overnight in the city of Wollongong. No list of such people exists, but if we confine ourselves to people staying in hotels or motels then it would be possible to construct a list of such establishments. We could then select a sample of hotels and motels from this list and select all guests from the selected establishments, in which case we have a cluster sample. It would probably be better to select a sample from the guests in each selected establishment allowing selection of more establishments, in which case we have a multi-stage sampling scheme. The probability of a particular guest being selected in the sample is the product of the probability of the establishment being selected and the probability the guest is selected given the establishment is selected. Provided the selection of establishments and guests within selected establishments is done using probability sampling, the sampling method is a valid probability sample. It would also be worthwhile stratifying according to the size of the establishment and its type.

Cluster and multistage sampling are used because a suitable sampling frame of population units does not exist but a list of PSUs does, or because they are less costly than a single stage sample of the same size in terms of population units. In multistage sampling the probability a population unit is selected is the probability the PSU containing the unit is selected multiplied by the conditional probability that the unit is selected given that the PSU it is in is selected.

Cluster and multistage sampling are often cheaper and more convenient than other methods but there is usually an increase in standard errors for the same sample size in terms of number of finally selected population units. It is important that the estimation of sampling error reflects the sample design used (See Lohr 1999, Chap. 9).

In many situations, the problems of compiling lists of population units and travel between selected population units are present even within selected PSUs. Consideration is then given to selecting the sample of population units within a selected PSU by grouping the population units into second stage units, a sample of which is selected. The population units are then selected from selected second stage units. This is called three-stage sampling. This process can be continued to any number of stages. The set of all selected population units in a selected PSU is called an ultimate cluster.

Multistage sampling is very flexible since many aspects of the design have to be chosen including the number of

stages and, for each stage, the unit of selection, the method of selection and number of units selected. Stratification and ratio or other estimation techniques may be used. This flexibility means that there is large scope for meeting the demands of a particular survey in an efficient way.

For a multistage sample the sampling variance of an estimator of a mean or total has a component arising from each stage of selection. The contribution of a stage of selection is determined by the number of units selected at that stage and the variation between the units at that stage, within the units at the next highest level. The precise formula depends on the selection and estimation methods used (See Lohr 1999, Chaps. 5–6; Cochran 1977, Chaps. 9, 9A, 10–11; Kish 1963, Chaps. 5–7, 9–10).

If PSUs vary appreciably in size then it can be useful to control the impact of this variation using ratio estimation or Probability Proportional to Size (PPS) sampling using the number of units in the PSU. For two-stage sampling a common design involves PPS selection of PSUs and selection of an equal number of units in each selected PSU. This gives each population unit the same chance of selection, which is usually a sensible feature for a sample of people, and an equal workload within each selected PSU, which has operational benefits. The first stage component of variance is determined by the variation of the PSU means. To use PPS sampling we need to know the population size of each PSU in the population. For ratio estimation we only need to know the total population size.

Optimal Design in Multistage Sampling

One of the main problems in designing multistage samples is to determine what size sample within selected PSUs to take to optimally balance cost and sampling error. In a two stage sampling scheme in which m PSUs are to be selected and the average number of units selected in each PSU is $\overline{n}$ the sampling variance is minimized for fixed sample size when $\overline{n} = 1$, since then the sample includes the largest number of PSUs. However, costs will be minimized when as few PSUs as possible are selected. Costs and variances are pulling in opposite directions and we must try to optimally balance them. In a two-stage sample several types of costs can be distinguished: overhead costs, costs associated with the selection of PSUs and costs associated with the selection of 2nd stage units. This leads to specifying a cost function of the form

$$C_0 + C_1 m + C_2 m\overline{n}.$$

For some of the common two-stage sampling and estimation methods used in practice the variance of the estimator of total or mean can be written as

$$V_0^2 + \frac{V_1^2}{m} + \frac{V_2^2}{m\overline{n}}.$$

For fixed cost the variance is minimized by choosing

$$\overline{n} = \sqrt{\frac{C_1}{C_2}\frac{V_2^2}{V_1^2}}.$$

The optimum choice of $\overline{n}$ thus depends on the ratios of costs and variances. As the first stage costs increase relative to the second stage costs the optimum $\overline{n}$ increase, so we are led to a more clustered sample. As the second stage component of variance increases relative to the first stage we are also led to a more clustered design.

The optimum value of $\overline{n}$ can be expressed in terms of the measure of homogeneity $\delta = \frac{V_1^2}{V_1^2 + V_2^2}$, as $\overline{n} = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}$. As δ increases the optimal choice of $\overline{n}$ decreases. For example if $C_1/C_2 = 10$ and $\delta = 0.05$ then the optimal $\overline{n} = 14$. To determine the optimal choice of $\overline{n}$ we only need to obtain an idea of the ratio of first stage to second stage cost coefficients and δ.

About the Author

Dr David Steel is a Professor in the School of Mathematics and Applied Statistics, University of Wollongong, Australia. He was the Head of the School of Mathematics and Applied Statistics (2000–2004) and Associate Dean (Research) for the Faculty of Informatics (2004–2006). He is foundation Director of the Center for Statistical and Survey Methodology (2007–). He has authored and co-authored more than 60 papers and books chapters. Professor Steel is currently an Associate Editor for the *Journal of the Royal Statistical Society (Series A)* and *Survey Methodology*. He is a foundation member of the Methodological Advisory Committee of the Australian Bureau of Statistics (1995–).

Cross References

►Cluster Sampling
►Sample Survey Methods
►Sampling From Finite Populations
►Stratified Sampling

References and Further Reading

Clark R, Steel DG (2002) The effect of using household as a sampling unit. Int Stat Rev 70:289–314
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Lohr S (1999) Sampling: design and analysis. Duxbury, Pacific Grove
Kish L (1965) Survey sampling. Wiley, New York

Multivariable Fractional Polynomial Models

Willi Sauerbrei[1], Patrick Royston[2]
[1]Professor
University Medical Center Freiburg, Freiburg, Germany
[2]Professor
University College London, London, UK

Fractional Polynomial Models

Suppose that we have an outcome variable, a single continuous covariate X, and a suitable regression model relating them. Our starting point is the straight line model, $\beta_1 X$ (for simplicity, we suppress the constant term, β_0). Often a straight line is an adequate description of the relationship, but other models must be investigated for possible improvements in fit. A simple extension of the straight line is a power transformation model, $\beta_1 X^p$. The latter model has often been used by practitioners in an *ad hoc* way, utilising different choices of p. Royston and Altman (1994) formalize the model slightly by calling it a first-degree fractional polynomial or FP1 function. The power p is chosen from a pragmatically chosen restricted set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where X^0 denotes $\log X$.

As with polynomial regression, extension from one-term FP1 functions to the more complex and flexible two-term FP2 functions follows immediately. Instead of $\beta_1 X^1 + \beta_2 X^2$, FP2 functions with powers (p_1, p_2) are defined as $\beta_1 X^{p_1} + \beta_2 X^{p_2}$ with p_1 and p_2 taken from S. If $p_1 = p_2$ Royston and Altman proposed $\beta_1 X^{p_1} + \beta_2 X^{p_1} \log X$, a so-called repeated-powers FP2 model.

For a more formal definition, we use the notation from Royston and Sauerbrei (2008). An FP1 function or model is defined as $\varphi_1(X, p) = \beta_0 + \beta_1 X^p$, the constant (β_0) being optional and context-specific. For example, β_0 is usually included in a normal-errors regression model but is always excluded from a Cox proportional-hazards model. An FP2 transformation of X with powers $\mathbf{p} = (p_1, p_2)$, or when $p_1 = p_2$ with repeated powers $\mathbf{p} = (p_1, p_1)$ is the vector $X^{\mathbf{p}}$ with

$$X^{\mathbf{p}} = X^{(p_1,p_2)} = \begin{cases} (X^{p_1}, X^{p_2}), & p_1 \neq p_2 \\ (X^{p_1}, X^{p_1} \log X), & p_1 = p_2 \end{cases}$$

An FP2 function (or model) with parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$ and powers $\mathbf{p}$ is $\varphi_2(X, \mathbf{p}) = \beta_0 + X^{\mathbf{p}}\boldsymbol{\beta}$. With the set S of powers as just given, there are 8 FP1 transformations, 28 FP2 transformations with distinct powers $(p_1 \neq p_2)$ and 8 FP2 transformations with equal powers $(p_1 = p_2)$. The best fit among the combinations of powers from S is defined as that with the highest likelihood.

The general definition of an FPm function with powers $\mathbf{p} = (p_1 \leq \cdots \leq p_m)$ is conveniently written as a recurrence relation. Let $h_0(X) = 1$ and $p_0 = 0$. Then

$$\varphi_m(X, \mathbf{p}) = \beta_0 + X^{\mathbf{p}}\beta = \beta_0 + \sum_{j=1}^{m} \beta_j h_j(X)$$

where for $j = 1, \ldots, m$

$$h_j(X) = \begin{cases} X^{p_j}, & p_{j-1} \neq p_j \\ h_{j-1}(X) \log X, & p_{j-1} = p_j \end{cases}$$

For example, for $m = 2$ and $\mathbf{p} = (-1, 2)$ we have $h_1(X) = X^{-1}$, $h_2(X) = X^2$. For $\mathbf{p} = (2, 2)$ we have $h_1(X) = X^2$, $h_2(X) = X^2 \log X$.

Figure 1 shows some FP2 curves, chosen to indicate the flexibility available with a few pairs of powers (p_1, p_2). The ability to fit a variety of curve shapes, some of which have asymptotes or which have both a sharply rising or falling portion and a nearly flat portion, to real data is a particularly useful practical feature of FP2 functions.

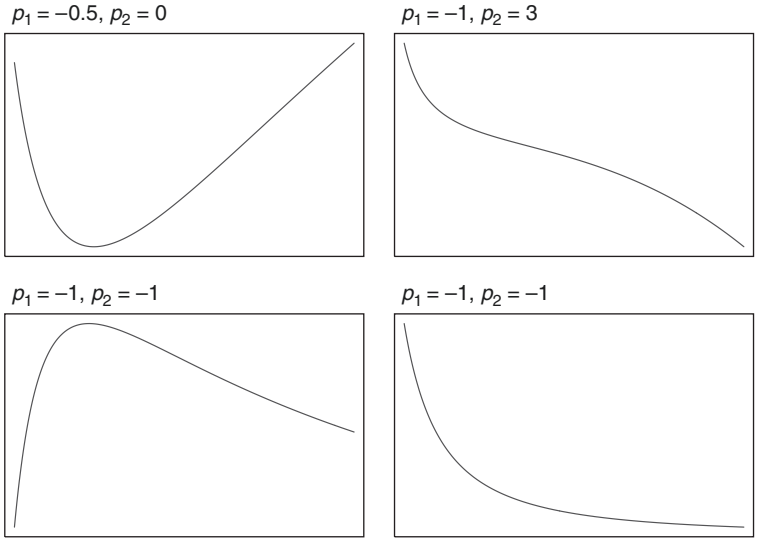
Function Selection Procedure (FSP)

Choosing the best FP1 or FP2 function by mininizing the deviance (minus twice the maximized log likelihood) is straightforward. However, having a sensible default function is important for increasing the parsimony, stability and general usefulness of selected functions. In most of the algorithms implementing FP modelling, the default function is linear – arguably, a natural choice. Therefore, unless the data support a more complex FP function, a straight line model is chosen. There are occasional exceptions; for example, in modelling time-varying regression coefficients in the Cox model, Sauerbrei et al. (2007a) chose a default time transformation of $\log t$ rather than t.

It is assumed in what follows that the null distribution of the difference in deviances between an FPm and an FP$(m-1)$ model is approximately central χ^2 on two degrees of freedom. Justification of this result is given in Sect. 4.9.1 of Royston and Sauerbrei (2008) and supported by simulation results (Ambler and Royston 2001).

For FP model selection, Royston and Sauerbrei (2008) proposed using the following closed test procedure (although other procedures are possible). It runs as follows:

1. Test the best FP2 model for X at the α significance level against the null model using four d.f. If the test is not significant, stop, concluding that the effect of X is "not significant" at the α level. Otherwise continue.



Multivariable Fractional Polynomial Models. Fig. 1 Examples of FP2 curves for different powers (p_1, p_2)

2. Test the best FP2 for X against a straight line at the α level using three d.f. If the test is not significant, stop, the final model being a straight line. Otherwise continue.
3. Test the best FP2 for X against the best FP1 at the α level using two d.f. If the test is not significant, the final model is FP1, otherwise the final model is FP2. End of procedure.

The test at step 1 is of overall association of the outcome with X. The test at step 2 examines the evidence for non-linearity. The test at step 3 chooses between a simpler or more complex non-linear model. Before applying the procedure, the analyst must decide on the nominal P-value (α) and on the degree (m) of the most complex FP model allowed. Typical choices are $\alpha = 0.05$ and FP2 ($m = 2$).

Multivariable Fractional Polynomial (MFP) Procedure

In many studies, a relatively large number of predictors is available and the aim is to derive an interpretable multivariable model which captures the important features of the data: the stronger predictors are included and plausible functional forms are found for continuous variables.

As a pragmatic strategy to building such models, a systematic search for possible non-linearity (provided by the FSP) is added to a backward elimination (BE) procedure. For arguments to combine FSP with BE, see Royston and Sauerbrei (2008). The extension is feasible with any type of regression model to which BE is applicable. Sauerbrei and Royston (1999) called it the multivariable fractional polynomial (MFP) procedure, or simply MFP. Using MFP successfully requires only general knowledge about building regression models.

The nominal significance level is the main tuning parameter required by MFP. Actually, two significance levels are needed: α_1 for selecting variables with BE, and α_2 for comparing the fit of functions within the FSP. Often, $\alpha_1 = \alpha_2$ is a good choice. A degree greater than 2 ($m > 2$) is rarely if ever needed in a multivariable context. Since the model is derived data-dependently, parameter estimates are likely to be somewhat biased.

As with any multivariable selection procedure checks of the underlying assumptions and of the influence of single observations are required and may result in model refinement. To improve robustness of FP models in the univariate and multivariable context Royston and Sauerbrei (2007) proposed a preliminary transformation of X. The transformation shifts the origin of X and smoothly pulls in extreme low and extreme high values towards the center of the distribution. The transformation is linear in the central bulk of the observations.

If available, subject-matter knowledge should replace data-dependent model choice. Only minor modifications are required to incorporate various types of subject-matter knowledge into MFP modelling. For the discussion of a detailed example, see Sauerbrei and Royston (1999).

For model-building by selection of variables and functional forms for continuous predictors, MFP has several advantages over spline-based models (the most important alternatives). For example, MFP models exhibit fewer

artefacts in fitted functions, and are more transportable, mathematically concise and generally more useful than spline models (Royston and Sauerbrei 2008; Sauerbrei et al. 2007b). Residual analysis with spline models may be used to check whether the globally defined functions derived by MFP analysis have missed any important local features in the functional form for a given continuous predictor (Binder and Sauerbrei 2010).

Recommendations for practitioners of MFP modelling are given in Royston and Sauerbrei (2008) and Sauerbrei et al. (2007b).

Extensions of MFP to Investigate for Interactions

MFP was developed to select main effects of predictors on the outcome. If a variable X_2 explains (at least partially) the relationship between a predictor X_1 and the outcome Y then confounding is present. Another important issue is interaction between two or more predictors in a multivariable model. An interaction between X_1 and X_2 is present if X_2 modifies the relationship between X_1 and the outcome. That means that the effect of X_1 is different in subgroups determined by X_2. Extensions of MFP have been proposed to handle two-way interactions involving at least one continuous covariate (Royston and Sauerbrei 2004). Higher order interactions, which typically play a role in factorial experiments, are a further extension, but not one that has yet been considered in the FP context.

To investigate for a possible interaction between a continuous predictor and two treatment arms in a randomized controlled trial, the multivariable fractional polynomial interaction (MFPI) procedure was introduced (Royston and Sauerbrei 2004). In a first step, the FP class is used to model the prognostic effect of the continuous variable separately in the two treatment arms, usually under some restrictions such as the same power terms in each arm. In a second step, a test for the equality of the prognostic functions is conducted. If significant, an interaction is present and the difference between two functions estimates the influence of the prognostic factor on the effect of treatment. The difference function is called a treatment effect function (and should be plotted). For interpretation, it is important to distinguish between the two cases of a predefined hypothesis and of searching for hypotheses (Royston and Sauerbrei 2004, 2008).

For more than two groups, extensions to investigate continuous by categorical interactions are immediate. Furthermore, MFPI allows investigation of treatment-covariate interactions in models with or without adjustment for other covariates. The adjustment for other covariates enables the use of the procedure in observational studies, where the multivariable context is more important than in an RCT.

Continuous-by-continuous interactions are important in observational studies. A popular approach is to assume linearity for both variables and test the multiplicative term for significance. However, the model may fit poorly if one or both of the main effects is non-linear. Royston and Sauerbrei (2008, Chap. 7) introduced an extension of MFPI, known as MFPIgen, in which products of selected main effect FP functions are considered as candidates for an interaction between a pair of continuous variables. Several continuous variables are usually available, and a test of interaction is conducted for each such pair. If more than one interaction is detected, interactions are added to the main-effects model in a step-up manner.

The MFPT(ime) algorithm (Sauerbrei et al. 2007a) combines selection of variables and of the functional form for continuous variables with determination of time-varying effects in a Cox proportional hazards model for ►survival data. A procedure analogous to the FSP was suggested for investigating whether the effect of a variable varies in time, i.e., whether a time-by-covariate interaction is present.

Further Contributions to Fractional Polynomial Modelling

Methods based on fractional polynomials have been reported recently, aiming to improve or extend the modelling of continuous covariates in various contexts. For example, Faes et al. (2007) applied model averaging to fractional polynomial functions to estimate a safe level of exposure; Lambert et al. (2005) considered time-dependent effects in regression models for relative survival; and Long and Ryoo (2010) used FPs to model non-linear trends in longitudinal data. For further topics and references, see Sect. 11.3 of Royston and Sauerbrei (2008).

About the Authors

Willi Sauerbrei, Ph.D., is a senior statistician and professor in medical biometry at the University Medical Center Freiburg. He has authored many research papers in biostatistics, and has published over 150 articles in leading statistical and clinical journals. He worked for more than 2 decades as an academic biostatistician and has extensive experience of cancer research. Together with Patrick Royston, he has written a book on modeling (*Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*, Wiley 2008).

Patrick Royston, D.Sc., is a senior statistician at the MRC Clinical Trials Unit, London, an honorary professor of statistics at University College London and a Fellow of the Royal Statistical Society. He has authored many research papers in biostatistics, including over 150 articles in leading statistical journals. He is co-author (with Willi Sauerbrei, see above) of a book on multivariable modeling. He is also an experienced statistical consultant, Stata programmer and software author.

Cross References

►Interaction
►Measurement Error Models
►Model Selection
►Nonparametric Regression Using Kernel and Spline Methods

References and Further Reading

Ambler G, Royston P (2001) Fractional polynomial model selection procedures: investigation of Type I error rate. J Stat Comput Simul 69:89–108

Binder H, Sauerbrei W (2010) Adding local components to global functions for continuous covariates in multivariable regression modeling. Stat Med 29:808–817

Faes C, Aerts M, Geys H, Molenberghs G (2007) Model averaging using fractional polynomials to estimate a safe level of exposure. Risk Anal 27:111–123

Lambert PC, Smith LK, Jones DR, Botha JL (2005) Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Stat Med 24:3871–3885

Long J, Ryoo J (2010) Using fractional polynomials to model nonlinear trends in longitudinal data. Br J Math Stat Psychol 63: 177–203

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Stat 43(3):429–467

Royston P, Sauerbrei W (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 23: 2509–2525

Royston P, Sauerbrei W (2007) Improving the robustness of fractional polynomial models by preliminary covariate transformation. Comput Stat Data Anal 51:4240–4253

Royston P, Sauerbrei W (2008) Multivariable model-building – a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, Chichester

Sauerbrei W, Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. J R Stat Soc A 162:71–94

Sauerbrei W, Royston P, Look M (2007a) A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biomet J 49:453–473

Sauerbrei W, Royston P, Binder H (2007b) Selection of important variables and determination of functional form for continuous predictors in multivariable model-building. Stat Med 26: 5512–5528

Multivariate Analysis of Variance (MANOVA)

Barbara G. Tabachnick, Linda S. Fidell
California State University, Northridge, CA, USA

ANOVA (►analysis of variance) tests whether mean differences among groups on a single DV (dependent variable) are likely to have occurred by chance. MANOVA (multivariate analysis of variance) tests whether mean differences among groups on a *combination* of DVs are likely to have occurred by chance. For example, suppose a researcher is interested in the effect of different types of treatment (the IV; say, desensitization, relaxation training, and a waiting-list control) on anxiety. In ANOVA, the researcher chooses one measure of anxiety from among many. With MANOVA, the researcher can assess several types of anxiety (say, test anxiety, anxiety in reaction to minor life stresses, and so-called free-floating anxiety). After random assignment of participants to one of the three treatments and a subsequent period of treatment, participants are measured for test anxiety, stress anxiety, and free-floating anxiety. Scores on all three measures for each participant serve as DVs. MANOVA is used to ask whether a combination of the three anxiety measures varies as a function of treatment. (MANOVA is statistically identical to discriminant analysis. The difference between the techniques is one of emphasis. MANOVA emphasizes the mean differences and statistical significance of differences among groups. Discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) emphasizes prediction of group membership and the dimensions on which groups differ.)

MANOVA developed in the tradition of ANOVA. Traditionally, MANOVA is applied to experimental situations where all, or at least some, IVs are manipulated and participants are randomly assigned to groups, usually with equal cell sizes. The goal of research using MANOVA is to discover whether outcomes, as reflected by the DVs, are changed by manipulation (or other action) of the IVs.

In MANOVA, a new DV is created from the set of DVs that maximizes group differences. The new DV is a linear combination of measured DVs, combined so as to separate the groups as much as possible. ANOVA is then performed on the newly created DV. As in ANOVA, hypotheses about means are tested by comparing variances between means relative to variances in scores within groups-hence multivariate analysis of variance.

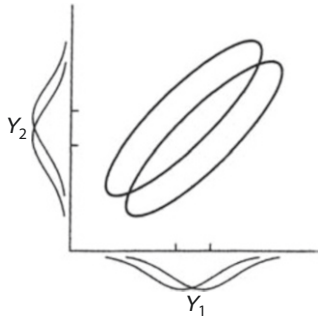
In factorial or more complicated MANOVA, a different linear combination of DVs is formed for each IV and

interaction. If gender of participant is added to type of treatment as a second IV, one combination of the three DVs maximizes the separation of the three treatment groups, a second combination maximizes separation of women and men, and a third combination maximizes separation of the six cells of the interaction. Further, if an IV has more than two levels, the DVs can be recombined in yet other ways to maximize the separation of groups formed by comparisons.

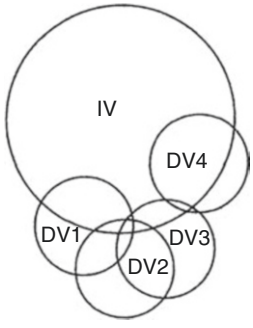
MANOVA has a number of advantages over ANOVA. First, by measuring several DVs instead of only one, the researcher improves the chance of discovering what it is that changes as a result of different IVs and their interactions. For instance, desensitization may have an advantage over relaxation training or waiting-list control, but only on test anxiety; the effect is missed in ANOVA if test anxiety is not chosen as the DV. A second advantage of MANOVA over a series of ANOVAs (one for each DV) is protection against inflated Type I error due to multiple tests of (likely) correlated DVs. (The linear combinations themselves are usually of interest in discriminant analysis, but not in MANOVA.)

Another advantage of MANOVA is that, under certain, probably rare conditions, it may reveal differences not shown in separate ANOVAs (Maxwell 2001). Such a situation is shown in Fig. 1 for a one-way design with two levels. In this figure, the axes represent frequency distributions for each of two DVs, Y_1 and Y_2. Notice that from the point of view of either axis, the distributions are sufficiently overlapping that a mean difference might not be found in ANOVA. The ellipses in the quadrant, however, represent the distributions of Y_1 and Y_2 for each group separately. When responses to two DVs are considered in combination, group differences become apparent. Thus, MANOVA, which considers DVs in combination, may occasionally be more powerful than separate ANOVAs.

The goal in MANOVA is to choose a small number of DVs where each DV is related to the IV, but the DVs are not related to each other. Good luck. In the usual situation there are correlations among the DVs, resulting in some ambiguity in interpretation of the effects of IVs on any single DV and loss of power relative to ANOVA. Figure 2 shows a set of hypothetical relationships between a single IV and four DVs. DV1is highly related to the IV and shares some variance with DV2 and DV3. DV2 is related to both DV1 and DV3 and shares very little unique variance with the IV. DV3 is somewhat related to the IV, but also to all of the other DVs. DV4 is highly related to the IV and shares only a little bit of variance with DV3. Thus, DV2 is completely redundant with the other DVs, and DV3 adds only a bit of unique variance to the set. (However, DV2 might be useful as a covariate if that use is conceptually viable because it reduces the total variances in DVs 1 and 3 that are not overlapping with the IV.)



Multivariate Analysis of Variance (MANOVA). Fig. 1 Advantage of MANOVA, which combines DVs, over ANOVA. Each axis represents a DV; frequency distributions projected to axes show considerable overlap, while ellipses, showing DVs in combination, do not



Multivariate Analysis of Variance (MANOVA). Fig. 2 Hypothetical relationships among a single IV and four DVs

Although computing procedures and programs for MANOVA and MANCOVA are not as well developed as for ANOVA and ANCOVA, there is in theory no limit to the generalization of the model. The usual questions regarding main effects of IVs, interactions among IVs, importance of DVs, parameter estimates (marginal and cell means), specific comparisons and trend analysis (for IVs with more than two levels), effect sizes of treatments, and effects of covariates, if any, are equally interesting with MANOVA as with ANOVA. There is no reason why all types of designs - one-way, factorial, repeated measures, nonorthogonal, and so on - cannot be extended to research with several DVs.

For example, multivariate analysis of covariance (MANCOVA) is the multivariate extension of ANCOVA. MANCOVA asks if there are statistically significant mean differences among groups after adjusting the newly created DV for differences on one or more covariates. To extend the example, suppose that before treatment participants are

pretested on test anxiety, minor stress anxiety, and free-floating anxiety; these pretest scores are used as covariates in the final analysis. MANCOVA asks if mean anxiety on the composite score differs in the three treatment groups, after adjusting for preexisting differences in the three types of anxieties.

MANOVA is also a legitimate alternative to repeated-measures ANOVA in which differences between pairs of responses to the levels of the within-subjects IV are simply viewed as separate DVs.

Univariate analyses are also useful following a MANOVA or MANCOVA. For example, if DVs can be prioritized, ANCOVA is used after MANOVA (or MANCOVA) in Roy-Bargmann stepdown analysis where the goal is to assess the contributions of the various DVs to a significant effect (Bock 1971; Bock and Haggard 1968). One asks whether, after adjusting for differences on higher-priority DVs serving as covariates, there is any significant mean difference among groups on a lower-priority DV. That is, does a lower-priority DV provide additional separation of groups beyond that of the DVs already used? In this sense, ANCOVA is used as a tool in interpreting MANOVA results. Results of stepdown analysis are reported in addition to individual ANOVAs.

However, MANOVA is a substantially more complicated analysis than ANOVA because there are several important issues to consider. MANOVA has all of the complications of ANOVA (e.g., homogeneity of variance; equality of sample sizes within groups; absence of ►outliers; power, cf. Woodward et al. 1990; normality of sampling distributions, independence of errors) and several more besides (homogeneity of variance-covariance matrices; multivariate normality, cf. Mardia 1971 and Seo et al. 1995; linearity, absence of ►multicollinearity and singularity; and choice among statistical criteria, cf. Olson 1979). These are not impossible to understand or test prior to analysis, but they are vital to an honest analysis.

Comprehensive statistical software packages typically include programs for MANOVA. The major SPSS module is GLM, however the older MANOVA module remains available through syntax and includes Roy-Bargmann stepdown analysis as an option. NCSS and SYSTAT have specific MANOVA modules, whereas SAS provides analysis of MANOVA through its GLM module. Analysis is also available through BMDP4V, STATA, and Statistica.

For more information about MANOVA, see Chaps. 7 and 8 of Tabachnick and Fidell (2007).

About the Authors

Dr Barbara Tabachnick is a Professor Emerita at California State University, Northridge. She has authored and co-authored more than 100 papers and chapters, as well as two books, including *Using Multivariate Statistics* (5th edition, Allyn & Bacon, 2007) and *Experimental Designs Using ANOVA* (Duxbury 2007), both with Dr. Linda Fidell. She continues to consult on research grants.

Dr. Linda Fidell is a Professor Emerita at California State University, Northridge. She has authored and co-authored more than 60 papers and chapters, as well as two books, including *Using Multivariate Statistics* (5th edition, Allyn & Bacon, 2007) and *Experimental Designs Using ANOVA* (Duxbury 2007), both with Dr. Barbara Tabachnick. She continues to consult on research grants.

Cross References

►Analysis of Variance
►Discriminant Analysis: An Overview
►Discriminant Analysis: Issues and Problems
►General Linear Models
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Nonparametric Models for ANOVA and ANCOVA Designs
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Bock RD, Haggard EA (l968) The use of multivariate analysis of variance in behavioral research. McGraw-Hill, New York

Mardia KV (1971) The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika 58(1):105–121

Maxwell S (2001) When to use MANOVA and significant MANOVAs and insignificant ANOVAs or vice versa. J Consum Psychol 10(1–2):29–30

Olson CL (1976) On choosing a test statistic in multivariate analysis of variance. Psychol Bull 83(4):579–586

Seo T, Kanda T, Fujikoshi Y (1995) The effects of nonnormality on tests for dimensionality in canonical correlation and MANOVA models. J Multivariate Anal 52:325–337

Tabachnick BG, Fidell LS (2007) Using multivariate statistics. Allyn & Bacon, Boston

Woodward JA, Overall JE (1975) Multivariate analysis of variance by multiple regression methods. Psychol Bull 82(1):21–32

Multivariate Data Analysis: An Overview

Joseph F. Hair
Professor of Marketing
Kennesaw State University, Kennesaw, GA, USA

Most business problems involve many variables. Managers look at multiple performance measures and related metrics

when making decisions. Consumers evaluate many characteristics of products or services in deciding which to purchase. Multiple factors influence the stocks a broker recommends. Restaurant patrons consider many factors in deciding where to dine. As the world becomes more complex, more factors influence the decisions managers and customers make. Thus, increasingly business researchers, as well as managers and customers, must rely on more sophisticated approaches to analyzing and understanding data.

Analysis of data has previously involved mostly univariate and bivariate approaches. Univariate analysis involves statistically testing a single variable, while bivariate analysis involves two variables. When problems involve three or more variables they are inherently multidimensional and require the use of multivariate data analysis. For example, managers trying to better understand their employees might examine job satisfaction, job commitment, work type (part-time vs. full-time), shift worked (day or night), age and so on. Similarly, consumers comparing supermarkets might look at the freshness and variety of produce, store location, hours of operation, cleanliness, prices, courtesy and helpfulness of employees, and so forth. Managers and business researchers need multivariate statistical techniques to fully understand such complex problems.

Multivariate data analysis refers to all statistical methods that simultaneously analyze multiple measurements on each individual respondent or object under investigation. Thus, any simultaneous analysis of more than two variables can be considered multivariate analysis. Multivariate data analysis is therefore an extension of univariate (analysis of a single variable) and bivariate analysis (cross-classification, correlation, and simple regression used to examine two variables).

Figure 1 displays a useful classification of statistical techniques. Multivariate as well as univariate and bivariate techniques are included to help you better understand the similarities and differences. As you can see at the top, we divide the techniques into dependence and interdependence depending on the number of dependent variables. If there is one or more dependent variables a technique is referred to as a dependence method. That is, we have both dependent and independent variables in our analysis. In contrast, when we do not have a dependent variable we refer to the technique as an interdependence method. That is, all variables are analyzed together and our goal is to form groups or give meaning to a set of variables or respondents.

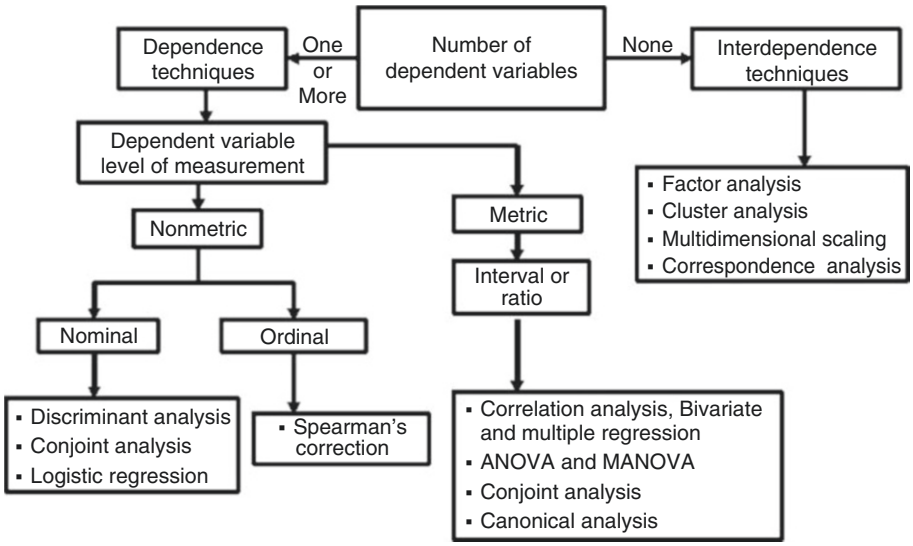
The classification can help us understand the differences in the various statistical techniques. If a research problem involves association or prediction using both dependent and independent variables, one of the dependence techniques on the left side of the diagram is appropriate. The choice of a particular statistical technique depends on whether the dependent variable is metric or nonmetric, and how many dependent variables are involved. With a nonmetric, ordinally measured dependent we would use the Spearman correlation. With a nonmetric, nominal dependent variable we use discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems), conjoint analysis or ►logistic regression. On the other hand, if our dependent variable is metric, we can use correlation, regression, ANOVA or MANOVA, canonical correlation, and conjoint analysis (the statistical technique of conjoint analysis can be formulated to handle both metric and nonmetric variables). The various statistical techniques are defined in Fig. 2. For more information on multivariate statistical techniques see Hair et al. (2010).

Concluding Observations

Today multivariate data analysis is being used by most medium and large sized businesses, and even some small businesses. Also, most business researchers rely on multivariate analysis to better understand their data. Thus, in today's business environment it's just as important to understand the relationship between variables, which often requires multivariate analysis, as it is to gather the information in the first place. The importance of multivariate statistical methods that help us to understand relationships has increased dramatically in recent years. What can we expect in the future as applications of multivariate data analysis expand: (1) data will continue to increase exponentially, (2) data quality will improve as will data cleaning techniques and data maintenance, (3) data analysis tools will be more powerful and easier to use, and (4) there will be many more career opportunities involving examining and interpreting data using multivariate data analysis.

About the Author

Professor Joe Hair is a member of the American Marketing Association, Academy of Marketing Science, and Society for Marketing Advances. He has authored 55 books, monographs, and cases, and over 80 articles in scholarly journals. He is a co-author (with William C. Black, Barry Babin and Rolph Anderson) of the well known applications-oriented introduction to multivariate analysis text *Multivariate Data Analysis* (Prentice Hall, 7th edition, 2010). He serves on the editorial review boards of several journals and was the 2009 Academy of Marketing Science/Harold Berkman Lifetime Service Award recipient, the KSU Coles College Foundation Distinguished Professor in 2009, the Marketing Management Association Innovative Marketer



Multivariate Data Analysis: An Overview. Fig. 1 Classification of statistical techniques

ANOVA – ANOVA stands for analysis of variance. It is used to examine statistical differences between the means of two or more groups. The dependent variable is metric and the independent variable(s) is nonmetric. One-way ANOVA has a single non-metric independent variable and two-way ANOVA can have two or more non-metric independent variables. ANOVA is bivariate while MANOVA is the multivariate extension of ANOVA.
Bivariate Regression – this is a type of regression that has a single metric dependent variable and a single metric independent variable.
Cluster Analysis – this type of analysis enables researchers to place objects (e.g., customers, brands, products) into groups so that objects within the groups are similar to each other. At the same time, objects in any particular group are different from objects in all other groups.
Correlation – correlation examines the association between two metric variables. The strength of the association is measured by the correlation coefficient. **Canonical correlation** analyzes the relationship between multiple dependent and multiple independent variables, most often using metric measured variables.
Conjoint Analysis – this technique enables researchers to determine the preferences individuals have for various products and services, and which product features are valued the most.
Discriminant Analysis – enables the researcher to predict group membership using two or more metric dependent variables. The group membership variable is a non-metric dependent variable.
Factor Analysis – this technique is used to summarize the information from a large number of variables into a much smaller number of variables or factors. This technique is used to combine variables whereas cluster analysis is used to identify groups with similar characteristics.
Logistic Regression – logistic regression is a special type of regression that involves a non-metric dependent variable and several metric independent variables.
Multiple Regression – this type of regression has a single metric dependent variable and several metric independent variables.
MANOVA – same technique as ANOVA but it can examine group differences across two or more metric dependent variables at the same time.
Perceptual Mapping – this approach uses information from other statistical techniques (e.g., multidimensional scaling) to map customer perceptions of products, brands, companies, and so forth.

Multivariate Data Analysis: An Overview. Fig. 2 Definitions of statistical techniques

of the Year in 2007, and the 2004 recipient of the Academy of Marketing Science Excellence in Teaching Award.

Cross References

►Canonical Correlation Analysis
►Cluster Analysis: An Introduction
►Correspondence Analysis
►Data Analysis
►Discriminant Analysis: An Overview
►Discriminant Analysis: Issues and Problems
►Factor Analysis and Latent Variable Modelling
►Linear Regression Models
►Logistic Regression
►Multidimensional Scaling
►Multidimensional Scaling: An Introduction
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Multivariate Statistical Process Control
►Principal Component Analysis
►Scales of Measurement
►Scales of Measurement and Choice of Statistical Methods
►Structural Equation Models

References and Further Reading

Esbensen KH (2006) Multivariate data analysis. IM Publications, Chichester
Hair J et al (2010) Multivariate data analysis, 7th edn. Prentice-Hall
Ho R (2006) Handbook of univariate and multivariate data analysis and interpretation with SPSS. Chapman & Hall, CRC, Boca Raton
Manly B (2005) Multivariate statistical methods a primer. Chapman & Hall, CRC, Boca Raton
Spicer J (2005) Making sense of multivariate data analysis: an intuitive approach. Sage Publications, Thousand Oaks

Multivariate Normal Distributions

Damir Kalpić[1] , Nikica Hlupić[2]
[1]Professor and Head, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[2]Assistant Professor, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia

The multivariate normal distribution is a generalization of the familiar univariate normal or Gaussian distribution (Hogg et al. 2005; Miller and Miller 1999) to $p \geq 2$ dimensions. Just as with its univariate counterpart, the importance of the multivariate normal distribution emanates from a number of its useful properties, and especially from the fact that, according to the central limit theorem (Anderson 2003; Johnson and Wichern 2007) under certain regularity conditions, sum of random variables generated from various (likely unknown) distributions tends to behave as if its underlying distribution were multivariate normal.

The need for generalization to the multivariate distribution naturally arises if we simultaneously investigate more than one quantity of interest. In that case, single observation (result of an experiment) is not value of a single variable, but the set of p values of $p \geq 2$ random variables. Therefore, we deal with $p \times 1$ random vector $\mathbf{X}$ and each single observation becomes $p \times 1$ vector $\mathbf{x}$ of single realizations of p random variables under examination. All these variables have their particular expected values that jointly constitute $p \times 1$ mean vector $\boldsymbol{\mu}$, which is expected value of random vector $\mathbf{X}$. Since analysis of collective behaviour of several quantities must take into account their mutual *correlations,* in multivariate analysis we also define $p \times p$ *variance-covariance matrix*

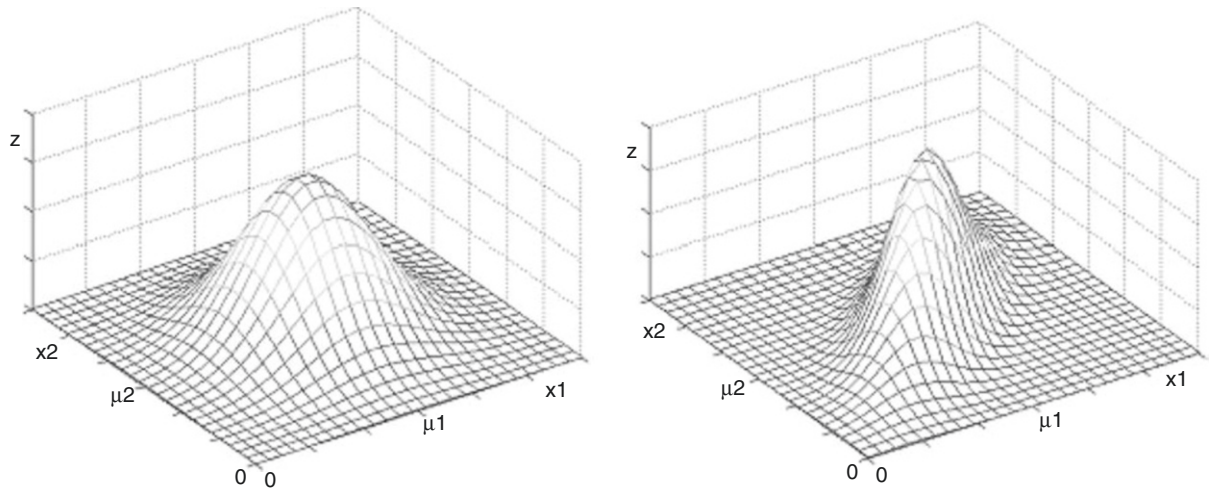
$$\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}) = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}, \quad (1)$$

where σ_{ij} are covariances between ith and jth component of $\mathbf{X}$ and σ_{ii} are variances of ith variable (more commonly denoted σ_i^2). This matrix is symmetric because $\sigma_{ij} = \sigma_{ji}$ and it is assumed to be *positive definite.*

Conceptually, the development of multivariate normal distribution starts from the univariate *probability* density function of a normal random variable X with the mean μ and variance σ^2. Common notation is $X \sim N(\mu, \sigma^2)$ and *probability density function* (pdf) of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2}; -\infty < x < +\infty. \quad (2)$$

Variable Z is so-called *standard normal variable* or *z-score* and it represents the square of the distance from a single observation (measurement) x to the population



Multivariate Normal Distributions. Fig. 1 Bivariate normal distribution with: *left* - $\sigma_1 = \sigma_2, \rho = 0$; *right* - $\sigma_1 = \sigma_2, \rho = 0,75$

mean μ, expressed in standard deviation units. It is this distance that directly generalizes to $p \geq 2$ dimensions, because in the univariate case we can write

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu), \tag{3}$$

and in the multivariate case, by analogy, we have the *Mahalanobis distance* (Johnson and Wichern 2007) expressed as

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{4}$$

The multivariate normal probability density function is obtained (Anderson 2003; Hogg et al. 2005; Johnson and Wichern 2007) by replacing (3) by (4) in the density function (2) and substituting the normalizing constant by $(2\pi)^{-p/2}|\Sigma|^{-1/2}$, so that the p-dimensional normal probability density for the random vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]^T$ is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \tag{5}$$

where $x_i \in (-\infty, \infty)$ and $i = 1, 2, \ldots, p$. Again analogously to the univariate case, we write $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

As an example, consider bivariate $(p = 2)$ distribution in terms of the individual parameters $\mu_1, \mu_2, \sigma_1^2 = \sigma_{11}, \sigma_2^2 = \sigma_{22}$ and $\sigma_{12} = \sigma_{21}$. If we also introduce *correlation coefficient* $\rho = \rho_{12} = \text{corr}(X_1, X_2) = \sigma_{12}^2/(\sigma_1 \cdot \sigma_2)$, density (5) becomes

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right]\right\}. \tag{6}$$

Formula (6) clearly indicates certain important general properties of multivariate normal distributions. First of all, if random variables X_1 and X_2 are uncorrelated, i.e., $\rho = 0$, it immediately follows that their joint density (6) can be factored as the product of two univariate normal densities of the form of (2). Since $f(x_1, x_2)$ factors as $f(x_1, x_2) = f(x_1) \cdot f(x_2)$, it follows that if X_1 and X_2 are uncorrelated, they are also *statistically independent*. This is a direct consequence of the general $(p \geq 2)$ multivariate normal property that uncorrelated variables are independent and have *marginal distributions* univariate normal. However, converse is not necessarily true for both of these statements and requires caution. Independent normal variables certainly are uncorrelated (this is true for any distribution anyway), but marginal distributions may be univariate normal without the joint distribution being multivariate normal. Similarly, marginally normal variables can be uncorrelated without being independent (Anderson 2003; Miller and Miller 1999).

Several other general properties of multivariate normal distribution are easier to conceive by studying the bivariate normal surface defined by (6) and illustrated in Fig. 1. Obviously, the bivariate (as well as multivariate) probability density function has a maximum at (μ_1, μ_2). Next, any intersection of this surface and a plane parallel to the z-axis has the shape of an univariate normal distribution, indicating that marginal distributions are univariate normal.

Finally, any intersection of this surface and a plane parallel to the x_1x_2 plane is an ellipse called *contour of constant probability density*. In the special case when variables are uncorrelated (independent) and $\sigma_1 = \sigma_2$ (Fig. 1 - left), contours of constant probability density are circles

and it is customary to refer to the corresponding joint density as a *circular normal density*. When variables are uncorrelated, but $\sigma_1 \neq \sigma_2$, contours are ellipses whose semi-axes are parallel to the x_1, x_2 axes of the coordinate system. In the presence of correlation, probability density concentrates along the line (Fig. 1 - right) determined by the coefficient of correlation and variances of variables, so the contours of constant probability density are ellipses rotated in a plane parallel to x_1x_2 plane (Anderson 2003; Miller and Miller 1999). All these properties are valid in p-dimensional spaces ($p > 2$) as well.

Here is the list of most important properties of the multivariate normal distribution (Anderson 2003; Johnson and Wichern 2007; Rao 2002).

1. Let $\mathbf{X}$ be a random vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{a}$ an arbitrary $p \times 1$ vector. Then the linear combination $\mathbf{a}^T\mathbf{X} = a_1X_1 + a_2X_2 + \ldots + a_pX_p$ is distributed as $N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$. In words, any linear combination of jointly normal random variables is normally distributed. Converse is also true: if $\mathbf{a}^T\mathbf{X}$ is $\sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$ for every $\mathbf{a}$, then $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. Generalization of property 1: Let $\mathbf{X}$ be a random vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let us form q linear combinations $\mathbf{AX}$, where $\mathbf{A}$ is an arbitrary $q \times p$ matrix. Then it is true that $\mathbf{AX} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$. Similarly, for any vector of constants $\mathbf{d}$ we have $\mathbf{X} + \mathbf{d} \sim N_p(\boldsymbol{\mu} + \mathbf{d}, \boldsymbol{\Sigma})$.
3. All subsets of variables constituting $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are (multivariate) normally distributed.
4. Multivariate normal $q_1 \times 1$ and $q_2 \times 1$ vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if they are uncorrelated, i.e., $\mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$ (a $q_1 \times q_2$ matrix of zeros).
5. If multivariate normal $q_1 \times 1$ and $q_2 \times 1$ vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and distributed as $N_{q1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $N_{q2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, then $(q_1 + q_2) \times 1$ vector $[\mathbf{X}_1^T\ \mathbf{X}_2^T]^T$ has multivariate normal distribution

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

6. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be mutually independent random vectors that are all multivariate normally distributed, each having its particular mean, but all having the same covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{X}_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Linear combination of these vectors $\mathbf{V}_1 = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \ldots + c_n\mathbf{X}_n$ is distributed as $N_p\left(\sum_{j=1}^{n} c_j\boldsymbol{\mu}_j, \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma}\right)$. Moreover, similarly to property 5, $\mathbf{V}_1$ and some other linear combination $\mathbf{V}_2 = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \ldots + b_n\mathbf{X}_n$ are jointly multivariate normally distributed with covariance matrix

$$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma} & (\mathbf{b}^T\mathbf{c})\boldsymbol{\Sigma} \\ (\mathbf{b}^T\mathbf{c})\boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\boldsymbol{\Sigma} \end{bmatrix}.$$

Thus, if $\mathbf{b}^T\mathbf{c} = 0$, i.e., vectors $\mathbf{b}$ and $\mathbf{c}$ are orthogonal, it follows that $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent and vice versa.

7. All conditional distributions are multivariate normal. Formally, let $\mathbf{X}_1$ and $\mathbf{X}_2$ be any two subsets of a multivariate normal vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, and $|\boldsymbol{\Sigma}_{22}| > 0$. The conditional distribution of $\mathbf{X}_1$, given a fixed $\mathbf{X}_2 = \mathbf{x}_2$, is multivariate normal with

$$\begin{aligned} \mathrm{mean}(\mathbf{X}_1|\mathbf{x}_2) &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \text{ and } \mathrm{cov}(\mathbf{X}_1|\mathbf{x}_2) \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

8. Generalized distance $(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ of observations $\mathbf{x}$ of a vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from the mean $\boldsymbol{\mu}$ has a chi squared distribution with p *degrees of freedom* denoted χ_p^2.
9. With $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a set of n observations from a (multivariate) normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we have the following results:
   (a) $\overline{\mathbf{X}}$ is distributed as $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$
   (b) $(n-1)\mathbf{S}$ has a *Wishart distribution;* with $n-1$ degrees of freedom
   (c) $\overline{\mathbf{X}}$ and $\mathbf{S}$ are independent.

Cross References

►Bivariate Distributions
►Central Limit Theorems
►Hotelling's T^2 Statistic
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Statistical Analysis
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation
►Normal Distribution, Univariate
►Statistical Distributions: An Overview
►Statistical Quality Control: Recent Advances

References and Further Reading

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, Hoboken

Ghurye SG, Olkin I (1962) A characterization of the multivariate normal distribution. Ann Math Stat 33:533–541

Green PE (1978) Analyzing multivariate data. Dryden Press, London
Hogg RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics, 6th edn. Pearson Prentice Hall, Upper Saddle River
Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice Hall, New York
Kagan A, Linnik YV, Rao CR (1972) Characterization problems of mathematical statistics. Wiley, New York
Miller I, Miller M (1999) John E. Freund's mathematical statistics, 6th edn. Pearson Prentice Hall, Upper Saddle River
Rao CR (2002) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Seal HL (1967) Studies in the history of probability and statistics. XV The historical development of the Gauss linear model, Biometrika, 54:1–24

Multivariate Outliers

Isabel M. Rodrigues[1], Graciela Boente[2]
[1]Assistant Professor
Technical University of Lisbon (TULisbon), Lisboa, Portugal
[2]Professor, Facultad de Ciencias Exactas and Naturales Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

In the statistical analysis of data one is often confronted with observations that "appear to be inconsistent with the remainder of that set of data" (Barnett and Lewis 1994). Although such observations (the ▶outliers) have been the subject of numerous investigations, there is no general accepted formal definition of outlyingness. Nevertheless, the outliers describe abnormal data behavior, i.e., data that are deviating from the natural data variability (see, e.g., Peña and Prieto 2001, Filzmoser 2004, and Filzmoser et al. 2008 for a discussion).

Sometimes outliers can grossly distort the statistical analysis, while at other times their influence may not be as noticeable. Statisticians have accordingly developed numerous algorithms for the detection and treatment of outliers, but most of these methods were developed for univariate data sets. They are based on the estimation of location and scale, or on quantiles of the data. Since in a univariate sample outliers may be identified as an exceptionally large or small value, a simple plot of the data, such as scatterplot, stem-and-leaf plot, and QQ-plot can often reveal which points are outliers.

In contrast, for multivariate data sets the problem of outliers identification gives challenges that do not occur with univariate data since there is no simple concept of ordering the data. Furthermore, the multivariate case introduces a different kind of outlier, a point that is not extreme component wise but departs from the prevailing pattern of correlation structure. This departs causes that the observations appear as univariate outliers in some direction not easily identifiable. In this context, to detect an observation as possible outlier not only the distance from the centroid of the data is important but also the data shape. Also, as Gnanadesikan and Kettenring (1972) pointed out the visual detection of multivariate outliers is virtually impossible because the outliers do not "stick out on the end."

Since most standard multivariate analysis techniques rely on the assumption of normality, in 1963, Wilks proposed identifying sets of outliers of size j from $\{1, 2, \ldots, n\}$, in normal multivariate data, by checking the minimum values of the ratios $|A_{(I)}|/|A|$, where $|A_{(I)}|$ is the internal scatter of a modified sample in which the set of observations I of size j has been deleted and $|A|$ is the internal scatter of the complete sample. For $j = 1$ this method is equivalent to the classical way to declare a multivariate observation as a possible outlier by using the squared Mahalanobis' distance defined as

$$MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V}) = ((\mathbf{x}_i - \mathbf{t})^T \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{t}))^{1/2}$$

where $\mathbf{t}$ is the estimated multivariate location and $\mathbf{V}$ the estimated scatter matrix. Usually $\mathbf{t}$ is the multivariate arithmetic mean, the centroid, and $\mathbf{V}$ the sample covariance matrix. Mahalanobis' distance identifies observations that lie far away from the center of the data cloud, giving less weight to variables with large variances or to groups of highly correlated variables. For a p-multivariate normally distributed data $MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V})$ converge to χ_p^2, a chi-square distribution with p degree of freedom. Points with large $MD_i^2 \equiv MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V})$, compared with some χ_p^2 quantile, are then considered outliers. Hence, to evaluate multivariate normality one may plot the ordered $MD_{(i)}^2$ against the expected order statistics of the ▶chi-square distribution with sample quantiles $\chi^2_{p[(i-1/2)/2]} = q_i$ where q_i $(i = 1, \ldots, n)$ is the $100(i - 1/2)/n$ sample quantile of χ_p^2. The plotted points $(MD_{(i)}, q_i)$ should be close to a line, so the points far from the line are potential outliers. Formal tests for multivariate outliers are considered by Barnett and Lewis (1994).

Clearly, the Mahalanobis distance relies on classical location and scatter estimators. The presence of outliers may distort arbitrarily the values of these estimators and render meaningless the results. This is particularly acute when there are several outliers forming a cluster, because

they will move the arithmetic mean toward them and inflate the classical tolerance ellipsoid in their direction. So this approach suffers from the *masking* and *swamping* effects by which multiple outliers do not have a large MD_i^2. A solution to this problem is well known in ►robust statistics: **t** and **V** have to be estimated in a robust manner, where the expression "robust"means resistance against the influence of outlying observations. Thus, the "robustified" ordered Mahalanobis distances, $RMD_{(i)}^2$ may be plotted to locate extreme outliers. This is the approach considered by Becker and Gather (2001), Filzmoser (2004), and Hardin and Rocke (2005) who studied outlier identification rules adapted to the sample size using different location and scatter robust estimators.

For a review on some of the robust estimators for location and scatter introduced in the literature see Maronna et al. (2006). The minimum covariance determinant (MCD) estimator – the procedure is due to Rousseeuw (1984) – is probably most frequently used in practice, partly because a computationally fast algorithm has been developped (Rousseeuw and Van Driessen 1999). The MCD estimator also benefits from the availability of software implementation in different languages, including R, S-Plus, Fortran, Matlab, and SAS. For these reasons the MCD estimator had gained much popularity, not only for outliers identification but also as an ingredient of many robust multivariate techniques.

Other currently popular multivariate outlier detection methods fall under projection pursuit techniques, originally proposed by Kruskal (1969). Projection pursuit searches for "interesting" linear projections of multivariate data sets, where a projection is deemed interesting if it minimizes or maximizes a projection index (typically a scale estimator). Therefore, the goal of projection pursuit methods is to find suitable projections of the data in which the outliers are readily apparent and may thus be downweighted to yield a estimator, which in turn can be used to identify the outliers. Since they do not assume the data to originate from a particular distribution but only search for useful projections, projection pursuit procedures are not affected by non-normality and can be widely applied in diverse data situations. The penalty for such freedom comes in the form of increased computational burden, since it is not clear which projections should be examined. An exact method would require to test over all possible directions.

The most well-known outlier identification method based upon the projection pursuit concept is the Stahel–Donoho (Stahel 1981; Donoho 1982) estimator. This was the first introduced high-breakdown and affine equivariant estimator of multivariate location and scatter that became better known after Maronna and Yohai (1995) published an analysis of it. It is based on a measure of the outlyingness of data points, which is obtained by projecting the observation on univariate directions. The Stahel–Donoho estimator then computes a weighted mean and covariance matrix, with weights inverse proportional to the outlyingness. This outlyingness measure is based upon the projection pursuit idea that if a point is a multivariate outlier, there must be some one-dimensional projection of the data in which this point is a univariate outlier. Using a particular observation as a reference point, the Stahel–Donoho algorithm determines which directions have optimal values for a pair of robust univariate location/scale estimators and then uses these estimators to assign weights to the other points. One way of reducing the computational cost of the Stahel–Donoho estimator is to reduce the number of projections that need to be examined.

In this direction, Peña and Prieto (2001) proposed a method, the Kurtosis1, which involves projecting the data onto a set of $2p$ directions. These directions are chosen to maximize and minimize the kurtosis coefficient of the data along them. A small number of outliers would cause heavy tails and lead to a larger kurtosis coefficient, while a larger number of outliers would start introducing bimodality and decrease the kurtosis coefficient. Viewing the data along projections that have maximum and minimum kurtosis values would therefore seem to display the outliers in a more recognizable representation.

For a much more detailed overview about outliers see Barnett and Lewis (1994) and also Rousseeuw et al. (2006) for a review on robust statistical methods and outlier detection.

Cross References

►Chi-Square Distribution
►Distance Measures
►Multivariate Normal Distributions
►Multivariate Technique: Robustness
►Outliers
►Robust Statistical Methods

References and Further Reading

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, Chichester

Becker C, Gather U (2001) The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. Comput Stat Data Anal 36:119–127

Donoho D (1982) Breakdown properties of multivariate location estimators. Ph.D. thesis, Harvard University

Filzmoser P (2004) A multivariate outlier detection method. In: Aivazian S, Filzmoser P, Kharin Yu (eds) Proceedings of the seventh international conference on computer data analysis and modeling, vol 1. Belarusian State University, Minsk, pp 18–22

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52:1694–1711
Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81–124
Hardin J, Rocke D (2005) The distribution of robust distances. J Comput Graph Stat 14:928–946
Kruskal JB (1969) Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In: Milton RC, Nelder JA (eds) Statistical computation. Academic, New York, pp 427–440
Maronna RA, Yohai VJ (1995) The behavior of the Stahel-Donoho robust multivariate estimator. J Am Stat Assoc 90:330–341
Maronna RA, Martin RD, Yohai V (2006) Robust statistics: theory and methods. Wiley, New York
Peña D, Prieto FJ (2001) Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43:286–310
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) Robustness and outlier detection in chemometrics. Cr Rev Anal Chem 36:221–242
Stahel WA (1981) Robust estimation: infinitesimal optimality and covariance matrix estimators. Ph.D. thesis in German, Swiss Federal Institute of Technology, Zurich, Switzerland
Wilks SS (1963) Multivariate statistical outliers. Sankhya 25:407–426

Multivariate Rank Procedures : Perspectives and Prospectives

Pranab K. Sen
Cary C. Boshamer Professor of Biostatistics and Professor of Statistics and Operations Research
University of North Carolina, Chapel Hill, NC, USA

Developments ►in multivariate statistical analysis have genesis in the parametrics surrounding the *multivariate normal* distribution (see ►Multivariate Normal Distributions) in the continuous case while the *product multinomial law* dominates in discrete multivariate analysis. Characterizations of multi-normal distributions have provided a wealth of rigid mathematical tools leading to a very systematic evolution of mathematical theory laying down the foundation of multivariate statistical methods. *Internal multivariate* analyses comprising of *principal component models, canonical correlation* and *factor analysis* are all based on appropriate *invariance structures* that exploit the underlying linearity of the interrelation of different characteristics, without depending much on underlying normality, and these tools are very useful in many areas of applied research, such as sociology, psychology, economics, and agricultural sciences. In the recent past, there has been a phenomenal growth of multivariate analysis in medical studies, clinical trials and ►bioinformatics, among others. The role of multinormality is being scrutinized increasingly in these contexts.

External multivariate analyses pertaining to ►*multivariate analysis of variance* (MANOVA) and covariance (MANOCOVA), *classification and discrimination,* among others, have their roots in the basic assumption of multinormal distribution, providing some optimal, or at least desirable, properties of statistical inference procedures. Such optimal statistical procedures generally exist only when the multinormality assumption holds. Yet, in real life applications, the postulation of multinormality may not be tenable in a majority of cases. Whereas in the univariate case, there are some other distributions, some belonging to the so-called *exponential family of densities* and some not, for which exact statistical inference can be drawn, often being confined to suitable subclass of statistical procedures. In the multivariate case, alternatives to multinormal distributions are relatively few and lack generality. As such, almost five decades ago, it was strongly felt that statistical procedures should be developed to bypass the stringent assumption of multinormality; this is the genesis of *multivariate nonparametrics*.

Whereas the classical normal theory likelihood based multivariate analysis exploited *affine invariance*, leading to some optimality properties, it has some shortcomings too. Affine invariance makes sense only when the different characteristics or variates are linearly combinable in a meaningful way. Further, such parametric procedures are quite vulnerable to even small departures from the assumed multinormality. Thus, they are generally *nonrobust* even in a local sense. Moreover, in many applications, different characteristics are recorded on different units and often on a relative scale (viz., ranking of n individuals on some multivariate traits) where linearly combinability may not be compatible. Rather, it is more important to have coordinatewise invariance under arbitrary strictly monotone transformations – a feature that favors ranks over actual measurements. Multivariate rank procedures have this basic advantage of invariance under coordinatewise arbitrary strictly monotone transformations, not necessarily linear. Of course, this way the emphasis on affine invariance is sacrificed, albeit, there are affine-invariant rank procedures too (see Oja 2010).

The basic difference between univariate and multivariate rank procedures is that for suitable *hypothesis of invariance*, in the univariate case, such procedures are genuinely distribution-free, whereas in the multivariate case,

even the hypothesis of invariance holds, these tests are usually *conditionally distribution-free*. This feature, known as the *rank-permutation principle*, was initially developed by Chatterjee and Sen (1964) and in a more general framework, compiled and reported in Puri and Sen (1971), the first text in multivariate nonparametrics. During the past four decades, a phenomenal growth of research literature in multivariate nonparametrics has taken place; specific entries in the *Encyclopedia of Statistical Science* and *Encyclopedia of Biostatistics* (both published from Wiley-Interscience, New York) provide detailed accounts of these developments.

In the recent past, *high-dimensional low sample size* (HDLSS) problems have cropped up in diverse fields of application. In this setup, the dimension is generally far larger than the number of sample observations, and hence, standard parametric procedures are untenable; nonparametrics fare much better. This is a new frontier of multivariate nonparametrics and there is a tremendous scope of prospective research with deep impact on fruitful applications. ►Data mining (or knowledge discovery and data mining) and statistical learning algorithms also rest on multivariate nonparametrics to a greater extent. Bioinformatics and environmetrics problems also involve such nonstandard multivariate nonparametric procedures. In a micro-array data model, an application of multivariate rank methods has been thoroughly explored in Sen (2008).

About the Author

Dr. Pranab Kumar Sen is a Cary C. Boshamer Professor of Biostatistics, University of North Carolina (1982–) and a lifelong Adjunct Professor, Indian Statistical Institute, Calcutta (1993–). He was born on November 7, 1937 in Calcutta, India. He had his school and college education (B.Sc. (1955), M.Sc. (1957) and Ph.D. (1962), all in Statistics) from Calcutta University. Professor Sen is Fellow of the Institute of Mathematical Statistics (1968), Fellow of the American Statistical Association (1969), and Elected Member of the International Statistical Institute (1973). Professor Sen has (co-)authored over 615 publications in Statistics, Probability Theory, Stochastic Processes, and Biostatistics in leading journals in these areas, and (co-)authored or (co-) edited 23 books and monographs in Statistics, Probability Theory and Biostatistics. He has (co-)supervised the Doctoral Dissertation of 82 students from University of North Carolina (1969–2009), many of whom have achieved distinction both nationally and internationally. In 1988 he was awarded the Boltzman Award in Mathematical Sciences from Charles University, Prague, and in 1998, the Commemoration Medal by the Czech Union of Mathematicians and Physicists, Prague. In 2002, he was awarded the Senior Noether Award from the American Statistical Association for his significant contributions to Nonparametrics, teaching as well as research. In 2010, Professor Sen has received the Wilks Medal, American Statistical Association. He was the Founding (joint) Editor of two international journals: *Sequential Analysis* (1982) and *Statistics and Decisions* (1983). Currently, he is the Chief Editor of *Sankhya* (Series A and B).

"Professor Sen's pioneering contributions have touched nearly every area of statistics. He is the first person who, in joint collaboration with Professor S. K. Chatterjee, developed multivariate rank tests as well as time-sequential nonparametric methods. He is also the first person who carried out in-depth research in sequential nonparametrics culminating in his now famous Wiley book *Sequential Nonparametrics: Invariance Principles and Statistical Inference* and SIAM monograph." (Malay Ghosh and Michael J. Schell, A Conversation with Pranab Kumar Sen, *Statistical Science*, Volume 23, Number 4 (2008), 548–564.

Cross References

►Data Mining
►Multivariate Data Analysis: An Overview
►Multivariate Normal Distributions
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Nonparametric Statistical Inference

References and Further Reading

Chatterjee SK, Sen PK (1964) Nonparametric testing for the bivariate two-sample location problem. Calcutta Stat Assoc Bull 13:18–58
Oja H (2010) Springer book on multivariate rank procedure, August 2010
Puri ML, Sen PK (1971) Nonparametric methods in multivariate analysis. Wiley, New York
Sen PK (2008) Kendall's tau in high dimensional genomics parsimony. Institute of Mathematical Statistics, Collection Ser. 3 pp 251–266

Multivariate Reduced-Rank Regression

Alan J. Izenman
Senior Research Professor of Statistics, Director of the Center for Statistical and Information Science
Temple University, Philadelphia, PA, USA

Multivariate reduced-rank regression is a way of constraining the multivariate linear regression model so that the rank of the regression coefficient matrix has less than full

rank. Without the constraint, multivariate linear regression has no true multivariate content.

To see this, suppose we have a random r-vector $\mathbf{X} = (X_1, \ldots, X_r)^\tau$ of predictor (or input) variables with mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_{XX}$, and a random s-vector $\mathbf{Y} = (Y_1, \ldots, Y_s)^\tau$ of response (or output) variables with mean vector $\boldsymbol{\mu}_Y$ and covariance matrix $\boldsymbol{\Sigma}_{YY}$. Suppose that the $(r+s)$-vector $\mathbf{Z} = (\mathbf{X}^\tau, \mathbf{Y}^\tau)^\tau$ has a joint distribution with mean vector and covariance matrix,

$$\boldsymbol{\mu}_Z = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{ZZ} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}, \tag{1}$$

respectively, where we assume that $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$ are both nonsingular. Now, consider the classical multivariate linear regression model,

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times r}{\boldsymbol{\Theta}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}, \tag{2}$$

where $\mathbf{Y}$ depends linearly on $\mathbf{X}$, $\boldsymbol{\mu}$ is the overall mean vector, $\boldsymbol{\Theta}$ is the multivariate regression coefficient matrix, and $\mathcal{E}$ is the error term. In this model, $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ are unknown and are to be estimated. The least-squares estimator of $(\boldsymbol{\mu}, \boldsymbol{\Theta})$ is given by

$$(\boldsymbol{\mu}^*, \boldsymbol{\Theta}^*) = arg \min_{\boldsymbol{\mu},\boldsymbol{\Theta}} \mathrm{E}\{(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})^\tau\}, \tag{3}$$

where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y - \boldsymbol{\Theta}^*\boldsymbol{\mu}_X, \quad \boldsymbol{\Theta}^* = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}. \tag{4}$$

In (3), the expectation is taken over the joint distribution of $(X^\tau, Y^\tau)^\tau$. The minimum achieved is $\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$. The $(s \times r)$-matrix $\boldsymbol{\Theta}^*$ is called the (full-rank) regression coefficient matrix. This solution is identical to that obtained by performing a sequence of s ordinary least-squares multiple regressions. For the jth such multiple regression, Y_j is regressed on the r-vector $\mathbf{X}$, where $j = 1, 2, \ldots, s$. Suppose the minimizing regression coefficient vectors are the r-vectors $\boldsymbol{\beta}_j^*$, $j = 1, 2, \ldots, s$. Arranging the coefficient vectors as the columns, $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_r^*)$, of an $(r \times s)$-matrix, and then transposing the result, it follows from (4) that

$$\boldsymbol{\Theta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_r^*)^\tau. \tag{5}$$

Thus, multivariate linear regression is equivalent to just carrying out a sequence of multiple regressions. This is why multivariate regression is often confused with multiple regression.

Now, rewrite the multivariate linear model as

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times r}{\mathbf{C}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}, \tag{6}$$

where the rank constraint is

$$\mathrm{rank}(\mathbf{C}) = t \le \min(r, s). \tag{7}$$

Equations (6) and (7) form the multivariate reduced-rank regression model. When the rank condition (7) holds, there exist two (nonunique) full-rank matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A}$ is an $(s \times t)$-matrix and $\mathbf{B}$ is a $(t \times r)$-matrix, such that

$$\underset{s\times r}{\mathbf{C}} = \underset{s\times t}{\mathbf{A}}\ \underset{t\times r}{\mathbf{B}}. \tag{8}$$

The multivariate reduced-rank regression model can now be written as

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times t}{\mathbf{A}}\ \underset{t\times r}{\mathbf{B}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}. \tag{9}$$

The rank condition has been embedded into the regression model. The goal is to estimate $\boldsymbol{\mu}$, $\mathbf{A}$, and $\mathbf{B}$ (and, hence, $\mathbf{C}$).

Let $\boldsymbol{\Gamma}$ be a positive-definite symmetric $(s \times s)$-matrix of weights. The weighted least-squares estimates of $(\boldsymbol{\mu}, A, B)$ are

$$(\boldsymbol{\mu}^*, \mathbf{A}^*, \mathbf{B}^*) = arg \min_{\boldsymbol{\mu},\mathbf{A},\mathbf{B}} \mathrm{E}\{(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})^\tau \boldsymbol{\Gamma} (\mathbf{Y} - \boldsymbol{\mu} - \mathbf{ABX})\} \tag{10}$$

where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y - \mathbf{AB}\boldsymbol{\mu}_X \tag{11}$$

$$\mathbf{A}^* = \boldsymbol{\Gamma}^{-1/2}\mathbf{V} \tag{12}$$

$$\mathbf{B}^* = \mathbf{V}^\tau \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1}, \tag{13}$$

and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_t)$ is an $(s \times t)$-matrix, where the jth column, $\mathbf{v}_j$, is the eigenvector corresponding to the jth largest eigenvalue, λ_j, of the $(s \times s)$ symmetric matrix,

$$\boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Gamma}^{1/2}. \tag{14}$$

The multivariate reduced-rank regression coefficient matrix $\mathbf{C}$ with rank t is, therefore, given by

$$\mathbf{C}^* = \boldsymbol{\Gamma}^{-1/2} \left(\sum_{j=1}^{t} \mathbf{v}_j \mathbf{v}_j^\tau \right) \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1}. \tag{15}$$

The minimum achieved is $tr\{\boldsymbol{\Sigma}_{YY}\boldsymbol{\Gamma}\} - \sum_{j=1}^{t} \lambda_j$.

The main reason that multivariate reduced-rank regression is so important is that it contains as special cases the classical statistical techniques of ▶principal component analysis, canonical variate and correlation analysis (see ▶Discriminant Analysis: An Overview, and ▶Discriminant Analysis: Issues and Problems), linear discriminant analysis, exploratory factor analysis, multiple correspondence analysis, and other linear methods of analyzing multivariate data. It is also closely related to artificial neural network models and to cointegration in the econometric literature.

For example, the special cases of principal component analysis, canonical variate and correlation analysis, and linear discriminant analysis are given by the following choices: For *principal component analysis*, set $\mathbf{X} \equiv \mathbf{Y}$

and $\mathbf{\Gamma} = \mathbf{I}_s$; for *canonical variate and correlation analysis*, set $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$; for *linear discriminant analysis*, use the canonical-variate analysis choice of $\mathbf{\Gamma}$ and set $\mathbf{Y}$ to be a vector of binary variables whose component values (0 or 1) indicate the group or class to which an observation belongs. Details of these and other special cases can be found in Izenman (2008). If the elements of $\mathbf{\Sigma}_{ZZ}$ in (1) are unknown, as will happen in most practical problems, they have to be estimated using sample data on $\mathbf{Z}$.

The relationships between multivariate reduced-rank regression and the classical linear dimensionality reduction techniques become more interesting when the meta-parameter t is unknown and has to be estimated. The value of t is called the *effective dimensionality* of the multivariate regression (Izenman 1980). Estimating t is equivalent to the classical problems of determining the number of principal components to retain, the number of canonical variate to retain, or the number of linear discriminant functions necessary for classification purposes. Graphical methods for estimating t include the scree plot, the rank trace plot, and heatmap plots. Formal hypothesis tests have also been developed for estimating t.

When the number of variables is greater than the number of observations, some adjustments to the results have to be made to ensure that $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ can be inverted. One simple way of doing this is to replace $\mathbf{\Sigma}_{XX}$ by $\mathbf{\Sigma}_{XX} + \delta\mathbf{I}_r$ and to replace $\mathbf{\Sigma}_{YY}$ by $\mathbf{\Sigma}_{YY} + \kappa\mathbf{I}_s$ as appropriate, where $\delta > 0$ and $\kappa > 0$. Other methods, including regularization, banding, tapering, and thresholding, have been studied for estimating large covariance matrices and can be used here as appropriate.

The multivariate reduced-rank regression model can also be developed for the case of nonstochastic (or fixed) predictor variables.

The multivariate reduced-rank regression model has its origins in Anderson (1951), Rao (1964, 1965), and Brillinger (1969), and its name was coined by Izenman (1972, 1975). For the asymptotic distribution of the estimated reduced-rank regression coefficient matrix, see Anderson (1999), who gives results for both the random-$\mathbf{X}$ and fixed-$\mathbf{X}$ cases. Additional references are the monographs by van der Leeden (1990) and Reinsel and Velu (1998).

About the Author

Professor Izenman was Director of the Statistics and Probability Program at the National Science Foundation (1992–1994). He has been an Associate Editor of the *Journal of the American Statistical Association*. He is Associate Editor of the journals *Law, Probability, and Risk* and *Statistical Analysis and Data Mining*. He is a Fellow of the American Statistical Association. He was Vice-President, ASA Philadelphia Chapter (1987–1988).

Cross References

►Canonical Correlation Analysis
►Discriminant Analysis: An Overview
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Statistical Analysis
►Principal Component Analysis

References and Further Reading

Anderson TW (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann Math Stat 22:327–351

Anderson TW (1999) Asymptotic distribution of the reduced-rank regression estimator under general conditions. Ann Stat 27:1141–1154

Brillinger DR (1969) The canonical analysis of stationary time series. In: Multivariate analysis II, Krishaiah PR (ed) Academic, New York, pp 331–350

Izenman AJ (1972) Reduced-rank regression for the multivariate linear model, its relationship to certain multivariate techniques, and its application to the analysis of multivariate data, Ph.D. dissertation, University of California, Berkeley

Izenman AJ (1975) Reduced-rank regression for the multivariate linear model. J Multivariate Anal 5:248–264

Izenman AJ (1980) Assessing dimensionality in multivariate regression. In: Handbook of statistics I, Krishnaiah PR (ed) North-Holland, Amsterdam, pp 571–591

Izenman AJ (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York

Rao CR (1964) The use and interpretation of principal components in applied research. Sankhya A 26:329–358

Rao CR (1965) Linear statistical inference and its applications. Wiley, New York

Reinsel GC, Velu RP (1998) Multivariate reduced-rank regression, Lecture notes in statistics, vol 136, Springer, New York

Van der Leeden R (1990) Reduced-rank regression with structured residuals. DSWO, Leiden

Multivariate Statistical Analysis

Nanny Wermuth
Professor of Statistics
Chalmers Technical University/University of Gothenburg, Gothenburg, Sweden

Classical multivariate statistical methods concern models, distributions and inference based on the Gaussian distribution. These are the topics in the first textbook for mathematical statisticians by T. W. Anderson that was published in 1958 and that appeared as a slightly expanded 3rd edition in 2003. Matrix theory and notation is used

there extensively to efficiently derive properties of the multivariate Gaussian or the Wishart distribution, of principal components, of canonical correlation and discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) and of the general multivariate linear model in which a Gaussian response vector variable Y_a has linear least-squares regression on all components of an explanatory vector variable Y_b.

In contrast, many methods for analyzing sets of observed variables have been developed first within special substantive fields and some or all of the models in a given class were justified in terms of probabilistic and statistical theory much later. Among them are factor analysis (see ►Factor Analysis and Latent Variable Modelling), path analysis, ►structural equation models, and models for which partial-least squares estimation have been proposed. Other multivariate techniques such as cluster analysis (see ►Cluster Analysis: An Introduction) and ►multidimensional scaling have been often used, but the result of such an analysis cannot be formulated as a hypothesis to be tested in a new study and satisfactory theoretical justifications are still lacking.

Factor analysis was proposed by psychologist C. Spearman (1904), (1926) and, at the time, thought of as a tool for measuring human intelligence. Such a model has one or several latent variables. These are hidden or unobserved and are to explain the observed correlations among a set of observed variables, called items in that context. The difficult task is to decide how many and which of a possibly large set of items to include into a model. But, given a set of latent variables, a classical factor analysis model specifies for a joint Gaussian distribution mutual independence of the observed variables given the latent variables. This can be recognized to be one special type of a graphical Markov model; see Cox and Wermuth (1996), Edwards (2000), Lauritzen (1996), Whittaker (1990).

Path analysis was developed by geneticist S. Wright (1923), (1934) for systems of linear dependence of variables with zero mean and unit variance. He used what we now call directed acyclic graphs to represent hypotheses of how the variables he was studying could have been generated. He compared correlations implied for missing edges in the graph with corresponding observed correlations to test the goodness of fit of such a hypothesis.

By now it is known, under which condition for these models in standardized Gaussian variables, maximum-likelihood estimates of correlations coincide with Wright's estimates via path coefficients. The condition on the graph is simple: there should be no three-node-two-edge subgraph of the following kind $\circ \longrightarrow \circ \longleftarrow \circ$. Then, the directed acyclic graph is said to be decomposable and captures the same independences as the concentration graph obtained by replacing each arrow by an undirected edge. In such Gaussian concentration graph models, estimated variances are matched to the observed variances so that estimation of correlations and variances is equivalent to estimation of covariances and variances.

Wright's method of computing implied path coefficients by "tracing paths" has been generalized via a so-called separation criterion. This criterion, given by Geiger, Verma and Pearl (1990), permits to read off a directed acyclic graph all independence statements that are implied by the graph. The criterion takes into account that not only ignoring (marginalizing over) variables might destroy an independence, but also conditioning on common responses may render two formerly independent variables to be dependent. In addition, the separation criterion holds for any distribution generated over the graph.

The separation criterion for directed acyclic graphs has been translated into conditions for the presence of edge-inducing paths in the graph; see Marchetti and Wermuth (2009). Such an edge-inducing path is also association-inducing in the corresponding model, given some mild conditions on the graph and on the distributions generated over it; see Wermuth (2010). In the special case of only marginalizing over linearly related variables, these induced dependences coincide with the path-tracing results given by Wright provided the directed acyclic graph model is decomposable and the variables are standardized to have zero means and unit variances. This applies not only to Gaussian distributions but also to special distributions of symmetric binary variables; see Wermuth et al. (2009).

Typically however, directed acyclic graph models are defined for unstandardized random variables of any type. Then, most dependences are no longer appropriately represented by linear regression coefficients or correlations, but maximum-likelihood estimates of all measures of dependence can still be obtained by separately maximizing each univariate conditional distribution, provided only that its parameters are variation-independent from parameters of distributions in the past.

Structural equation models, developed in econometrics, can be viewed as another extension of Wright's path analyses. The result obtained by T. Haavelmo (1943) gave an important impetus. For his insight that separate linear least-squares estimation may be inappropriate for equations having strongly correlated residuals, Haavelmo received a Nobel prize in 1989. It led to a class of models defined by linear equations with correlated residuals and to responses called endogenous. Other variables conditioned on and considered to be predetermined were named

exogenous. Vigorous discussions of estimation methods for structural equations occurred during the first few Berkeley symposia on mathematical statistics and probability from 1945 to 1965.

Path analysis and structural equation models were introduced to sociological research via the work by O.D. Duncan (1966, 1975). Applications of structural equation models in psychological and psychometric research resulted from cooperations between A. Goldberger and K. Jöreskog; see Goldberger (1971, 1972) and Jöreskog (1973, 1981). The methods became widely used once a corresponding computer program for estimation and tests was made available; see also Kline (2010).

In 1962, A. Zellner published his results on seemingly unrelated regressions. He points out that two simple regression equations are not separate if the two responses are correlated and that two dependent endogenous variables need to be considered jointly and require simultaneous estimation methods. These models are now recognized as special cases of both linear structural equations and of multivariate regression chains, a subclass of graphical Markov models; see Cox and Wermuth (1993), Drton (2009), Marchetti and Lupparelli (2010).

But it was not until 40 years later, that a maximum-likelihood solution for the Gaussian distribution in four variables, split into a response vector Y_a and vector variable Y_b, was given and an example of a poorly fitting data set with very few observations for which the likelihood equations have two real roots; see Drton and Richardson (2004). For well-fitting data and reasonably large sample sizes, this is unlikely to happen; see Sundberg (2010). For such situations, a close approximation to the maximum-likelihood estimate has been given in closed form for the seemingly unrelated regression model, exploiting that it is a reduced model to the covering model that has closed-form maximum-likelihood estimates, the general linear model of Y_a given Y_b; see Wermuth et al. (2006), Cox and Wermuth (1990).

For several discrete random variables of equal standing, i.e., without splits into response and explanatory variables, maximum-likelihood estimation was developed under different conditional independence constraints in a path-breaking paper by M. Birch (1963). This led to the formulation of general log-linear models, which were studied intensively among others by Haberman (1974), Bishop et al. (1975), Sundberg (1975) and by L. Goodman, as summarized in a book of his main papers on this topic, published in 1978. His work was motivated mainly by research questions from the social and medical sciences.

For several Gaussian variables of equal standing, two different approaches to reducing the number of parameters in a model, were proposed at about the same time. T. W. Anderson put structure on the covariances, the moment parameters of a joint Gaussian distribution and called the resulting models, hypotheses linear in covariances; see Anderson (1973), while A. P. Dempster put structure on the canonical parameters with zero constraints on concentrations, the off-diagonal elements of the inverse of a covariance matrix, and called the resulting models covariance selection models; see Dempster (1972).

Nowadays, log-linear models and covariance selection models are viewed as special cases of concentration graph models and zero constraints on the covariance matrix of a Gaussian distribution as special cases of covariance graph models. Covariance and concentration graph models are graphical Markov models with undirected graphs capturing independences. A missing edge means marginal independence in the former and conditional independence given all remaining variables in the latter; see also Wermuth and Lauritzen (1990), Wermuth and Cox (1998), (2004), Wermuth (2010).

The largest known class of Gaussian models that is in common to structural equation models and to graphical Markov models are the recursive linear equations with correlated residuals. These include linear summary graph models of Wermuth (2010), linear maximal ancestral graph of Richardson and Spirtes (2002), linear multivariate regression chains, and linear directed acyclic graph models. Deficiencies of some formulations start to be discovered by using algebraic methods. Identification is still an issue to be considered for recursive linear equations with correlated residuals, since so far only necessary or sufficient conditions are known but not both. Similarly, maximum-likelihood estimation still needs further exploration; see Drton et al. (2009).

For several economic time series, it became possible to judge whether such fluctuating series develop nevertheless in parallel, that is whether they represent cointegrating variables because they have a common stochastic trend. Maximum-likelihood analysis for cointegrating variables, formulated by Johansen (1988, 2009), has led to many important applications and insights; see also Hendry and Nielsen (2007).

Algorithms and corresponding programs are essential for any widespread use of multivariate statistical methods and for successful analyses. In particular, iterative proportional fitting, formulated by Bishop (1964) for log-linear models, and studied further by Darroch and Ratcliff (1972), was adapted to concentration graph models for CG (conditional Gaussian)-distributions (Lauritzen and Wermuth 1989) of mixed discrete and continuous variables by Frydenberg and Edwards (1989).

The EM (expectation-maximization)-algorithm of Dempster et al. (1977) was adapted to Gaussian directed

acyclic graph models with latent variables by Kiiveri (1987) and to discrete concentration graph models with missing observation by Lauritzen (1995).

With the TM-algorithm of Edwards and Lauritzen (2001), studied further by Sundberg (2002), maximum-likelihood estimation became feasible for all chain graph models called blocked concentration chains in the case these are made up of CG (conditional Gaussian)-regressions (Lauritzen and Wermuth 1989).

For multivariate regression chains of discrete random variables, maximum-likelihood estimation has now been related to the multivariate logistic link function by Marchetti and Lupparelli (2010), where these link functions provide a common framework and corresponding algorithm for ▶generalized linear models, which include among others linear, logistic and probit regressions as special cases; see McCullagh and Nelder (1989), Glonek and McCullagh (1995).

Even in linear models, estimation may become difficult when some of the explanatory variables are almost linear functions of others, that is if there is a problem of ▶multicollinearity. This appears to be often the case in applications in chemistry and in the environmental sciences. Thus, in connection with consulting work for chemists, Hoerl and Kennard (1970) proposed the use of ridge-regression (see ▶Ridge and Surrogate Ridge Regressions) instead of linear least-squares regression. This means for regressions of vector variable Y on X, to add to X^TX some positive constant k along the diagonal before matrix inversion to give as estimator $\tilde{\beta} = (kI + X^TX)^{-1}X^TY$.

Both ridge-regression and partial-least-squares, (see ▶Partial Least Squares Regression Versus Other Methods) proposed as an estimation method in the presence of latent variables by Wold (1980), have been recognized by Björkström and Sundberg (1999) to be shrinkage estimators and as such special cases of Tykhonov (1963) regularization.

More recently, a number of methods have been suggested which combine adaptive skrinkage methods with variable selection. A unifying approach which includes the least-squares estimator, shrinkage estimators and various combinations of variable selection and shrinkage has recently been given via a least squares approximation by Wang and Leng (2007). Estimation results depend necessarily on the chosen formulations and the criteria for shrinking dependences and for selecting variables.

Many more specialized algorithms and programs have been made available within the open access programming environment R, also those aiming to analyze large numbers of variables for only few observed individuals. It remains to be seen, whether important scientific insights will be gained by their use.

About the Author

Dr Nanny Wermuth is Professor of Statistics, at the joint Department of Mathematical Sciences of Chalmers Technical University and the University of Gothenburg. She is a Past President, Institute of Mathematical Statistics (2008–2009) and Past President of the International Biometric Society (2000–2001). In 1992 she was awarded a Max Planck-Research Prize, jointly with Sir David Cox. She chaired the Life Science Committee of the International Statistical Institute (2001–2005) and was an Associate editor of the *Journal of Multivariate Analysis* (1998–2001) and *Bernoulli* (2007–2010). Professor Wermuth is an Elected member of the German Academy of Sciences and of the International Statistical Institute (1982), an elected Fellow of the American Statistical Association (1989), and of the Institute of Mathematical Statistics (2001). She is a co-author (with David R. Cox) of the text *Multivariate dependencies: models, analysis and interpretation* (Chapman and Hall, 1996).

Cross References

▶Canonical Correlation Analysis
▶Cluster Analysis: An Introduction
▶Correspondence Analysis
▶Discriminant Analysis: An Overview
▶Discriminant Analysis: Issues and Problems
▶Factor Analysis and Latent Variable Modelling
▶General Linear Models
▶Likelihood
▶Logistic Regression
▶Multidimensional Scaling
▶Multidimensional Scaling: An Introduction
▶Multivariate Analysis of Variance (MANOVA)
▶Multivariate Data Analysis: An Overview
▶Multivariate Normal Distributions
▶Multivariate Rank Procedures: Perspectives and Prospectives
▶Multivariate Reduced-Rank Regression
▶Multivariate Statistical Process Control
▶Multivariate Technique: Robustness
▶Partial Least Squares Regression Versus Other Methods
▶Principal Component Analysis
▶R Language
▶Ridge and Surrogate Ridge Regressions
▶Structural Equation Models

References and Further Reading

Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York; (2003) 3rd edn. Wiley, New York

Anderson TW (1973) Asymptotically efficient estimation of covariance matrices with linear structure. Ann Stat 1:135–141

Birch MW (1963) Maximum likelihood in three-way contingency tables. J Roy Stat Soc B 25:220–233

Bishop YMM (1967) Multidimensional contingency tables: cell estimates. Ph.D. dissertation, Department of Statistics, Harvard University

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Björkström A, Sundberg R (1999) A generalized view on continuum regression. Scand J Stat 26:17–30

Cox DR, Wermuth N (1990) An approximation to maximum-likelihood estimates in reduced models. Biometrika 77:747–761

Cox DR, Wermuth N (1993) Linear dependencies represented by chain graphs (with discussion). Stat Sci 8:204–218; 247–277

Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis, and interpretation. Chapman & Hall, London

Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. Ann Math Stat 43:1470–1480

Dempster AP (1972) Covariance selection Biometrics 28:157–175

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–38

Drton M (2009) Discrete chain graph models. Bernoulli 15:736–753

Drton M, Richardson TS (2004) Multimodality of the likelihood in the bivariate seemingly unrelated regression model. Biometrika 91:383–392

Drton M, Eichler M, Richardson TS (2009) Computing maximum likelihood estimates in recursive linear models. J Mach Learn Res 10:2329–2348

Duncan OD (1966) Path analysis: sociological examples. Am J Sociol 72:1–12

Duncan OD (1975) Introduction to structural equation models. Academic, New York

Edwards D (2000) Introduction to graphical modelling, 2nd edn. Springer, New York

Edwards D, Lauritzen SL (2001) The TM algorithm for maximising a conditional likelihood function. Biometrika 88:961–972

Frydenberg M, Edwards D (1989) A modified iterative proportional scaling algorithm for estimation in regular exponential families. Comput Stat Data Anal 8:143–153

Frydenberg M, Lauritzen SL (1989) Decomposition of maximum likelihood in mixed interaction models. Biometrika 76:539–555

Geiger D, Verma TS, Pearl J (1990) Identifying independence in Bayesian networks. Networks 20:507–534

Glonek GFV, McCullagh P (1995) Multivariate logistic models. J Roy Stat Soc B 57:533–546

Goldberger AS (1971) Econometrics and psychometrics: a survey of communalities. Psychometrika 36:83–107

Goldberger AS (1972) Structural equation methods in the social sciences. Econometrica 40:979–1002

Goodman LA (1978) Analyzing qualitative/categorical data. Abt Books, Cambridge

Haberman SJ (1974) The analysis of frequency data. University of Chicago Press, Chicago

Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11:1–12; Reprinted in: Hendry DF, Morgan MS (eds) (1995) The foundations of econometric analysis. Cambridge University Press, Cambridge

Hendry DF, Nielsen B (2007) Econometric modeling: a likelihood approach. Princeton University Press, Princeton

Hoerl AE, Kennard RN (1970) Ridge regression. Biased estimation for non-orthogonal problems. Technometrics 12:55–67

Johansen S (1988) Statistical analysis of cointegration vectors. J Econ Dyn Contr 12:231–254; Reprinted in: Engle RF, Granger CWJ (eds) (1991) Long-run economic relationships, readings in cointegration. Oxford University Press, Oxford, pp 131–152

Johansen S (2009) Cointegration: overview and development. In: Handbook of financial time series, Andersen TG, Davis R, Kreiss J-P, Mikosch T (eds), Springer, New York, pp 671–693

Jöreskog KG (1973) A general method for estimating a linear structural equation system. In: Structural equation models in the social sciences, Goldberger AS, Duncan OD (eds), Seminar, New York, pp 85–112

Jöreskog KG (1981) Analysis of covariance structures. Scan J Stat 8:65–92

Kiiveri HT (1987) An incomplete data approach to the analysis of covariance structures. Psychometrika 52:539–554

Kline RB (2010) Principles and practice of structural equation modeling, 3rd edn. Guilford, New York

Lauritzen SL (1995) The EM-algorithm for graphical association models with missing data. Comp Stat Data Anal 1:191–201

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lauritzen SL, Wermuth N (1989) Graphical models for association between variables, some of which are qualitative and some quantitative. Ann Stat 17:31–57

Marchetti GM, Lupparelli M (2010) Chain graph models of multivariate regression type for categorical data. Bernoulli, to appear and available on ArXiv, http://arxiv.org/abs/0906.2098v2

Marchetti GM, Wermuth N (2009) Matrix representations and independencies in directed acyclic graphs. Ann Stat 47:961–978

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton

Richardson TS, Spirtes P (2002) Ancestral Markov graphical models. Ann Stat 30:962–1030

Spearman C (1904) General intelligence, objectively determined and measured. Am J Psych 15:201–293

Spearman C (1926) The abilities of man. Macmillan, New York

Sundberg R (1975) Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. Scand J Stat 2:71–79

Sundberg R (2002) The convergence rate of the TM algorithm of Edwards and Lauritzen. Biometrika 89:478–483

Sundberg R (2010) Flat and multimodal likelihoods and model lack of fit in curved exponential families. Scand J Stat, published online: 28 June 2010

Tikhonov AN (1963) Solution of ill-posed problems and the regularization method (Russian). Dokl Akad Nauk SSSR 153:49–52

Wang H, Leng C (2007) Unified lasso estimation via least square approximation. J Am Stat Assoc 102:1039–1048

Wermuth N (2010) Probability distributions with summary graph structure. Bernoulli, to appear and available on ArXiv, http://arxiv.org/abs/1003.3259

Wermuth N, Cox DR (1998) On association models defined over independence graphs. Bernoulli 4:477–495

Wermuth N, Cox DR (2004) Joint response graphs and separation induced by triangular systems. J Roy Stat Soc B 66:687–717
Wermuth N, Lauritzen SL (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discusssion). J Roy Stat Soc B 52:21–75
Wermuth N, Marchetti GM, Cox DR (2009) Triangular systems for symmetric binary variables. Electr J Stat 3:932–955
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester
Wold HOA (1954) Causality and econometrics. Econometrica 22:162–177
Wold HOA (1980) Model construction and evaluation when theoretical knowledge is scarce: theory and application of partial least squares. In: Evaluation of econometric models, Kmenta J, Ramsey J (eds), Academic, New York, pp 47–74
Wright S (1923) The theory of path coefficients: a reply to Niles' criticism. Genetics 8:239–255
Wright S (1934) The method of path coefficients. Ann Math Stat 5:161–215
Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57:348–368

Multivariate Statistical Distributions

Donald R. Jensen
Professor Emeritus
Virginia Polytechnic Institute and State University,
Blacksburg, VA, USA

Origins and Uses

Multivariate distributions (MDs) are defined on finite-dimensional spaces. Origins trace to early studies of ►*multivariate normal distributions* as models for dependent chance observations (Adrian 1808; Bravais 1846; Dickson 1886; Edgeworth 1892; Galton 1889; Gauss 1823; Helmert 1868; Laplace 1811; Pearson 1896; Plana 1813; Schols 1875; Spearman 1904; Student 1908); for two and three dimensions in Bravais (1846) and Schols (1875); and for finite dimensions in Edgeworth (1892) and Gauss (1823), advancing such now–familiar concepts as regression and partial correlation. Let $\boldsymbol{Y} = [Y_1, \ldots, Y_5]$ designate chance observations; in pharmacology as systolic (Y_1) and diastolic (Y_2) pressures, pulse rate (Y_3), and gross (Y_4) and fine (Y_5) motor skills. Strengths of materials may register moduli of elasticity (Y_1) and of rupture (Y_2), specific gravity (Y_3), coefficient of linear expansion (Y_4), and melting point (Y_5). A complete probabilistic description of each vector observation entails the joint distribution of $[Y_1, \ldots, Y_5]$. A sample of n such k–vectors, arranged as rows, yields a random matrix $\boldsymbol{Y} = [Y_{ij}]$ of order $(n \times k)$, its distribution supporting much of ►*multivariate statistical analysis*.

Beyond modeling chance outcomes, MDs describe probabilistic features of data–analytic operations, to include statistical inference, decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview), and other evidentiary analyses. In inference the frequentist seeks joint distributions (1) of multiparameter estimates, and (2) of statistics for testing multiple hypotheses, both parametric and nonparametric. Such distributions derive from observational models. Similarly, multiparameter Bayesian methods require MDs in modeling prior, contemporary, and posterior distributions for the parameters. In addition, MDs serve to capture dependencies owing to repeated measurements on experimental subjects. MDs derive from other distributions through transformations, projections, conditioning, convolutions, extreme values, mixing, compounding, truncating, and censoring. Specifically, experiments modeled conditionally in a random environment yield unconditional distributions as mixtures; see Everitt and Hand (1981), Lindsay (1995), McLachlan and Basford (1988), and Titterington et al. (1985). Random processes, to include such concepts as stationarity, are characterized through MDs as their finite–dimensional projections. Beyond probability, MD–theory occasionally supports probabilistic proofs for purely mathematical theorems. In short, MDs arise throughout statistics, applied probability, and beyond, and their properties are essential to understanding those fields.

In what follows $\mathbb{R}^k$, $\mathbb{R}^k_+$, $\mathbb{F}_{n\times k}$, $\mathbb{S}_k$, and $\mathbb{S}^+_k$ respectively designate Euclidean k–space, its positive orthant, the real $(n \times k)$ matrices, the real symmetric $(k \times k)$ matrices, and their positive definite varieties. Special arrays are $\boldsymbol{I}_k$, the $(k \times k)$ identity, and the diagonal matrix $\mathrm{Diag}(a_1, \ldots, a_k)$. The transpose, inverse, trace, and determinant of $\boldsymbol{A} \in \mathbb{F}_{k\times k}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$, $\mathrm{tr}(\boldsymbol{A})$, and $|\boldsymbol{A}|$, with $\boldsymbol{a}' = [a_1, \ldots, a_k]$ as the transpose of $\boldsymbol{a} \in \mathbb{R}^k$. For $\boldsymbol{Y} \in \mathbb{R}^k$ random, its expected vector, dispersion matrix, and law of distribution are $E(\boldsymbol{Y}) \in \mathbb{R}^k$, $V(\boldsymbol{Y}) \in \mathbb{S}^+_k$, and $\mathcal{L}(\boldsymbol{Y})$. Abbreviations include *pdf*, *pmf*, *cdf*, and *chf*, for probability density, probability mass, cumulative distribution, and ►characteristic functions, respectively.

Some Properties

MDs merit scrutiny at several levels. At one extreme are weak assumptions on existence of low-order moments, as in Gauss–Markov theory. At the other extremity are rigidly parametric models, having MDs of specified functional forms to be surveyed subsequently. In between are

Multivariate Statistical Distributions. Table 1 Examples of spherical distributions on $\mathbb{R}^n$ having density $f(\boldsymbol{x})$ or characteristic function $\xi(\boldsymbol{t})$; see Chmielewski (1981)

Density or *chf*	Comments	
Normal	$f(\boldsymbol{x}) = c_1 \exp(-\boldsymbol{x}'\boldsymbol{x}/2)$	$N_n(\boldsymbol{0}, \boldsymbol{I}_n)$
Pearson Type II	$f(\boldsymbol{x}) = c_2(1-\boldsymbol{x}'\boldsymbol{x})^{\gamma-1}$	$\gamma > 1$
Pearson Type VII	$f(\boldsymbol{x}) = c_3(1+\boldsymbol{x}'\boldsymbol{x})^{-\gamma}$	$\gamma > n/2$
Student t	$f(\boldsymbol{x}) = c_4(1+\nu^{-1}\boldsymbol{x}'\boldsymbol{x})^{-(\nu+n)/2}$	ν a positive integer
Cauchy	$f(\boldsymbol{x}) = c_5(1+\boldsymbol{x}'\boldsymbol{x})^{-(n+1)/2}$	Student t $\nu = 1$
Scale mixtures	$f(\boldsymbol{x}) = c_6 \int_0^\infty t^{-n/2} \exp(-\boldsymbol{x}'\boldsymbol{x}/2t) dG(t)$	$G(t)$ a *cdf*
Stable laws	$\xi(\boldsymbol{t}) = c_7 \exp[\gamma(\boldsymbol{t}'\boldsymbol{t})^{\alpha/2}]$	$0 < \alpha < 2; \gamma > 0$

classes of MDs exhibiting such common structural features as symmetry or unimodality, giving rise to *semiparametric* models of note. Of particular relevance are derived distributions that are unique to all members of an underlying class.

Specifically, distributions on $\mathbb{F}_{n\times k}$ in the class $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ have *pdf*s as given in Table 3. Here $\boldsymbol{\Theta} \in \mathbb{F}_{n\times k}$ comprise location parameters; $\boldsymbol{\Gamma} \in \mathbb{S}_n^+$ and $\boldsymbol{\Sigma} \in \mathbb{S}_k^+$ are scale parameters; $\phi(\cdot)$ is a function on $\mathbb{S}_k^+$; and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is a factor of $\boldsymbol{\Sigma}^{-1}$. These distributions are invariant for $\boldsymbol{\Gamma} = \boldsymbol{I}_n$ in that $\mathcal{L}(\boldsymbol{Y} - \boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{\Theta}))$ for every real orthogonal matrix $\boldsymbol{Q}(n \times n)$. A subclass, taking $\phi(\boldsymbol{A}) = \psi(\mathrm{tr}(\boldsymbol{A}))$, with ψ defined on $[0,\infty)$, is $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ as in Table 3. Here independence among rows of $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]'$ and multinormality are linked: If $\mathcal{L}(\boldsymbol{Y}) \in S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$, then $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ are mutually independent if and only if $\boldsymbol{Y}$ is matrix normal, namely $N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$ on $\mathbb{F}_{n\times k}$; see James (1954). A further subclass on $\mathbb{R}^n$, with $k = 1$ and $\boldsymbol{\Sigma}(1\times 1) = 1$, are the *elliptical distributions* on $\mathbb{R}^n$, namely, $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \psi); \psi \in \Psi\}$, with location-scale parameters $(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ and the typical *pdf* $f(\boldsymbol{y}) = |\boldsymbol{\Gamma}|^{-\frac{1}{2}} \psi((\boldsymbol{y} - \boldsymbol{\theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{\theta}))$. The foregoing all contain multivariate normal and heavy–tailed Cauchy models as special cases, and all have served as observational models *in lieu of* multivariate normality. In particular, $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ often serve as semiparametric surrogates for $N_n(\boldsymbol{\theta}, \boldsymbol{I}_n)$ in univariate samples, and $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ in the analysis of multivariate data. Examples from $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ are listed in Table 1, cross-referenced as in Chmielewski (1981) to well-known distributions on $\mathbb{R}^1$.

Inferences built on these models often remain exact as for normal models, certifying their use as semiparametric surrogates. This follows from the invariance of stipulated derived distributions as in Jensen and Good (1981). Further details, for their use as observational models on $\mathbb{R}^k$ and $\mathbb{F}_{n\times k}$, for catalogs of related and derived distributions, and for the robustness of various inferential procedures, are found in Cambanis et al. (1981), Chmielewski (1981), Devlin et al. (1976), Fang and Anderson (1990), Fang et al. (1990), Fang and Zhang (1990), James (1954), and Kariya and Sinha (1989). Regarding $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ and its extensions, see Dawid (1977), Dempster (1969), and Jensen and Good (1981). These facts bear heavily on the robustness and validity of normal-theory procedures for use with nonnormal data, including distributions having heavy tails. The cited distributions all exhibit symmetries, including symmetries under reflections. Considerable recent work addresses skewed MDs, often resulting from truncation; see Arnold and Beaver (2000), for example.

Properties of distributions on $\mathbb{R}^1$ often extend nonuniquely to the case of MDs. Concepts of unimodality on $\mathbb{R}^k$ are developed in Dharmadhikari and Joag-Dev (1988), some enabling a sharpening of joint Chebyshev bounds. Stochastic ordering on $\mathbb{R}^1$ likewise admits a multiplicity of extensions. These in turn support useful probability inequalities on $\mathbb{R}^k$ as in Tong (1980), many pertaining to distributions cited here. Let $\mu(\cdot)$ and $\nu(\cdot)$ be probability measures on $\mathbb{R}^k$, and $\mathcal{C}_k$ the compact convex sets in $\mathbb{R}^k$ symmetric under reflection about $\boldsymbol{0} \in \mathbb{R}^k$. The concentration ordering (Birnbaum 1948) on $\mathbb{R}^1$ is extended in Sherman (1904): $\mu(\cdot)$ is said to be *more peaked about* $\boldsymbol{0} \in \mathbb{R}^k$ than $\nu(\cdot)$ if and only if $\mu(A) \geq \nu(A)$ for every $A \in \mathcal{C}_k$. Specifically, let $P_\Sigma(\cdot; \psi)$ and $P_\Omega(\cdot; \psi)$ be probability measures for $\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{\Sigma}, \psi)$ and $\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{\Omega}, \psi)$. Then a necessary and sufficient condition that $P_\Sigma(\cdot; \psi)$ should be more peaked about $\boldsymbol{0}$ than $P_\Omega(\cdot; \psi)$, is that $(\boldsymbol{\Omega} - \boldsymbol{\Sigma}) \in \mathbb{S}_n^+$, sufficiency in Fefferman et al. (1972), necessity in Jensen (1984). Similar orderings apply when both $(\boldsymbol{\Sigma}, \psi)$ are allowed to vary (Jensen 1984), extending directly to include distributions in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$. Numerous further notions of stochastic orderings for MDs are treated in Shaked and Shanthikumar (2007).

Interest in MDs often centers on their dependencies. A burgeoning literature surrounds ▶*copulas*, expressing a joint distribution function in terms of its marginals, together with a finite–dimensional parameter quantifying the degree of dependence; see Nelsen (1998) for example. Further concepts of dependence, including notions rooted in the geometry of $\mathbb{R}^k$, are developed in Joe (1997).

The Basic Tools

Let $(\Omega, \mathfrak{B}, P)$ be a probability space, Ω an event set, $\mathfrak{B}$ a field of subsets of Ω, and P a probability measure. Given a set $\mathfrak{X}_0$, an $\mathfrak{X}_0$–valued random element is a measurable mapping $\boldsymbol{X}(\omega)$ from Ω to $\mathfrak{X}_0$, multivariate when $\mathfrak{X}_0$ is finite-dimensional, as $\mathbb{R}^k$, its *cdf* then given by $F(x_1, \ldots, x_k) = P(\omega : X_1(\omega) \le x_1, \ldots, X_k(\omega) \le x_k)$. To each *cdf* corresponds a P_X on $(\mathbb{R}^k, \mathfrak{B}_k, P_X)$ and conversely, with $\mathfrak{B}_k$ as a field of subsets of $\mathbb{R}^k$. Moreover, $\{P_X = a_1P_1 + a_2P_2 + a_3P_3; a_i \ge 0, a_1 + a_2 + a_3 = 1\}$ decomposes as a mixture: P_1 assigns positive probability to the mass points of P_X; P_2 is absolutely continuous with respect to Lebesgue (volume) measure on $(\mathbb{R}^k, \mathfrak{B}_X, \cdot)$; and P_3 is purely singular. Corresponding to $\{P_1, P_2, P_3\}$ are *cdf*s $\{F_1, F_2, F_3\}$: F_1 has a mass function (*pmf*) $p(x_1, \ldots, x_k) = P(X_1 = x_1, \ldots, X_k = x_k)$, giving jumps of F_1 at its mass points; F_2 has a *pdf* $f_2(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_2(x_1, \ldots, x_k)$ for almost all $\{x_1, \ldots, x_k\}$. The marginal *cdf* of $\boldsymbol{X}_1' = [X_1, \ldots, X_r]$ is $F_{m1}(x_1, \ldots, x_r) = F(x_1, \ldots, x_r, \infty, \ldots, \infty)$. With $\boldsymbol{X}_2' = [X_{r+1}, \ldots, X_k]$ and $\boldsymbol{x}_2' = [x_{r+1}, \ldots, x_k]$, the conditional *pmf* for $\mathcal{L}(\boldsymbol{X}_1 \mid \boldsymbol{x}_2)$, given that $\{\boldsymbol{X}_2 = \boldsymbol{x}_2\}$, is $p_{1\cdot 2}(x_1, \ldots, x_r) = \frac{p(x_1, \ldots, x_k)}{p_2(x_{r+1}, \ldots, x_k)}$ with $p_2(x_{r+1}, \ldots, x_k)$ as the marginal *pmf* for $\boldsymbol{X}_2$. A similar expression holds for P_2 in terms of the joint and marginal *pdf*s $f(x_1, \ldots, x_k)$ and $f_2(x_{r+1}, \ldots, x_k)$. As noted, F_1 is discrete and F_2 absolutely continuous, pure types to warrant their separate cataloging in the literature. On the other hand, P_3 is singular on a set in $\mathbb{R}^k$ having Lebesgue measure zero, often illustrated as a linear subspace. In contrast, P_3 is known to originate in practice through pairs $(\boldsymbol{X}, \boldsymbol{Y})$ as in Olkin and Tate (1961), such that $\boldsymbol{X}$ is multinomial and $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ is multivariate normal. Related studies are reported in a succession of articles including the recent (Bedrick et al. 2000).

The study of MDs draws heavily on the calculus of $\mathbb{R}^k$. Distributions not expressible in closed form may admit series expansions, asymptotic expansions of Cornish-Fisher and Edgeworth types, or large-sample approximations via central limit theory. Accuracy of the latter is gauged through Berry–Esséen bounds on rates of convergence, as developed extensively in Bhattacharya and Ranga Rao (1976) under moments of order greater than 2. Moreover, the integral transform pairs of Fourier, Laplace, and Mellin, including *chf*s on $\mathbb{R}^k$, are basic. Elementary operations in the space of transforms carry back to the space of distributions through inversion. Affine data transformations are intrinsic to the use of *chf*s of MDs, as treated extensively in Lukacs and Laha (1964). On the other hand, Mellin transforms couple nicely with such nonlinear operations as powers, products, and quotients of random variables, as treated in Epstein (1948) and Subrahmaniam (1970) and subsequently. In addition, functions generating joint moments, cumulants, factorial moments, and probabilities are used routinely. Projection methods determine distributions on $\mathbb{R}^k$ completely, via the one-dimensional distributions of every linear function. To continue, a property is said *to characterize* a distribution if unique to that distribution. A general treatise is Kagan et al. (1973), including reference to some MDs reviewed here.

We next undertake a limited survey of continuous and discrete MDs encountered with varying frequencies in practice. Developments are cited for random vectors and matrices. Continuing to focus on semiparametric models, we identify those distributions derived and unique to underlying classes of models, facts not widely accessible otherwise. The principal reference for continuous MDs is the encyclopedic (Kotz et al. 2000), coupled with monographs on multivariate normal (Tong 1990) and Student t (Kotz and Nadarajah 2004) distributions. For discrete MDs, encyclopedic accounts are archived in Johnson et al. (1997) and Patil and Joshi (1968).

Continuous Distributions

Central to classical **multivariate statistical analysis** are $\{N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma}); n > k\}$ for $\mathcal{L}(\boldsymbol{Y})$, and the essential derived distribution $\mathcal{L}(\boldsymbol{W}) = W_k(n, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, with $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$, as noncentral Wishart having n degrees of freedom, scale matrix $\boldsymbol{\Sigma}$, and noncentrality matrix $\boldsymbol{\Lambda} = \boldsymbol{\Theta}'\boldsymbol{\Theta}$, with central *pdf* as in Table 3.

Student tDistributions

Vector distributions. There are two basic types. Let $[Y_1, \ldots, Y_k]$ be multivariate normal with means $[\mu_1, \ldots, \mu_k]$, unit variances, and correlation matrix $\boldsymbol{R}(k \times k)$. A *Type I t distribution* is that of $\{T_j = Y_j/S; 1 \le j \le k\}$ such that $\mathcal{L}(\nu S^2) = \chi^2(\nu)$ independently of $[Y_1, \ldots, Y_k]$. Its central *pdf* is listed in Table 2. To continue, suppose that $\boldsymbol{S} = [S_{ij}]$ and $\mathcal{L}(\nu \boldsymbol{S}) = W_k(\nu, \boldsymbol{R})$, independently of $[Y_1, \ldots, Y_k]$. A *Type II t distribution* is that of $\{T_j = Y_j/S_{jj}; 1 \le j \le k\}$. Both types are central if and only if $\{\mu_1 = \cdots = \mu_k = 0\}$. These distributions arise in multiple comparisons, in the construction of rectangular confidence sets for means, in the Bayesian analysis of multivariate normal data, and in various multistage procedures. For further details see Kotz et al. (2000) and Tong (1990).

More generally, if $\mathcal{L}(X_1, \ldots, X_k, Z_1, \ldots, Z_\nu) = \mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ with $\boldsymbol{\theta}' = [\mu_1, \ldots, \mu_k, 0, \ldots, 0]$ and $\boldsymbol{\Gamma} = \mathrm{Diag}(\boldsymbol{R}, \boldsymbol{I}_\nu)$, then with $\nu S^2 = \left(Z_1^2 + \ldots + Z_\nu^2\right)$, the central distribution of $\{T_j = X_j/S; 1 \le j \le k\}$ is Type I multivariate t for all distributions in $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \psi); \psi \in \Psi\}$ as structured. Multiple comparisons using $\{T_1, \ldots, T_k\}$ under normality thus are

Multivariate Statistical Distributions. Table 2 Standard *pdf*s for some continuous distributions on $\mathbb{R}^k$

Type	Density	Comments
Student t	$k_1\left[1+\nu^{-1}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{R}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+k)/2}$	$\boldsymbol{t}\in\mathbb{R}^k$
Dirichlet	$k_2(1-\Sigma_1^k u_j)^{\alpha_0-1}\Pi_1^k u_j^{\alpha_j-1}$	$\{0\le u_j\le 1;\Sigma_1^k u_j\le 1\}$
Inv. Dirichlet	$k_3\Pi_1^k v_j^{\alpha_j-1}/[1+\Sigma_1^k v_j]^{\alpha/2}$	$\{0\le v_j<\infty;\alpha=\Sigma_0^k\alpha_j$
$\lvert\boldsymbol{W}-w\boldsymbol{\Sigma}\rvert=0$	$k_4\Pi_1^k w_i^{(\nu-k-1)/2}\Pi_{i<j}(w_i-w_j)e^{-\frac{1}{2}(\Sigma_1^k w_i)}$	$\{w_1>\cdots>w_k>0\}$
$\lvert\boldsymbol{S}_1-\ell\boldsymbol{S}_0\rvert=0$	$k_5\Pi_1^k \ell_i^{\frac{1}{2}(m-k-1)}\Pi_1^k(\ell_i+1)^{-(m+n)/2}\Pi_{i<j}(\ell_i-\ell_j)$	$\{\ell_1>\cdots>\ell_k>0\}$

Multivariate Statistical Distributions. Table 3 Standard *pdf*s for some continuous distributions on $\mathbb{R}^k$

Type	Density	Comments
$N_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_1\exp[-\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}]$	$\boldsymbol{Y}\in\mathbb{F}_{n\times k}$
$L_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_2\lvert\boldsymbol{\Gamma}\rvert^{-\frac{k}{2}}\lvert\boldsymbol{\Sigma}\rvert^{-\frac{n}{2}}\phi(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-\frac{1}{2}})$	$\boldsymbol{Y}\in\mathbb{F}_{n\times k},\phi\in\Phi$
$S_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_3\lvert\boldsymbol{\Gamma}\rvert^{-\frac{k}{2}}\lvert\boldsymbol{\Sigma}\rvert^{-\frac{n}{2}}\psi(\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1})$	ψ on $[0,\infty)$
Wishart	$\kappa_4\lvert\boldsymbol{W}\rvert^{(\nu-k-1)/2}\exp(-\frac{1}{2}\mathrm{tr}\boldsymbol{W}\boldsymbol{\Sigma}^{-1})$	$\boldsymbol{W}\in\mathbb{S}_k^+$
Gamma Hsu (l940)	$\kappa_5\lvert\boldsymbol{W}\rvert^{(n-k-1)/2}\phi(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Sigma}^{-\frac{1}{2}})$	$\phi\in\Phi,\boldsymbol{W}\in\mathbb{S}_k^+$
Gamma Lukacs and Laha (1964)	$\kappa_6\lvert\boldsymbol{W}\rvert^{\lambda-1}\exp(-\mathrm{tr}\boldsymbol{W}\boldsymbol{\Sigma}^{-1})$	$\lambda>0,\boldsymbol{W}\in\mathbb{S}_k^+$
Matric T	$\kappa_7\lvert\boldsymbol{I}_k-\nu^{-1}\boldsymbol{T}'\boldsymbol{T}\rvert^{-(\nu+r)/2}$	$\boldsymbol{T}\in\mathbb{F}_{r\times k}$
Dirichlet	$\kappa_8\Pi_1^k\lvert\boldsymbol{W}_j\rvert^{(\nu_j-k-1)/2}\lvert\boldsymbol{I}_k-\Sigma_1^k\boldsymbol{W}_j\rvert^{(\nu_0-k-1)/2}$	$f(\boldsymbol{W}_1,\ldots,\boldsymbol{W}_k)$
Inv. Dirichlet	$\kappa_9\Pi_1^k\lvert\boldsymbol{V}_j\rvert^{(\nu_j-k-1)/2}\lvert\boldsymbol{I}_k+\Sigma_1^k\boldsymbol{V}_j\rvert^{(\nu_T-k-1)/2}$	$f(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_k)$

exact in level for linear models having spherical errors (Jensen 1979). Similarly, if $\mathcal{L}(\boldsymbol{Y})=S_{n,k}(\boldsymbol{\Theta},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ with parameters $\boldsymbol{\Theta}=[\boldsymbol{\theta},\ldots,\boldsymbol{\theta}]'$, $\boldsymbol{\theta}\in\mathbb{R}^k$; if $X_j=n^{1/2}\bar{Y}_j$ with $\{\bar{Y}_j=(Y_{1j}+\cdots+Y_{nj})/n;1\le j\le k\}$; and if $\boldsymbol{S}$ is the sample dispersion matrix; then the central distribution of $\left\{T_j=X_j/S_{jj}^{\frac{1}{2}};1\le j\le k\right\}$ is Type II multivariate t for every $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma},\psi);\psi\in\Psi\}$. Noncentral distributions generally depend on the particular distribution $\mathcal{S}_n(\boldsymbol{\theta},\boldsymbol{\Gamma})$ or $S_{n,k}(\boldsymbol{\Theta},\boldsymbol{I}_n,\boldsymbol{\Sigma})$.

Matric T distributions. Let $\boldsymbol{Y}$ and $\boldsymbol{W}$ be independent, $\mathcal{L}(\boldsymbol{Y})=N_{r,k}(\boldsymbol{0},\boldsymbol{I}_r,\boldsymbol{\Sigma})$ and $\mathcal{L}(\boldsymbol{W})=W_k(\nu,\boldsymbol{\Sigma})$ such that $\nu\ge k$, and let $\boldsymbol{T}=\boldsymbol{Y}\boldsymbol{W}^{-\frac{1}{2}}$ using any factorization $\boldsymbol{W}^{-\frac{1}{2}}$ of $\boldsymbol{W}^{-1}$. Then $\mathcal{L}(\boldsymbol{T})$ is *matric t* with *pdf* as in Table 3. Alternatively, consider $\boldsymbol{X}'=[\boldsymbol{Y}',\boldsymbol{Z}']$ with distribution $S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ such that $n=r+\nu$ and $\nu\ge k$, and again let $\boldsymbol{T}=\boldsymbol{Y}\boldsymbol{W}^{-\frac{1}{2}}$ but now with $\boldsymbol{W}=\boldsymbol{Z}'\boldsymbol{Z}$. These variables arise from distributions $S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ in the same manner as for $N_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$. Then $\boldsymbol{T}$ has a matric t distribution for every distribution $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma},\psi);\psi\in\Psi\}$. This property transfers directly to $\mathcal{L}(\boldsymbol{ATB})$ as in Dickey (1967) with $\boldsymbol{A}$ and $\boldsymbol{B}$ nonsingular.

Gamma Distributions

Vector Distributions. Extract $\mathrm{Diag}(W_{11},\ldots,W_{kk})$ from $\boldsymbol{W}=[W_{ij}]$. Their joint distributions arise in the analysis of nonorthogonal designs, in time-series, in multiple comparisons, in the analysis of multidimensional contingency tables, in extensions of Friedman's χ^2 test in two-way data based on ranks, and elsewhere. There is a gamma distribution on $\mathbb{R}_+^k$ for diagonals of the matrix Gamma (Lukacs and Laha 1964) of Table 3; k–variate χ^2 when $\boldsymbol{W}$ is Wishart; see Kibble (1941) for $k=2$; and a k–variate exponential distribution for the case $n=2$. Rayleigh distributions $\mathcal{L}\left(W_{11}^{\frac{1}{2}},W_{22}^{\frac{1}{2}},\ldots,W_{kk}^{\frac{1}{2}}\right)$ on $\mathbb{R}_+^k$ support the detection of signals from noise (Miller 1975); more general such distributions are known (Jensen 1970a); as are more general χ^2 distributions on $\mathbb{R}^k$ having differing marginal degrees of freedom (Jensen 1970b). Densities here are typically intractable, often admitting multiple series expansions in special functions. Details are given in Kotz et al. (2000). As $n\to\infty$, the χ^2 and Rayleigh distributions on $\mathbb{R}_+^k$ are multinormal in the limit, for central and noncentral cases alike, whereas for fixed n, the limits as noncentrality parameters

grow again are multivariate normal (Jensen 1969). Alternative approximations, through normalizing Wilson-Hilferty transformations, are given in Jensen (1976) and Jensen and Solomon (1994).

Matric distributions. Let $\mathcal{L}(\boldsymbol{Y}) \in L_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi)$ with $n \geq k$; the *pdf* of $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$ is given in Table 3 under Gamma (Hsu 1940) as in that reference. The *pdf* under Gamma (Lukacs and Laha 1964), with $\lambda > 0$, reduces to that of a scaled Wishart matrix when 2λ is an integer. The noncentral Wishart *pdf* with $\boldsymbol{\Lambda} \neq \boldsymbol{0}$ admits series expansions in special polynomials. Moreover, as $n \to \infty$, for fixed $\boldsymbol{\Lambda}$ its limit distribution is multinormal, and for fixed n, its ▶asymptotic normality attains as the noncentrality parameters grow in a specified manner (Jensen 1976). Wishart matrices arise in matrix normal samples, e.g., as scaled sample dispersion matrices, and otherwise throughout multivariate distribution theory. Parallel remarks apply for Gamma (Hsu 1940) of Table 3 when the underlying observational model belongs to $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi); \phi \in \Phi\}$.

Dirichlet Distributions

If X and Y are independent gamma variates having a common scale, then $U = X/(X + Y)$ and $V = X/Y$ have *beta* and *inverted beta* distributions, respectively, the scaled Snedecor-Fisher F specializing from the latter. This section treats vector and matrix versions of these.

Vector distributions. Let $\{Z_0, \ldots, Z_k\}$ be independent gamma variates with common scale and the shape parameters $\{\alpha_0, \ldots, \alpha_k\}$, and let $T = (Z_0 + \cdots + Z_k)$. Then the joint distribution of $\{U_j = Z_j/T; 1 \leq j \leq k\}$ is the k-dimensional *Dirichlet distribution* $D(\alpha_0, \ldots, \alpha_k)$ with *pdf* as given in Table 2. An important special case is that $\{\alpha_j = \nu_j/2; 0 \leq j \leq k\}$ with $\{\nu_0, \ldots, \nu_k\}$ as positive integers and with $\{Z_0, \ldots, Z_k\}$ as independent χ^2 variates. However, in this case neither χ^2 nor independence is required. For if $\boldsymbol{y} = \left[\boldsymbol{y}_0', \boldsymbol{y}_1', \ldots, \boldsymbol{y}_k'\right]' \in \mathbb{R}^n$ with $\{\boldsymbol{y}_j \in \mathbb{R}^{\nu_j}; 0 \leq j \leq k\}$ and $n = \nu_0 + \cdots + \nu_k$ such that $\mathcal{L}(\boldsymbol{y}) = \mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, then $\left\{U_j = \boldsymbol{y}_j'\boldsymbol{y}_j/T; 1 \leq j \leq k\right\}$, but now with $T = \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{y}_1'\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_k'\boldsymbol{y}_k$, has the distribution $D(\nu_0/2, \nu_1/2, \ldots, \nu_k/2)$ for all such $\mathcal{L}(\boldsymbol{y}) \in \{\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$.

The *inverted* Dirichlet is that of $\{V_j = Z_j/Z_0; 1 \leq j \leq k\}$, with $\{Z_0, \ldots, Z_k\}$ as before, having *pdf* as listed in Table 2. The scaled $\{V_j = \nu_0 Z_j/\nu_j Z_0; 1 \leq j \leq k\}$ then have a *multivariate F distribution* whenever $\{\alpha_j = \nu_j/2; 0 \leq j \leq k\}$ with $\{\nu_0, \ldots, \nu_k\}$ as positive integers. This arises in the ▶analysis of variance in conjunction with ratios of independent mean squares to a common denominator (Finney 1941). As before, neither χ^2 nor independence is required in the latter; take $\left\{V_j = \nu_0 \boldsymbol{y}_j'\boldsymbol{y}_j/\nu_j \boldsymbol{y}_0'\boldsymbol{y}_0; 1 \leq j \leq k\right\}$ with $\mathcal{L}(\boldsymbol{y}) \in \{\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ as for Dirichlet distributions.

Matric distributions. Take $\{\boldsymbol{S}_0, \ldots, \boldsymbol{S}_k\}$ in $\mathbb{S}_k^+$ as independent Wishart matrices with $\{\mathcal{L}(\boldsymbol{S}_j) = W_k(\nu_j, \boldsymbol{\Sigma}); \nu_j \geq k; 0 \leq j \leq k\}$. Let $\boldsymbol{T} = \boldsymbol{S}_0 + \cdots + \boldsymbol{S}_k$ and $\left\{\boldsymbol{W}_j = \boldsymbol{T}^{-\frac{1}{2}}\boldsymbol{S}_j\boldsymbol{T}^{-\frac{1}{2}}; 1 \leq j \leq k\right\}$. A matric Dirichlet distribution (Olkin and Rubin 1964), taking the lower triangular square root, has *pdf* as listed in Table 3, such that $\boldsymbol{W}_j$ and $\left(\boldsymbol{I}_k - \Sigma_1^k \boldsymbol{W}_j\right)$ are positive definite, and $\nu_T = \nu_0 + \cdots + \nu_k$. Neither independence nor the Wishart character is required. If instead $\boldsymbol{Y} = \left[\boldsymbol{Y}_0', \boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_k'\right] \in \mathbb{F}_{n \times k}$, $n = \nu_0 + \cdots + \nu_k$, $\nu_j \geq k$, and $\left\{\boldsymbol{S}_j = \boldsymbol{Y}_j'\boldsymbol{Y}_j; j = 0, 1, \ldots, k\right\}$, then for $\mathcal{L}(\boldsymbol{Y}) = S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$, invariance properties assure that $f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k)$ is identical to that given in Table 3, for every distribution $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$.

An *inverted* matric Dirichlet distribution (Olkin and Rubin 1964) takes $\{\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_k\}$ as before, and defines $\left\{\boldsymbol{V}_j = \boldsymbol{S}_0^{-\frac{1}{2}}\boldsymbol{S}_j\boldsymbol{S}_0^{-\frac{1}{2}}; 1 \leq j \leq k\right\}$ using the symmetric root of $\boldsymbol{S}_0$. Its *pdf* $f(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k)$ is known allowing $\boldsymbol{S}_0$ to be noncentral. The central *pdf* is given in Table 3. The special case $k = 1$ is sometimes called a *Type II multivariate beta distribution*. Again neither independence nor the Wishart character is required. To see this, again take $\left\{\boldsymbol{S}_j = \boldsymbol{Y}_j'\boldsymbol{Y}_j; 0 \leq j \leq k\right\}$ as for matric Dirichlet distributions, and conclude that $f(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k)$, as in Table 3, is identical for every $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$.

Distributions of Latent Roots

Topics in multivariate statistics, to include reduction by invariance, tests for hypotheses regarding dispersion parameters, and the study of energy levels in physical systems, all entail the latent roots of random matrices. Suppose that $\mathcal{L}(\boldsymbol{W}) = W_k(\nu, \boldsymbol{\Sigma})$, and consider the ordered roots $\{w_1 > \ldots > w_k > 0\}$ of $|\ \boldsymbol{W} - w\boldsymbol{\Sigma}\ | = 0$. Their joint *pdf* is listed in Table 2. On occasion ratios of these roots are required, including simultaneous inferences for dispersion parameters, for which invariance in distribution holds. For if $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$, then the joint distributions of ratios of the roots of $|\ \boldsymbol{W} - w\boldsymbol{\Sigma}\ | = 0$ are identical for all $\mathcal{L}(\boldsymbol{Y}) \in \{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$ such that $n \geq k$.

To continue, consider $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ as independent Wishart matrices having $W_k(\nu_0, \boldsymbol{\Sigma})$ and $W_k(\nu_1, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, respectively. Then central ($\boldsymbol{\Lambda} = \boldsymbol{0}$) and noncentral joint distributions of the roots of $|\ \boldsymbol{S}_1 - \ell\boldsymbol{S}_0\ | = 0$ are known, as given in Table 2 for the case $\boldsymbol{\Lambda} = \boldsymbol{0}$. An invariance result holds for the central case. For if $\boldsymbol{Y} = \left[\boldsymbol{Y}_0', \boldsymbol{Y}_1'\right]'$ with $n = \nu_0 + \nu_1$ such that $\nu_0 \geq k$ and $\nu_1 \geq k$, $\boldsymbol{S}_0 = \boldsymbol{Y}_0'\boldsymbol{Y}_0$ and $\boldsymbol{S}_1 = \boldsymbol{Y}_1'\boldsymbol{Y}_1$, then by invariance the latent root *pdf* $f(\ell_1, \ldots, \ell_k)$ is the same for all $\mathcal{L}(\boldsymbol{Y})$ in $\{L_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi) : \phi \in \Phi\}$, as given in Table 3.

Multivariate Statistical Distributions. Table 4 Some discrete multivariate compound distributions

Basic distribution	Mixing parameters	Compounding distribution	Source	Resulting distribution
Bivariate binomial $(n, \pi_{01}, \pi_{10}, \pi_{11})$	n	Poisson	Papageorgiou (1983)	bivariate Poisson
Multinomial $(n, \pi_1, \ldots, \pi_s)$	$(\pi_1, \ldots, \pi_s)$	Dirichlet	Johnson et al. (1997) and Patil and Joshi (1968)	s–variate negative hypergeometric
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Logarithmic series	Patil and Joshi (1968)	s–variate modified logarithmic series
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Negative binomial	Patil and Joshi (1968)	s–variate negative multinomial
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Poisson	Patil and Joshi (1968)	multiple Poisson
Multiple Poisson $(u\lambda_1, \ldots, u\lambda_s)$	u	Gamma	Patil and Joshi (1968)	s–variate negative multinomial
Multiple Poisson $(\lambda_1, \ldots, \lambda_s)$	$(\lambda_1, \ldots, \lambda_s)$	Multinormal	Steyn (1976)	s–variate Poisson–normal
Multiple Poisson $\{\lambda_i = \alpha + (\beta - \alpha)u$	u	Rectangular on (0,1)	Patil and Joshi (1968)	s–variate Poisson–rectangular
Multivariate Poisson $(u\lambda_1, u\lambda_{12}, \ldots, u\lambda_{12\cdot s})$	u	Gamma	Patil and Joshi (1968)	s–variate negative binomial
Negative multinomial $(k, \pi_1, \ldots, \pi_s)$	$(\pi_1, \ldots, \pi_s)$	Dirichlet	Johnson et al. (1997) Patil and Joshi (1968)	s–variate negative multinomial-Dirichlet
Convolution of multinomials $(\gamma_1, \ldots, \gamma_{2^k}, \theta_1, \ldots, \theta_s)$	$(\gamma_1, \ldots, \gamma_{2^k})$	Multivariate hypergeometric	Kotz and Johnson (1983)	numbers judged defective of k types in lot inspection

Other Distributions

Numerous other continuous multivariate distributions are known; a compendium is offered in Kotz et al. (2000). Multivariate versions of *Burr distributions* arise through gamma mixtures of independent Weibull distributions. Various *multivariate exponential distributions* are known; some properties and examples are found on specializing multivariate Weibull distributions. Various *multivariate stable distributions*, symmetric and asymmetric, are characterized through the structure of their *chf*s, as are types of symmetric MDs surveyed earlier. *Multivariate extreme-value distributions* are treated in Kotz et al. (2000), with emphasis on the bivariate case. The *Beta-Stacy distributions* yield a *multivariate Weibull distribution* as a special case. *Multivariate Pareto distributions* have their origins in econometrics. *Multivariate logistic distributions* model binary data in the analysis of quantal responses. Properties of *chf*s support a bivariate distribution having normal and gamma marginals (Kibble 1941).

Discrete Distributions

A guided tour is given with special reference to Johnson et al. (1997) and Patil and Joshi (1968). Inequalities for selected multivariate discrete distributions are offered in Jogdeo and Patil (1975).

Binomial, Multinomial, and Related

The outcome of a random experiment is classified as having or not having each of s attributes $\{A_1, \ldots, A_s\}$. If $\{X_1, \ldots, X_s\}$ are the numbers having these attributes in n independent trials, then theirs is a *multivariate binomial distribution* with parameters

$$\{\pi_i = \Pr(A_i), \pi_{ij} = \Pr(A_iA_j), \ldots, \pi_{12\cdot s} = \Pr(A_1A_2 \ldots A_s); i \in [1, 2, \ldots s]; i \neq j \neq k \neq \ldots\}$$

where ι takes successive values $\{i, j, k, \ldots\}$. The ►*binomial* ►distribution $B(n, \pi)$ obtains at $s = 1$. For bivariate binomial distributions see Hamdan (1972), Hamdan and Al-Bayyati (1971), and Hamdan and Jensen (1976). The limit as $n \to \infty$ and $\pi \to 0$ such that $n\pi \to \lambda$ is Poisson, the distribution of "rare events". More generally, as $n \to \infty$ and $\pi_i \to 0$, such that $\{n\theta_i \to \lambda_i, n\theta_{ij} \to \lambda_{ij}, \ldots, n\pi_{12\cdot s} \to \lambda_{12\cdot s}\}$, where $\{\theta_i, \theta_{ij}, \ldots\}$ are specified functions of $\{\pi_i, \pi_{ij}, \ldots\}$, then the limit of the multivariate binomial distribution is *multivariate Poisson*.

Suppose that independent trials are continued until exactly k trials exhibit none of the s attributes. The joint distribution of the numbers $\{Y_1, \ldots, Y_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ during these trials is a *multivariate Pascal distribution*.

To continue, let $\{A_0, \ldots, A_s\}$ be exclusive and exhaustive outcomes having probabilities $\{\pi_0, \ldots, \pi_s\}$, with $\{0 < \pi_i < 1; \pi_0 + \cdots + \pi_s = 1\}$. The numbers $\{X_1, \ldots, X_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ in n independent trials have the ►*multinomial distribution* with parameters $(n, \pi_1, \ldots, \pi_s)$. If independent trials are repeated until A_0 occurs exactly k times, the numbers of occurrences of $\{A_1, \ldots, A_s\}$ during these trials have a *negative multinomial distribution* with parameters $(k, \pi_1, \ldots, \pi_s)$.

In a multiway contingency table an outcome is classified according each of k criteria having the exclusive and exhaustive classes $\{A_{io}, A_{i1}, \ldots A_{is_i}; i = 1, \ldots, k\}$. If in n independent trials $\{X_{i1}, \ldots, X_{is_i}; i = 1, \ldots, k\}$ are the numbers occurring in $\{A_{i1}, \ldots, A_{is_i}; i = 1, \ldots, k\}$, then their joint distribution is called a *multivariate multinomial distribution* (also multivector multinomial). These are the joint distributions of marginal sums of the contingency table, to include the k–variate binomial distribution when $\{s_1 = s_2 = \cdots = s_k = 1\}$.

Hypergeometric and Related

A collection of N items consists of $s + 1$ types: N_0 of type A_0, N_1 of type A_1, …, N_s of type A_s, with $N = N_0 + \cdots + N_s$. Random samples are taken from this collection. If n items are drawn without replacement, the joint distribution of the numbers of items of types $\{A_1, \ldots, A_s\}$ is a *multivariate hypergeometric distribution* with parameters $(n, N, N_1, \ldots, N_s)$. With replacement, their distribution is multinomial with parameters $(n, N_1/N, \ldots, N_s/N)$.

If successive items are drawn without replacement until exactly k items of type A_0 are drawn, then the numbers of types $\{A_1, \ldots, A_s\}$ thus drawn have a *multivariate inverse hypergeometric distribution* with parameters $(k, N, N_1, \ldots, N_s)$.

To continue, sampling proceeds in two stages. First, m items are drawn without replacement, giving $\{x_1, \ldots, x_s\}$ items of types $\{A_1, \ldots, A_s\}$. Without replacing the first sample, n additional items are drawn without replacement at the second stage, giving $\{Y_1, \ldots, Y_s\}$ items of types $\{A_1, \ldots, A_s\}$. The conditional distribution of $(Y_1, \ldots, Y_s)$, given that $\{X_1 = x_1, \ldots, X_s = x_s\}$, is a *multivariate negative hypergeometric distribution.*

Multivariate Series Distributions

Further classes of discrete multivariate distributions are identified by types of their *pmf*s. Some arise through truncation and limits. If $[X_1, \ldots, X_s]$ has the s–variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$, then the conditional distribution of $[X_1, \ldots, X_s]$, given that $[X_1, \ldots, X_s] \neq [0, \ldots, 0]$, converges as $k \to 0$ to the s–variate *logarithmic series distribution* with parameters $(\theta_1, \ldots, \theta_s)$ where $\{\theta_i = 1 - \pi_i; i = 1, \ldots, s\}$. See Patil and Joshi (1968) for details. A modified multivariate logarithmic series distribution arises as a mixture, on n, of the multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$, where the mixing distribution is a logarithmic series distribution (Patil and Joshi 1968).

A class of distributions with parameters $(\theta_1, \ldots, \theta_s) \in \Theta$, derived from convergent power series, has *pmf*s of the form $p(x_1, \ldots, x_s) = \frac{a(x_1, \ldots, x_s)\theta_1^{x_1} \cdots \theta_s^{x_s}}{f(\theta_1, \ldots, \theta_s)}$ for $\{x_i = 0, 1, 2, \ldots; i = 1, \ldots, s\}$. The class of such distributions, called *multivariate power series distributions,* contains the s–variate multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$; the s-variate logarithmic series distribution with parameters $(\theta_1, \ldots, \theta_s)$; the s-variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$; and others. See Patil and Joshi (1968) for further properties. Other discrete multivariate distributions are described next.

Other Distributions

A typical *Borel-Tanner* distribution refers to the number of customers served before a queue vanishes for the first time. If service in a single-server queue begins with r customers of type I and s of type II with different arrival rates and service needs for each type, then the joint distribution of the numbers served is the *bivariate Borel-Tanner* distribution as in Shenton and Consul (1973).

In practice *compound distributions* often arise from an experiment undertaken in a random environment; the compounding distribution then describes variation of parameters of the model over environments. Numerous bivariate and multivariate discrete distributions have been obtained through compounding, typically motivated by the structure of the problem at hand. Numerous examples are cataloged in references Johnson et al. (1997) and Patil

and Joshi (1968); examples are listed in Table 4 from those and other sources.

About the Author

Donald Jensen received his Ph.D. from Iowa State University in 1962, and joined Virginia Polytechnic Institute and State University in 1965, attaining the rank of Professor in 1973. He has published over 140 journal articles in distribution theory, multivariate analysis, linear inference, robustness, outlier detection and influence diagnostics, regression design, and quality control. Dr. Jensen served as Associate editor of *The American Statistician* for a decade (1971–1980), and has been a reviewer for Mathematical Reviews for the last 30 years. He is an elected member of the International Statistical Institute. Professor Jensen received an early five-year Research Career Development Award from the US National Institutes of Health.

Cross References

►Binomial Distribution
►Bivariate Distributions
►Gamma Distribution
►Hypergeometric Distribution and Its Application in Statistics
►Multinomial Distribution
►Multivariate Normal Distributions
►Multivariate Statistical Analysis
►Multivariate Statistical Simulation
►Multivariate Technique: Robustness
►Poisson Distribution and Its Application in Statistics
►Statistical Distributions: An Overview
►Student's *t*-Distribution
►Weibull Distribution

References and Further Reading

Adrian R (1808) Research concerning the probabilities of errors which happen in making observations, etc. Analyst Math 1: 93–109
Arnold BC, Beaver RJ (2000) Some skewed multivariate distributions. Am J Math Manage Sci 20:27–38
Bedrick EJ, Lapidus J, Powell JF (2000) Estimating the Mahalanobis distance from mixed continuous and discrete data. Biometrics 56:394–401
Bhattacharya RN, Ranga Rao R (1976) Normal apprximations and asymptotic expansions. Wiley, New York
Birnbaum ZW (1948) On random variables with comparable peakedness. Ann Math Stat 19:76–81
Bravais A (1846) Analyse mathématique sur les probabilites des erreurs de situation d'un point. Mémoires Présentés par Divers Savants a l'Académie Royale des Sciences de l'Institut de France, Paris 9:255–332
Cambanis S, Huang S, Simons G (1981) On the theory of elliptically contoured distributions. J Multivariate Anal 11:368–385
Chmielewski MA (1981) Elliptically symmetric distributions: a review and bibliography. Int Stat Rev 49:67–74 (Excellent survey article on elliptical distributions)
Dawid AP (1977) Spherical matrix distributions and a multivariate model. J Roy Stat Soc B 39:254–261 (Technical source paper on the strucure of distributions)
Dempster AP (1969) Elements of continuous multivariate analysis. Addison-Wesley, London (General reference featuring a geometric approach)
Devlin SJ, Gnanadesikan R, Kettenring JR (1976) Some multivariate applications of elliptical distributions. In: Ikeda S et al (eds) Essays in probability and statistics. Shinko Tsusho, Tokyo, pp 365–394 (Excellent survey article on ellipsoidal distributions)
Dharmadhikari S, Joag-Dev K (1988) Unimodality, convexity, and applications. Academic, New York
Dickey JM (1967) Matrix variate generalizations of the multivariate *t* distribution and the inverted multivariate t distribution. Ann Math Stat 38:511–518 (Source paper on matric *t* distributions and their applications)
Dickson IDH (1886) Appendix to "Family likeness in stature" by F. Galton. Proc Roy Soc Lond 40:63–73
Edgeworth FY (1892) Correlated averages. Philos Mag 5 34:190–204
Epstein B (1948) Some applications of the Mellin transform in statistics. Ann Math Stat 19:370–379
Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman & Hall, New York
Fang KT, Anderson TW (eds) (1990) Statistical inference in elliptically contoured and related distributions. Allerton, New York
Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman & Hall, London
Fang KT, Zhang YT (1990) Generalized multivariate analysis. Springer, New York
Fefferman C, Jodeit M, Perlman MD (1972) A spherical surface measure inequality for convex sets. Proc Am Math Soc 33: 114–119
Finney DJ (1941) The joint distribution of variance ratios based on a common error mean square. Ann Eugenic 11:136–140 (Source paper on dependent *F* ratios in the analysis of variance)
Galton F (1889) Natural inheritance. MacMillan, London, pp 134–145
Gauss CF (1823) Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. Muster-Schmidt, Göttingen
Hamdan MA (1972) Canonical expansion of the bivariate binomial distribution with unequal marginal indices. Int Stat Rev 40: 277–280 (Source paper on bivariate binomial distributions)
Hamdan MA, Al-Bayyati HA (1971) Canonical expansion of the compound correlated bivariate Poisson distribution. J Am Stat Assoc 66:390–393 (Source paper on a compound bivariate Poisson distribution)
Hamdan MA, Jensen DR (1976) A bivariate binomial distribution and some applications. Aust J Stat 18:163–169 (Source paper on bivariate binomial distributions)
Helmert FR (1868) Studien über rationelle Vermessungen, im Gebeite der höheren Geodäsie. Zeitschrift für Mathematik und Physik 13:73–129
Hsu PL (1940) An algebraic derivation of the distribution of rectangular coordinates. Proc Edinburgh Math Soc 2 6:185–189 (Source paper on generalizations of Wishart's distribution)
James AT (1954) Normal multivariate analysis and the orthogonal group. Ann Math Stat 25:40–75
Jensen DR (1969) Limit properties of noncentral multivariate Rayleigh and chi-square distributions. SIAM J Appl Math

17:807–814 (Source paper on limits of certain noncentral distributions)
Jensen DR (1970a) A generalization of the multivariate Rayleigh distribution. Sankhya A 32:192–208 (Source paper on generalizations of Rayleigh distributions)
Jensen DR (1970b) The joint distribution of traces of Wishart matrices and some applications. Ann Math Stat 41:133–145 (Source paper on multivariate chi–squared and F distributions)
Jensen DR (l972) The limiting form of the noncentral Wishart distribution. Aust J Stat 14:10–16 (Source paper on limits of noncentral Wishart distributions)
Jensen DR (1976) Gaussian approximation to bivariate Rayleigh distributions. J Stat Comput Sim 4:259–268 (Source paper on normalizing bivariate transformations)
Jensen DR (1979) Linear models without moments. Biometrika 66:611–617 (Source paper on linear models under symmetric errors)
Jensen DR (1984) Ordering ellipsoidal measures: scale and peakedness orderings. SIAM J Appl Math 44:1226–1231
Jensen DR, Good IJ (1981) Invariant distributions associated with matrix laws under structural symmetry. J Roy Stat Soc B 43: 327–332 (Source paper on invariance of derived distributions under symmetry)
Jensen DR, Solomon H (1994) Approximations to joint distributions of definite quadratic forms. J Am Stat Assoc 89:480–486
Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall/CRC, Boca Raton
Jogdeo K, Patil GP (1975) Probability inequalities for certain multivariate discrete distributions. Sankhya B 37:158–164 (Source paper on probability inequalities for discrete multivariate distributions)
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York (An excellent primary source wlth extensive bibliography)
Kagan AM, Linnik YV, Rao CR (1973) Characterization problems in mathematical statistics. Wiley, New York
Kariya T, Sinha BK (1989) Robustness of statistical tests. Academic, New York
Kibble WF (1941) A two-variate gamma type distribution. Sankhya 5:137–150 (Source paper on expansions of bivariate distributions)
Kotz S, Balakrishnan N, Johnson NL (2000) Continuous multivariate distributions, 2nd edn. Wiley, New York (An excellent primary source with extensive bibliography)
Kotz S, Johnson NL (1983) Some distributions arising from faulty inspection with multitype defectives, and an application to grading. Commun Stat A Theo Meth 12:2809–2821
Kotz S, Nadarajah S (2004) Multivariate t distributions and their applications. Cambridge University Press, Cambridge
Laplace PS (1811) Memoir sur les integrales definies et leur application aux probabilites. Memoires de la classes des Sciences Mathématiques et Physiques l'Institut Impérial de France Année 1810:279–347
Lindsay BG (1995) Mixture models: theory, geometry and applications. NSF–CBMS regional conference series in probability and statistics, vol 5. Institute of Mathematical Statistics, Hayward
Lukacs E, Laha RG (1964) Applications of characteristic functions. Hafner, New York (Excellent reference with emphasis on multivariate distributions)
McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
Miller KS (1975) Multivariate distributions. Krieger, Huntington (An excellent reference with emphasis on problems in engineering and communications theory)
Nelsen R (1998) An introduction to copulas. Springer, New York
Olkin I, Rubin H (1964) Multivariate beta distributions and independence properties of the Wishart distribution. Ann Math Stat 35:261–269; Correction, 37:297 (Source paper on matric Dirichlet, beta, inverted beta, and related distributions)
Olkin I, Tate RF (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465; Correction 36:343–344
Papageorgiou H (1983) On characterizing some bivariate discrete distributions. Aust J Stat 25:136–144
Patil GP, Joshi SW (1968) A dictionary and bibliography of discrete distributions. Hafner, New York (An excellent primary source with extensive bibliography)
Pearson K (1896) Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philos Trans Roy Soc Lond A 187:253–318
Plana GAA (1813) Mémoire sur divers problémes de probabilitié. Mémoires de l'Académie Impériale de Turin 20:355–408
Schols CM (1875) Over de theorie der fouten in de ruimte en in het platte vlak. Verh Nederland Akademie Wetensch 15:1–75
Shaked M, Shanthikumar JG (2007) Stochastic orders. Springer, New York
Shenton LR, Consul PC (1973) On bivariate Lagrange and Borel–Tanner distributions and their use in queueing theory. Sankhya A 35:229–236 (Source paper on bivariate Lagrange and Borel–Tanner distributions and their applications)
Sherman S (1904) A theorem on convex sets with applications. Ann Math Stat 25:763–766
Spearman C (1904) The proof and measurement of associaton between two things. Am J Psychol 15:72–101
Steyn HS (1976) On the multivariate Poisson normal distribution. J Am Stat Assoc 71:233–236 (Source paper on multivariate Poisson–normal distributions)
Student (1908) The probable error of a mean. Biometrika 6:1–25
Subrahmaniam K (1970) On some applications of Mellin transformations to statistics: dependent random variables. SIAM J Appl Math 19:658–662
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York
Tong YL (1980) Probability inequalities in multivariate distributions. Academic, New York
Tong YL (1990) The multivariate normal distribution. Springer-Verlag, New York

Multivariate Statistical Process Control

Robert L. Mason[1], John C. Young[2]
[1]Southwest Research Institute, San Antonio, TX, USA
[2]Lake Charles, LA, USA

Statistical process control (SPC) includes the use of statistical techniques and tools, such as ▶control charts, to

monitor change in a process. These are typically applied separately to each process variable of interest. Statistical process control procedures help provide an answer to the question: "Is the process in control?" When an out-of-control event is identified as a signal in a control chart, procedures often are available for locating the specific process variables that are the cause of the problem.

In multivariate statistical process control (MVSPC), multivariate statistical control procedures are used to simultaneously monitor many process variables that are interrelated and form a correlated set that move together (see Mason and Young 2002). The relationships that exist between and among the variables of the multivariate process are used in developing the procedure. Assume that the observation vectors obtained from a process are independent random variables that can be described by a multivariate normal distribution (see ►Multivariate Normal Distributions) with a mean vector and a covariance matrix. Any change in the mean vector and/or the covariance matrix of this distribution is considered an out-of-control situation and should be detectible with an appropriate multivariate control chart.

Implementation of a multivariate control procedure is usually divided into two parts: Phase I and Phase II. Phase I includes the planning, development, and construction phase. In this phase, the practitioner studies the process in great detail. Preliminary data are collected under good operational conditions and examined for statistical control and other potential problems. The major problems include statistical ►outliers, variable collinearities, and autocorrelated observations, i.e., time-dependent observations. After statistical control of the preliminary data is established, the data is used as the process history and referred to as the historical data set (HDS). If the parameters of the process are unknown, parameter estimates of the mean vector and covariance matrix are obtained from the data of the HDS for use in monitoring the process.

Phase II is the monitoring stage. In this phase, new observations are examined in order to determine if the process has deviated from the in-control situation specified by the HDS. Note that, in MVSPC, deviations from the HDS can occur through a mean vector change, a covariance matrix change, or both a mean vector and covariance matrix change in the process. In certain situations a change in one parameter can also induce a change in the other parameter.

Process control is usually determined by examining a control statistic based on the observed value of an individual observation and/or a statistic related to a rational subgroup (i.e., sample) of the observations such as the sample mean. Easy monitoring is accomplished by charting the value of the multivariate control statistic on a univariate chart. Depending on the charted value of this statistic, one can determine if control is being maintained or if the process has moved to an out-of-control situation.

For detecting both large and small shifts in the mean vector, there are three popular multivariate control chart methods. An implicit assumption when using these charts is that the underlying population covariance matrix is constant over the time period of interest. Various forms of ►Hotelling's T^2 statistic are generally chosen when the detection of large mean shifts is of interest (e.g., see Mason and Young 2002). For detecting small shifts in the process mean, the multivariate exponential weighted moving average (MEWMA) statistic (e.g., see Lowry et al. 1992) or the multivariate cumulative sum (MCUSUM) statistic (e.g., Woodall and Ncube 1985) can be utilized. These statistics each have advantages and disadvantages, and they can be used together or separately.

All of the above procedures were developed under the assumption that the data are independent and follow a multivariate normal distribution. Autocorrelated data can present a serious problem for both the MCUSUM and MEWMA statistics, but seems to have lesser influence on the behavior of the T^2 statistic. A main reason for the influence of autocorrelation on the MEWMA and MCUSUM statistics is that both of them are dependent on a subset of past-observed observation vectors, whereas the T^2 statistic depends only on the present observation.

A related problem in MVSPC is monitoring shifts in the covariance matrix for a multivariate normal process when the mean vector is assumed to be stable. A useful review of procedures for monitoring multivariate process variability is contained in Yeh et al. (2006). The methods for detecting large shifts in the covariance matrix include charts based on the determinant of the sample covariance matrix (Djauhari 2005), while the methods for detecting small shifts include charts based on a likelihood-ratio EWMA statistic (Yeh et al. 2004) and on related EWMA-type statistics (Yeh et al. 2003). A recent charting method that is applicable in monitoring the change in covariance matrix for a multivariate normal process is based on a form of Wilks' ratio statistic (Wilks 1963). It consists of taking the ratio of the determinants of two estimators of the process covariance matrix (Mason et al. 2009). One estimator is obtained using the HDS and the other estimator is computed using an augmented data set consisting of the newest observed sample and the HDS. The Wilks' chart statistic is particularly helpful when the number of variables is large relative to the sample size.

Current attention in the MVSPC literature is focused on procedures that simultaneously monitor both the mean vector and the covariance matrix in a multivariate process (e.g., see Reynolds and Cho 2006 or Chen et al. 2005). These charts are based on EWMA procedures and can be very useful in detecting small-to-moderate changes in a process. Several papers also exist that present useful overviews of MVSPC (e.g., see Woodall and Montgomery 1999 and Bersimis et al. 2007). These papers are valuable for their insights on the subject and their extensive reference lists.

About the Authors

Dr. Robert L. Mason is an Institute Analyst at Southwest Research Institute in San Antonio, Texas. He was President of the American Statistical Association in 2003, Vice-President in 1992–1994, and a Member of its Board of Directors in 1987–1989. He is a Fellow of both the American Statistical Association and the American Society for Quality, and an Elected Member of the International Statistical Institute. He has been awarded the Founder's Award and the Don Owen Award from the American Statistical Association and the W.J. Youden Award (twice) from the American Society for Quality. He is on the Editorial Board of the *Journal of Quality Technology*, and is an Associate Editor of *Communications in Statistics*. He has published over 130 research papers and coauthored 6 textbooks including *Statistical Design and Analysis of Experiments with Applications to Engineering and Science* (Wiley, 1989; 2nd ed. 2003). He also is the coauthor (with John C. Young) of *Multivariate Statistical Process Control with Industrial Applications* (ASA-SIAM; 2002).

Prior to his retirement in 2007, Dr. John C. Young was Professor of Statistics for 40 years at McNeese State University in Lake Charles, Louisiana. He has published approximately 100 papers in the statistical, medical, chemical, and environmental literature, and is coauthor of numerous book chapters and three textbooks.

Cross References

►Control Charts
►Hotelling's T^2 Statistic
►Multivariate Normal Distributions
►Outliers
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Bersimis S, Psarakis S, Panaretos J (2007) Multivariate statistical process control charts: an overview. Qual Reliab Eng Int 23:517–543

Chen G, Cheng SW, Xie H (2005) A new multivariate control chart for monitoring both location and dispersion. Commun Stat Simulat 34:203–218

Djauhari MA (2005) Improved monitoring of multivariate process variability. J Qual Technol 37:32–39

Lowry CA, Woodall WH, Champ CW, Rigdon SE (1992) A multivariate exponentially weighted moving average control chart. Technometrics 34:46–53

Mason RL, Young JC (2002) Multivariate statistical process control with industrial applications. ASA-SIAM, Philadelphia, PA

Mason RL, Chou YM, Young JC (2009) Monitoring variation in a multivariate process when the dimension is large relative to the sample size. Commun Stat Theory 38:939–951

Reynolds MR, Cho GY (2006) Multivariate control charts for monitoring the mean vector and covariance matrix. J Qual Technol 38:230–253

Wilks SS (1963) Multivariate statistical outliers. Sankhya A 25: 407–426

Woodall WH, Montgomery DC (1999) Research issues and ideas in statistical process control. J Qual Technol 31:376–386

Woodall WH, Ncube MM (1985) Multivariate CUSUM quality control procedures. Technometrics 27:285–292

Yeh AB, Lin DK, Zhou H, Venkataramani C (2003) A multivariate exponentially weighted moving average control chart for monitoring process variability. J Appl Stat 30: 507–536

Yeh AB, Huwang L, Wu YF (2004) A likelihood-ratio-based EWMA control chart for monitoring variability of multivariate normal processes. IIE Trans 36:865–879

Yeh AB, Lin DK, McGrath RN (2006) Multivariate control charts for monitoring covariance matrix: a review. Qual Technol Quant Manage 3:415–436

Multivariate Statistical Simulation

Mark E. Johnson
Professor
University of Central Florida, Orlando, FL, USA

Multivariate statistical simulation comprises the computer generation of multivariate probability distributions for use in statistical investigations. These investigations may be robustness studies, calibrations of small sample behavior of estimators or confidence intervals, power studies, or other Monte Carlo studies. The distributions to be generated may be continuous, discrete or a combination of both types. Assuming that the n-dimensional distributions have independent components, the problem of variate generation is reduced to simulating from univariate distributions for which, fortunately, there is a vast literature (Devroye 1986; L'Eculer 2010; and international standard ISO 28640, for

example). Thus, the real challenge of multivariate statistical simulation is in addressing the dependence structure of the multivariate distributions.

For a few situations, the dependence structure is readily accommodated from a generation standpoint. Consider the usual n-dimensional multivariate normal distribution (see ▶Multivariate Normal Distributions) with mean vector $\underline{\mu}$ and covariance matrix Σ. For a positive definite covariance matrix, there exists a lower triangular (Cholesky) decomposition $LL' = \Sigma$. Assuming a source of independent univariate normal variates to occupy the vector $\underline{X}$, the random vector $\underline{Y} = L\,\underline{X} + \underline{\mu}$ has the desired multivariate normal distribution. Having been able to generate multivariate normal random vectors, componentwise transformations provide the capability to generate the full Johnson translation system (1949a), of which the lognormal distribution may be the most familiar. In using the multivariate Johnson system, it is possible to specify the covariance matrix of the transformed distribution. Some researchers transform the multivariate normal distribution without noting the severe impact on the covariance matrix of the transformed distribution. This oversight makes it difficult to interpret the results of simulation studies involving the Johnson translation system (see Johnson 1987 for further elaboration).

In expanding to distributions beyond the Johnson translation system, it is natural to consider generalizations of the normal distribution at the core of this system. The exponential power distribution with density function $f(x)$ proportional to $\exp(-|x|^{\tau})$ is a natural starting point since it includes the double exponential distribution ($\tau = 1$), the normal distribution ($\tau = 2$) and the uniform distribution in the limit ($\tau \to \infty$) and is easy to simulate (Johnson 1979). A further generalization of the exponential power distribution amenable to variance reduction simulation designs was developed by Johnson, Beckman and Tietjen (1980) who noted that the normal distribution arises as the product of ZU where Z is distributed as the square root of a chi-squared(3) distribution and is independent of U which is uniform on the interval $(-1, 1)$. Their generalization occurs by considering arbitrary degrees of freedom and powers other than 0.5. Since by Khintchine's unimodality theorem, any unimodal distribution can be represented as such a product there are many possibilities that could be pursued for other constructions ultimately for use in multivariate simulation contexts.

Multivariate distribution *families* are appealing for simulation purposes. A useful extension of the Johnson translation system has been developed by Jones and Pewsey (2009). The family is defined implicitly via the equation

$$Z = \sinh[\delta \sinh^{-1}(X_{\varepsilon,\delta}) - \varepsilon]$$

where Z has the standard normal distribution, $X_{\delta,\varepsilon}$ has a sinh-arcsinh distribution, ε is a skewness parameter and δ relates to the tail weight of the distribution. This family of distributions is attractive for use in Monte Carlo studies, since it includes the normal distribution as a special intermediate (non-limiting) case and covers a variety of skewness and tailweight combinations. Extensions of the Jones-Pewsey family to the multivariate case can follow the approach originally taken by Johnson (1949b), with adaptations by Johnson et al. (1982) to better control impacts of the covariance structure and component distributions.

Variate generation for multivariate distributions is readily accomplished (at least, in principle) for a specific multivariate distribution provided certain conditional distributions are identified. Suppose $\underline{X}$ is a random vector to be generated. A direct algorithm is to first generate X_1 as the marginal distribution of the first component of $\underline{X}$, say x_1. Second, generate from the conditional distribution of X_2 given $X_1 = x_1$ to obtain x_2. Third, generate from the conditional distribution X_3 given, $X_1 = x_1$ and $X_2 = x_2$ and then continue until all n components have been generated. This conditional distribution approach converts the multivariate generation problem into a series of univariate generation problems. For cases in which the conditional distributions are very complicated or not particularly recognizable, there may be alternative formulae for generation, typically involving a transformation to $n+1$ or more independent random variables. Examples include a multivariate Cauchy distribution and the multivariate Burr-Pareto-logistic distributions (see Johnson 1987).

The general challenge in multivariate statistical simulation is to incorporate the dependence structure as it exists in a particular distribution. As noted earlier, the multivariate normal distribution is particularly convenient since dependence is introduced to independent normal components through appropriate linear transformations. Further transformations to the components of the multivariate normal distribution give rise to skewed, light tailed or heavy tailed marginal distributions while retaining some semblance of the dependence structure. An important approach to grappling with the dependence structure is to recognize that marginal distributions are not terribly relevant in that the components can be transformed to the uniform distribution via $U_i = F_i(X_i)$, where F_i is the distribution function of X_i. In other words, in comparing multivariate distributions, the focus can be on the transformed distribution having uniform marginal's. This multivariate distribution is known as a "copula." Examining the

►copulas associated with the Burr, Pareto and logistic distributions led Cook and Johnson to recognize the essential similarity of these three multivariate distributions. A very useful introduction to copulas is Nelsen (2006) while Genest and MacKay (1986) deserve credit for bringing copulas to the attention of the statistical community.

This entry does not cover all possible distributions or families of distributions that could be considered for use in multivariate simulation studies. Additional possibilities (most notably elliptically contoured distributions) are reviewed in Johnson (1987).

About the Author

For biography *see* the entry ►Statistical Aspects of Hurricane Modeling and Forecasting.

Cross References

►Copulas
►Monte Carlo Methods in Statistics
►Multivariate Normal Distributions
►Multivariate Statistical Distributions

References and Further Reading

Cook RD, Johnson ME (1981) A family of distributions for modelling non-elliptically symmetric multivariate data. Technometrics 28:123–131

Devroye L (1986) Non-uniform variate generation. Springer, New York. Available for free pdf download at http://cg.scs.carleton.ca/~luc/mbookindex.html

Genest C, MacKay RJ (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

International Standard 28640 (2010) Random variate generation methods. International Standards Organization (to appear), Geneva

Johnson ME (1987) Multivariate statistical simulation. Wiley, New York

Johnson ME (1979) Computer generation of the exponential power distribution. J Stat Comput Sim 9:239–240

Johnson ME, Beckman RJ, Tietjen GL (1980) A new family of probability distributions with applications to monte carlo studies. JASA 75:276–279

Johnson ME, Ramberg JS, Wang C (1982) The johnson translation system in monte carlo studies. Commun Stat Comput Sim 11:521–525

Johnson NL (1949a) Systems of frequency curves generated by methods of translation. Biometrika 36:149–176

Johnson NL (1949b) Bivariate distributions based on simple translation systems. Biometrika 36:297–304

Jones MC, Pewsey A (2009) Sinh-arcsinh distributions. Biometrika 96:761–780

L'Ecuyer P (2010) Non-uniform random variate generation. Encyclopedia of statistical science. Springer, New York

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Multivariate Techniques: Robustness

Mia Hubert[1], Peter J. Rousseeuw[2]
[1]Associate Professor
Katholieke Universiteit Leuven, Leuven, Belgium
[2]Senior Researcher
Renaissance Technologies, New York, NY, USA

The usual multivariate analysis techniques include location and scatter estimation, ►principal component analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems), ►canonical correlation analysis, multiple regression and cluster analysis (see ►Cluster Analysis: An Introduction). These methods all try to describe and discover structure in the data, and thus rely on the correlation structure between the variables. Classical procedures typically assume normality (i.e. gaussianity) and consequently use the sample mean and sample covariance matrix to estimate the true underlying model parameters.

Below are three examples of multivariate settings used to analyze a data set with n objects and p variables, forming an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ the ith observation.

1. ►Hotelling's T^2 statistic for inference about the center of the (normal) underlying distribution is based on the sample mean $\bar{\boldsymbol{x}} = \sum_{i=1}^{n} \boldsymbol{x}_i$ and the sample covariance matrix $\boldsymbol{S}_x = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$.
2. Classical principal component analysis (PCA) uses the eigenvectors and eigenvalues of $\boldsymbol{S}_x$ to construct a smaller set of uncorrelated variables.
3. In the multiple regression setting, also a response variable $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is measured. The goal of linear regression is to estimate the parameter $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})' = (\beta_0, \beta_1, \ldots, \beta_p)'$ relating the response variable and the predictor variables in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

The least squares slope estimator can be written as $\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{S}_x^{-1} \boldsymbol{s}_{xy}$ with $\boldsymbol{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ the cross-covariance vector. The intercept is given by $\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}_{LS}' \bar{\boldsymbol{x}}$.

These classical estimators often possess optimal properties under the Gaussian model assumptions, but they can be strongly affected by even a few ►outliers. Outliers are data points that deviate from the pattern suggested by

the majority of the data. Outliers are more likely to occur in datasets with many observations and/or variables, and often they do not show up by simple visual inspection. When the data contain nasty outliers, typically two things happen:

- The multivariate estimates differ substantially from the "right" answer, defined here as the estimates we would have obtained without the outliers.
- The resulting fitted model does not allow to detect the outliers by means of their residuals, Mahalanobis distances, or the widely used "leave-one-out" diagnostics.

The first consequence is fairly well-known (although the size of the effect is often underestimated). Unfortunately the second consequence is less well-known, and when stated many people find it hard to believe or paradoxical. Common intuition says that outliers must "stick out" from the classical fitted model, and indeed some of them do so. But the most harmful types of outliers, especially if there are several of them, may affect the estimated model so much "in their direction" that they are now well-fitted by it.

Once this effect is understood, one sees that the following two problems are essentially equivalent:

- Robust estimation: find a "robust" fit, which is similar to the fit we would have found without the outliers.
- Outlier detection: find all the outliers that matter.

Indeed, a solution to the first problem allows us, as a by-product, to identify the outliers by their deviation from the robust fit. Conversely, a solution to the second problem would allow us to remove or downweight the outliers followed by a classical fit, which yields a robust estimate.

It turns out that the more fruitful approach is to solve the first problem and to use its result to answer the second. This is because from a combinatorial viewpoint it is more feasible to search for *sufficiently many* "good" data points than to find *all* the "bad" data points.

Many robust multivariate estimators have been constructed by replacing the empirical mean and covariance matrix with a robust alternative. Currently the most popular estimator for this purpose is the *Minimum Covariance Determinant* (MCD) estimator (Rousseeuw 1984). The MCD method looks for the h observations (out of n) whose classical covariance matrix has the lowest possible determinant. The raw MCD estimate of location is then the average of these h points, whereas the raw MCD estimate of scatter is a multiple of their covariance matrix. Based on these raw estimates one typically carries out a reweighting step, yielding the reweighted MCD estimates (Rousseeuw and Van Driessen 1999).

The MCD location and scatter estimates are affine equivariant, which means that they behave properly under affine transformations of the data. Computation of the MCD is non-trivial, but can be performed efficiently by means of the FAST-MCD algorithm (Rousseeuw and Van Driessen 1999) which is available in standard SAS, S-Plus, and R.

A useful measure of robustness is the *finite-sample breakdown value* (Donoho and Huber 1983; Hampel et al. 1986). The breakdown value is the smallest amount of contamination that can have an arbitrarily large effect on the estimator. The MCD estimates of multivariate location and scatter have breakdown value $\approx (n-h)/n$. The MCD has its highest possible breakdown value of 50% when $h = [(n+p+1)/2]$. Note that no affine equivariant estimator can have a breakdown value above 50%.

Another measure of robustness is the *influence function* (Hampel et al. 1986), which measures the effect on an estimator of adding a small mass of data in a specific place. The MCD has a bounded influence function, which means that a small contamination at any position can only have a small effect on the estimator (Croux and Haesbroeck 1999).

In regression, a popular estimator with high breakdown value is the *Least Trimmed Squares* (LTS) estimator (Rousseeuw 1984; Rousseeuw and Van Driessen 2006). The LTS is the fit that minimizes the sum of the h smallest squared residuals (out of n). Other frequently used robust estimators include S-estimators (Rousseeuw and Yohai 1984) and MM-estimators (Yohai 1987), which can achieve a higher finite-sample efficiency than the LTS.

Robust multivariate estimators have been used to robustify the Hotelling T^2 statistic (Willems et al. 2002), PCA (Croux and Haesbroeck 2000; Salibian-Barrera et al. 2006), multiple regression with one or several response variables (Rousseeuw et al. 2004; Agulló et al. 2008), discriminant analysis (Hawkins and McLachlan 1997; Hubert and Van Driessen 2004; Croux and Dehon 2001), factor analysis (Pison et al. 2003), canonical correlation (Croux and Dehon 2002), and cluster analysis (Hardin and Rocke 2004).

Another important group of robust multivariate methods are based on projection pursuit (PP) techniques. They are especially useful when the dimension p of the data is larger than the sample size n, in which case the MCD is no longer well-defined. Robust PP methods project the data on many univariate directions and apply robust estimators of location and scale (such as the median and the median absolute deviation) to each projection. Examples include the Stahel-Donoho estimator of location and scatter (Maronna and Yohai 1995) and generalizations (Zuo et al. 2004), robust

PCA (Li and Chen 1985; Croux and Ruiz-Gazen 2005; Hubert et al. 2002; Boente et al. 2006), discriminant analysis (Pires 2003), canonical correlation (Branco et al. 2005), and outlier detection in skewed data (Brys et al. 2005; Hubert and Van der Veeken 2008). The hybrid ROBPCA method (Hubert et al. 2005; Debruyne and Hubert 2009) combines PP techniques with the MCD and has led to the construction of robust principal component regression (Hubert and Verboven 2003), partial least squares (Hubert and Vanden Branden 2003), and classification for high-dimensional data (Vanden Branden and Hubert 2005).

A more extensive description of robust multivariate methods and their applications can be found in (Hubert et al. 2008; Hubert and Debruyne 2010).

About the Author

Dr. Peter Rousseeuw was Professor and Head (since 1992) of the Division of Applied Mathematics, Universiteit Antwerpen, Belgium. Currently he is a Senior Researcher at Renaissance Technologies in New York. He has (co-)authored over 160 papers, two edited volumes and three books, including *Robust Regression and Outlier Detection* (with A.M. Leroy, Wiley-Interscience, 1987). In 2003 ISI-Thompson included him in their list of Highly Cited Mathematicians. His paper *Least Median of Squares Regression* (1984), Journal of the American Statistical Association, 79, 871–880) which proposed new robust methods for regression and covariance, has been reprinted in *Breakthroughs in Statistics III* (the three-volume collection consists of the 60 most influential publications in statistics from 1850 to 1990), Kotz and Johnson 1997, Springer-Verlag, New York. He is an Elected Member, International Statistical Institute (1991) and an Elected Fellow of Institute of Mathematical Statistics (elected 1993) and American Statistical Association (elected 1994). He was Associate Editor, *Journal of the American Statistical Association* (1988–1993), and *Computational Statistics and Data Analysis* (1988–1998). He has supervised 20 Ph.D. students.

Cross References

►Eigenvalue, Eigenvector and Eigenspace
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Hotelling's T^2 Statistic
►Multivariate Data Analysis: An Overview
►Multivariate Outliers
►Multivariate Statistical Analysis
►Outliers
►Principal Component Analysis
►Robust Inference
►Robust Statistics

References and Further Reading

Agulló J, Croux C, Van Aelst S (2008) The multivariate least trimmed squares estimator. J Multivariate Anal 99:311–318
Boente G, Pires AM, Rodrigues I (2006) General projection-pursuit estimates for the common principal components model: Influence functions and Monte Carlo study. J Multivariate Anal 97:124–147
Branco JA, Croux C, Filzmoser P, Oliviera MR (2005) Robust canonical correlations: a comparative study. Comput Stat 20:203–229
Brys G, Hubert M, Rousseeuw PJ (2005) A robustification of independent component analysis. J Chemometr 19:364–375
Croux C, Dehon C (2001) Robust linear discriminant analysis using *S*-estimators. Can J Stat 29:473–492
Croux C, Dehon C (2002) Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. La Revue de Statistique Appliquée 2:5–26
Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J Multivariate Anal 71:161–190
Croux C, Haesbroeck G (2000) Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87:603–618
Croux C, Ruiz-Gazen A (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. J Multivariate Anal 95:206–226
Debruyne M, Hubert M (2009) The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. Stat Probab Lett 79:275–282
Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel P, Doksum K, Hodges JL (eds) A Festschrift for Erich Lehmann. Wadsworth, Belmont, pp 157–184
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley-Interscience, New York
Hardin J, Rocke DM (2004) Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput Stat Data Anal 44:625–638
Hawkins DM, McLachlan GJ (1997) High-breakdown linear discriminant analysis. J Am Stat Assoc 92:136–143
Hubert M, Debruyne M (2010) Minimum covariance determinant. Wiley Interdisciplinary Rev Comput Stat 2:36–43
Hubert M, Van der Veeken S (2008) Outlier detection for skewed data. J Chemometr 22:235–246
Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Comput Stat Data Anal 45:301–320
Hubert M, Vanden Branden K (2003) Robust methods for partial least squares regression. J Chemometr 17:537–549
Hubert M, Verboven S (2003) A robust PCR method for high-dimensional regressors. J Chemometr 17:438–452
Hubert M, Rousseeuw PJ, Verboven S (2002) A fast robust method for principal components with applications to chemometrics. Chemomet Intell Lab 60:101–111
Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: a new approach to robust principal components analysis. Technometrics 47:64–79
Hubert M, Rousseeuw PJ, Van Aelst S (2008) High breakdown robust multivariate methods. Stat Sci 23:92–119

Li G, Chen Z (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. J Am Stat Assoc 80:759–766

Maronna RA, Yohai VJ (1995) The behavior of the Stahel-Donoho robust multivariate estimator. J Am Stat Assoc 90: 330–341

Pires AM (2003) Robust discriminant analysis and the projection pursuit approach: practical aspects. In: Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds) Developments in robust statistics. Physika Verlag, Heidelberg, pp 317–329

Pison G, Rousseeuw PJ, Filzmoser P, Croux C (2003) Robust factor analysis. J Multivariate Anal 84:145–172

Rousseeuw PJ, Yohai V (1984) Robust regression based on S-estimators. In: Franke J, Haerdle W, Martin RD (eds) Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics No. 26, Springer Verlag, New York, pp 256–272

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880

Rousseeuw PJ, Yohai AM (1987) Robust regression and outlier detection. Wiley-Interscience, New York

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223

Rousseeuw PJ, Van Driessen K (2006) Computing LTS regression for large data sets. Data Min Knowl Disc 12:29–45

Rousseeuw PJ, Van Aelst S, Van Driessen K, Agulló J (2004) Robust multivariate regression. Technometrics 46:293–305

Salibian-Barrera M, Van Aelst S, Willems G (2006) PCA based on multivariate MM-estimators with fast and robust bootstrap. J Am Stat Assoc 101:1198–1211

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometr Intell Lab 79:10–21

Willems G, Pison G, Rousseeuw PJ, Van Aelst S (2002) A robust Hotelling test. Metrika 55:125–138

Yohai VJ (1987) High breakdown point and high efficiency robust estimates for regression. Ann Stat 15:642–656

Zuo Y, Cui H, He X (2004) On the Stahel-Donoho estimator and depth-weighted means of multivariate data. Annals Stat 32: 167–188