A

Absolute Penalty Estimation

Ejaz S. Ahmed[1], Enayetur Raheem[2], Shakhawat Hossain[2]
[1]Professor and Department Head of Mathematics and Statistics
University of Windsor, Windsor, ON, Canada
[2]University of Windsor, Windsor, ON, Canada

In statistics, the technique of ▶least squares is used for estimating the unknown parameters in a linear regression model (see ▶Linear Regression Models). This method minimizes the sum of squared distances between the observed responses in a set of data, and the fitted responses from the regression model. Suppose we observe a collection of data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ on n units, where y_is are responses and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of predictors. It is convenient to write the model in matrix notation, as,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\boldsymbol{y}$ is $n \times 1$ vector of responses, $\boldsymbol{X}$ is $n \times p$ matrix, known as the design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the unknown parameter vector and ε is the vector of random errors. In ordinary least squares (OLS) regression, we estimate $\boldsymbol{\beta}$ by minimizing the residual sum of squares, $RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, giving $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. This estimator is simple and has some good statistical properties. However, the estimator suffers from lack of uniqueness if the design matrix $\boldsymbol{X}$ is less than full rank, and if the columns of $\boldsymbol{X}$ are (nearly) collinear. To achieve better prediction and to alleviate ill conditioning problem of $\boldsymbol{X}^T\boldsymbol{X}$, Hoerl and Kernard (1970) introduced ridge regression (see ▶Ridge and Surrogate Ridge Regressions), which minimizes the RSS subject to a constraint, $\sum \beta_j^2 \le t$, in other words

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2}$$

where $\lambda \ge 0$ is known as the complexity parameter that controls the amount of shrinkage. The larger the value of λ, the greater the amount of shrinkage. The quadratic penalty term makes $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ a linear function of $\boldsymbol{y}$. Frank and Friedman (1993) introduced bridge regression, a generalized version of penalty (or absolute penalty type) estimation, which includes ridge regression when $\gamma = 2$. For a given penalty function $\pi(\cdot)$ and regularization parameter λ, the general form can be written as

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\pi(\boldsymbol{\beta}),$$

where the penalty function is of the form

$$\pi(\beta) = \sum_{j=1}^{p} |\beta_j|^\gamma, \ \gamma > 0. \tag{3}$$

The penalty function in (3) bounds the L_γ norm of the parameters in the given model as $\sum_{j=1}^{m} |\beta_j|^\gamma \le t$, where t is the tuning parameter that controls the amount of shrinkage. We see that for $\gamma = 2$, we obtain ridge regression. However, if $\gamma \ne 2$, the penalty function will not be rotationally invariant. Interestingly, for $\gamma < 2$, it shrinks the coefficient toward zero, and depending on the value of λ, it sets some of them to be exactly zero. Thus, the procedure combines variable selection and shrinkage of coefficients of penalized regression. An important member of the penalized least squares (PLS) family is the L_1 penalized least squares estimator or the *lasso* [*least absolute shrinkage and selection operator*, Tibshirani (1996)]. In other words, the *absolute penalty estimator* (APE) arises when the absolute value of penalty term is considered, i.e., $\gamma = 1$ in (3). Similar to the ridge regression, the lasso estimates are obtained as

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

The lasso shrinks the OLS estimator toward zero and depending on the value of λ, it sets some coefficients to exactly zero. Tibshirani (1996) used a quadratic programming method to solve (4) for $\hat{\boldsymbol{\beta}}^{\text{lasso}}$. Later, Efron et al. (2004) proposed least angle regression (LAR), a type of stepwise regression, with which the

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

lasso estimates can be obtained at the same computational cost as that of an ordinary least squares estimation Hastie et al. (2009). Further, the lasso estimator remains numerically feasible for dimensions m that are much higher than the sample size n. Zou and Hastie (2005) introduced a hybrid PLS regression with the so called *elastic net penalty* defined as $\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)\,|\beta_j|)$. Here the penalty function is a linear combination of the ridge regression penalty function and lasso penalty function. A different type of PLS, called *garotte* is due to Breiman (1993). Further, PLS estimation provides a generalization of both nonparametric least squares and weighted projection estimators, and a popular version of the PLS is given by Tikhonov regularization (Tikhonov 1963). Generally speaking, the ridge regression is highly efficient and stable when there are many small coefficients. The performance of lasso is superior when there are a small-to-medium number of moderate-sized coefficients. On the other hand, shrinkage estimators perform well when there are large known zero coefficients.

Ahmed et al. (2007) proposed an APE for partially linear models. Further, they reappraised the properties of shrinkage estimators based on Stein-rule estimation. There exists a whole family of estimators that are better than OLS estimators in regression models when the number of predictors is large. A partially linear regression model is defined as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where $t_i \in [0,1]$ are design points, $g(\cdot)$ is an unknown real-valued function defined on $[0,1]$, and y_i, $\boldsymbol{x}$, $\boldsymbol{\beta}$, and ε_i's are as defined in the context of (1). We consider experiments where the vector of coefficients $\boldsymbol{\beta}$ in the linear part of (5) can be partitioned as $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ is the coefficient vector of order $p_1 \times 1$ for main effects (e.g., treatment effects, genetic effects) and $\boldsymbol{\beta}_2$ is a vector of order $p_2 \times 1$ for "nuisance" effects (e.g., age, laboratory). Our relevant hypothesis is $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}_1$ be a semiparametric least squares estimator of $\boldsymbol{\beta}_1$, and we let $\tilde{\boldsymbol{\beta}}_1$ denote the restricted semiparametric least squares estimator of $\boldsymbol{\beta}_1$. Then the semiparametric Stein-type estimator (see ►James-Stein Estimator and Semiparametric Regression Models), $\hat{\boldsymbol{\beta}}_1^S$, of $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1^S = \tilde{\boldsymbol{\beta}}_1 + \{1 - (p_2 - 2)T^{-1}\}(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1), \quad p_2 \geq 3 \tag{6}$$

where T is an appropriate test statistic for the H_0. A positive-rule shrinkage estimator (PSE) $\hat{\boldsymbol{\beta}}_1^{S+}$ is defined as

$$\hat{\boldsymbol{\beta}}_1^{S+} = \tilde{\boldsymbol{\beta}}_1 + \{1 - (p_2 - 2)T^{-1}\}^+(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1), \quad p_2 \geq 3 \tag{7}$$

where $z^+ = max(0, z)$. The PSE is particularly important to control the over-shrinking inherent in $\hat{\boldsymbol{\beta}}_1^S$. The shrinkage estimators can be viewed as a competitor to the APE approach. Ahmed et al. (2007) finds that, when p_2 is relatively small with respect to p, APE performs better than the shrinkage method. On the other hand, the shrinkage method performs better when p_2 is large, which is consistent with the performance of the APE in linear models. Importantly, the shrinkage approach is free from any tuning parameters, easy to compute and calculations are not iterative. The shrinkage estimation strategy can be extended in various directions to more complex problems. It may be worth mentioning that this is one of the two areas Bradley Efron predicted for the early twenty-first century (RSS News, January 1995). Shrinkage and likelihood-based methods continue to be extremely useful tools for efficient estimation.

About the Author

The author S. Ejaz Ahmed is Professor and Head Department of Mathematics and Statistics. For biography, *see* entry ►Optimal Shrinkage Estimation.

Cross References

►Estimation
►Estimation: An Overview
►James-Stein Estimator
►Linear Regression Models
►Optimal Shrinkage Estimation
►Residuals
►Ridge and Surrogate Ridge Regressions
►Semiparametric Regression Models

References and Further Reading

Ahmed SE, Doksum KA, Hossain S, You J (2007) Shrinkage, pretest and absolute penalty estimators in partially linear models. Aust NZ J Stat 49(4):435–454

Breiman L (1993) Better subset selection using the non-negative garotte. Technical report, University of California, Berkeley

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Ann Stat 32(2):407–499

Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35:109–148

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Tikhonov An (1963) Solution of incorrectly formulated problems and the regularization method. Soviet Math Dokl 4:1035–1038, English translation of Dokl Akad Nauk SSSR 151, 1963, 501–504

Zou H, Hastie T (2005) Regularization and variable selction via the elastic net. J R Stat Soc B 67(2):301–320

Accelerated Lifetime Testing

Francisco Louzada-Neto
Associate Professor
Universidade Federal de São Carlos, Sao Paulo, Brazil

Accelerated life tests (ALT) are efficient industrial experiments for obtaining measures of a device reliability under the usual working conditions.

A practical problem for industries of different areas is to obtain measures of a device reliability under its usual working conditions. Typically, the time and cost of such experimentation are long and expensive. The ALT are efficient for handling such situation, since the information on the device performance under the usual working conditions are obtained by considering a time and cost-reduced experimental scheme. The ALT are performed by testing items at higher stress covariate levels than the usual working conditions, such as temperature, pressure and voltage.

There is a large literature on ALT and interested readers can refer to Mann et al. (1974), Nelson (1990), Meeker and Escobar (1998) which are excellent sources for ALT. Nelson (2005a, b) provides a brief background on accelerated testing and test plans and surveys the related literature point out more than 150 related references.

A simple ALT scenario is characterized by putting k groups of n_i items each under constant and fixed stress covariate levels, X_i (hereafter stress level), for $i = 1, \ldots, k$, where $i = 1$ generally denotes the usual stress level, that is, the usual working conditions. The experiment ends after a certain pre-fixed number $r_i < n_i$ of failures, $t_{i1}, t_{i2}, \ldots, t_{ir_i}$, at each stress level, characterizing a type II censoring scheme (Lawless 2003; see also ▶Censoring Methodology). Other stress schemes, such as step (see ▶Step-Stress Accelerated Life Tests) and progressive ones, are also common in practice but will not be considered here. Examples of those more sophisticated stress schemes can be found in Nelson (1990).

The ALT models are composed by two components. One is a probabilistic component, which is represented by a lifetime distribution, such as exponential, Weibull, log-normal, log-logistic, among others. The other is a stress-response relationship (SRR), which relates the mean lifetime (or a function of this parameter) with the stress levels. Common SRRs are the power law, Eyring and Arrhenius models (Meeker and Escobar 1998) or even a general log-linear or log-non-linear SRR which encompass the formers. For sake of illustration, we shall assume an exponential distribution as the lifetime model and a general log-linear SRR. Here, the mean lifetime under the usual working conditions shall represent our device reliability measure of interesting.

Let $T > 0$ be the lifetime random variable with an exponential density

$$f(t, \lambda_i) = \lambda_i \exp\{-\lambda_i t\}, \tag{1}$$

where $\lambda_i > 0$ is an unknown parameter representing the constant failure rate for $i = 1, \ldots, k$ (number of stress levels). The mean lifetime is given by $\theta_i = 1/\lambda_i$.

The likelihood function for λ_i, under the i-th stress level X_i, is given by

$$L_i(\lambda_i) = \left(\prod_{j=1}^{r_i} f(t_{ij}, \lambda_i)\right) (S(t_{ir_i}, \lambda_i))^{n_i - r_i} = \lambda_i^{r_i} \exp\{-\lambda_i A_i\},$$

where $S(t_{ir_i}, \lambda_i)$ is the survival function at t_{ir_i} and $A_i = \sum_{j=1}^{r_i} t_{ij} + (n_i - r_i) t_{ir_i}$ denotes the total time on test for the i-th stress level.

Considering data under the k random stress levels, the likelihood function for the parameter vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ is given by

$$L(\lambda) = \prod_{i=1}^{k} \lambda_i^{r_i} \exp\{-\lambda_i A_i\}. \tag{2}$$

We consider a general log-linear SRR defined as

$$\lambda_i = \exp(-Z_i - \beta_0 - \beta_1 X_i), \tag{3}$$

where X is the covariate, $Z = g(X)$ and β_0 and β_1 are unknown parameters such that $-\infty < \beta_0, \beta_1 < \infty$.

The SRR (3) has several models as particular cases. The Arrhenius model is obtained if $Z_i = 0$, $Xi = 1/Vi$, $\beta_0 = -\alpha_1$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the temperature variable. If $Z_i = 0$, $Xi = -log(Vi)$, $\beta_0 = log(\alpha)$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the voltage variable we obtain the power model. Following Louzada-Neto and Pardo-Fernandéz (2001), the Eyring model is obtained if $Z_i = -\log V_i$, $Xi = 1/Vi$, $\beta_0 = -\alpha_1$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the temperature variable. Interested readers can refer to Meeker and Escobar (1998) for more information about the physical models considered here.

From (2) and (3), the likelihood function for β_0 and β_1 is given by

$$L(\beta_0, \beta_1) = \prod_{i=1}^{k} \{\exp(-Z_i - \beta_0 - \beta_1 X_i)^{r_i} \exp(-\exp(-Z_i - \beta_0 - \beta_1 X_i) A_i)\}. \quad (4)$$

The maximum likelihood estimates (MLEs) of β_0 and β_1 can be obtained by direct maximization of (4), or by solving the system of nonlinear equations, $\partial \log L/\partial\theta = 0$, where $\theta' = (\beta_0, \beta_1)$. Obtaining the score function is conceptually simple and the expressions are not given explicitly. The MLEs of θ_i can be obtained, in principle, straightforwardly by considering the invariance property of the MLEs.

Large-sample inference for the parameters can be based on the MLEs and their estimated variances, obtained by inverting the expected information matrix (Cox and Hinkley 1974). For small or moderate-sized samples however we may consider simulation approaches, such as the bootstrap confidence intervals (see ►Bootstrap Methods) that are based on the empirical evidence and are therefore preferred (Davison and Hinkley 1997). Formal goodness-of-fit tests are also feasible since, from (3), we can use the likelihood ratio statistics (LRS) for testing goodness-of-fit of hypotheses such as $H_0 : \beta_1 = 0$.

Although we considered only an exponential distribution as our lifetime model, more general lifetime distributions, such as the Weibull (see ►Weibull Distribution and Generalized Weibull Distributions), log-normal, log-logistic, among others, could be considered in principle. However, the degree of difficulty in the calculations increase considerably. Also we considered only one stress covariate, however this is not critical for the overall approach to hold and the multiple covariate case can be handle straightforwardly.

A study on the effect of different reparametrizations on the accuracy of inferences for ALT is discussed in Louzada-Neto and Pardo-Fernandéz 2001). Modeling ALT with a log-non-linear SRR can be found in Perdoná et al. (2004). Modeling ALT with a threshold stress, below which the lifetime of a product can be considered to be infinity or much higher than that for which it has been developed is proposed by Tojeiro et al. (2004).

We only considered ALT in presence of constant stress loading, however non-constant stress loading, such as step stress and linearly increasing stress are provided by Miller and Nelson (1983) and Bai, Cha and Chung (1992), respectively. A comparison between constant and step stress tests is provided by Khamis (1997). A log-logistic step stress model is provided by Srivastava and Shukla (2008).

Two types of software for ALT are provided by Meeker and Escobar (2002) and ReliaSoft Corporation (2004).

About the Author

Francisco Louzada-Neto is an associate professor of Statistics at Universidade Federal de São Carlos (UFSCar), Brazil. He received his Ph.D in Statistics from University of Oxford (England). He is Director of the Centre for Hazard Studies (2004–2010, UFSCar, Brazil) and Editor in Chief of the *Brazilian Journal of Statistics* (2004–2010, Brazil). He is a past-Director for Undergraduate Studies (1992–1994, UFSCar, Brazil) and was Director for Graduate Studies in Statistics (1999–2008, UFSCar, Brazil). Louzada-Neto is single and joint author of more than 100 publications in statistical peer reviewed journals, books and book chapters, He has supervised more than 50 assistant researches, Ph.Ds, masters and undergraduates.

Cross References

►Degradation Models in Reliability and Survival Analysis
►Modeling Survival Data
►Step-Stress Accelerated Life Tests
►Survival Data

References and Further Reading

Bai DS, Cha MS, Chung SW (1992) Optimum simple ramp tests for the Weibull distribution and type-I censoring. IEEE T Reliab 41:407–413

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Khamis IH (1997) Comparison between constant- and step-stress tests for Weibull models. Int J Qual Reliab Manag 14:74–81

Lawless JF (2003) Statistical models and methods for lifetime data, 2nd end. Wiley, New York

Louzada-Neto F, Pardo-Fernandéz JC (2001) The effect of reparametrization on the accuracy of inferences for accelerated lifetime tests. J Appl Stat 28:703–711

Mann NR, Schaffer RE, Singpurwalla ND (1974) Methods for statistical analysis of reliability and life test data. Wiley, New York

Meeker WQ, Escobar LA (1998) Statistical methods for reliability data. Wiley, New York

Meeker WQ, Escobar LA (2002) SPLIDA (S-PLUS Life Data Analysis) software–graphical user interface. http://www.public.iastate.edu/~splida

Miller R, Nelson WB (1983) Optimum simple step-stress plans for accelerated life testing. IEEE T Reliab 32:59–65

Nelson W (1990) Accelerated testing – statistical models, test plans, and data analyses. Wiley, New York

Nelson W (2005a) A bibliography of accelerated test plans. IEEE T Reliab 54:194–197

Nelson W (2005b) A bibliography of accelerated test plans part II – references. IEEE T Reliab 54:370–373

Perdoná GSC, Louzada Neto F, Tojeiro CAV (2004) Bayesian modelling of log-non-linear stress-response relationships in accelerated lifetime tests. J Stat Theory Appl 3(1):5–12

Reliasoft Corporation (2004) Optimum allocations of stress levels and test units in accelerated tests. Reliab EDGE 5:10–17. http://www.reliasoft.com

Srivastava PW, Shukla R (2008) A log-logistic step-stress model. IEEE T Reliab 57:431–434

Tojeiro CAV, Louzada Neto F, Bolfarine H (2004) A Bayesian analysis for accelerated lifetime tests under an exponential power law model with threshold stress. J Appl Stat 31(6):685–691

Acceptance Sampling

M. Ivette Gomes
Professor
Universidade de Lisboa, DEIO and CEAUL, Lisboa, Portugal

Introduction

Acceptance sampling (AS) is one of the oldest statistical techniques in the area of ▶statistical quality control. It is performed out of the line production, most commonly before it, for deciding on incoming batches, but also after it, for evaluating the final product (see Duncan 1986; Stephens 2001; Pandey 2007; Montgomery 2009; and Schilling and Neubauer 2009, among others). Accepted batches go into the production line or are sold to consumers; the rejected ones are usually submitted to a *rectification process*. A *sampling plan* is defined by the *size of the sample* (samples) taken from the batch and by the associated *acceptance–rejection* criterion. The most widely used plans are given by the Military Standard tables, developed during the *World War* II, and first issued in 1950. We mention MIL STD 105E (1989) and the civil version ANSI/ASQC Z1.9 (1993) of the *American National Standards Institution* and the *American Society for Quality Control*.

At the beginning, all items and products were inspected for the identification of nonconformities. At the late 20s, Dodge and Romig (see Dodge and Romig 1959), in the Bell Laboratories, developed the area of AS, as an alternative to 100% inspection. The aim of AS is to lead producers to a decision (acceptance or rejection of a batch) and not to the estimation or improvement of the quality of a batch. Consequently, AS does not provide a direct form of *quality control*, but its indirect effects in *quality* are important: if a batch is rejected, either the supplier tries improving its production methods or the consumer (producer) looks for a better supplier, indirectly increasing quality.

Regarding the decision on the batches, we distinguish three different approaches: (1) *acceptance without inspection*, applied when the supplier is highly reliable; (2) 100% *inspection*, which is expensive and can lead to a sloppy attitude towards quality; (3) *an intermediate decision*, i.e., an *acceptance sampling program*. This increases the interest on quality and leads to the lemma: *make things right in the first place*. The type of inspection that should be applied depends on the quality of the last batches inspected. At the beginning of inspection, a so-called *normal inspection* is used, but there are two other types of inspection, a *tightened inspection* (for a history of low quality), and a *reduced inspection* (for a history of high quality). There are special and empirical switching rules between the three types of inspection, as well as for discontinuation of inspection.

Factors for Classifications of Sampling Plans

Sampling plans by attributes versus sampling plans by variables. If the item inspection leads to a binary result (conforming or nonconforming), we are dealing with *sampling by attributes*, detailed later on. If the item inspection leads to a continuous measurement X, we are *sampling by variables*. Then, we generally use sampling plans based on the sample mean and standard deviation, the so-called *variable sampling plans*. If X is normal, it is easy to compute the number of items to be collected and the criteria that leads to the rejection of the batch, with chosen risks α and β. For different *sampling plans by variables*, see Duncan (1986), among others.

Incoming versus outgoing inspection. If the batches are inspected before the product is sent to the consumer, it is called *outgoing inspection*. If the inspection is done by the consumer (producer), after they were received from the supplier, it is called *incoming inspection*.

Rectifying versus non-rectifying sampling plans. All depends on what is done with nonconforming items that were found during the inspection. When the cost of replacing faulty items with new ones, or reworking them is accounted for, the sampling plan is rectifying.

Single, double, multiple and sequential sampling plans.

- **Single sampling**. This is the most common sampling plan: we draw a random sample of n items from the batch, and count the number of nonconforming items (or the number of nonconformities, if more than one nonconformity is possible on a single item). Such a

plan is defined by n and by an associated *acceptance-rejection* criterion, usually a value c, the so-called *acceptance number*, the number of nonconforming items that cannot be exceeded. If the number of nonconforming items is greater than c, the batch is rejected; otherwise, the batch is accepted. The number r, defined as the minimum number of nonconforming items leading to the rejection of the batch, is the so-called *rejection number*. In the most simple case, as above, $r = c + 1$, but we can have $r > c + 1$.

- **Double sampling**. A *double sampling plan* is characterized by four parameters: $n_1 << n$, the size of the first sample, c_1 the acceptance number for the first sample, n_2 the size of the second sample and c_2 ($> c_1$) the acceptance number for the joint sample. The main advantage of a double sampling plan is the reduction of the total inspection and associated cost, particularly if we proceed to a *curtailment* in the second sample, i.e. we stop the inspection whenever c_2 is exceeded. Another (psychological) advantage of these plans is the way they give a second opportunity to the batch.
- **Multiple sampling**. In the *multiple plans* a predetermined number of samples are drawn before taking a decision.
- ►**Sequential sampling**. The *sequential plans* are a generalization of multiple plans. The main difference is that the number of samples is not pre-determined. If, at each step, we draw a sample of size *one*, the plan, based on Wald's test, is called *sequential item-to-item*; otherwise, it is *sequential by groups*. For a full study of multiple and sequential plans see, for instance, Duncan (1986) (see also the entry ►Sequential Sampling).

Special sampling plans. Among the great variety of special plans, we distinguish:

- **Chain sampling**. When the inspection procedures are destructive or very expensive, a small n is recommendable. We are then led to acceptance numbers equal to zero. This is dangerous for the supplier and if rectifying inspection is used, it is expensive for the consumer. In 1955, Dodge suggested a procedure alternative to this type of plans, which uses also the information of preceding batches, the so-called *chain sampling method* (see Dogdge and Romig 1959).
- **Continuous sampling plans** (CSP). There are continuous production processes, where the raw material is not naturally provided in batches. For this type of production it is common to alternate sequences of sampling inspection with 100% inspection – they are in a certain sense rectifying plans. The simplest plan of this type, the CSP-1, was suggested in 1943 by Dodge. It begins with a 100% inspection. When a pre-specified number i of consecutive nonconforming items is achieved, the plan changes into sampling inspection, with the inspection of f items, randomly selected, along the continuous production. If *one* nonconforming item is detected (the reason for the terminology CSP-1), 100% inspection comes again, and the nonconforming item is replaced. For properties of this plan and its generalizations see Duncan (1986).

A Few Characteristics of a Sampling Plan

OCC. The *operational characteristic curve* (OCC) is $P_a \equiv P_a(p) = \mathbb{P}(\text{acceptance of the batch} \mid p)$, where p is the probability of a nonconforming item in the batch.

AQL and LTPD (or RQL). The sampling plans are built taken into account the wishes of both the supplier and the consumer, defining two quality levels for the judgment of the batches: the *acceptance quality level* (AQL), the worst operating quality of the process which leads to a high probability of acceptance of the batch, usually 95% – for the protection of the supplier regarding high quality batches, and the *lot tolerance percent defective* (LTPD) or *rejectable quality level* (RQL), the quality level below which an item cannot be considered acceptable. This leads to a small acceptance of the batch, usually 10% – for the protection of the consumer against low quality batches. There exist two types of decision, acceptance or rejection of the batch, and two types of risks, to reject a "good" (high quality) batch, and to accept a "bad" (low quality) batch. The probabilities of occurrence of these risks are the so-called *supplier risk* and *consumer risk*, respectively. In a *single sampling plan*, the *supplier risk* is $\alpha = 1 - P_a(\text{AQL})$ and the *consumer risk* is $\beta = P_a(\text{LTPD})$. The sampling plans should take into account the specifications AQL and LTPD, i.e. we are supposed to find a single plan with an OCC that passes through the points (AQL, 1-α) and (LTPD, β). The construction of double plans which protect both the supplier and the consumer are much more difficult, and it is no longer sufficient to provide indication on two points of the OCC. There exist the so-called *Grubbs' tables* (see Montgomery 2009) providing (c_1, c_2, n_1, n_2), for $n_2 = 2n_1$, as an example, $\alpha = 0.05$, $\beta = 0.10$ and several rates RQL/AQL.

AOQ, AOQL and ATI. If there is a *rectifying inspection program* – a corrective program, based on a 100% inspection and replacement of nonconforming by conforming items, after the rejection of a batch by an AS plan –, the most relevant *characteristics* are the *average outgoing quality* (AOQ), $\text{AOQ}(p) = p\,(1 - n/N)\,P_a$, which attains

a maximum at the so-called *average output quality limit* (AOQL), the worst average quality of a product after a rectifying inspection program, as well as the *average total inspection* (ATI), the amount of items subject to inspection, equal to n if there is no rectification, but given by $\mathrm{ATI}(p) = nP_a + N(1 - P_a)$, otherwise.

Acknowledgments

Research partially supported by FCT/OE, POCI 2010 and PTDC/FEDER.

About the Author

For biography of M. Ivette Gomes *see* the entry ►Statistical Quality Control.

Cross References

►Industrial Statistics
►Sequential Sampling
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Dodge HF, Romig HG (1959) Sampling inspection tables, single and double sampling, 2nd edn. Wiley, New York
Duncan AJ (1986) Quality control and industrial statistics, 5th edn. Irwin, Homehood
Montgomery DC (2009) Statistical quality control: a modern introduction, 6th edn. Wiley, Hoboken, NJ
Pandey BN (2007) Statistical techniques in life-testing, reliability, sampling theory and quality control. Narosa, New Delhi
Schilling EG, Neubauer DV (2009) Acceptance sampling in quality control, 2nd edn. Chapman and Hall/CRC, New York
Stephens KS (2001) The handbook of applied acceptance sampling: plans, principles, and procedures. ASQ Quality, Milwaukee

Actuarial Methods

Vassiliy Simchera
Director
Rosstat's Statistical Research Institute, Moscow, Russia

A specific (and relatively new) type of financial calculations are actuarial operations, which represent a special (in majority of countries they are usually licensed) sphere of activity related to identifications of risks outcomes and market assessment of future (temporary) borrowed current assets and liabilities costs for their redemption.

The broad range of existing and applicable actuarial calculations require use of various methods and inevitably predetermines a necessity of their alteration depending on concrete cases of comparison analysis and selection of most efficient of them.

The condition of success is a typology of actuarial calculations methods, based on existing typology fields and objects of their applications, as well as knowledge of rule for selection of most efficient methods, which would provide selection of target results with minimum costs or high accuracy.

Regarding the continuous character of financial transactions, the actuarial calculations are carried out permanently. The aim of actuarial calculations in every particular case is probabilistic determination of profit sharing (transaction return) either in the form of financial liabilities (interest, margin, agio, etc.) or as commission charges (such as royalty).

The subject of actuarial calculations can be distinguished in the narrow and in the broad senses.

The given subject in the broad sense covers financial and actuarial accounts, budgeting, balance, audit, assessment of financial conditions and financial provision for all categories and types of borrowing institutions, basis for their preferential financial decisions and transactions, conditions and results of work for different financial and credit institutions; financial management of cash flows, resources, indicators, mechanisms, instruments, as well as financial analysis and audit of financial activity of companies, countries, nations their groups and unions, including national system of financial account, financial control, engineering, and forecast. In other words, the subject of actuarial calculations is a process of determination of any expenditures and incomes from any type of transactions in the shortest way.

In the narrow sense it is a process of determination, in the same way, of future liabilities and their comparison with present assets in order to estimate their sufficiency, deficit of surplus.

We can define general and efficient actuarial calculations, the principals of which are given below.

Efficient actuarial calculations imply calculations of any derivative indicators, which are carried out through conjugation (comparison) of two or more dissimilar initial indicators, the results of which are presented as different relative numbers (coefficients, norms, percents, shares, indices, rates, tariffs, etc.), characterizing differential (effect) of anticipatory increment of one indicator in comparison with another one.

In some cases similar values are called gradients, derivatives (of different orders), elasticity coefficients, or

anticipatory coefficients and can be determined by reference to more complex statistical and mathematical methods including geometrical, differential, integral, and correlation and regression multivariate calculations.

Herewith in case of application of nominal comparison scales for two or more simple values (so called scale of simple interests, which are calculated and represented in terms of current prices) they are determined and operated as it was mentioned by current nominal financial indicators, but in case of real scales application, i.e. scales of so called compound interests, they are calculated and represented in terms of future or current prices, that is real efficient financial indicators.

In case of insurance scheme the calculation of efficient financial indicators signify the special type of financial calculations i.e. actuarial calculations, which imply additional profit (discounts) or demanding compensation of loss (loss, damage or loss of profit) in connection with occurrence of contingency and risks (risk of legislation alteration, exchange rates, devaluation or revaluation, inflation or deflation, changes in efficiency coefficients).

Actuarial calculations represent special branch of activity (usually licensed activity) dealing with market assessment of compliance of current assets of insurance, joint-stock, investment, pension, credit and other financial companies (i.e. companies engaged in credit relations) with future liabilities to the repayment of credit in order to prevent insolvency of a debtor and to provide efficient protection for investors-creditors.

Actuarial calculations assume the comparison of assets (ways of use or allocation of obtained funds) with liabilities (sources of gained funds) for borrowing companies of all types and forms, which are carried out in aggregate by particular items of their expenses under circumstances of mutual risks in order to expose the degree of compliance or incompliance (surplus or deficit) of borrowed assets with future liabilities in term of repayment, in other words to check the solvency of borrowing companies.

Borrowing companies – insurance, stock, broker and auditor firms, banks, mutual, pension, and other specialized investment funds whose accounts payable two or more times exceeds their own assets and appear to be a source of high risk, which in turn affects interests of broad groups of business society as well as population – are considered as companies that are subjects to obligatory insurance and actuarial assessment.

Actuarial calculations assume the construction of balances for future assets and liabilities, probabilistic assessment of future liabilities repayment (debts) at the expense of disposable assets with regard to risks of changes of their amount on hand and market prices. The procedures of documentary adoption, which include construction of actuarial balances and preparation of actuarial reports and conclusions, are called actuarial estimation; the organizations that are carrying out such procedures are called actuarial organizations.

Hence, there is a necessity to learn the organization and technique of actuarial methods (estimations) in aggregate; as well as to introduce the knowledge of actuarial subjects to any expert who is involved in direct actuarial estimations of future assets and liabilities costs of various funds, credit, insurance, and similarly financial companies. This is true for assets and liabilities of any country.

The knowledge of these actuarial assessments and practical use is a significant reserve for increasing not only efficiency but (more important today) legitimate, transparent, and protected futures for both borrowing and lending companies.

Key Terms

Actuary (actuarius – Latin) – profession, appraiser of risks, certified expert on assessment of documentary insurance (and wider – financial) risks; in insurance – insurer; in realty agencies – appraiser; in accounting – auditor; in financial markets – broker (or bookmaker); in the past registrar and holder of insurance documents; in England – adjuster or underwriter.

Actuarial transactions – special field of activity related to determination of insurance outcomes in circumstances of uncertainty that require knowledge of probability theory and actuarial statistics methods and mathematics, including modern computer programs.

Actuarial assessment – type of practical activity, licensed in the majority of countries, related to preparation of actuarial balances, market assessment of current and future costs of assets and liabilities of insurer (in case of pension insurance assets and liabilities of non-governmental pension funds, insurances companies and specialized mutual trust funds); completed with preparation of actuarial report according to standard methodologies and procedures approved, as a rule in conventional (sometimes in legislative) order.

Actuarial estimations – documentary estimations of chance outcomes (betting) of any risk (gambling) actions (games) with participation of two or more parties with fixed (registered) rates of repayment of insurance premium and compensations premium for possible losses. They differ by criteria of complexity – that is elementary (simple or initial) and complex. The most widespread cases of elementary actuarial estimations are bookmaker estimations of profit and loss from different types of gambling including playing cards, lottery, and casinos, as well as risk

taking on modern stock exchange, foreign exchange markets, commodity exchanges, etc. The complex estimations assume determination of profit from second and consequent derived risks (outcomes over outcomes, insurance over insurance, repayment on repayment, transactions with derivatives, etc.). All of these estimations are carried out with the help of various method of high mathematics (first of all, numeric methods of probability theory and mathematical statistics). They are also often represented as methods of high actuarial estimations.

Generally due to ignorance about such estimations, current world debt (in 2008 approximately 700 trillion USD, including 300 trillion USD in the USA) has drastically exceeded real assets, which account for about 65 trillion USD, which is actually causing the enormous financial crisis everywhere in the world.

Usually such estimations are being undertaken towards future insurance operations, profits and losses, and that is why they are classified as strictly approximate and represented in categories of probabilistic expectations.

The fundamental methods of actuarial estimations are the following: methods for valuing investments, selecting portfolios, pricing insurance contracts, estimating reserves, valuing portfolios, controlling pension scheme, finances, asset management, time delays and underwriting cycle, stochastic approach to life insurance mathematics, pension funding and feed back, multiple state and disability insurance, and methods of actuarial balances.

The most popular range of application for actuarial methods are: 1) investments, (actuarial estimations) of investments assets and liabilities, internal and external, real and portfolio types their mathematical methods and models, investments risks and management; 2) life insurance (various types and methods, insurance bonuses, insurance companies and risks, role of the actuarial methods in management of insurance companies and reduction of insurance risks); 3) general insurance (insurance schemes, premium rating, reinsurance, reserving); 4) actuarial provision of pension insurance (pension investments – investment policy, actuarial databases, meeting the cost, actuarial researches).

Scientist who have greatly contributed to actuarial practices: William Morgan, Jacob Bernoulli, A. A. Markov, V. Y. Bunyakovsky, M. E. Atkinson, M. H. Amsler, B. Benjamin, G. Clark, C. Haberman, S. M. Hoem, W. F. Scott, and H. R. Watson.

World's famous actuary's schools and institutes: The Institute of Actuaries in London, Faculty of Actuaries in Edinburgh (on 25 May 2010, following a ballot of Fellows of both institutions, it was announced that the Institute and Faculty would merge to form one body – the "Institute and Faculty of Actuaries"), Charted Insurance Institute, International Association of Actuaries, International Forum of Actuaries Associations, International Congress of Actuaries, and Groupe Consultatif Actuariel Européen.

About the Author

Professor Vassiliy M. Simchera received his PhD at the age of 24 and his Doctor's degree when he was 35. He has been Vice-president of the Russian Academy of Economical Sciences (RAES), Chairman of the Academic Council and Counsel of PhDs dissertations of RAES, Director of Russian State Scientific and Research Statistical Institute of Rosstat (Moscow, from 2000). He was also Head of Chair of statistics in the All-Russian Distant Financial and Statistical Institute (1983–2000), Director of Computer Statistics Department in the State Committee on statistics and techniques of the USSR (1973–1983), and Head of Section of Statistical Researches in the Science Academy of the USSR (1965–1973). He has supervised 8 Doctors and over 50 PhD's. He has (co-) authored over 50 books and 350 articles, including the following books: *Encyclopedia of Statistical Publications* (2001, 991 p., in co-authorship), *Financial and Actuarial Calculations* (2002), *Organization of State Statistics in Russian Federation* (2004) and *Development of Russia's Economy for 100 Years*, 1900–2005 (2006). Professor Simchera was founder and executive director (1987–1991) of Russian Statistical Association, member of various domestic and foreign academies, as well as scientific councils and societies. He has received numerous honors and awards for his work, including Honored Scientist of Russian Federation (2001) (Decree of the President of the Russian Federation) and Saint Nicolay Chudotvoretz honor of III degree (2006). He is a full member of the International Statistical Institute (from 2001).

Cross References

►Careers in Statistics
►Insurance, Statistics in
►Kaplan-Meier Estimator
►Life Table
►Population Projections
►Probability, History of
►Quantitative Risk Management
►Risk Analysis
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts
►Survival Data

References and Further Reading

Benjamin B, Pollard JH (1980) The analysis of mortality and other actuarial statistics, 2nd edn. Heinemann, London

Black K, Skipper HD (1987) Life insurance. Prentice Hall, Englewood Cliffs, New Jersey

Booth P, Chadburn R, Cooper D, Haberman S and James D (1999) Modern actuarial theory and practice. Chapman and Hall/CHC, London, New York

Simchera VM (2003) Introduction to financial and actuarial calculations. Financy and Statistika Publishing House, Moscow

Teugels JL, Sundt B (2004) The encyclopedia of actuarial science, 3 vols. Wiley, Hoboken, NJ

Transactions of International Congress of Actuaries, vol. 1–10; J Inst Actuar, vol. 1–150

Adaptive Linear Regression

Jana Jurečková
Professor
Charles University in Prague, Prague, Czech Republic

Consider a set of data consisting of n observations of a response variable Y and of vector of p explanatory variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$. Their relationship is described by the *linear regression model* (see ►Linear Regression Models)

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e.$$

In terms of the observed data, the model is

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, 2, \ldots, n.$$

The variables $e_1, \ldots, e_n$ are unobservable model errors, which are assumed being independent and identically distributed random variables with a distribution function F and density f. The density is unknown, we only assume that it is symmetric around 0. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is an unknown parameter, and the problem of interest is to estimate $\boldsymbol{\beta}$ based on observations $Y_1, \ldots, Y_n$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, $i = 1, \ldots, n$.

Besides the classical ►least squares estimator, there exists a big variety of *robust estimators* of $\boldsymbol{\beta}$. Some are distributionally robust (less sensitive to deviations from the assumed shape of f), others are resistant to the leverage points in the design matrix and have a high breakdown point [introduced originally by Hampel (1968), the finite sample version is studied in Donoho and Huber (1983)].

The last 40 years brought a host of statistical procedures, many of them enjoying excellent properties and being equipped with a computational software (see ►Computational Statistics and ►Statistical Software: An Overview). On the other hand, this progress has put an applied statistician into a difficult situation: If one needs to fit the data with a regression hyperplane, he (she) is hesitating which procedure to use. If there is more information on the model, then the estimation procedure can be chosen accordingly. If the data are automatically collected by a computer and the statistician is not able to make any diagnostics, then he (she) might use one of the high breakdown-point estimators. However, many decline this idea due to the difficult computation. Then, at the end, the statistician can prefer the simplicity to the optimality and uses either the classical least squares (LS), LAD-method or other reasonably simple method.

Instead of to fix ourselves on one fixed method, one can try to combine two convenient estimation methods, and in this way diminish eventual shortages of both. Taylor (1973) suggested to combine the LAD (minimizing the L_1 norm) and the least squares (minimizing the L_2 norm) methods. Arthanari and Dodge (1981) considered a convex combination of LAD- and LS-methods. Simulation study by Dodge and Lindstrom (1981) showed that this procedure is robust to small deviations from the normal distribution (see ►Normal Distribution, Univariate). Dodge (1984) extended this method to a convex combination of LAD and Huber's M-estimation methods (see ►Robust Statistics and Robust Statistical Methods). Dodge and Jurečková (1987) observed that the convex combination of two methods could be adapted in such a way that the resulted estimator has the minimal asymptotic variance in the class of estimators of a similar kind, no matter what is the unknown distribution. The first numerical study of this procedure was made by Dodge et al. (1991). Dodge and Jurečková (1988, 1991) then extended the adaptive procedure to the combinations of LAD- with M-estimation and with the trimmed least squares estimation. The results and examples are summarized in monograph of Dodge and Jurečková (2000), where are many references added.

Let us describe the general idea, leading to a construction of an adaptive convex combination of two estimation methods: We consider a family of symmetric densities indexed by an suitable measure of scale s :

$$\mathcal{F} = \left\{ f : f(z) = s^{-1} f_0(z/s),\ s > 0 \right\}.$$

The shape of f_0 is generally unknown; it only satisfies some regularity conditions and the unit element $f_0 \in \mathcal{F}$ has the scale $s_0 = 1$. We take $s = 1/f(0)$ when we combine L_1-estimator with other class of estimators.

The scale characteristic s is estimated by a consistent estimator $\hat{s}_n$ based on $Y_1, \ldots, Y_n$, which is regression-invariant and scale-equivariant, i.e.,

(a) $\hat{s}_n(\mathbf{Y}) \xrightarrow{p} s$ as $n \to \infty$
(b) $\hat{s}_n(\mathbf{Y} + \mathbf{Xb}) = \hat{s}_n(\mathbf{Y})$ for any $\mathbf{b} \in \mathbb{R}^p$ (regression-invariance)
(c) $\hat{s}_n(c\mathbf{Y}) = c\hat{s}_n(\mathbf{Y})$ for $c > 0$ (scale-equivariance).

Such estimator based on the regression quantiles was constructed e.g., by Dodge and Jurečková (1995). Other estimators are described in the monograph by Koenker (2005).

The adaptive estimator $\mathbf{T}_n(\delta)$ of $\boldsymbol{\beta}$ is defined as a solution of the minimization problem

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_i^\top \mathbf{t}}{\hat{s}_n}\right) := \min$$

with respect to $\mathbf{t} \in \mathbb{R}^p$, where

$$\rho(z) = \delta\rho_1(z) + (1-\delta)\rho_2(z) \tag{1}$$

with a suitable fixed δ, $0 \le \delta \le 1$, where $\rho_1(z)$ and $\rho_2(z)$ are symmetric (convex) discrepancy functions defining the respective estimators. For instance, $\rho_1(z) = |z|$ and $\rho_2(z) = z^2$ if we want to combine LAD and LS estimators. Then $\sqrt{n}(\mathbf{T}_n(\delta) - \boldsymbol{\beta})$ has an asymptotically normal distribution (see ►Asymptotic Normality) $\mathcal{N}_p(\mathbf{0}, \mathbf{Q}^{-1}\sigma^2(\delta,\rho,f))$ with the variance dependent on δ, ρ and f, where

$$\mathbf{Q} = \lim_{n\to\infty} \mathbf{Q}_n, \quad \mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top.$$

Using $\delta = \delta_0$ which minimizes $\sigma^2(\delta,\rho,f)$ with respect to δ, $0 \le \delta \le 1$, we get an estimator $\mathbf{T}_n(\delta_0)$ minimizing the asymptotic variance for a fixed distribution shape. Typically, $\sigma^2(\delta,\rho,f)$ depends on f only through two moments of f_0. However, these moments should be estimated on the data.

Let us illustrate the procedure on the combination of the least squares and the L_1 procedures. Set

$$\sigma^2 = \int z^2 f(z)dz, \quad \sigma_0^2 = \int z^2 f_0(z)dz \tag{2}$$
$$E_1^0 = \int |z| f_0(z)dz, \quad E_1 = \int |z| f(z)dz.$$

Then

$$\sigma^2 = \int z^2 f(z)dz = s^2\sigma_0^2, \qquad E_1 = \int |z| f(z)dz = sE_1^0$$

and the corresponding asymptotic variance of $\mathbf{T}_n(\delta)$ is

$$\sigma^2(\delta,f,s) = \frac{s^2}{4}\{4(1-\delta)^2\sigma_0^2 + 4\delta(1-\delta)E_1^0 + \delta^2\}. \tag{3}$$

If we know all moments in (2), we minimize the variance (3) with respect to δ, under the restriction $0 \le \delta \le 1$. It is minimized for $\delta = \delta_0$ where

$$\delta_0 = \begin{cases} 0 & \text{if } 2\sigma_0^2 \le E_1^0 < 1/2 \\ \dfrac{4\sigma_0^2 - 2E_1^0}{4\sigma_0^2 - 4E_1^0 + 1} & \text{if } E_1^0 < 1/2 \text{ and } E_1^0 < 2\sigma_0^2 \\ 1 & \text{if } 1/2 \le E_1^0 < 2\sigma_0^2. \end{cases}$$

The estimator $\mathbf{T}_n(\delta_0)$ of $\boldsymbol{\beta}$ is a solution of the minimization

$$\sum_{i=1}^{n} \rho((Y_i - \mathbf{x}_i^\top \mathbf{t})/\hat{s}_n) := \min, \quad \mathbf{t} \in \mathbb{R}^p,$$
$$\rho(z) = (1-\delta_0)z^2 + \delta_0|z|, \; z \in \mathbb{R}^1. \tag{4}$$

But δ_0 is unknown, because the entities in (2) depend on the unknown distribution f. Hence, we should replace δ_0 by an appropriate estimator based on $\mathbf{Y}$. We shall proceed in the following way:

First estimate $E_1^0 = E_1/s = f(0)\int_{\mathbb{R}} |z| f(z)dz$ by

$$\widehat{E}_1^0 = \hat{s}_n^{-1}(n-p)^{-1}\sum_{i=1}^{n}\left|Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n\left(\frac{1}{2}\right)\right| \tag{5}$$

where $\widehat{\boldsymbol{\beta}}_n(1/2)$ is the LAD-estimator of $\boldsymbol{\beta}$. The choice of optimal $\widehat{\delta}_{0n}$ is then based on the following decision procedure (Table 1).

It can be proved that $\widehat{\delta}_{0n} \xrightarrow{p} \delta_0$ as $n \to \infty$ and $\mathbf{T}_n(\widehat{\delta}_{0n})$ is a consistent estimator of $\boldsymbol{\beta}$ and is asymptotically normally distributed with the minimum possible variance.

Adaptive Linear Regression. Table 1 Decision procedure

Compute $\widehat{E}_1^0$ as in (5).

(1) If $\widehat{E}_{1n}^0 < 1/2$, calculate

$$\widehat{\sigma}_{0n}^2 = \frac{1}{\hat{s}_n^2(n-p)}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}(1/2)\right)^2$$

and go to (2). If not, go to (4).

(2) If $\widehat{E}_{1n}^0 \ge 2\widehat{\sigma}_{0n}^2$, put $\widehat{\delta}_{0n} = 0$. Then $\mathbf{T}_n$ is the ordinary LS estimator of $\boldsymbol{\beta}$. If not, go to (3).

(3) If $\widehat{E}_{1n}^0 < 2\widehat{\sigma}_{0n}^2$, calculate

$$\widehat{\delta}_{0n} = \frac{4\widehat{\sigma}_{0n}^2 - 2\widehat{E}_{1n}^0}{4\widehat{\sigma}_{0n}^2 - 4\widehat{E}_{1n}^0 + 1}$$

and perform the minimization (4) with the function ρ equal to

$$(1-\widehat{\delta}_{0n})\sum_{i=1}^{n}\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{\hat{s}_n}\right)^2 + \widehat{\delta}_{0n}\sum_{i=1}^{n}\left|\frac{Y_i - \mathbf{x}_i^\top\mathbf{t}}{\hat{s}_n}\right|.$$

(4) Put $\widehat{\delta}_{0n} = 1$; then $\mathbf{T}_n$ is the LAD-estimate of $\boldsymbol{\beta}$.

Many numerical examp les based on real data can be find in the monograph Dodge and Jurečková (2000).

Acknowledgments

The research was supported by the Czech Republic Grant 201/09/0133 and by Research Projects MSM 0021620839 and LC 06024.

About the Author

Jana Jurečková was born on September 20, 1940 in Prague, Czechoslovakia. She has her Ph.D. in Statistics from Czechoslovak Academy of Sciences in 1967; some twenty years later, she was awarded the DrSc from Charles University. Her dissertation, under the able supervision of late Jaroslav Hajek, related to "uniform asymptotic linearity of rank statistics" and this central theme led to significant developments in nonparametrics, robust statistics, time series, and other related fields. She has extensively collaborated with other leading statisticians in Russia, USA, Canada, Australia, Germany, Belgium, and of course, Czech Republic, among other places. A (co-)author of several advanced monographs and texts in Statistics, Jana has earned excellent international reputation for her scholarly work, her professional accomplishment and her devotion to academic teaching and councelling. She has been with the Mathematics and Physics faculty at Charles University, Prague, since 1967, where she earned the Full Professor's rank in 1992. She has over 100 publications in the leading international journals in statistics and probability, and she has supervised a number of Ph.D. students, some of them have acquired international reputation on their own. (Communicated by P. K. Sen.)

Cross References

►Robust Regression Estimation in Generalized Linear Models
►Robust Statistical Methods
►Robust Statistics

References and Further Reading

Arthanari TS, Dodge Y (1981) Mathematical programming in statistics. Wily, Interscience Division, New York; (1993) Wiley Classic Library

Dodge Y (1984) Robust estimation of regression coefficient by minimizing a convex combination of least squares and least absolute deviations. Comp Stat Quart 1:139–153

Dodge Y, Jurečková J (1987) Adaptive combination of least squares and least absolute deviations estimators. In: Dodge Y (ed) Statistical data analysis based on L_1 – norm and related methods. North-Holland, Amsterdam, pp 275–284

Dodge Y, Jurečková J (1988) Adaptive combination of M-estimator and L_1 – estimator in the linear model. In: Dodge Y, Fedorov VV, Wynn HP (eds) Optimal design and analysis of experiments. North-Holland, Amsterdam, pp 167–176

Dodge Y, Jurečková J (1991) Flexible L-estimation in the linear model. Comp Stat Data Anal 12:211–220

Dodge Y, Jurečková J (1995) Estimation of quantile density function based on regression quantiles. Stat Probab Lett 23: 73–78

Dodge Y, Jurečková J (2000) Adaptive regression. Springer, New York. ISBN 0-387-98965-X

Dodge Y, Lindstrom FT (1981) An alternative to least squares estimations when dealing with contaminated data. Technical report No 79, Oregon State University, Corvallis

Dodge Y, Antoch J, Jurečková J (1991) Adaptive combination of least squares and least absolute deviation estimators. Comp State Data Anal 12:87–99

Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds) A festschrift for Erich Lehmann. Wadsworth, Belmont, California

Hampel FR (1968) Contributions to the theory of robust estimation. PhD Thesis, University of California, Berkely

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge. ISBN 0-521-84573-4

Taylor LD (1973) Estimation by minimizing the sum of absolute errors. In: Zarembka P (ed) Frontiers in econometrics. Academic, New York, pp 189–190

Adaptive Methods

Saïd El Melhaoui
Professor Assistant
Université Mohammed Premier, Oujda, Morocco

Introduction

Statistical procedures, the efficiencies of which are optimal and invariant with regard to the knowledge or not of certain features of the data, are called adaptive statistical methods.

Such procedures should be used when one suspects that the usual inference assumptions, for example, the normality of the error's distribution, may not be met. Indeed, traditional methods have a serious defect. If the distribution of the error is non-normal, the power of classical tests, as *pseudo-Gaussian tests*, can be much less than the optimal power. So, the variance of the classical least squares estimator is much bigger than the smallest possible variance.

What Is Adaptivity?

The adaptive methods deal with the problem of estimating and testing hypotheses about a parameter of interest θ in the presence of nuisance parameter ν. The fact that ν remains unspecified induces, in general, a loss of efficiency

with the situation where ν is exactly specified. *Adaptivity* occurs when the loss of efficiency is null, i.e., when we can estimate (testing hypotheses about) θ as when not knowing ν as well as when knowing ν. The method used in this respect is called *adaptive*.

Adaptivity is a property of the model under study, the best known of which is undoubtedly the symmetric location model; see Stone (1975). However, under a totally unspecified density, possibly non-symmetric, the mean can not be adaptively estimated.

Approaches to Adaptive Inference

Approaches to adaptive inference mainly belong to one of two types: either to estimate the unknown parameters ν in some way, or to use the data itself to determine which statistical procedure is the most appropriate to these data. These two approaches are the starting points of two rather distinct strands of the statistical literature. *Nonparametric adaptive inference*, on one hand, where ν is estimated from the sample, and on the other hand, *data-driven methods*, where the shape of ν is identified via a selection statistic to distinguish the effective statistical procedure suitable at the current data.

Nonparametric Methods

The first approach is often used for the *semiparametric model*, where θ is a Euclidean parameter and the nuisance parameter is an infinite dimensional parameter f - often, the unspecified density of some white noise underlying the data generating process.

Stein (1956) introduced the notion of adaptation and gave a simple necessary condition for adaptation in semiparametric models. A comprehensive account of adaptive inference can be found in the monograph by Bickel et al. (1993) for semiparametric models with independent observations. Adaptive inference for dependent data have been studied in a series of papers, e.g., Kreiss (1987), Drost et al. (1997), and Koul and Schick (1997). The current state of the art is summarized in Grenwood et al. (2004).

The basic idea in this literature is to estimate the underlying f using a portion of the sample, and to reduce locally and asymptotically the semiparametric problem to a simpler parametric one, through the so-called "*least favorable parametric submodel*" argument. In general, the resulting computations are non-trivial.

An alternative technique is the use of *adaptive rank based statistics*. Hallin and Werker (2003) proposed a sufficient condition for adaptivity; that is, adaptivity occurs if a parametrically efficient method based on rank statistics can be derived. Then, it suffices, to substitute f in the rank statistics by an estimate $\hat{f}$ measurable on the ►*order statistics*. Some results in this direction have been obtained by Hájek (1962), Beran (1974), and Allal and El Melhaoui (2006).

Finally, these nonparametric adaptive methods, when they exist, are robust in efficiency: they cannot be outperformed by any non-adaptive method. However, these methods have not been widely used in practice, because the estimation of density, typically, requires a large number of observations.

Data-Driven Methods

The second strand of literature addresses the same problem of constructing adaptive inference, and consists of the use of the data to determine which statistical procedure should be used and then using the data again to carry out the procedure.

The was first proposed by Randles and Hogg (1973). Hogg et al. (1975) used the measure of symmetry and tailweight as selection statistics in and adaptive two-sample test. If the selection fell into one of the regions defined by the adaptive procedure, then a certain set of rank scores was selected, whereas if the selection statistic fell into a different region, then different rank scores would be used in the test. Hogg and Lenth (1984) proposed an adaptive estimator of the mean of symmetric distribution. They used selection statistics to determine if a mean, a 25% trimmed mean, or median should be used as an estimate of the mean of population. O'Gorman (2002) proposed an adaptive procedure that performs the commonly used tests of significance, including the two-sample test, a test for a slope in linear regression, and a test for interaction in two-way factorial design. A comprehensive account of this approach can be found in the monograph by O'Gorman (2004).

The advantage of the data-driven methods is that if an adaptive method is properly constructed, it automatically downweight outliers and could easily be applied in practice. However, and contrary to the nonparametric approach, the adaptive data-driven method is the best among the existing procedures, but not the best that can be built. As a consequence, the method so built is not definitively optimal.

Cross References

►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Robust Inference
►Robust Statistical Methods
►Robust Statistics

References and Further Reading

Allal J, El Melhaoui S (2006) Tests de rangs adaptatifs pour les modèles de régression linéaires avec erreurs ARMA. Annales des Sciences Mathématiques du Québec 30:29–54

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Annals of Statistics 2:63–74

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press, Baltimore, New York

Drost FC, Klaassen CAJ, Ritov Y, Werker BJM (1997) Adaptive estimation in time-series models. Ann Math Stat 29: 786–818

Greenwood PE, Muller UU, Wefelmeyer W (2004) An introduction to efficient estimation for semiparametric time series. In: Nikulin MS, Balakrishnan N, Mesbah M, Limnios N (eds) Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Statistics for Industry and Technology, Birkhäuser, Boston, pp. 253–269

Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147

Hallin M, Werker BJM (2003) Semiparametric Efficiency Distribution-Freeness, and Invariance. Bernoulli 9:137–165

Hogg RV, Fisher DM, Randles RH (1975) A two simple adaptive distribution-free tests. J Am Stat Assoc 70:656–661

Hogg RV, Lenth RV (1984) A review of some adaptive statistical techniques. Commun Stat – Theory Methods 13:1551–1579

Koul HL, Schick A (1997) Efficient estimation in nonlinear autoregressive time-series models. Bernoulli 3:247–277

Kreiss JP (1987) On adaptive estimation in stationary ARMA processes. Ann Stat 15:112–133

O'Gorman TW (2002) An adaptive test of significance for a subset of regression coefficients. Stat Med 21:3527–3542

O'Gorman TW (2004) Applied adaptive statistical methods: tests of significance and confidence intervals. Society for Industrial and Applied Mathematics, Philadelphia

Randles RH, Hogg RV (1973) Adaptive distribution-free tests. Commun Stat 2:337–356

Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of Third Berkeley Symposium on Mathametical Statistics and Probability, University of California Press, Berkeley, vol 1, pp. 187–195

Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284

Adaptive Sampling

George A. F. Seber[1], Mohammad Salehi M.[2]
[1]Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand
[2]Professor
Isfahan University of Technology, Isfahan, Iran

Adaptive sampling is particularly useful for sampling populations that are sparse but clustered. For example, fish can form large, widely scattered schools with few fish in between. Applying standard sampling methods such as simple random sampling (SRS, see ▶Simple Random Sample) to get a sample of plots from such a population could yield little information, with most of the plots being empty. The idea can be simply described follows. We go fishing in a lake using a boat and, assuming complete ignorance about the population, we select a location at random and fish. If we don't catch anything we select another location at random and try again. If we do catch something we fish in a specific neighborhood of that location and keep expanding the neighborhood until we catch no more fish. We then move on to a second location. This process continues until we have, for example, fished at a fixed number of locations or until our total catch has exceeded a certain number of fish. This kind of technique where the sampling is adapted to what turns up at each stage has been applied to a variety of diverse populations such as marine life, birds, mineral deposits, animal habitats, forests, and rare infectious diseases, and to pollution studies.

We now break down this process into components and introduce some general notation. Our initial focus will be on adaptive ▶cluster sampling, the most popular of the adaptive methods developed by Steven Thompson in the 1990s. Suppose we have a population of N plots and let y_i be a variable that we measure on the ith plot ($i = 1, 2, \ldots, N$). This variable can be continuous (e.g., level of pollution or biomass), discrete (e.g., number of animals or plants), or even just an indicator variable taking the value 1 for presence and zero for absence. Our aim is to estimate some function of the population y values such as, for example, the population total $\tau = \sum_{i=1}^{N} y_i$, the population mean $\mu = \tau/N$, or the population density $D = \tau/A$, where A is the population area.

The next step is to determine the nature of the neighborhood of each initially chosen plot. For example, we could choose all the adjacent units with a common boundary which, together with unit i, form a "cross" Neighborhoods can be defined to have a variety of patterns and the units in a neighborhood do not have to be contiguous (next to each other). We then specify a condition C such as $y_i > c$ which determines when we sample the neighborhood of the ith plot; typically $c = 0$ if y is a count. If C for the ith plot or unit is satisfied, we sample all the units in the neighborhood and if the rule is satisfied for any of those units we sample their neighborhoods as well, and so on, thus leading to a cluster of units. This cluster has the property that all the units on its "boundary" (called "edge units") do not satisfy C. Because of a dual role played by the edge units, the underlying theory is based on the concept of a network, which is a cluster minus its edge units.

It should be noted that if the initial unit selected is any one of the units in the cluster except an edge unit, then

all the units in the cluster end up being sampled. Clearly, if the unit is chosen at random, the probability of selecting the cluster will depend on the size of the cluster. For this reason adaptive cluster sampling can be described as unequal probability cluster sampling – a form of biased sampling.

The final step is to decide how we choose both the size and the method of selecting the initial sample size. Focusing on the second of these for the moment, one simple approach would be to use SRS to get a sample of size n_1, say. If a unit selected in the initial sample does not satisfy C, then there is no augmentation and we have a cluster of size one. We note that even if the units in the initial sample are distinct, as in SRS, repeats can occur in the final sample as clusters may overlap on their edge units or even coincide. For example, if two non-edge units in the same cluster are selected in the initial sample, then that whole cluster occurs twice in the final sample. The final sample then consists of n_1 (not necessarily distinct) clusters, one for each unit selected in the initial sample. We finally end up with a total of n units, which is random, and some units may be repeated.

There are many modifications of the above scheme depending on the nature of the population and we mention just a few. For example, the initial sample may be selected by sampling with replacement, or by using a form of systematic sampling (with a random start) or by using unequal probability sampling, as in sampling a tree with probability proportional to its basal area. Larger initial sampling units other than single plots can be used, for example a strip transect (primary unit) commonly used in both aerial and ship surveys of animals and marine mammals. Other shaped primary units can also be used and units in the primary unit need not be contiguous. If the population is divided into strata, then adaptive cluster sampling can be applied within each stratum, and the individual estimates combined. How they are combined depends on whether clusters are allowed to cross stratum boundaries or not. If instead of strata, we simply have a number of same-size primary units and choose a sample of primary units at random, and then apply the adaptive sampling within each of the chosen primary units, we have two-stage sampling with its appropriate theory.

In some situations, the choice of c in condition C is problematical as, with a wrong choice, we may end up with a feast or famine of plots. Thompson suggested using the data themselves, in fact the ►order statistics for the y_i values in the initial sample. Sometimes animals are not always detected and the theory has been modified to allow for incomplete detectability. If we replace y_i by a vector, then the scheme can be modified to allow for multivariate data.

We now turn our attention to sample sizes. Several ways of controlling sample sizes have been developed. For example, to avoid duplication we can remove a network once it has been selected by sampling networks without replacement. Sequential methods can also be used, such as selecting the initial sample sequentially until n exceeds some value. In fact Salehi, in collaboration with various other authors has developed a number of methods using both inverse and sequential schemes. One critical question remains: How can we use a pilot survey to design an experiment with a given efficiency or expected cost? One solution has been provided using the two-stage sampling method mentioned above (Salehi and Seber 1997).

We have not said anything about actual estimates as this would take several pages. However, a number of estimates associated with the authors Horvitz-Thompson (see ►Horvitz–Thompson Estimator), Hansen-Hurwitz, and Murthy have all been adapted to provide unbiased estimates for virtually all the above schemes and modifications. Salehi (1999) has also used the famous ►Rao-Blackwell theorem to provide more efficient unbiased estimates in a number of cases. The mentioned estimators based on small samples under adaptive cluster sampling often have highly skewed distributions. In such situations, confidence intervals (see ►Confidence Interval) based on traditional normal approximation can lead to unsatisfactory results, with poor coverage properties; for another solution see Salehi et al. (2009a).

As you can see, the topic is rich in applications and modifications and we have only told part of the story! For example, there is a related topic called adaptive allocation that has been used in fisheries; for a short review of adaptive allocation designs see Salehi et al. (2009b). Extensive references to the above are Thompson and Seber (1996) and Seber and Salehi (2004).

About the Author

Professor Seber was appointed to the foundation Chair in Statistics and Head of a newly created Statistics Unit within the Mathematics Department at the University of Auckland in 1973. He was involved in forming a separate Department of Statistics in 1994. He was awarded the Hector Medal by the Royal Society of New Zealand for fundamental contributions to statistical theory, for the development of the statistics profession in New Zealand, and for the advancement of statistics education through his teaching and writing (1999). He has authored or coauthored ten books as well as several second editions, and numerous research papers. However, despite the breadth of his contribution from linear models, multivariate statistics, linear regression, non-linear models, to adaptive sampling, he is perhaps still best known internationally for his research

on the estimation of animal abundance. He is the author of the internationally recognized text *Estimation of Animal Abundance and Related Parameters* (Wiley, 2nd edit., 1994; paperback reprint, Blackburn, 2002). The third conference on Statistics in Ecology and Environmental Monitoring was held in Dunedin (1999) "to mark and recapture the contribution of Professor George Seber to Statistical Ecology."

Cross References

►Cluster Sampling
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Statistical Ecology

References and Further Reading

Salehi MM (1999) Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling. J Environ Ecol Stat 6:183–195
Salehi MM, Seber GAF (1997) Two stage adaptive cluster sampling. Biometrics 53:959–970
Salehi MM, Mohammadi M, Rao JNK, Berger YG (2010a) Empirical Likelihood confidence intervals for adaptive cluster sampling. J Environ Ecol Stat 17:111–123
Salehi MM, Moradi M, Brown JA, Smith DR (2010b) Efficient estimators for adaptive two-stage sequential sampling. J Stat Comput Sim, DOI: 10.1080/00949650903005664
Seber GAF, Salehi MM (2004) Adaptive sampling. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 1, 2nd edn. Wiley, New York
Thompson SK, Seber GAF (1996) Adaptive sampling. Wiley, New York

Advantages of Bayesian Structuring: Estimating Ranks and Histograms

Thomas A. Louis
Professsor
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Introduction

Methods developed using the Bayesian formalism can be very effective in addressing both Bayesian and frequentist goals. These advantages are conferred by full probability modeling are most apparent in the context of ►nonlinear models or in addressing non-standard goals. Once the likelihood and the prior have been specified and data observed, ►Bayes' Theorem maps the prior distribution into the posterior. Then, inferences are computed from the posterior, possibly guided by a ►loss function. This last step allows proper processing for complicated, nonintuitive goals. In this context, we show how the Bayesian approach is effective in estimating ►ranks and CDFs (histograms). We give the basic ideas; see Lin et al. (2006, 2008); Paddock et al. (2006) and the references thereof for full details and generalizations.

Importantly, as Carlin and Louis (2009) and many authors caution, the Bayesian approach is not a panacea. Indeed, the requirements for an effective procedure are more demanding than those for a frequentist approach. However, the benefits are many and generally worth the effort, especially now that ►Markov Chain Monte Carlo (MCMC) and other computing innovations are available.

A Basic Hierarchical Model

Consider a basic, compound sampling model with parameters of interest $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and data $\mathbf{Y} = (Y_1, \ldots, Y_K)$. The θ_k are *iid* and conditional on the θs, the Y_k are independent.

$$\begin{aligned} \theta_k &\overset{iid}{\sim} G(\cdot) \\ Y_k|\theta_k &\overset{indep}{\sim} f_k(Y_k|\theta_k) \end{aligned} \tag{1}$$

in practice, the θ_k might be the true differential expression of the *kth* gene, the true standardized mortality ratio for the *kth* dialysis clinic, or the true, underlying region-specific disease rate. Generalizations of (1) include adding a third stage to represent uncertainty in the prior, a regression model in the prior, or a priori association among the θs.

Assume that the θ_k and $\boldsymbol{\eta}$ are continuous random variables. Then, their posterior distribution is,

$$g(\boldsymbol{\theta} \mid \mathbf{Y}) = \prod_1^K g(\theta_k \mid Y_k) \tag{2}$$

$$g(\theta_k \mid Y_k) = \frac{f_k(Y_k \mid \theta_k) g(\theta_k)}{\int f_k(Y_k \mid s) g(s) ds} = \frac{f_k(Y_k \mid \theta_k) g(\theta_k)}{f_G(Y_k)}$$

Ranking

The ranking goal nicely shows the beauty of Bayesian structuring. Following Shen and Louis (1998), if the θ_k were directly observed, then their ranks (R_k) and percentiles (P_k) are:

$$R_k(\boldsymbol{\theta}) = \text{rank}(\theta_k) = \sum_{j=1}^K I_{\{\theta_k \geq \theta_j\}}; \; P_k(\boldsymbol{\theta}) = R_k(\boldsymbol{\theta})/(K+1). \tag{3}$$

The smallest θ has rank 1 and the largest has rank K. Note that the ranks are monotone transform invariant (e.g., ranking the logs of parameters produces the original ranks) and estimated ranks should preserve this invariance. In practice, we don't get to observe the θ_k, but can use their posterior distribution (2) to make inferences. For example, minimizing posterior squared-error loss *for the ranks* produces,

$$\bar{R}_k(\mathbf{Y}) = E_{\boldsymbol{\theta}|\mathbf{Y}}[R_k(\boldsymbol{\theta}) \mid \mathbf{Y}] = \sum_{j=1}^{K} \text{pr}(\theta_k \geq \theta_j \mid \mathbf{Y}). \qquad (4)$$

The $\bar{R}_k$ are shrunk towards the mid-rank, $(K+1)/2$, and generally are not integers. Optimal integer ranks result from ranking the $\bar{R}_k$, producing,

$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})); \; \hat{P}_k = \hat{R}_k/(K+1). \qquad (5)$$

Unless the posterior distributions of the θ_k are stochastically ordered, ranks based on maximum likelihood estimates or those based on hypothesis test statistics perform poorly. For example, if all θ_k are equal, MLEs with relatively high variance will tend to be ranked at the extremes; if Z-scores testing the hypothesis that a θ_k is equal to the typical value are used, then the units with relatively small variance will tend to be at the extremes. Optimal ranks compromise between these two extremes, a compromise best structured by minimizing posterior expected loss in the Bayesian context.

Example: The basic Gaussian-Gaussian model

We specialize (1) to the model with a Gaussian prior and Gaussian sampling distributions, with possibly different sampling variances. Without loss of generality assume that the prior mean is $\mu = 0$ and the prior variance is $\tau^2 = 1$. We have,

$$\begin{aligned} \theta_k \; &iid \; N(0,1), \\ Y_k|\theta_k &\sim N(\theta_k, \sigma_k^2) \\ \theta_k \mid Y_k \; &ind \; N\left(\theta_k^{pm}, (1-B_k)\sigma_k^2\right) \\ \theta_k^{pm} &= (1-B_k)Y_k; \; B_k = \sigma_k^2/(\sigma_k^2+1). \end{aligned}$$

The σ_k^2 are an ordered, geometric sequence with ratio of the largest σ^2 to the smallest $rls = \sigma_K^2/\sigma_1^2$ and ▶geometric mean $gmv = GM(\sigma_1^2, \ldots, \sigma_K^2)$. When $rls = 1$, the σ_k^2 are all equal. The quantity gmv measures the typical sampling variance and here we consider only $gmv = 1$.

Table 1 documents *SEL* performance for $\hat{P}_k$ (the optimal approach), Y_k (the MLE), ranked θ_k^{pm} and ranked $\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$ (the posterior mean of e^{θ_k}). We present this last to assess performance for a monotone, non-linear transform of the target parameters. For $rls = 1$, the posterior distributions are stochastically ordered and the four sets of percentiles are identical, as is their performance. As rls increases, performance of Y_k-derived percentiles degrades, those based on the θ_k^{pm} are quite competitive with $\hat{P}_k$, but performance for percentiles based on the posterior mean of e^{θ_k} degrades as rls increases. Results show that though the posterior mean can perform well, in general it is not competitive with the optimal approach.

Advantages of Bayesian Structuring: Estimating Ranks and Histograms. Table 1 Simulated preposterior 10, 000 × *SEL* for $gmv = 1$. As a baseline for comparison, if the data provided no information on the θ_k ($gmv = \infty$), all entries would be 833. If the data provided perfect information ($gmv = 0$), all entries would be 0

	Percentiles based on			
rls	$\hat{P}_k$	θ_k^{pm}	$\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$	Y_k
1	516	516	516	516
25	517	517	534	582
100	522	525	547	644

Estimating the CDF or Histogram

Similar advantages of the Bayesian approach apply to estimating the empirical distribution function (EDF) of the θ_k,

$$G_K(t \mid \boldsymbol{\theta}) = K^{-1} \sum I_{\{\theta_k \leq t\}}.$$

As shown by Shen and Louis (1998), The optimal SEL estimate is

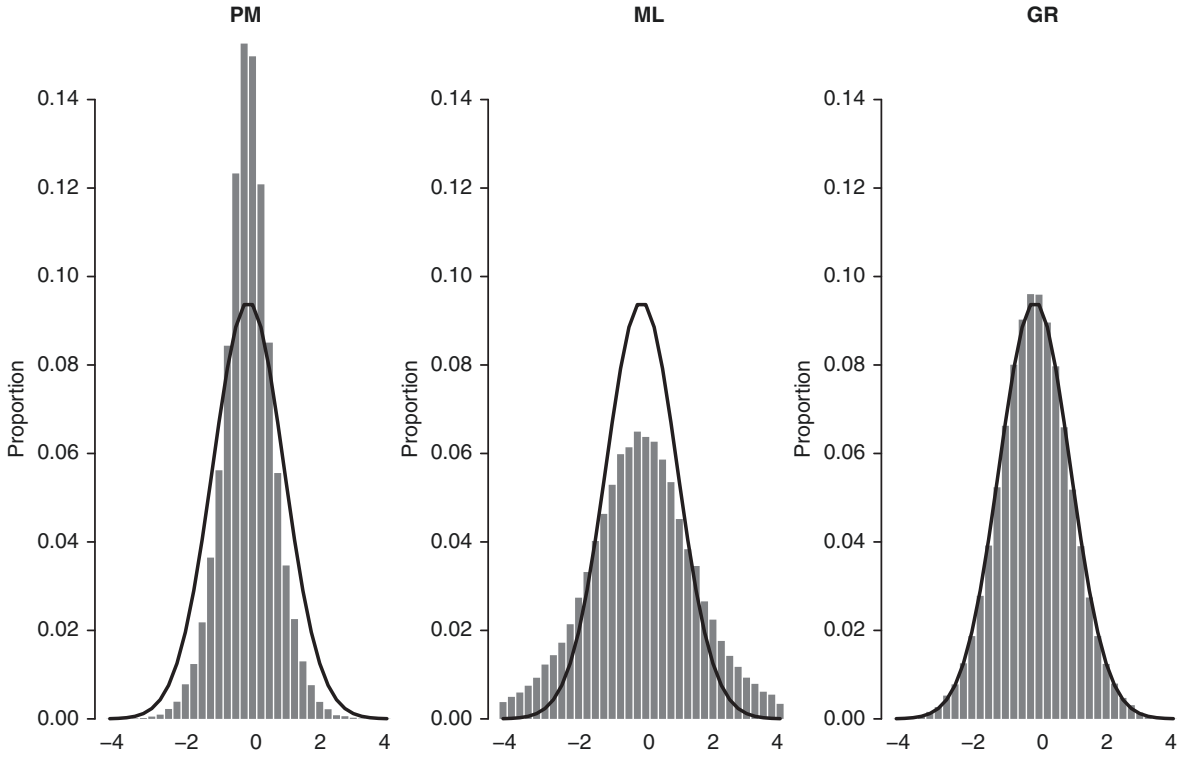
$$\bar{G}_K(t|\mathbf{Y}) = E[G_K(t \mid \boldsymbol{\theta})|\mathbf{Y}] = K^{-1} \sum Pr(\theta_k \leq t|\mathbf{Y}).$$

The optimal discrete distribution estimate with at most K mass points is $\hat{G}_K$, with mass K^{-1} at

$$\hat{U}_j = \bar{G}_K^{-1}\left(\frac{2j-1}{2K}\middle|\mathbf{Y}\right), \qquad j = 1, \ldots, K$$

The EDF is easy to compute from MCMC sampling. After burn-in, pool all θs, order them and set U_j equal to the $(2j-1)th$ order statistic.

Bayesian structuring to estimate G_K pays big dividends. As shown in Fig. 1, for the basic Gaussian model it produces the correct spread, whereas the histogram of the θ_k^{pm} (the posterior means) is under-dispersed and that of the Y_k (the MLEs) is over dispersed. More generally, when the true EDF is asymmetric or multi-modal,



Advantages of Bayesian Structuring: Estimating Ranks and Histograms. Fig. 1 Histogram estimates using θ^{pm}, ML, and $\bar{G}_K$ for the basic Gaussian/Gaussian model. $GM(\{\sigma_k^2\}) = 1$, $rls = 100$

the Bayesian approach also produces the correct shape Paddock et al. (2006).

Discussion

The foregoing are but two examples of the effectiveness of Bayesian structuring. Many more are available in the cited references and in other literature. In closing, we reiterate that the Bayesian approach needs to be used with care; there is nothing automatic about realizing its benefits.

Acknowledgments

Research supported by NIH/NIDDK Grant 5R01DK061662.

About the Author

Dr. Thomas Louis is Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health. He was President, International Biometric Society (IBS), Eastern North American Region (1992) and President, International Biometric Society (2006–2007). He is a Fellow of the American Statistical Association (1988), American Association for the Advancement of Science (1996), and Elected member, International Statistical Institute (1985). He was Editor, *JASA Applications and Case Studies* (2001–2003), Currently he is Co-editor, *Biometrics* (2009–2011). He is principal or co-advisor for 65 doctoral students and more than 40 masters students. He has delivered more than 450 invited presentations. Professor Louis has (co-)authored about 170 refereed papers and books, including *Bayesian Methods for Data Analysis* (with B.P. Carlin, Chapman & Hall/CRC, 3rd edition, 2009).

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Prior Bayes: Rubin's View of Statistics

References and Further Reading

Carlin BP, Louis TA (2009) Bayesian methods for data analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton

Lin R, Louis TA, Paddock SM, Ridgeway G (2006) Loss function based ranking in two-stage, hierarchical models. Bayesian Anal 1:915–946

Lin R, Louis TA, Paddock SM, Ridgeway G (2009) Ranking of USRDS, provider-specific SMRs from 1998–2001. Health Serv Out Res Methodol 8:22–48

Paddock S, Ridgeway G, Lin R, Louis TA (2006) Flexible distributions for triple-goal estimates in two-stage hierarchical models. Comput Stat Data An 50(11):3243–3262
Shen W, Louis TA (1998) Triple-goal estimates in two-stage, hierarchical models. J Roy Stat Soc B 60:455–471

African Population Censuses

James P. M. Ntozi
Professor of Demographic Statistics
Makerere University, Kampala, Uganda

Definition

A Population ►census is the total process of collecting, compiling, evaluating, analyzing and disseminating demographic, economic and social data related to a specified time, to all persons in a country or a well defined part of a country.

History of Population Censuses

Population censuses are as old as human history. There are records of census enumerations as early as in 4000 BC in Babylonia, in 3000 BC in China and in 2500 BC in Egypt. The Roman Empire conducted population censuses and one of the most remembered censuses was the one held around AD 1 when Jesus Christ was born as his parents had moved from Nazareth to Bethlehem for the purpose of being counted. However, modern censuses did not start taking place until one was held in Quebec, Canada in 1666. This was followed by one in Sweden in 1749, USA in 1790, UK in 1801 and India 1871.

African Population Censuses

In the absence of complete civil registration systems in Africa, population censuses provide one of the best sources of socioeconomic and demographic information for the continent. Like in other parts of the world, censuses in Africa started as headcounts and assemblies until after the Second World War. The British were the first to introduce modern censuses in their colonial territories in west, east and southern Africa. For example in East Africa, the first modern census was conducted in 1948 in what was being referred to as British East Africa consisting of Kenya and Uganda. This was followed by censuses in 1957 in Tanzania, in 1959 in Uganda and 1962 in Kenya to prepare the countries for their political independence in 1961, 1962 and 1963, respectively. Other censuses have followed in these three countries. Similarly, the British West African countries of Ghana, Gambia, Nigeria and Sierra Leone were held in 1950s, 1960s and 1970s. In Southern Africa, similar censuses were held in Botswana, Lesotho, Malawi, Swaziland, Zambia and Zimbabwe in 1960s and 1970s, long before the Francophone and Lusophone countries did so. It was not until in 1970s and 1980s that the Francophone and Lusophone African countries started doing censuses instead of sample surveys which they preferred.

To help African countries do population censuses, United Nations set up an African census programme in late 1960s. Out of 41 countries, 22 participated in the programme. This programme closed in 1977 and was succeeded by the Regional Advisory Services in the demographic statistics set up as a section of Statistics Division at the United Nations Economic Commission for Africa. This section supported many African countries in conducting the 1980 and 1990 rounds of censuses. The section was superseded by the UNFPA sub-regional country support teams stationed in Addis Ababa, Cairo, Dakar and Harare. Each of these teams had census experts to give advisory services to countries in the 2000 round of censuses. These teams have now been reduced to three teams stationed in Pretoria, Cairo and Dakar and are currently supporting the African countries in population censuses.

There were working group committees on census on each round of censuses to work on the content of census ►questionnaire. For instance, in the 1980 round of censuses the working group recommended that the census questionnaire should have geographic characteristics, demographic characteristics, economic characteristics, community level variables and housing characteristics. In 1990 round of censuses, questions on the disabled persons were recommended to be added to the 1980 round questions. Later in the 2000 round of censuses, questions on economic establishments, agricultural sector and deaths in households were added. In the current round of 2010 censuses, the questions on disability were sharpened to capture the data better. New questions being asked include those on child labour, age at first marriage, ownership of mobile phone, ownership of email address, access to internet, distance to police post, access to salt in household, most commonly spoken language in household and cause of death in household.

In the 1960 and 1970s round of censuses, Post enumeration surveys (PES) to check on the quality of the censuses were attempted in Ghana. However, the experience with and results from PES were not encouraging, which discouraged most of the African countries from conducting them. Recently, the Post enumeration surveys have been revived and conducted in several African

countries like South Africa, Tanzania and Uganda. The challenges of PES have included: poor cartographic work, neglecting operational independence, inadequate funding, fatigue after the census, matching alternative names, lack of qualified personnel, useless questions in PES, probability sample design and selection, field reconciliation, lack of unique physical addresses in Africa and neglect of pretest of PES.

The achievements of the African censuses include supplying the needed sub-national data to the decentralized units for decision making processes, generating data for monitoring poverty reduction programmes, providing information for measuring indicators of most MDGs, using the data for measuring the achievement of indicators of International Conference on Population and Development (ICP), meeting the demand for data for emerging issues of socioeconomic concerns, accumulating experience in the region of census operations and capacity building at census and national statistical offices.

However, there are still several limitations associated with the African censuses. These have included inadequate participation of the population of the region; only 57% of the African population was counted in the 2000 round of censuses, which was much below to what happened in other regions: Oceania – 100%, Europe and North America – 99%, Asia – 97%, South America – 80% and the world – 91%. Other shortcomings were weak organizational and managerial skills, inadequate funding, non-conducive political environment, civil conflicts, weak technical expertise at NSOs and lack of data for gender indicators.

About the Author

Dr. James P. M. Ntozi is a Professor of demographic statistics at the Institute of Statistics, Makerere University, Kampala, Uganda. He is a founder and Past president of Uganda Statistical Society and Population Association of Uganda. He was a Council member of the International Statistical Institute and Union for African Population Studies, currently a Fellow and Chartered Statistician of the Royal Statistical Society and Council member of the Uganda National Academy of Sciences. He has authored, coauthored, and presented over 100 scientific papers as well as 6 books on fertility and censuses in Africa. He was an Editor of *African Population Studies*, co-edited 4 books, and is currently on the editorial board of *African Statistical Journal* and the *Journal of African Health Sciences*. He has received awards from Population Association of America, Uganda Statistical Society, Makerere University, Bishop Stuart University, Uganda and Ankole Diocese, Church of Uganda. James has been involved in planning and implementation of past Uganda censuses of population and housing of 1980, 1991, and 2002. He is currently helping the Liberian Statistical office to analyze the 2008 census data. Professor Ntozi is a past Director of the Institute of Statistics and Applied Economics, a regional statistical training center based at Makerere University, Uganda, and responsible for training many leaders in statistics and demography in sub-Saharan Africa for over 40 years. His other professional achievements have been research and consultancies in fertility, HIV/AIDS, Human Development Reports, and strategic planning.

Cross References

►Census
►Economic Statistics
►Integrated Statistical Databases
►Population Projections
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics: Developing Country Perspective
►Selection of Appropriate Statistical Methods in Developing Countries

References and Further Reading

Onsembe JO (2009) Postenumeration surveys in Africa. Paper presented at the 57th ISI session, Durban, South Africa

Onsembe JO, Ntozi JPM (2006) The 2000 round of censuses in Africa: achievements and challenges. Afr Stat J 3, November 6

Aggregation Schemes

Devendra Chhetry
President of the Nepal Statistical Association (NEPSA),
Professor and Head
Tribhuvan University, Kathmandu, Nepal

Given a data vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and a weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, there exist three aggregation schemes in the area of statistics that, under certain assumptions, generate three well-known measures of location: arithmetic mean (*AM*), ►geometric mean (*GM*), and ►harmonic mean (*HM*), where it is implicitly understood that the data vector $\mathbf{x}$ contains values of a single variable. Among all these three measures, *AM* is more frequently used in statistics for some theoretical reasons. It is well known that $AM \geq GM \geq HM$ where equality holds only when all components of $\mathbf{x}$ are equal.

In recent years, some of these three and a new aggregation scheme are being practiced in the aggregation of development or deprivation indicators by extending the definition of data vector to a vector of indicators, in the sense that it contains measurements of development or deprivation of several sub-population groups or measurements of several dimensions of development or deprivation. The measurements of development or deprivation are either available in the form of percentages or need to be transformed in the form of unit free indices. Physical Quality of Life Index (Morris 1979), Human Development Index (UNDP 1991), Gender-related Development Index (UNDP 1995), Gender Empowerment Measure (UNDP 1995), and Human Poverty Index (UNDP 1997) are some of the aggregated indices of several dimensions of development or deprivation.

In developing countries, aggregation of development or deprivation indicators is a challenging task, mainly due to two reasons. First, indicators usually display large variations or inequalities in the achievement of development or in the reduction of deprivation across the sub-populations or across the dimensions of development or deprivation within a region. Second, during the process of aggregation it is desired to incorporate the public aversion to social inequalities or, equivalently, public preference for social equalities. Public aversion to social inequalities is essential for development workers or planners of developing countries for bringing marginalized sub-populations into the mainstream by monitoring and evaluation of the development works. Motivated by this problem, Anand and Sen (UNDP 1995) introduced the notion of the gender-equality sensitive indicator (GESI).

In societies of equal proportion of female and male population, for example, the AM of 60 and 30 percent of male and female literacy rate is the same as that of 50 and 40 percent, showing that AM fails to incorporate the public aversion to gender inequality due to the *AM's built-in problem of perfect substitutability*, in the sense that a 10 percentage point decrease in female literacy rate in the former society as compared to the latter one is substituted by the 10 percentage point increase in male literacy rate. The *GM* or *HM*, however, incorporates the public aversion to gender inequality because they do not posses the perfect substitutability property. Instead of *AM*, Anand and Sen used *HM* in the construction of GESI.

In the above example consider that society perceives the social problem from the perspective of deprivation; that is, instead of gender-disaggregated literacy rates society considers gender-disaggregated illiteracy rates. Arguing as before, it immediately follows that *AM* fails to incorporate the public aversion to gender inequality. It also follows that neither *GM* nor *HM* incorporates the public aversion to gender inequality. A new aggregation scheme is required for aggregating indicators of deprivation.

So far, currently practiced aggregation schemes are accommodated within a slightly modified version of the following single mathematical function due to Hardy et al. (1952) under the assumption that components of $\mathbf{x}$ and $\mathbf{w}$ are positive and the sum of the components of $\mathbf{w}$ is unity.

$$\mu(\mathbf{x}, \mathbf{w}, r) = \begin{cases} \left(\sum_{i=1}^{n} w_i {x_i}^r \right)^{1/r} & \text{if } r \neq 0, \\ \prod_{i=1}^{n} {x_i}^{w_i} & \text{if } r = 0. \end{cases} \tag{1}$$

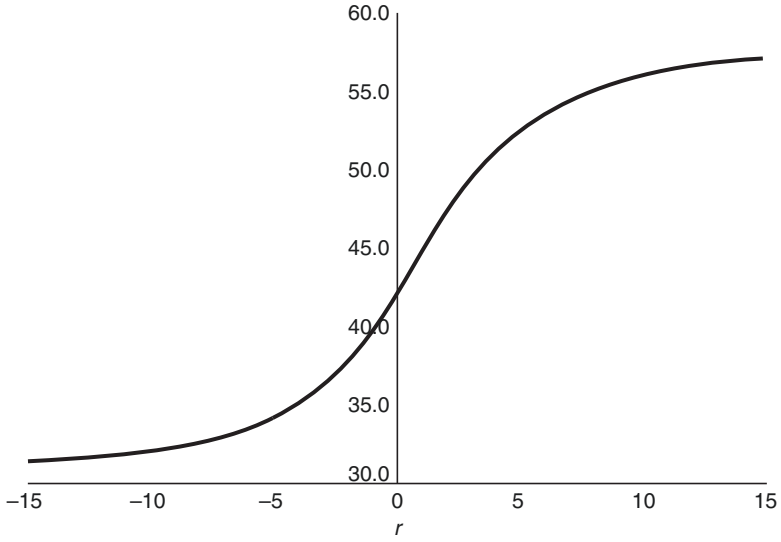
For fixed $\mathbf{x}$ and $\mathbf{w}$, the function (1) is defined for all real numbers, implying that the function (1) yields an infinite number of aggregation schemes. In particular, it yields *AM* when $r = 1$, *HM* when $r = -1$, and obviously *GM* when $r = 0$, and a new aggregation scheme suggested by Anand and Sen in constructing Human Poverty Index when $n = 3$, $w_1 = w_2 = w_3 = 1/3$ and $r = 3$ (UNDP 1997). It is well known that the values of the function are bounded between $x_{(1)}$ and $x_{(n)}$, where $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$, and the function is strictly increasing with respect to r if all the components of data vector are not equal (see Fig. 1 when $w_1 = w_2 = 0.5$, $x_1 = 60\%$ and $x_2 = 30\%$).

The first two partial derivatives of the function with respect to the k^{th} component of the vector $\mathbf{x}$ yield the following results where $g(\mathbf{x}, \mathbf{w})$ is *GM*.

$$\frac{\partial \mu(\mathbf{x}, \mathbf{w}, r)}{\partial x_k} = \begin{cases} w_k \left(\frac{x_k}{\mu(\mathbf{x}, \mathbf{w}, r)} \right)^{r-1} & \text{if } r \neq 0, \\ w_k \, g(\mathbf{x}, \mathbf{w}) x_k^{-1} & \text{if } r = 0. \end{cases} \tag{2}$$

$$\frac{\partial^2 \mu(\mathbf{x}, \mathbf{w}, r)}{\partial x_k^2} = \begin{cases} (r-1) w_k \left[\frac{x_k}{\mu(\mathbf{x}, \mathbf{w}, r)} \right]^{r-2} \sum_{i \neq k} \mathbf{w_i} \mathbf{x_i^r} & \text{if } r \neq 0, \\ w_k \, (w_k - 1) g(\mathbf{x}, \mathbf{w}) x_k{}^{-}2 & \text{if } r = 0. \end{cases} \tag{3}$$

For fixed $\begin{pmatrix} r < 1 \\ r > 1 \end{pmatrix}$ and $\mathbf{w}$, (2) and (3) imply that the function (1) is increasing and $\begin{pmatrix} \text{concave} \\ \text{convex} \end{pmatrix}$ with



Aggregation Schemes. Fig. 1 Nature of the function in a particular case

respect to each x_k, implying that the aggregated value increases at $\begin{pmatrix}\text{decreasing}\\\text{increasing}\end{pmatrix}$ rate with respect to each component of **x**. These properties are desirable for aggregating the $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ indicators, since the aggregated value of $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ is expected to $\begin{pmatrix}\text{rise}\\\text{fall}\end{pmatrix}$ from the $\begin{pmatrix}\text{floor to the ceiling value}\\\text{ceiling to the floor value}\end{pmatrix}$ at decreasing rate with respect to each component of **x**. For given **x** and **w**, the function (1) with any value of r, $\begin{pmatrix}r<1\\r>1\end{pmatrix}$, could be used to aggregate the $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ indicators if the public aversion to social inequalities should be incorporated.

What value of r should one use in practice? There is no simple answer to this question, since the answer depends upon the society's degree of preference for social equality. If a society has no preference for social equality, then one can use $r = 1$ in aggregating development or deprivation indicators, which is still a common practice in developing countries, even though the public efforts for bringing marginalized sub-populations into the mainstream has become a major agenda of development.

If a society has preference for social equality, then subjective judgment in the choice of r seems to be unavoidable. For the purpose of monitoring and evaluation, such judgment does not seem to be a serious issue as long as a fixed value of r is decided upon. In this context, Anand and Sen suggested using $r = -1$ for aggregating the indicators of development when $n = 2$ (UNDP 1995), and $r = 3$ for aggregating the indicators of deprivation when $n = 3$ (UNDP 1997). A lot of research work still needs to be done in this area for producing social-equality sensitive indicators of development or deprivation.

Cross References

►Composite Indicators
►Lorenz Curve
►Role of Statistics: Developing Country Perspective

References and Further Reading

Hardy GH, Littlewood JE, Polya G (1952) Inequalities. Cambridge University Press, London
Morris MD (1979) Measuring the condition of the world's poor: the physical quality of life index. Frank Case, London
UNDP (1991) Human Development Report 1991, Financing Human Development Oxford University Press, New York
UNDP (1995) Human Development Report 1995, Gender and Human Development. Oxford University Press, New York
UNDP (1997) Human Development Report 1997, Human Development to Eradicate Poverty. Oxford University Press, New York

Agriculture, Statistics in

Gavin J. S. Ross
Rothamsted Research, Harpenden, UK

The need to collect information on agricultural production has been with us since the dawn of civilization. Agriculture was the main economic activity, supplying both food for growing populations and the basis for taxation. The Sumerians of Mesopotamia before 3000 BC developed writing systems in order to record crop yields and livestock numbers. The Ancient Egyptians recorded the extent and productivity of arable land on the banks of the Nile. Later conquerors surveyed their new possessions, as in the Norman conquest of England which resulted in the Domesday Book of 1086, recording the agricultural potential of each district in great detail.

The pioneers of scientific agriculture, such as J.B. Lawes and J.H.Gilbert at Rothamsted, England, from 1843 onwards, insisted on accurate measurement and recording as the first requirement for a better understanding of the processes of agricultural production. The Royal Statistical Society (RSS) was founded in 1834 with its symbol of a sheaf of corn, implying that the duty of statisticians was to gather numerical information, but for others to interpret the data. Lawes published numerous papers on the variability of crop yields from year to year, and later joined the Council of the RSS. By 1900 agricultural experiments were conducted in several countries, including Germany, the Netherlands and Ireland, where W.S. Gosset, publishing under the name of "Student," conducted trials of barley varieties for the brewing industry.

In 1919 R.A. Fisher was appointed to analyze the accumulated results of 70 years of field experimentation at Rothamsted, initiating a revolution in statistical theory and practice. Fisher had already published the theoretical explanation of Student's t-distribution and the sampling distribution of the correlation coefficient, and challenged Karl Pearson's position that statistical analysis was only possible with large samples. His first task was to study the relationship between rainfall and crop yields on the long-term experiments, for which he demanded a powerful mechanical calculator, the "Millionaire." Introducing orthogonal polynomials to fit the yearly weather patterns and to eliminate the long-term trend in crop yield, he performed multiple regressions on the rainfall components, and developed the variance ratio test (later the F-distribution) to justify which terms to include using what became the ►analysis of variance. If the results were of minor interest to farmers, the methods used were of enormous importance in establishing the new methodology of curve fitting, regression analysis and the analysis of variance.

Fisher's work with agricultural scientists brought him a whole range of statistical challenges. Working with small samples he saw the role of the statistician as one who extracts the information in a sample as efficiently as possible. Working with non-normally distributed data he proposed the concept of likelihood, and the method of maximum likelihood to estimate parameters in a model. The early field experiments at Rothamsted contained the accepted notion of comparison of treatments with controls at the same location, and some plots included factorial combinations of fertilizer sources. Fisher saw that in order to apply statistical methods to assess the significance of observed effects it was necessary to introduce ►randomization and replication. Local control on land of varying fertility could be improved by blocking, and for trends in two directions he introduced Latin Square designs. The analysis of factorial experiments could be expressed in terms of main effects and interaction effects, with the components of interaction between blocks and treatments regarded as the basic residual error variance.

Fisher's ideas rapidly gained attention and his ideas and methods were extended to many fields beyond agricultural science. George Snedecor in Iowa, Mahalanobis and C.R. Rao in India, were early disciples, and his assistants included L.H.C. Tippett, J. Wishart and H. Hotelling. He was visited in 1926 by J. Neyman, who was working with agricultural scientists in Poland. In 1930 he was joined by Frank Yates who had experience of ►least squares methods as a surveyor in West Africa. Fisher left Rothamsted in 1933 to pursue his interests in genetics, but continued to collaborate with Yates. They introduced Balanced Incomplete Blocks and Lattice designs, and Split Plot designs with more than one component of error variance. Their *Statistical Tables*, first published in 1938, were widely used for many decades later.

Yates expanded his department to provide statistical analysis and consulting to agricultural departments and institutes in Britain and the British Empire. Field experimentation spread to South America with W.L Stevens, and his assistants W.G. Cochran, D.J. Finney and O. Kempthorne became well-known statistical innovators in many applications. During World War II Yates persuaded the government of the value of sample surveys to provide information about farm productivity, pests and diseases and fertilizer use. He later advised Indian statisticians on

the design and analysis of experiments in which small farmers in a particular area might be responsible for one plot each.

In 1954 Yates saw the potential of the electronic computer in statistical research, and was able to acquire the first computer devoted to civilian research, the Elliott 401. On this computer the first statistical programs were written for the analysis of field experiments and surveys, for bioassay and ►probit analysis, for multiple regression and multivariate analysis, and for model fitting by maximum likelihood. All the programs were in response to the needs of agricultural scientists, at field or laboratory level, including those working in animal science. Animal experiments typically had unequal numbers of units with different treatments, and iterative methods were needed to fit parameters by least squares or maximum likelihood. Animal breeding data required lengthy computing to obtain components of variance from which to estimate heritabilities and selection indices. The needs of researcher workers in fruit tree research, forestry, glasshouse crops and agricultural engineering all posed different challenges to the statistical profession.

In 1968 J.A. Nelder came to Rothamsted as head of the Statistics Department, having been previously at the National Vegetable Research Station at Wellesbourne, where he had explored the used of systematic designs for vegetable trials, and had developed the well-used Simplex Algorithm with R. Mead to fit ►nonlinear models. With more powerful computers it was now possible to combine many analyses into one system, and he invited G.N. Wilkinson from Adelaide to include his general algorithm for the analysis of variance in a more comprehensive system that would allow the whole range of nested and crossed experimental designs to be handled, along with facilities for regression and multivariate analysis. The program GENSTAT is now used world-wide in agricultural and other research settings.

Nelder worked with R.M. Wedderburn to show how the methodology of Probit Analysis (fitting binomial data to a transformed regression line) could be generalized to a whole class of ►Generalized Linear Models. These methods were particularly useful for the analysis of multiway contingency tables, using logit transformations for binomial data and log transformations for positive data with long-tailed distributions. The applications may have been originally in agriculture but found many uses elsewhere, such as in medical and pharmaceutical research.

The needs of soil scientists brought new classes of statistical problems. The classification of soils was complicated by the fact that overlapping horizons with different properties did not occur at the same depth, although samples were essential similar but displaced. The method of Kriging, first used by South African mining engineers, was found to be useful in describing the spatial variability of agricultural land, with its allowance for differing trends and sharp boundaries.

The need to model responses to fertilizer applications, the growth of plants and animals, and the spread of weeds, pests and diseases led to developments in fitting non-linear models. While improvements in the efficiency of numerical optimization algorithms were important, attention to the parameters to be optimized helped to show the relationship between the model and the data, and which observations contributed most to the parameters of interest. The limitations of agricultural data, with many unknown or unmeasurable factors present, makes it necessary to limit the complexity of the models being fitted, or to fit common parameters to several related samples.

Interest in spatial statistics, and in the use of models with more than one source of error, has led to developments such as the powerful REML algorithm. The use of intercropping to make better use of productive land has led to appropriate developments in experimental design and analysis.

With the increase in power of computers it became possible to construct large, complex models, incorporating where possible known relationships between growing crops and all the natural and artificial influences affecting their growth over the whole cycle from planting to harvest. These models have been valuable in understanding the processes involved, but have not been very useful in predicting final yields. The statistical ideas developed by Fisher and his successors have concentrated on the choices which farmers can make in the light of information available at the time, rather than to provide the best outcomes for speculators in crop futures. Modeling on its own is no substitute for continued experimentation.

The challenge for the 21st century will be to ensure sustainable agriculture for the future, taking account of climate change, resistance to pesticides and herbicides, soil degradation and water and energy shortages. Statistical methods will always be needed to evaluate new techniques of plant and animal breeding, alternative food sources and environmental effects.

About the Author

Gavin J.S. Ross has worked in the Statistics Department at Rothamsted Experimental Station since 1961, now as a retired visiting worker. He served under Frank Yates,

John Nelder and John Gower, advising agricultural workers, and creating statistical software for nonlinear modelling and for cluster analysis and multivariate analysis, contributing to the GENSTAT program as well as producing the specialist programs MLP and CLASP for his major research interests. His textbook *Nonlinear Estimation* (Springer 1990) describes the use of stable parameter transformations to fit and interpret nonlinear models. He served as President of the British Classification Society.

Cross References

►Analysis of Multivariate Agricultural Data
►Farmer Participatory Research Designs
►Spatial Statistics
►Statistics and Climate Change

References and Further Reading

Cochran WG, Cox GM (1957) Experimental designs, 2nd edn. Wiley, New York
Finney DJ (1962) An introduction to statistical science in agriculture. Edinburgh, Oliver and Boyd
Fisher RA (1924) The influence of rainfall on the yield of wheat at Rothamsted. Phil Trans Roy Soc London B 213:89–142
Mead R, Curnow RM (1983) Statistical methods in agriculture and experimental biology, 2nd edn. Chapman and Hall, London
Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58(3): 545–554
Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, New York
Yates F (1981) Sampling methods for censuses and surveys, 4th edn. Griffin, London

Akaike's Information Criterion

Hirotugu Akaike†
Former Director General of the Institute of Statistical Mathematics and a Kyoto Prize Winner
Tokyo, Japan

The Information Criterion $I(g:f)$ that measures the deviation of a model specified by the probability distribution f from the true distribution g is defined by the formula

$$I(g:f) = E\log g(X) - E\log f(X).$$

Here E denotes the expectation with respect to the true distribution g of X. The criterion is a measure of the deviation of the model f from the true model g, or the best possible model for the handling of the present problem.

The following relation illustrates the significant characteristic of the log likelihood:

$$I(g:f_1) - I(g:f_2) = -E(\log f_1(X) - \log f_2(X)).$$

This formula shows that for an observation x of X the log likelihood $\log f(x)$ provides a relative measure of the closeness of the model f to the truth, or the goodness of the model. This measure is useful even when the true structure g is unknown.

For a model $f(X/\mathbf{a})$ with unknown parameter $\mathbf{a}$ the maximum likelihood estimate $\mathbf{a}(x)$ is defined as the value of $\mathbf{a}$ that maximizes the likelihood $f(x/\mathbf{a})$ for a given observation x. Due to this process the value of $\log f(x/\mathbf{a}(x))$ shows an upward bias as an estimate of $\log f(X/\mathbf{a})$. Thus to use $\log f(x/\mathbf{a}(x))$ as the measure of the goodness of the model $f(X/\mathbf{a})$, it must be corrected for the expected bias.

In typical application of the method of maximum likelihood this expected bias is equal the dimension, or the number of components, of the unknown parameter $\mathbf{a}$. Thus the relative goodness of a model determined by the maximum likelihood estimate is given by

AIC = −2 (log maximum likelihood − (number of parameters)).

Here log denotes natural logarithm. The coefficient −2 is used to make the quantity similar to the familiar chi-square statistic in the test of dimensionality of the parameter.

AIC is the abbreviation of An Information Criterion.

About the Author

Professor Akaike died of pneumonia in Tokyo on 4th August 2009, aged 81. He was the Founding Head of the first Department of Statistical Science in Japan. "Now that he has left us forever, the world has lost one of its most innovative statisticians, the Japanese people have lost the finest statistician in their history and many of us a most noble friend" (Professor Howell Tong, from "The Obituary of Professor Hirotugu Akaike." *Journal of the Royal Statistical Society, Series A,* March, 2010). Professor Akaike had sent his Encyclopedia entry on May 14 2009, adding the following sentence in his email: "This is all that I could do under the present physical condition."

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Cp Statistic
►Kullback-Leibler Divergence
►Model Selection

Akaike's Information Criterion: Background, Derivation, Properties, and Refinements

Joseph E. Cavanaugh[1], Andrew A. Neath[2]
[1]Professor
The University of Iowa, Iowa City, IA, USA
[2]Professor
Southern Illinois University Edwardsville, Edwardsville, IL, USA

Introduction

The ▶Akaike Information Criterion, AIC, was introduced by Hirotogu Akaike in his seminal 1973 paper "Information Theory and an Extension of the Maximum Likelihood Principle." AIC was the first model selection criterion to gain widespread attention in the statistical community. Today, AIC continues to be the most widely known and used model selection tool among practitioners.

The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure. Akaike extended this paradigm by considering a framework in which the model dimension is also unknown, and must therefore be determined from the data. Thus, Akaike proposed a framework wherein both model estimation and selection could be simultaneously accomplished.

For a parametric candidate model of interest, the likelihood function reflects the conformity of the model to the observed data. As the complexity of the model is increased, the model becomes more capable of adapting to the characteristics of the data. Thus, selecting the fitted model that maximizes the empirical likelihood will invariably lead one to choose the most complex model in the candidate collection. ▶Model selection based on the likelihood principle, therefore, requires an extension of the traditional likelihood paradigm.

Background

To formally introduce AIC, consider the following model selection framework. Suppose we endeavor to find a suitable model to describe a collection of response measurements y. We will assume that y has been generated according to an unknown density $g(y)$. We refer to $g(y)$ as the *true* or *generating model*.

A model formulated by the investigator to describe the data y is called a *candidate* or *approximating model*. We will assume that any candidate model structurally corresponds to a parametric class of distributions. Specifically, for a certain candidate model, we assume there exists a k-dimensional parametric class of density functions

$$\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\},$$

a class in which the parameter space $\Theta(k)$ consists of k-dimensional vectors whose components are functionally independent.

Let $L(\theta_k\,|\,y)$ denote the likelihood corresponding to the density $f(y|\,\theta_k)$, i.e., $L(\theta_k\,|\,y) = f(y|\,\theta_k)$. Let $\hat{\theta}_k$ denote a vector of estimates obtained by maximizing $L(\theta_k\,|\,y)$ over $\Theta(k)$.

Suppose we formulate a collection of candidate models of various dimensions k. These models may be based on different subsets of explanatory variables, different mean and variance/covariance structures, and even different specifications for the type of distribution for the response variable. Our objective is to search among this collection for the fitted model that "best" approximates $g(y)$.

In the development of AIC, optimal approximation is defined in terms of a well-known measure that can be used to gauge the similarity between the true model $g(y)$ and a candidate model $f(y|\,\theta_k)$: the *Kullback–Leibler information* (Kullback and Leibler 1951; Kullback 1968). The Kullback–Leibler information between $g(y)$ and $f(y|\,\theta_k)$ with respect to $g(y)$ is defined as

$$I(\theta_k) = \mathrm{E}\left\{\log\frac{g(y)}{f(y|\,\theta_k)}\right\},$$

where $\mathrm{E}(\cdot)$ denotes the expectation under $g(y)$. It can be shown that $I(\theta_k) \geq 0$ with equality if and only if $f(y|\,\theta_k)$ is the same density as $g(y)$. $I(\theta_k)$ is not a formal metric, yet we view the measure in a similar manner to a distance: i.e., as the disparity between $f(y|\,\theta_k)$ and $g(y)$ grows, the magnitude of $I(\theta_k)$ will generally increase to reflect this separation.

Next, define

$$d(\theta_k) = \mathrm{E}\{-2\log f(y|\,\theta_k)\}.$$

We can then write

$$2I(\theta_k) = d(\theta_k) - \mathrm{E}\{-2\log g(y)\}.$$

Since $\mathrm{E}\{-2\log g(y)\}$ does not depend on θ_k, any ranking of a set of candidate models corresponding to values of $I(\theta_k)$ would be identical to a ranking corresponding to values of $d(\theta_k)$. Hence, for the purpose of discriminating among various candidate models, $d(\theta_k)$ serves as a valid substitute for $I(\theta_k)$. We will refer to $d(\theta_k)$ as the *Kullback discrepancy*.

To measure the separation between between a fitted candidate model $f(y|\,\hat{\theta}_k)$ and the generating model

$g(y)$, we consider the Kullback discrepancy evaluated at $\hat{\theta}_k$:

$$d(\hat{\theta}_k) = \mathrm{E}\{-2\log f(y|\,\theta_k)\}|_{\theta_k=\hat{\theta}_k}.$$

Obviously, $d(\hat{\theta}_k)$ would provide an attractive means for comparing various fitted models for the purpose of discerning which model is closest to the truth. Yet evaluating $d(\hat{\theta}_k)$ is not possible, since doing so requires knowledge of the true distribution $g(\cdot)$. The work of Akaike (1973, 1974), however, suggests that $-2\log f(y|\,\hat{\theta}_k)$ serves as a biased estimator of $d(\hat{\theta}_k)$, and that the bias adjustment

$$\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta}_k)\} \tag{1}$$

can often be asymptotically estimated by twice the dimension of θ_k.

Since k denotes the dimension of θ_k, under appropriate conditions, the expected value of

$$\mathrm{AIC} = -2\log f(y|\,\hat{\theta}_k) + 2k$$

will asymptotically approach the expected value of $d(\hat{\theta}_k)$, say

$$\Delta(k) = \mathrm{E}\{d(\hat{\theta}_k)\}.$$

Specifically, we will establish that

$$\mathrm{E}\{\mathrm{AIC}\} + o(1) = \Delta(k). \tag{2}$$

AIC therefore provides an asymptotically unbiased estimator of $\Delta(k)$. $\Delta(k)$ is often called the *expected Kullback discrepancy*.

In AIC, the empirical log-likelihood term $-2\log f(y|\,\hat{\theta}_k)$ is called the *goodness-of-fit term*. The bias correction $2k$ is called the *penalty term*. Intuitively, models which are too simplistic to adequately accommodate the data at hand will be characterized by large goodness-of-fit terms yet small penalty terms. On the other hand, models that conform well to the data, yet do so at the expense of containing unnecessary parameters, will be characterized by small goodness-of-fit terms yet large penalty terms. Models that provide a desirable balance between fidelity to the data and parsimony should correspond to small AIC values, with the sum of the two AIC components reflecting this balance.

Derivation

To justify AIC as an asymptotically unbiased estimator of $\Delta(k)$, we will focus on a particular candidate class $\mathcal{F}(k)$. For notational simplicity, we will suppress the dimension index k on the parameter vector θ_k and its estimator $\hat{\theta}_k$.

The justification of (2) requires the strong assumption that the true density $g(y)$ is a member of the candidate class $\mathcal{F}(k)$. Under this assumption, we may define a parameter vector θ_o having the same size as θ, and write $g(y)$ using the parametric form $f(y|\,\theta_o)$. The assumption that $f(y|\,\theta_o) \in \mathcal{F}(k)$ implies that the fitted model is either correctly specified or overfit.

To justify (2), consider writing $\Delta(k)$ as indicated:

$$\begin{aligned}
&\Delta(k)\\
&= \mathrm{E}\{d(\hat{\theta})\}\\
&= \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad + \left[\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\right] \qquad (3)\\
&\quad + \left[\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\}\right]. \qquad (4)
\end{aligned}$$

The following lemma asserts that (3) and (4) are both within $o(1)$ of k.

We assume the necessary regularity conditions required to ensure the consistency and ▸asymptotic normality of the maximum likelihood vector $\hat{\theta}$.

Lemma

$$\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\} = k + o(1), \tag{5}$$

$$\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\} = k + o(1). \tag{6}$$

Proof

Define

$$\mathrm{I}(\theta) = \mathrm{E}\left[-\frac{\partial^2 \log f(y|\,\theta)}{\partial\theta\partial\theta'}\right]$$

$$\text{and } \mathcal{I}(\theta, y) = \left[-\frac{\partial^2 \log f(y|\,\theta)}{\partial\theta\partial\theta'}\right].$$

$\mathrm{I}(\theta)$ denotes the *expected Fisher information matrix* and $\mathcal{I}(\theta, y)$ denotes the *observed Fisher information matrix*.

First, consider taking a second-order expansion of $-2\log f(y|\,\theta_o)$ about $\hat{\theta}$, and evaluating the expectation of the result. Since $-2\log f(y|\,\theta)$ is minimized at $\theta = \hat{\theta}$, the first-order term disappears, and we obtain

$$\begin{aligned}
\mathrm{E}\{-2\log f(y|\,\theta_o)\} &= \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad + \mathrm{E}\left\{(\hat{\theta} - \theta_o)'\{\mathcal{I}(\hat{\theta}, y)\}(\hat{\theta} - \theta_o)\right\}\\
&\quad + o(1).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad = \mathrm{E}\left\{(\hat{\theta} - \theta_o)'\{\mathcal{I}(\hat{\theta}, y)\}(\hat{\theta} - \theta_o)\right\} + o(1). \qquad (7)
\end{aligned}$$

Next, consider taking a second-order expansion of $d(\hat{\theta})$ about θ_o, again evaluating the expectation of the

result. Since $d(\theta)$ is minimized at $\theta = \theta_o$, the first-order term disappears, and we obtain

$$\begin{aligned}\mathrm{E}\{d(\hat{\theta})\} &= \mathrm{E}\{-2\log f(y|\,\theta_o)\} \\ &\quad + \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} \\ &\quad + o(1).\end{aligned}$$

Thus,

$$\begin{aligned}&\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\} \\ &\quad = \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} + o(1). \qquad (8)\end{aligned}$$

Recall that by assumption, $\theta_o \in \Theta(k)$. Therefore, the quadratic forms

$$(\hat{\theta}-\theta_o)'\{\mathcal{I}(\hat{\theta},y)\}(\hat{\theta}-\theta_o) \text{ and } (\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)$$

both converge to centrally distributed chi-square random variables with k degrees of freedom. Thus, the expectations of both quadratic forms are within $o(1)$ of k. This fact along with (7) and (8) establishes (5) and (6). □

Properties

The previous lemma establishes that AIC provides an asymptotically unbiased estimator of $\Delta(k)$ for fitted candidate models that are correctly specified or overfit. From a practical perspective, AIC estimates $\Delta(k)$ with negligible bias in settings where n is large and k is comparatively small. In settings where n is small and k is comparatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment, making AIC substantially negatively biased as an estimator of $\Delta(k)$.

If AIC severely underestimates $\Delta(k)$ for higher dimensional fitted models in the candidate collection, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large. Examples illustrating this phenomenon appear in Linhart and Zucchini (1986, 86–88), who comment (p. 78) that "in some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased."

AIC is *asymptotically efficient* in the sense of Shibata (1980, 1981), yet it is not *consistent*. Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A consistent criterion will asymptotically select the fitted candidate model having the correct structure with probability one. On the other hand, suppose that the generating model is of an infinite dimension, and therefore lies outside of the candidate collection under consideration. An asymptotically efficient criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction.

From a theoretical standpoint, asymptotic efficiency is arguably the strongest optimality property of AIC. The property is somewhat surprising, however, since demonstrating the asymptotic unbiasedness of AIC as an estimator of the expected Kullback discrepancy requires the assumption that the candidate model of interest subsumes the true model.

Refinements

A number of AIC variants have been developed and proposed since the introduction of the criterion. In general, these variants have been designed to achieve either or both of two objectives: (1) to relax the assumptions or expand the setting under which the criterion can be applied, (2) to improve the small-sample performance of the criterion.

In the Gaussian linear regression framework, Sugiura (1978) established that the bias adjustment (1) can be exactly evaluated for correctly specified or overfit models. The resulting criterion, with a refined penalty term, is known as "corrected" AIC, or AICc. Hurvich and Tsai (1989) extended AICc to the frameworks of Gaussian nonlinear regression models and time series autoregressive models. Subsequent work has extended AICc to other modeling frameworks, such as autoregressive moving average models, vector autoregressive models, and certain ►generalized linear models and ►linear mixed models.

The Takeuchi (1976) information criterion, TIC, was derived by obtaining a general, large-sample approximation to each of (3) and (4) that does not rely on the assumption that the true density $g(y)$ is a member of the candidate class $\mathcal{F}(k)$. The resulting approximation is given by the trace of the product of two matrices: an information matrix based on the score vector, and the inverse of an information matrix based on the Hessian of the log likelihood. Under the assumption that $g(y) \in \mathcal{F}(k)$, the information matrices are equivalent. Thus, the trace reduces to k, and the penalty term of TIC reduces to that of AIC.

Bozdogon (1987) proposed a variant of AIC that corrects for its lack of consistency. The variant, called CAIC, has a penalty term that involves the log of the determinant of an information matrix. The contribution of this term leads to an overall complexity penalization that increases with the sample size at a rate sufficient to ensure consistency.

Pan (2001) introduced a variant of AIC for applications in the framework of generalized linear models fitted

using generalized estimating equations. The criterion is called QIC, since the goodness-of-fit term is based on the empirical quasi-likelihood.

Konishi and Kitagawa (1996) extended the setting in which AIC has been developed to a general framework where (1) the method used to fit the candidate model is not necessarily maximum likelihood, and (2) the true density $g(y)$ is not necessarily a member of the candidate class $\mathcal{F}(k)$. Their resulting criterion is called the generalized information criterion, GIC. The penalty term of GIC reduces to that of TIC when the fitting method is maximum likelihood.

AIC variants based on computationally intensive methods have also been proposed, including cross-validation (Stone 1977; Davies et al. 2005), bootstrapping (Ishiguro et al. 1997; Cavanaugh and Shumway 1997; Shibata 1997), and Monte Carlo simulation (Hurvich et al. 1990; Bengtsson and Cavanaugh 2006). These variants tend to perform well in settings where the sample size is small relative to the complexity of the models in the candidate collection.

About the Authors

Joseph E. Cavanaugh is Professor of Biostatistics and Professor of Statistics and Actuarial Science at The University of Iowa. He is an associate editor of the *Journal of the American Statistical Association* (2005–present) and the *Journal of Forecasting* (1999–present). He has published over 60 refereed articles.

Andrew Neath is a Professor of Mathematics and Statistics at Southern Illinois University Edwardsville. He has been recognized for his work in science education. He is an author on numerous papers, merging Bayesian views with model selection ideas. He wishes to thank Professor Miodrag Lovric for the honor of an invitation to contribute to a collection containing the works of so many notable statisticians.

Cross References

►Akaike's Information Criterion
►Cp Statistic
►Kullback-Leibler Divergence
►Model Selection

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) Proceedings of the 2nd International symposium on information theory. Akadémia Kiadó, Budapest, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE T Automat Contra AC-19:716–723

Bengtsson T, Cavanaugh JE (2006) An improved Akaike information criterion for state-space model selection. Comput Stat Data An 50:2635–2654

Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345–370

Cavanaugh JE, Shumway RH (1997) A bootstrap variant of AIC for state-space model selection. Stat Sinica 7:473–496

Davies SL, Neath AA, Cavanaugh JE (2005) Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. Stat Methodol 2:249–266

Hurvich CM, Shumway RH, Tsai CL (1990) Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika 77:709–719

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Ishiguro M, Sakamoto Y, Kitagawa G (1997) Bootstrapping log likelihood and EIC, an extension of AIC. Ann I Stat Math 49:411–434

Konishi S, Kitagawa G (1996) Generalised information criteria in model selection. Biometrika 83:875–890

Kullback S (1968) Information Theory and Statistics. Dover, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:76–86

Linhart H, Zucchini W (1986) Model selection. Wiley, New York

Pan W (2001) Akaike's information criterion in generalized estimating equations. Biometrics 57:120–125

Shibata R (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann Stat 80:147–164

Shibata R (1981) An optimal selection of regression variables. Biometrika 68:45–54

Shibata R (1997)Bootstrap estimate of Kullback-Leibler information for model selection. Stat Sinica 7:375–394

Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J R Stat Soc B 39:44–47

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat A7:13–26

Takeuchi K (1976) Distribution of information statistics and criteria for adequacy of models. Mathematical Sciences 153:12–18(in Japanese)

Algebraic Statistics

Sonja Petrović[1], Aleksandra B. Slavković[2]
[1]Research Assistant Professor
University of Illinois at Chicago, Chicago, IL, USA
[2]Associate Professor
The Pennsylvania State University, University Park, PA, USA

Algebraic statistics applies concepts from algebraic geometry, commutative algebra, and geometric combinatorics to better understand the structure of statistical models, to

improve statistical inference, and to explore new classes of models. Modern algebraic geometry was introduced to the field of statistics in the mid 1990s. Pistone and Wynn (1996) used Gröbner bases to address the issue of confounding in design of experiments, and Diaconis and Sturmfels (1998) used them to perform exact conditional tests. The term *algebraic statistics* was coined in the book by Pistone et al. (2001), which primarily addresses experimental design. The current algebraic statistics literature includes work on contingency tables, sampling methods, graphical and latent class models, and applications in areas such as statistical disclosure limitation (e.g., Dobra et al. (2009)), and computational biology and phylogenetics (e.g., Pachter and Sturmfels (2005)).

Algebraic Geometry of Statistical Models

Algebraic geometry is a broad subject that has seen an immense growth over the past century. It is concerned with the study of algebraic varieties, defined to be (closures of) solution sets of systems of polynomial equations. For an introduction to computational algebraic geometry and commutative algebra, see Cox et al. (2007).

Algebraic statistics studies statistical models whose parameter spaces correspond to real positive parts of algebraic varieties. To demonstrate how this correspondence works, consider the following simple example of the independence model of two binary random variables, X and Y, such that joint probabilities are arranged in a 2×2 matrix $p := [p_{ij}]$. The model postulates that the joint probabilities factor as a product of marginal distributions: $p_{ij} = p_{i+}p_{+j}$, where $i, j \in \{1, 2\}$. This is referred to as an *explicit* algebraic statistical model. Equivalently, the matrix p is of rank 1, that is, its 2×2 determinant is zero: $p_{11}p_{22} - p_{12}p_{21} = 0$. This is referred to as an *implicit* description of the independence model. In algebraic geometry, the set of rank-1 matrices, where we allow p_{ij} to be arbitrary complex numbers, is a classical object called a *Segre variety*. Thus, the independence model is the real positive part of the Segre variety. Exponential family models, in general, correspond to *toric varieties*, whose implicit description is given by a set of binomials. For a broad, general definition of algebraic statistical models, see Drton and Sullivant (2007).

By saying that "we understand the algebraic geometry of a model," we mean that we understand some basic information about the corresponding variety, such as: degree, dimension and codimension (i.e., degrees of freedom); the defining equations (i.e., the implicit description of the model); the singularities (i.e., degeneracy in the model). The current algebraic statistics literature demonstrates that understanding the geometry of a model can be useful for statistical inference (e.g., exact conditional inference, goodness-of-fit testing, parameter identifiability, and maximum likelihood estimation). Furthermore, many relevant questions of interest in statistics relate to classical open problems in algebraic geometry.

Algebraic Statistics for Contingency Tables

A paper by Diaconis and Sturmfels (1998) on algebraic methods for discrete probability distributions stimulated much of the work in algebraic statistics on contingency tables, and has led to two general classes of problems: (1) algebraic representation of a statistical model, and (2) conditional inference. The algebraic representation of the independence model given above generalizes to any k-way table and its corresponding hierarchical log-linear models (e.g., see Dobra et al. (2009)). A standard reference on log-linear models is Bishop et al. (1975).

Most of the algebraic work for contingency tables has focused on geometric characterizations of log-linear models and estimation of cell probabilities under those models. Algebraic geometry naturally provides an explicit description of the closure of the parameter space. This feature has been utilized, for example, by Eriksson et al. (2006) to describe polyhedral conditions for the nonexistence of the MLE for log-linear models. More recently, Petrović et al. (2010) provide the first study of algebraic geometry of the p_1 random graph model of Holland and Leinhardt (1981).

Conditional inference relies on the fact that data-dependent objects are a convex bounded set, $P_t = \{\mathbf{x} : x_i \in \mathbb{R}_{\geq 0}, \mathbf{t} = \mathbf{A}\mathbf{x}\}$, where x is a table, A is a design matrix, and $\mathbf{t}$ a vector of constraints, typically margins, that is, sufficient statistics of a log-linear model. The set of all integer points inside P_t is referred to as a *fiber*, which is the support of the conditional distribution of tables given $\mathbf{t}$, or the so-called *exact distribution*. Characterization of the fiber is crucial for three statistical tasks: counting, sampling and optimization. Diaconis and Sturmfels (1998) provide one of the fundamental results in algebraic statistics regarding sampling from exact distributions. They define a Markov basis, a set of integer valued vectors in the kernel of A, which is a smallest set of moves needed to perform a ▶random walk over the space of tables and to guarantee connectivity of the chain. In Hara et al. (2010), for example, the authors use Markov bases for exact tests in a multiple logistic regression. The earliest application of Markov bases, counting and optimization was in the area of statistical disclosure limitation for exploring issues of confidentiality with the release of contingency table data; for an overview,

see Dobra et al. (2009), and for other related topics, see Chen et al. (2006), Onn (2006), and Slavković and Lee (2009).

Graphical and Mixture Models

Graphical models (e.g., Lauritzen (1996)) are an active research topic in algebraic statistics. Non-trivial problems, for example, include complete characterization of Markov bases for these models, and counting the number of solutions of their likelihood equations. Geiger et al. (2006) give a remarkable result in this direction: *decomposable* graphical models are precisely those whose Markov bases consist of squarefree quadrics, or, equivalently, those graphical models whose maximum likelihood degree is 1. More recently, Feliz et al. (2010) made a contribution to the mathematical finance literature by proposing a new model for analyzing default correlation.

►Mixture models, including latent class models, appear frequently in statistics, however, standard asymptotics theory often does not apply due to the presence of singularities (e.g., see Watanabe (2009)). Singularities are created by marginalizing (smooth) models; geometrically, this is a projection of the corresponding variety. Algebraically, mixture models correspond to *secant varieties*. The complexity of such models presents many interesting problems for algebraic statistics; e.g., see Fienberg et al. (2009) for the problems of maximum likelihood estimation and parameter identifiability in latent class models. A further proliferation of algebraic statistics has been supported by studying mixture models in phylogenetics (e.g., see Allman et al. (2010)), but many questions about the geometry of these models still remain open.

Further Reading

There are many facets of algebraic statistics, including generalizations of classes of models discussed above: experimental design, continuous multivariate problems, and new connections between algebraic statistics and information geometry. For more details see Putinar and Sullivant (2008), Drton et al. (2009), Gibilisco et al. (2009), and references given therein. Furthermore, there are many freely available algebraic software packages (e.g., 4ti2 (4ti2 team), CoCoA (CoCoATeam)) that can be used for relevant computations alone, or in combination with standard statistical packages.

Acknowledgments

Supported in part by National Science Foundation grant SES-0532407 to the Department of Statistics, Pennsylvania State University.

Cross References

►Categorical Data Analysis
►Confounding and Confounder Control
►Degrees of Freedom
►Design of Experiments: A Pattern of Progress
►Graphical Markov Models
►Logistic Regression
►Mixture Models
►Statistical Design of Experiments (DOE)
►Statistical Inference
►Statistical Inference: An Overview

References and Further Reading

4ti2 team. 4ti2 – a software package for algebraic, geometric and combinatorial problems on linear spaces. http://WWW.4ti2.de

Allman E, Petrović S, Rhodes J, Sullivant S (2010) Identifiability of two-tree mixtures under group-based models. IEEE/ACM Trans Comput Biol Bioinfor. In press

Bishop YM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Cambridge, MA (Reprinted by Springer, 2007)

Chen Y, Dinwoodie I, Sullivant S (2006) Sequential importance sampling for multiway tables. Ann Stat 34(1):523–545

CoCoATeam. CoCoA: a system for doing computations in commutative algebra. http://cocoa.dima.unige.it

Cox D, Little J, O'Shea D (2007) Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra, 3rd edn. Springer, New York

Diaconis P, Sturmfels B (1998) Algebraic algorithms for sampling from conditional distributions. Ann Stat 26:363–397

Dobra A, Fienberg SE, Rinaldo A, Slavković A, Zhou Y(2009) Algebraic statistics and contingency table problems: estimations and disclosure limitation. In: Emerging Applications of Algebraic Geometry: IMA volumes in mathematics and its applications, 148:63–88

Drton M, Sturmfels B, Sullivant S (2009) Lectures on algebraic statistics, vol39. Oberwolfach seminars, Birkhäuser

Eriksson N, Fienberg SE, Rinaldo A, Sullivant S (2006) Polyhedral conditions for the nonexistence of the mle for hierarchical log-liner models. J Symb Comput 41(2):222–233

Feliz I, Guo X, Morton J, Sturmfels B (2010) Graphical models for correlated default. Math Financ (in press)

Fienberg SE, Hersh P, Zhou Y (2009) Maximum likelihood estimation in latent class models for contingency table data. In: Gibilisco P, Riccomagno E, Rogantin M, Wynn H (eds) Algebraic and geometric methods in statistics. Cambridge University Press, London, pp 27–62

Geiger D, Meek C, Sturmfels B (2006) On the toric algebra of graphical models. Ann Stat 34(3):1463–1492

Gibilisco P, Riccomagno E, Rogantin M, Wynn H (2009) Algebraic and geometric methods in statistics, Cambridge University press

Hara H, Takemura A, Yoshida R (2010) On connectivity of fibers with positive marginals in multiple logistic regression. J Multivariate Anal 101(4):909–925

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65

Lauritzen SL (2006) Graphical models. Clarendon, Oxford
Onn S (2006) Entry uniqueness in margined tables. Lect Notes Comput Sci 4302:94–101
Pachter L, Sturmfels B (2005) Algebraic statistics for computational biology. Cambridge University Press, New York, NY
Petrović S, Rinaldo A, Fienberg SE (2010) Algebraic statistics for a directed random graph model with reciprocation. In: Viana MAG, Wynn H (eds) Algebraic methods in statistics and probability, II, Contemporary Mathematics. Am Math Soc 516
Pistone G, Wynn H (1996) Generalised confounding with Gröbner bases. Biometrika 83(3):653–666
Pistone G, Riccomagno E, Wynn H (2001) Algebraic statistics: computational commutative algebra in statistics. CRC, Boca Raton
Putinar M, Sullivant S (2008) Emerging applications of algebraic geometry. Springer, Berlin
Slavković AB, Lee J (2010) Synthetic two-way contingency tables that preserve conditional frequencies. Stat Methodal 7(3): 225–239
Watanabe S (2009) Algebraic geometry and statistical learning theory: Cambridge monographs on applied and computational mathematics, 25, NewYork, Cambridge University Press

Almost Sure Convergence of Random Variables

Herold Dehling
Professor
Ruhr-Universität Bochum, Bochum, Germany

Definition and Relationship to Other Modes of Convergence

Almost sure convergence is one of the most fundamental concepts of convergence in probability and statistics. A sequence of random variables $(X_n)_{n\geq 1}$, defined on a common probability space $(\Omega, \mathcal{F}, P)$, is said to converge almost surely to the random variable X, if

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

Commonly used notations are $X_n \xrightarrow{a.s.} X$ or $\lim_{n\to\infty} X_n = X\ (a.s.)$. Conceptually, almost sure convergence is a very natural and easily understood mode of convergence; we simply require that the sequence of numbers $(X_n(\omega))_{n\geq 1}$ converges to $X(\omega)$ for almost all $\omega \in \Omega$. At the same time, proofs of almost sure convergence are usually quite subtle.

There are rich connections of almost sure convergence with other classical modes of convergence, such as convergence in probability, defined by $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$ for all $\epsilon > 0$, convergence in distribution, defined by $\lim_{n\to\infty} Ef(X_n) = Ef(X)$ for all real-valued bounded, continuous functions f, and convergence in L_p, defined by $\lim_{n\to\infty} E|X_n - X|^p = 0$. Almost sure convergence implies convergence in probability, which again implies convergence in distribution, but not vice versa. Almost sure convergence neither implies nor is it implied by convergence in L_p. A standard counterexample, defined on the probability space $[0,1]$, equipped with the Borel σ-field and Lebesgue measure, is the sequence $X_n(\omega) = 1_{[\frac{j}{2^k}, \frac{j+1}{2^k}]}(\omega)$, if $n = 2^k + j$, $k \geq 0$, $0 \leq j < 2^k$. The sequence $(X_n)_{n\geq 1}$ converges to zero in probability and in L_p, but not almost surely. On the same probability space, the sequence defined by $X_n = n^{1/p} 1_{[0,\frac{1}{n}]}$ provides an example that converges to zero almost surely, but not in L_p.

Although convergence in probability does not imply almost sure convergence, there is a partial result in this direction. If $(X_n)_{n\geq 1}$ converges in probabilty to X, one can find a subsequence $(n_k)_{k\geq 1}$ such that $X_{n_k} \xrightarrow{a.s.} X$.

Skorohod's almost sure representation theorem is a partial converse to the fact that almost sure convergence implies convergence in distribution. If $(X_n)_{n\geq 1}$ converges in distribution to X, one can find a sequence of random variables $(Y_n)_{n\geq 1}$ and a random variable Y such that X_n and Y_n have the same distribution, for each n, X and Y have the same distribution, and $\lim_{n\to\infty} Y_n = Y$ almost surely. Originally proved by Skorohod (1956) for random variables with values in a separable metric space, this representation theorem has been extended by Dudley (1968) to noncomplete spaces and later by Wichura (1970) to nonseparable spaces.

By some standard arguments, one can show that almost sure convergence of $(X_n)_{n\geq 1}$ to X is equivalent to

$$\lim_{n\to\infty} P(\sup_{k\geq n} |X_k - X| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

Thus almost sure convergence holds, if the series $\sum_{k\geq 1} P(|X_k - X| \geq \epsilon)$ converges. In this case, the sequence $(X_n)_{n\geq 1}$ is said to *converge completely* to X.

Important Almost Sure Convergence Theorems

Historically the earliest and also the best known almost sure convergence theorem is the *Strong Law of Large Numbers*, established originally by Borel (1909). Given an i.i.d. sequence $(X_k)_{k\geq 1}$ of random variables that are uniformly distributed on $[0,1]$, Borel showed that

$$\frac{1}{n} S_n \xrightarrow{a.s.} E(X_1),$$

where $S_n := \sum_{k=1}^n X_k$ denotes the partial sum. Later, this was generalized to sequences with arbitrary distributions. Finally, Kolmogorov (1930) could show that the existence of first moments is a necessary and sufficient condition for the strong law of large numbers for i.i.d. random variables.

Hsu and Robbins (1947) showed complete convergence in the law of large numbers, provided the random variables have finite second moments; Baum and Katz (1965) showed that this condition is also necessary.

Birkhoff (1931) proved the *Ergodic Theorem*, i.e., the validity of the strong law of large numbers for stationary ergodic sequences $(X_k)_{k\geq 1}$ with finite first moments. Kingman (1968) generalized this to the *Subadditive Ergodic Theorem*, valid for doubly indexed subadditive process $(X_{s,t})$ satisfying a certain moment condition. Doob (1953) established the *Martingale Convergence Theorem*, which states that every L_1-bounded submartingale converges almost surely.

The *Marcinkiewicz-Zygmund Strong Law of Large Numbers* (1937) is a sharpening of the law of large numbers for partial sums of i.i.d. random variables, stating that for $1 \leq p < 2$ we have

$$\lim_{n\to\infty} \frac{1}{n^{1/p}} \sum_{k=1}^{n} (X_k - E(X_k)) = 0 \text{ a.s.},$$

if and only if the random variables have finite p-th moments. Note that for $p = 2$ this result is false as it would contradict the central limit theorem (see ▶Central Limit Theorems).

For i.i.d. random variables with finite variance $\sigma^2 \neq 0$, Hartman and Wintner (1941) proved the *Law of the Iterated Logarithm*, stating that

$$\limsup_{n\to\infty} \frac{1}{\sqrt{2\sigma^2 n \log\log n}} \sum_{k=1}^{n} (X_i - E(X_1)) = 1 \text{ a.s.},$$

and that the corresponding lim inf equals −1. In the special case of a symmetric ▶random walk, this theorem had been established earlier by Khintchin (1924). The law of the iterated logarithm gives a very precise information about the behavior of the centered partial sum.

Strassen (1964) proved the *Functional Law of the Iterated Logarithm*, which concerns the normalized partial sum process, defined by

$$f_n\left(\frac{k}{n}\right) := \frac{1}{\sqrt{2\sigma^2 n \log\log n}} \sum_{i=1}^{k} (X_i - E(X_i)), 0 \leq k \leq n,$$

and linearly interpolated in between. The random sequence of functions $(f_n)_{n\geq 1}$ is almost surely relatively compact and has the following set of limit points

$$K = \{x \in C[0,1] : x \text{ is absolutely continuous and } \int_0^1 (x'(t))^2 dt \leq 1\}.$$

The functional law of the iterated logarithm gives a remarkably sharp information about the behavior of the partial sum process.

The *Almost Sure Invariance Principle*, originally established by Strassen (1964) is an important technical tool in many limit thorems. Strassen's theorem states that for i.i.d. random variables with finite variance, one can define a standard Brownian motion (see ▶Brownian Motion and Diffusions) $W(k)$ satisfying

$$\sum_{k=1}^{n} (X_k - E(X_1)) - \sigma W(n) = o(\sqrt{n \log\log n}), \text{ a.s.}.$$

Komlos et al. (1975) gave a remarkable sharperning of the error term in the almost sure invariance principle, showing that for $p > 2$ one can find a standard Brownian motion $(W_t)_{t\geq 0}$ satisfying

$$\sum_{k=1}^{n} (X_k - E(X_1)) - \sigma W(n) = o(n^{1/p}), \text{ a.s.}.$$

if and only if the random variables have finite p-th moments. In this way, results that hold for Brownian motion can be carried over to the partial sum process. E.g., many limit theorems in the statistical analysis of change-points are proved by a suitable application of strong approximations.

In the 1980s, Brosamler, Fisher and Schatte independently discovered the *Almost Sure Central Limit Theorem*, stating that for partial sums $S_k := \sum_{i=1}^{k} X_i$ of an i.i.d. sequence $(X_i)_{i\geq 1}$ with mean zero and variance σ^2

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k} 1_{\{S_k/\sigma\sqrt{k} \leq x\}} = \Phi(x),$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ denotes the standard normal distribution function. The remarkable feature of this theorem is that one can observe the central limit theorem, which in principle is a distributional limit theorem, along a single realization of the process.

In 1933, Glivenko and Cantelli independently discovered a result that is now known as the *Glivenko–Cantelli Theorem* (see ▶Glivenko-Cantelli Theorems). Given a sequence $(X_k)_{k\geq 1}$ of i.i.d random variables with distribution function $F(x) := P(X \leq x)$, we define the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_k \leq x\}}$. The Glivenko–Cantelli theorem states that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

This theorem is sometimes called the fundamental theorem of statistics, as it shows that it is possible to recover the

distribution of a random variable from a sequence of observations.

Almost sure convergence has been established for U-statistics, a class of sample statistics of great importance in mathematical statistics. Given a symmetric kernel $h(x, y)$, we define the bivariate U-statistic

$$U_n := \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h(X_i, X_j).$$

Hoeffding (1961) proved the *U-Statistic Strong Law of Large Numbers*, stating that for any integrable kernel and i.i.d. random variables $(X_i)_{i \ge 1}$,

$$U_n \xrightarrow{a.s.} Eh(X_1, X_2).$$

Aaronson et al. (1996) established the corresponding *U-Statistic Ergodic Theorem*, albeit under extra conditions. The *U-statistic Law of the Iterated Logarithm*, in the case of i.i.d. data (X_i) was established by Sen (1972). In the case of degenerate kernels, i.e., kernels satisfying $Eh(x, X_1) = 0$, for all x, this was sharpened by Dehling et al. (1985) and Dehling (1989). Their *Degenerate U-Statistic Law of the Iterated Logarithm* states that

$$\limsup_{n \to \infty} \frac{1}{n \log\log n} \sum_{1 \le i < j \le n} h(X_i, X_j) = c_h, \quad \text{a.s.},$$

where c_h is the largest eigenvalue (see ►Eigenvalue, Eigenvector and Eigenspace) of the integral operator with kernel $h(x, y)$. A functional version as well as an almost sure invariance principle were established by the same authors.

Proofs of Almost Sure Convergence

In most situations, especially in applications in Statistics, almost sure convergence is proved by identifying a given sequence as a a continuous function of a sequence of a type studied in one of the basic theorems on almost sure convergence.

The proofs of the basic almost sure convergence theorems are quite subtle and require a variety of technical tools, such as exponential inequalities, maximal inequalities, truncation techniques and the Borel-Cantelli lemma (see ►Borel–Cantelli Lemma and Its Generalizations).

About the Author

Herold Dehling (born 1954 in Westrhauderfehn, Germany) is Professor of Mathematics at the Ruhr-Universität Bochum, Germany. From 1988 to 2000, he was on the faculty of the University of Groningen, The Netherlands. Prior to that he held postdoc positions at Boston University and at the University of Göttingen. Herold Dehling studied Mathematics at the University of Göttingen and the University of Illinois at Urbana-Champaign. He obtained his Ph.D. in 1981 at Göttingen. Herold Dehling is an elected member of the International Statistical Institute. In 2005 he was awarded the Prix Gay-Lussac–Humboldt of the Republic of France. Herold Dehling conducts research in the area of asymptotic methods in probability and statistics, with special emphasis on dependent processes. He has published more than 75 research papers in probability and statistics. Herold Dehling is co-author of three books, *Kansrekening* (Epsilon Uitgaven, Utrecht 2005, with J. N. Kalma), *Einführung in die Wahrscheinlichkeitsrechnung und Statistik* (Springer, Heidelberg 2004, with B. Haupt) and *Stochastic modelling in process technology* (Elsevier Amsterdam 2007, with T. Gottschalk and A. C. Hoffmann). Moreover, he is coeditor of the books *Empirical Process Techniques for Dependent Data* (Birkhäuser, Boston 2002, with T. Mikosch and M. Sorensen) and *Weak Dependence in Probability, Analysis and Number Theory* (Kendrick Press, Utah 2010, with I. Berkes, R. Bradley, M. Peligrad and R. Tichy).

Cross References

►Brownian Motion and Diffusions
►Convergence of Random Variables
►Ergodic Theorem
►Random Variable
►Weak Convergence of Probability Measures

References and Further Reading

Aaronson J, Burton RM, Dehling H, Gilat D, Hill T, Weiss B (1996) Strong laws for L- and U-statistics. Trans Am Math Soc 348:2845–2866

Baum LE, Katz M (1965) Convergence rates in the law of large numbers. Trans Am Math Soc 120:108–123

Birkhoff GD (1931) Proof of the Ergodic theorem. Proc Nat Acad Sci USA 17:656–660

Borel E (1909) Les probabilités dénombrables et leurs application arithmétique. Rendiconti Circolo Mat Palermo 27: 247—271

Brosamler G (1988) An almost everywhere central limit theorem. Math Proc Cambridge Philos Soc 104:561–574

Cantelli FP (1933) Sulla determinazione empirica della leggi di probabilita. Gior Ist Ital Attuari 4:421–424

Csörgo M, Révész P (1981) Strong approximations in probability and statistics. Academic, New York

Dehling H, Denker M, Philipp W (1985) Invariance principles for von Mises and U-Statistics. Z Wahrsch verw Geb 67: 139–167

Dehling H (1989) The functional law of the iterated logarithm for von-Mises functionals and multiple Wiener integrals. J Multiv Anal 28:177–189

Dehling H (1989) Complete convergence of triangular arrays and the law of the iterated logarithm for U-statistics. Stat Prob Lett 7:319–321

Doob JL (1953) Stochastic processes. Wiley, New York

Dudley RM (1968) Distances of probability measures and random variables. Ann Math Stat 39:1563–1572
Fisher A (1989) Convex invariant means and a pathwise central limit theorem. Adv Math 63:213–246
Glivenko VI (1933) Sulla determinazione empirica della leggi di probabilita. Gior Ist Ital Attuari 4:92–99
Hartmann P, Wintner A (1941) On the law of the iterated logarithm. Am J Math 63:169–176
Hoeffding W (1961) The strong law of large numbers for U-statistics. University of North Carolina, Institute of Statistics Mimeograph Series 302
Hsu PL, Robbins H (1947) Complete convergence and the law of large numbers. Proc Nat Acad Sci USA 33:25–31
Khintchin A (1924) Über einen Satz der Wahrscheinlichkeitsrechnung. Fund Math 6:9–20
Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. J R Stat Soc B 30:499–510
Kolmogorov AN (1930) Sur la loi forte des grandes nombres. Comptes Rendus Acad Sci Paris 191:910–912
Komlos J, Major P, Tusnady G (1975) An approximation of partial sums of independent RVs and the sample DF I. Z Wahrsch verw Geb 32:111–131
Marcinkiewicz J, Zygmund A (1937) Sur les fonctions indépendantes. Fund Math 29:60–90
Schatte P (1988) On strong versions of the central limit theorem. Math Nachr 137:249–256
Sen PK (1972) Limiting behavior of regular functionals of empirical distributions for stationary mixing processes. Z Wahrsch verw Geb 25:71–82
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York
Skorohod AV (1956) Limit theorems for stochastic processes. Theory Prob Appl 1:261–290
Stout WF (1974) Almost sure convergence. Academic, New York
Strassen V (1964) An invariance principle for the law of the iterated logarithm. Z Wahrsch verw Geb 3:211–226
Van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wichura MJ (1970) On the construction of almost uniformly convergent random variables with given weakly convergent image laws. Ann Math Stat 41:284–291

Analysis of Areal and Spatial Interaction Data

Gunter Spöck[1], Jürgen Pilz[2]
[1]Associate Professor
University of Klagenfurt, Klagenfurt, Austria
[2]Professor
University of Klagenfurt, Klagenfurt, Austria

Areal Data

Areal data y_i are data that are assigned to spatial regions A_i, $i = 1, 2, \ldots, n$. Such data and spatial areas naturally arise at different levels of spatial aggregation, like data assigned to countries, counties, townships, political districts, constituencies or other spatial regions that are featured by more or less natural boundaries. Examples for data y_i might be the number of persons having a certain chronic illness, number of enterprises startups, average income, population density, number of working persons, area of cultivated land, air pollution, etc. Like all spatial data, also areal data are marked by the fact that they exert spatial correlation to the data from neighboring areas. Tobler (1970) expresses this in his first law of geography: "everything is related to everything else, but near things are more related than distant things." It is this spatial correlation which is investigated, modeled and taken into account in the analysis of areal data.

Spatial proximity matrix. A mathematical tool that is common to almost all areal analysis methods is the so-called $(n \times n)$ spatial proximity matrix $\mathbf{W}$, each of whose elements, w_{ij}, represents a measure of spatial proximity of area A_i and area A_j. According to Bailey and Gatrell (1995) some possible criteria might be:

- $w_{ij} = 1$ if A_j shares a common boundary with A_i and $w_{ij} = 0$ else.
- $w_{ij} = 1$ if the centroid of A_j is one of the k nearest centroids to that of A_i and $w_{ij} = 0$ else.
- $w_{ij} = d_{ij}^{\gamma}$ if the inter-centroid distance $d_{ij} < \delta$ $(\delta > 0, \gamma < 0)$; and $w_{ij} = 0$ else.
- $w_{ij} = \frac{l_{ij}}{l_i}$, where l_{ij} is the length of common boundary between A_i and A_j and l_i is the perimeter of A_i.

All diagonal elements w_{ii} are set to 0. The spatial proximity matrix $\mathbf{W}$ must not be symmetric. For instance, case 2 and case 4 above lead to asymmetric proximity matrices. For more proximity measures we refer to Bailey and Gatrell (1995) and any other textbook on areal spatial analysis like Anselin (1988).

Spatial Correlation Measures

Global measures of spatial correlation. The global Moran index I, first derived by Moran (1950), is a measure for spatial correlation of areal data having proximity matrix $\mathbf{W}$. Defining $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and $\bar{y}$, the mean of the data y_i, $i = 1, 2, \ldots, n$, the global Moran index may be written

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \qquad (1)$$

Thus the global Moran index may be be interpreted as measuring correlation between $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the spatial lag-variable $\mathbf{Wy}$. But the Moran index does not necessarily take values between -1 and 1. Its expectation for independent data y_i is $\mathrm{E}[I] = -\frac{1}{n-1}$. Values of the Moran index larger than this value thus are an indication of

positive global spatial correlation; values smaller than this value indicate negative spatial correlation.

A global correlation measure similar to the variogram known from classical geostatistics is the Geary-index (Geary's c, Geary 1954):

$$c = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2}$$

Under the independence assumption for the $\mathbf{y_i}$ its expectation is $\mathrm{E}[c] = 1$. Values of c larger than 1 indicate negative correlation and values smaller than 1 positive correlation.

The significance for Moran's I and Geary's c may be tested by means of building all $n!$ permutations of the y_i, $i = 1, 2, \ldots, n$, assigning them to the the different areas A_j, $j = 1, 2, \ldots, n$, calculating for each permutation Moran's I or Geary's c and then considering the distributions of these permuted spatial correlation statistics. True correlation statistics at the lower or upper end of these distributions are an indication of significance of the global correlation measures.

A map often useful for detecting spatial clusters of high or low values is the so-called LISA map. It may be shown that Moran's I is exactly the upward slope of the regression line between the regressors $(\mathbf{y} - \mathbf{1_n}\bar{y})$ and the spatial lag-variables $\mathbf{W}(\mathbf{y} - \mathbf{1_n}\bar{y})$ as responses, where the matrix $\mathbf{W}$ is here standardized to have rows which sum up to one. The corresponding scatterplot has four quadrants PP, NN, PN and NP, with P and N indicating positive and negative values for the regressors and responses. If one codes these four classes into which the pairs $[y_i - \bar{y}, \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})]$ may fall with colors and visualizes these colors in a map of the areas one can easily detect clusters of areas that are surrounded by low or high neighboring values.

Both statistics, the Moran I and Geary's c make a global assumption of second order stationarity, meaning that the y_i, $i = 1, 2, \ldots, n$ all have the same constant mean and variance. If one doubts that this condition is fully met one has to rely on local measures of spatial correlation, for local versions of Moran's I and Geary's c see Anselin (1995).

Spatial Linear Regression

A problem frequently occuring in areal data analysis is the regression problem. Response variables y_i and corresponding explanatory vectors $\mathbf{x_i}$ are observed in spatial areas A_i, $i = 1, 2, \ldots, n$ and one is interested in the linear regression relationship $y_i \approx \mathbf{x_i}^T \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown regression parameter vector to be estimated. Subsuming all row vectors $\mathbf{x_i}^T$ in the $(n \times p)$ design matrix $\mathbf{X}$ and writing $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ the ordinary ▶least squares solution to this regression problem, which does not take account of spatial correlation, is known to be $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. If the data in $\mathbf{y}$ are known to be correlated the above ordinary least squares estimator is known to be inefficient and statistical significance tests in this regression model are known to be misleading. Problems may be resolved by considering the generalized least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$, where the covariance matrix Σ is measuring the correlation between the data in $\mathbf{y}$. All regression procedures used in areal data analysis deal more or less with the modeling and estimation of this covariance structure Σ and the estimation of β. In all subsequent sections we will assume that the spatial proximity matrix $\mathbf{W}$ is standardized such that its rows sum up to one.

Simultaneous autoregressive model (SAR). The SAR model is given as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},\ \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}. \tag{3}$$

Here λ is an unknown parameter, $-1 < \lambda < 1$, measuring spatial correlation; the parameters λ and $\boldsymbol{\beta}$ are to be estimated. The error vector $\boldsymbol{\epsilon}$ has uncorrelated components with constant unknown variances σ^2, like $\mathbf{u}$ it has expectation zero. The two equations may be combined to get

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} - \lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Obviously $\mathbf{y}$ is modeled as being influenced also by the spatial lag-variables $\mathbf{W}\mathbf{y}$ and the spatial lag-regression $\mathbf{W}\mathbf{X}\boldsymbol{\beta}$. The coefficient λ is measuring the strength of this influence. The covariance matrix of $\mathbf{u}$ may be shown to be $\mathrm{cov}[\mathbf{u}] = \sigma^2((\mathbf{I_n} - \lambda\mathbf{W})^T(\mathbf{I_n} - \lambda\mathbf{W}))^{-1}$. An estimation procedure for the SAR model is implemented in the R-package spdep, Bivand (2006). It is based on the Gaussian assumption for $\mathbf{y}$ and iteratively calculates maximum (profile) likelihood estimates for σ^2 and λ and generalized least squares estimates for $\boldsymbol{\beta}$ based on the covariance matrix $\mathrm{cov}[\mathbf{u}]$ and the estimates for σ^2 and λ calculated a step before.

Spatial lag model. The so-called spatial lag model may be written

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4}$$

It is simpler in structure than the SAR model because the lag-regression term $-\lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ is missing. For its estimation, again, an iterative profile likelihood procedure similar to the SAR procedure may be used.

Spatial Durbin model. The spatial Durbin model is a generalization of the SAR model and given as

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{5}$$

with $\mathbf{W}\mathbf{X}\boldsymbol{\gamma}$ having its own regression parameter vector $\boldsymbol{\gamma}$. By means of the restriction $\boldsymbol{\gamma} = -\lambda\boldsymbol{\beta}$ the Durbin model

becomes equivalent to a SAR model. The so-called common factor test (Florax and de Graaf 2004), a likelihood ratio test, can be used to decide between the two hypotheses, - SAR-model and spatial Durbin model. As an alternative to the above models one may also use a SAR model with a lag-error component

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \lambda \mathbf{W}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}. \tag{6}$$

Deciding between models. For the investigation whether a SAR model, a spatial lag model or ordinary least squares give the best fit to the data one may adopt Lagrange multiplier tests as described in Florax and de Graaf (2004). Interestingly, these tests are based on ordinary least squares residuals and for this reason are easily calculable. Breitenecker (2009) gives a nice overview on all the possibilities related to testing models.

Geographically weighted regression. Fotheringham et al. (2002) propose, as an alternative to the above mentioned regression models, geographically weighted regression. The proposed methodology is particularly useful when the assumption of stationarity for the response and explanatory variables is not met and the regression relationship changes spatially. Denoting by (u_i, v_i) the centroids of the spatial areas A_i, $i = 1, 2, \ldots, n$, where the responses y_i and explanatory vectors $\mathbf{x_i}$ are observed, the model for geographically weighted regression may be written

$$y_i = \mathbf{x_i}^T \boldsymbol{\beta}(u_i, v_i) + \epsilon_i, i = 1, 2, \ldots, n. \tag{7}$$

The regression vector $\boldsymbol{\beta}(u_i, v_i)$ is thus dependent on the spatial location (u_i, v_i) and is estimated by means of a weighted least squares estimator that is locally dependent on a diagonal weight matrix $\mathbf{C_i}$:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = (\mathbf{X}^T \mathbf{C_i} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C_i} \mathbf{y}$$

The diagonal elements $c_{jj}^{(i)}$ of $\mathbf{C_i}$ are defined by means of a kernel function, e.g. $c_{jj}^{(i)} = \exp(-d_{ij}/h)$. Here d_{ij} is a value representing the distance beetween A_i and A_j; d_{ij} may either be Euclidean distance or any other metric measuring distance between areas. Further, h is the bandwidth measuring how related areas are and can be determined by means of crossvalidating the residuals from the regression or based on the ▶Akaike's information criterion (Brunsdon et al. 1998). Selecting the bandwidth h too large results in oversmoothing of the data. On the other hand a bandwidth too small allows for too less data during estimation.

All areal analysis methods discussed so far are implemented in the R-packages spdep and spgwr, (Bivand 2006, 2009). Methods for counting data, as they frequently appear in epidemiology, and Bayesian methods are not dealt with here; for those methods the interested reader is referred to Lawson (2009).

Spatial Interaction Data

This is a further category of spatial data which is related to modeling the "flow" of people and/or objects between a set of origins and a set of destinations. In contrast with areal (and geostatistical) data, which are located at points or in areas, spatial interaction data are related to pairs of points, or pairs of areas. Typical examples arise in health services (e.g., flow to hospitals), transport of freight goods, population migration and journeys-to-work. Good introductory material on spatial interaction models can be found in Haynes and Fotheringham (1984).

The primary objective is to model *aggregate* spatial interaction, i.e. the volume of flows, not the flows at an individual level. Having m origins and n destinations with associated flow data considered as random variables Y_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$, the general spatial interaction model is of the form

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij};\ i = 1, \ldots, m;\ j = 1, \ldots, n \tag{8}$$

where $E(Y_{ij}) = \mu_{ij}$ and ε_{ij} are error terms with $E(\varepsilon_{ij}) = 0$. The goal is then to find suitable models for μ_{ij} involving flow propensity parameters of the origins i, attractiveness parameters of the destinations j, and the effects of the "distances" d_{ij} between them. Here, the quantities d_{ij} may be real (Euclidean) distances, travel times, costs of travel or any other measure of the separation between origins and destinations. One of the most widely used classes of models for μ_{ij} is the so-called *gravity model*

$$\mu_{ij} = \alpha_i \beta_j \exp(\gamma\ d_{ij}) \tag{9}$$

involving origin parameters α_i, destination parameters β_j and a scaling parameter γ. Under the assumption that the Y_{ij} are independent Poisson random variables with mean μ_{ij}, this model can be treated simply as a particular case of a generalised linear model with a logarithmic link. Model fitting can then proceed by deriving maximum likelihood estimates of the parameters using iteratively weighted least squares (IRLS) techniques. The above gravity models can be further enhanced when replacing the parameters β_j by some function of observed covariates $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jk})^T$ characterising the attractiveness of each of the destinations $j = 1, \ldots, n$. Again, this is usually done in a log-linear way, and the model becomes

$$\mu_{ij} = \alpha_i \exp(g(\boldsymbol{x}_j, \boldsymbol{\theta}) + \gamma\ d_{ij}) \tag{10}$$

where g is some function (usually linear) of the vector of destination covariates and a vector of associated parameters $\boldsymbol{\theta}$. Contrary to (9), which reproduces both the total flows from any origin and the total observed flows to each destination, the new model (10) is only *origin-constrained*. The obvious counterpart to (10) is one which is *destination-constrained*:

$$\mu_{ij} = \beta_j \exp(h(\boldsymbol{z}_i, \boldsymbol{\omega}) + \gamma\, d_{ij})$$

where h is some function of origin characteristics $\boldsymbol{z}_i$ and a vector of associated parameters $\boldsymbol{\omega}$. Finally, when modeling both α_i and β_j as functions of observed characteristics at origins and destinations, we arrive at the *unconstrained model*

$$\log \mu_{ij} = h(\boldsymbol{z}_i, \boldsymbol{\omega}) + g(\boldsymbol{x}_j, \boldsymbol{\theta}) + \gamma\, d_{ij} \qquad (11)$$

In population migration one often uses a particular form of (11), where $\boldsymbol{z}_i$ and $\boldsymbol{x}_j$ are taken to be univariate variables meaning the logarithms of the population P_i and P_j at origin i and destination j, respectively. Adding an overall scaling parameter τ to reflect the general tendency for migration, the following simple model results:

$$Y_{ij} = \tau P_i^{\omega} P_j^{\theta} \exp(\gamma\, d_{ij}) + \varepsilon_{ij} \qquad (12)$$

Likewise, in all the above models one can introduce more complex distance functions than $\exp(\gamma\, d_{ij})$. Also, as mentioned before, d_{ij} could be replaced by a general separation term s_{ij} embracing travel time, actual distance and costs of overcoming distance.

The interaction models considered so far are only models for μ_{ij}, the mean flow from i to j. Thus, they are only first order models, no second order effects are included and the maximum likelihood methods for estimating the parameters of the gravity models rest on the explicit assumption that fluctuations about the mean are independent. Up to now, there has been only little work done on taking account of spatially correlated errors in interaction modeling. To address such problems, pseudo-likelihood-methods are in order. Good references for further reading on spatial interaction models are Upton and Fingleton (1989), Bailey and Gatrell (1995) and Anselin and Rey (2010).

Spatial interaction models have found broad attention among (economic) geographers and within the GIS community, but have received only little attention in the spatial statistics community. The book by Anselin and Rey (2010) forms a bridge between the two different worlds. It contains a reprint of the original paper by Getis (1991), who first suggested that the family of spatial interaction models is a special case of a general model of spatial autocorrelation. Fischer et al. (2010) present a generalization of the Getis-Ord statistic which enables to detect local non-stationarity and extend the log-additive model of spatial interaction to a general class of spatial econometric origin-destination flow models, with an error structure that reflects origin and/or destination autoregressive spatial dependence. They finally arrive at the general spatial econometric model (3), where the design matrix $\boldsymbol{X}$ includes the observed explanatory variables as well as the origin, destination and separation variables, and $\boldsymbol{W}$ is a row-standardized spatial weights matrix.

About the Author

For biography of the author Jürgen Pilz *see* the entry ►Statistical Design of Experiments (DOE).

Cross References

►Data Depth
►Gaussian Processes
►Geostatistics and Kriging Predictors
►Model-Based Geostatistics
►Spatial Point Pattern
►Spatial Statistics
►Statistical Ecology

References and Further Reading

Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht

Anselin L (1955) Local indicators of spatial association – LISA. Geogr Anal 27:93–115

Anselin L, Rey SJ (eds) (2010) Perspectives on spatial data analysis. Springer, Berlin

Bailey T, Gatrell A (1995) Interactive spatial data analysis. Longman Scientific and Technical, New York

Breitenecker R (2009) Raeumliche lineare modelle und autokorrelationsstrukturen in der gruendungsstatistik. Ibidem, Stuttgart

Bivand R (2006) SPDEP: spatial dependence: weighting schemes, statistics and models. R-package Version 0.3-12

Bivand R (2009) SPGWR: geographically weighted regression. R-package Version 0.6-2

Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression – modelling spatial non-stationary. The Statistician 47:431–443

Fischer MM, Reismann M, Scherngell Th (2010) Spatial interaction and spatial autocorrelation. In: Rey SJ, Anselin A (eds) perspective on spatial data analysis. Springer, Berlin, pp 61–79

Florax R, de Graaf T (2004) The performance of diagnostic tests for spatial dependence in linear regression models: a meta-analysis of simulation studies. In: Anselin L et al (eds) Advances in spatial econometrics: methodology, *tools and applications*. Springer, Berlin, pp 29–77

Fotheringham S, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Geary R (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145
Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environ plann A23:1269–1277
Haynes KF, Fotheringham AS (1984) Gravity and spatial models. Sage, London
Lawson A (2009) Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC, Chapman and Hall, New York
Moran P (1950) Notes on continuous stochastic phenomena. Biometrica 37:17–23
Tobler W (1970) A computer model simulating urban growth in the Detroit region. Econ Geogr 46:234–240
Upton GJG, Fingleton B (1989) Spatial data analysis by example, vol. 2. Wiley, Chichester

Analysis of Covariance

James J. Cochran
Associate Professor
Louisiana Tech University, Ruston, LA, USA

Introduction

The Analysis of Covariance (generally known as ANCOVA) is a statistical methodology for incorporating quantitatively measured independent observed (not controlled) variables in a designed experiment. Such a quantitatively measured independent observed variable is generally referred to as a covariate (hence the name of the methodology – analysis of covariance). Covariates are also referred to as concomitant variables or control variables.

If we denote the general linear model (GLM) associated with a completely randomized design as

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m$$

where
Y_{ij} = the ith observed value of the response variable at the jth treatment level
μ = a constant common to all observations
τ_j = the effect of the jth treatment level
ε_{ij} = the random variation attributable to all uncontrolled influences on the ith observed value of the response variable at the jth treatment level

For this model the within group variance is considered to be the experimental error, and this implies that the treatments have similar effects on all experimental units. However, in some experiments the effect of the treatments on the experimental units varies systematically with some characteristic that varies across the experimental units. For example, one may test for a difference in the efficacy of a new medical treatment and an existing treatment protocol by randomly assigning the treatments to patients (experimental units) and testing for a difference in the outcomes. However, if the ▶randomization results in the placement of a disproportionate number of young patients in the group that receives the new treatment and/or placement of a disproportionate number of elderly patients in the group that receives the existing treatment, the results will be biased if the treatment is more (or less) effective on young patients than it is on elderly patients. Under such conditions one could collect additional information on the patients' ages and include this variable in the model. The resulting general linear model

$$Y_{ij} = \mu + \tau_j + \beta X_{ij} + \varepsilon_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m.$$

where
X_{ij} = the ith observed value of the covariate at the jth treatment level,
β = the estimated change in the response that corresponds to a one unit increase in the value of the covariate at a fixed level of the treatment

is said to be a completely randomized design ANCOVA model and describes an experimental design GLM one factor experiment with a single covariate.

Note that the addition of covariate(s) can accompany many treatment and design structures. This article focuses on the simple one way treatment structure in a completely randomized design for the sake of simplicity and brevity.

Purpose of ANCOVA

There are three primary purposes for including a covariate in the ▶analysis of variance of an experiment:

1. To increase the precision of estimates of treatment means and inferences on differences in the response between treatment levels by accounting for concomitant variation on quantitative but uncontrollable variables. In this respect covariates are the quantitative analogies to blocks (which are qualitative/categorical) in that they are (1) not controlled and (2) used to remove a systematic source of variation from the experimental error. Note that while the inclusion of a covariate will result in a decrease in the experimental error, it will also reduce the degrees of freedom associated with the experimental error, and so inclusion of a covariate in an experimental model will not always result in greater precision and power.

2. To allow for the assessment of the nature of the relationship between the covariate(s) and the response variable after taking into consideration the treatment effects. In this respect covariates are analogous to independent variables in linear regression, and their associated slopes can provide important insight into the nature of the relationship between the response and the covariate.
3. To statistically adjust comparisons of the response between groups for imbalances in quantitative but uncontrollable variables. In this respect covariates are analogous to stratification and are of particular importance in situations where stratification on the covariate is impractical or infeasible.

Applications of ANCOVA

Typical applications of analysis of covariance include:

- Clinical trials in which quantitative but uncontrollable variables such as the weight, height, and age of the patients may influence the effectiveness of a treatment protocol.
- Marketing research in which quantitative but uncontrollable variables such as the pretest rating of a product given by a respondent may influence the respondent's posttest rating (i.e., after exposure to the test condition) of the product.
- Education experiments in which quantitative but uncontrollable variables such as the age, intelligence (if this can be measured), and prior scholastic performance of the students may influence the effectiveness of a pedagogical approach.
- Agricultural experiments in which quantitative but uncontrollable variables such as rainfall and historical yield of fruit bearing trees may influence the yield during an experiment.

Comparing Treatments in ANCOVA

Least squares means (or *LS* means) are generally used to compare treatment effects in experiments that include one or more covariates. *LS* means (which are sometimes referred to as marginal means, estimated marginal means, or adjusted treatment means) are the group means when the covariate is set equal to its grand mean $\overline{X}_m$ (mean of the covariate over all observations across all treatments). These are easily calculated by substituting the grand mean of the covariate into the estimated general linear model, i.e.,

$$\hat{Y}_j = \mu + \tau_j + \beta\overline{X}_m,\ j = 1, \ldots, m$$

Standard errors for *LS* means are typically calculated and used (in conjunction with the ►asymptotic normality of *LS* means) to conduct inference on individual *LS* means and contrasts based on the *LS* means.

Assumptions of ANCOVA

In addition to the standard ANOVA assumptions:

- Independence of error terms
- Homogeneity of variance of the error terms across treatments
- Normality of the error terms across treatments

One must also consider the *regression assumptions* when performing statistical inference with ANCOVA. The regression assumptions include:

- A linear relationship exists between the covariate and the response variable.

  If no relationship exists between the covariate and response, there is no reason to include the covariate in the experiment or resulting model. If the relationship between the covariate and the response variable is nonlinear, the inclusion of a covariate in the model will not remove all variation in the observed values of the response that can potentially be accounted for by the covariate. The nature of the relationship between the covariate and the response can be assessed with scatter plots of these two variables by treatment. If a nonlinear relationship exists between the covariate and the response, one can utilize a polynomial ANCOVA model.
- Homogeneity of the regression slopes associated with the covariate (i.e., parallel regression lines across treatments).

  The calculations of the *LS* means are predicated on the lack of existence of a response by covariate interaction. If this assumption is violated, the adjustment to the response variable for a common value of the covariate is misleading. This assumption can be assessed through either scatter plots of the covariate and the response by treatment or through the inclusion of a treatment-covariate interaction in the model.

If the sample results suggest that any of these assumptions are not satisfied, inference based on the model may not be valid.

Alternatives to ANCOVA

Bonate (2000) provides a good discussion of alternatives to ANCOVA in pretest-posttest designs; he considers the relative merits of difference scores, relative change functions, various blocking methods, and repeated-measures analysis. Several authors have suggested more general nonparametric alternatives to ANCOVA based on an analysis

of covariance of the ranks of the response and covariance. Some notable examples of these approaches have been suggested by Quade (1967, 1982), Puri and Sen (1969), McSweeney and Porter (1971), Burnett and Barr (1977), Shirley (1981), Conover and Iman (1982), Chang (1993), Lesaffre and Senn (2003), and Tsangari and Akritas (2004).

About the Author

For biography *see* the entry ►Role of Statistics in Advancing Quantitative Education.

Cross References

►Analysis of Variance
►General Linear Models
►Multivariate Analysis of Variance (MANOVA)
►Nonparametric Models for ANOVA and ANCOVA Designs
►Rank Transformations
►Statistical Fallacies: Misconceptions, and Myths
►Statistics: An Overview

References and Further Reading

Bonate PL (2000) Analysis of pretest-posttest designs. Chapman and Hall/CRC, Boca Raton

Burnett TD, Barr DR (1977) A nonparametric analogy of analysis of covariance. Educ Psychol Meas 37(2):341–348

Chang GH (1993) Nonparametric analysis of covariance in block designs. Dissertation (Texas Tech University)

Conover WJ, Iman RL (1982) Analysis of covariance using the rank transformation. Biometrics 38:715–724

Doncaster CP, Davey AJH (2007) Analysis of variance and covariance: how to choose and construct models for the life sciences. Cambridge University Press, Cambridge

Huitema BE (1980) The analysis of covariance and alternatives. Wiley, New York

Lesaffre E, Senn S (2003) A note on non-Parametric ANCOVA for covariate adjustment in randomized clinical trials. Stat Med 22(23):3583–3596

McSweeney M, Porter AC (1971) Small sample properties of nonparametric index of response and rank analysis of covariance. Presented at the Annual Meeting of the American Educational Research Association, New York

Milliken GA, Johnson DE (2002) Analysis of messy data vol.3: analysis of covariance. Chapman and Hall, New York

Puri ML, Sen PK (1969) Analysis of covariance based on general rank scores. Ann Math Stat 40:610–618

Quade D (1967) Rank analysis of covariance. J Am Stat Assoc 62:1187–1200

Quade D (1982) Nonparametric analysis of covariance by matching. Biometrics 38:597–611

Rutherford A (2001) Introducing ANOVA and ANCOVA: a GLM approach. Sage, Los Angeles

Shirley EA (1981) a Distribution-Free Method for Analysis of Covariance Based on Ranked Data. J Appl Stats 30:158–162

Tsangari H, Akritas MG (2004) Nonparametric ANCOVA with two and three covariates. J Multivariate Anal 88(2):298–319

Analysis of Multivariate Agricultural Data

Asghar Ali
Professor and Chairman
Bahauddin Zakariya University, Multan, Pakistan

Agricultural research is most oftenly based on observational studies and experimentation resulting in multi-response variables. The selection of appropriate variety to grow; amount and types of fertilizers, insecticides and pesticides to apply; the irrigation system to use; the plant sowing technology to apply and to assess the soil fertility through chemical analysis of macro and micro nutrients available in the soil are the major areas of interest for the researcher to work on for the improvement of the agricultural productivity in terms of quality and quantity. The role of Statistics in planning agricultural research, designing experiments, data collection, analysis, modeling and interpretation of agricultural results is very well established. The basic principles and theoretical development of experimental designs pioneered by R. A. Fisher are the result of collaborative work of agricultural scientists and statisticians. In the process of experimentation and observational studies, the researcher is keen to have as many data information as possible so that nothing is left unattended related to the phenomenon under study as there will be no chance to repeat the experiment till the next season of the crop and it will not be less than a miracle if data from one year of the crop is consistent with the results of second year, no matter how much care is taken to keep the experimental conditions identical.

Agricultural data obtained through experimentation is initially analyzed using ►analysis of variance technique and then depending on the nature of treatments/factors applied, either the approach of multiple comparisons or ►response surface methodology is used to explore further the hidden features of the data. For example, the experimenter might be interested to compare different varieties of a particular crop such that there are two local varieties (V_1, V_2) in practice; three varieties are imported (V_3, V_4, V_5) and two new varieties (V_6, V_7) are developed by a local agricultural institute. If results obtained from analysis of variance conclude that performance of the varieties is significantly different from each other then obvious questions arise are to test the difference between the following variety comparisons: $[V_1$ and $V_2]$; $[V_6$ and $V_7]$; $[(V_1, V_2)$ and $(V_6, V_7)]$; $[(V_1, V_2, V_6, V_7)$ and $(V_3, V_4, V_5)]$; and if V_4 is a hybrid variety, then one has two more comparisons to test i.e., $[V_4$ with $(V_3, V_5)]$

and $[V_3$ with $V_5]$. These contrasts are orthogonal to each other but it will not always be the case, other techniques of multiple comparisons will have to be used then, which are available in almost all the books on experimental designs. On the other hand, if multifactor experiments are conducted to determine the appropriate levels of the applied factors on which optimum response is achieved. For this purpose data sets are modeled in adequate functional forms and the researcher is intended to fit simple functional form. Ordinary polynomials are the most popular functional forms which are used to model experimental data from many fields of scientific research. If first order polynomial is fitted, the researcher very simply states that the concerned factor has linear effect and the interpretation is made accordingly that with increase of levels of factor will result in increase (or decrease) in the response. The second order polynomials are used with the expectation that it will be possible to identify the levels of the applied factors to get the optimum response. A number of response functions that have been widely used by the agricultural and biological researchers have been discussed by Mead and Pike (1975). It should not be taken as granted that one response function considered applicable to one sort of situation will also be applicable to other similar situations; it is advisable that graphical approach be used to guess the appropriate functional form of the response under consideration. An extremely useful concept that is revealed by Nelder (1966) is known as Inverse Polynomial Response Functions (IPRF). It emphasizes that in agricultural research the effect of increasing a factor indefinitely is either to produce a saturation effect, in which case the response does not exceed a finite amount, or to produce a toxic effect, in which case the response eventually fall to zero and the response curve has no build-in-symmetry.

Nelder (1966) and Pike (1977) advocated these surfaces as giving responses that are nonnegative and bounded if regression coefficients are constrained to be positive and it is further assumed that $Var(Y) \propto [E(Y)]^2$. Extension in the ideas has been developed by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) for the response variables that may not be normal and that the expected response may be a function of the linear predictors rather than just the linear predictors itself. Ali (1983) and Ali et al. (1986) have objected on placing constrains on the parameters as it will violate all the properties of good estimators and will no longer follow the distributional structure required for valid inferences. Their experience of examining many sets of data leads them not to expect all regression coefficients to be positive. Taking into account the error structure and functional form used for IPRF, Ali (1983) proposed the form of a response function called as Log Linear Response Functions (LLRF) based on the logarithmic transformation of the response variable and assuming that $\log Y \sim N(E(\log Y), \sigma^2)$. The estimation of regression coefficients achieved by carrying out a multiple regression of $\log Y$ on the terms required fitting the data adequately; the resulting estimators are therefore Minimum Variance Linear Unbiased Estimators. It is simple to estimate the variance-covariance matrix of these estimators and to test hypotheses concerning parameters by the usual linear regression methods. On the theoretical grounds the LLRF model therefore has much to commend it. The assumption that $\log Y_i$ follows the normal distribution may not always be true; in such cases it is recommended that Box-Cox family of transformation may be used under the same structure of the response function as has been used for LLRF and IPRF.

In order to produce an adequate prediction the researcher is usually uncertain as to which of the large number of terms should be included in the final model. The main point to bear in mind is that it should have as many terms as necessary so that maximum variation of the data is explained and as few terms as possible so that it can easily be interpreted. Ali (1983) argued that for summarizing the data from agricultural experiments the terms in the final model are required to be selected in a conforming order by preferring main effect terms over the interactions and lower order terms over the higher. It is further to remember that the inverse terms describe the rising ridge of the surfaces, the linear terms describe the optimum region and the higher degree terms contribute in explaining the falling portion of the surfaces. It is therefore recommended that for building the appropriate model one should concentrate on selection the inverse and linear terms along with their associated interaction terms. One who is not convinced with such types of model building method has the option to use the approach established by Nelder (1977).

The methods used for selection of final model are mainly based on the Minimum Mean Square Error criterion. It is possible to find more than one models which fulfills this criterion. In such cases one should select the one which has reasonable shape of the response surface, capable to determine the values of quantitative factors at which the response is an optimum, statistically significant regression coefficients and simple functional form.

There is no ambiguity to recognize the agricultural research as multifactor and multi-response and that these responses are measured at different stages of the maturity of the crop and that these are interrelated with one another. The univariate analyses of these variables therefore have partial impact on the true findings of research. Multivariate analyses are therefore natural and essential to consider the

data by giving due weight to the interrelationships among the variables under study. One possible approach which is widely used by the researchers is to study the correlation matrix of the variables. This approach only facilitates to assess the relationship among the pairs of variables and it can be extended to triplets of variables by considering the partial correlations and the multiple correlations among those. As a result there would be $\frac{1}{6}p(p^2 + 1)$ pairs and triplets to consider and it will certainly be confusing if the number p of variables under consideration is large.

To overcome this difficulty, Principal Component Analysis (PCA) can be used. It is a multivariate technique that has its aim the explanation of relationships among several difficult-to-interpret, correlated variables in terms of a few conceptually meaningful components which are uncorrelated with each other and are capable of accounting for nearly all the variation present in the observed data. PCA therefore finds a linear transformation of original variables into a new set of variables called as principal components which are uncorrelated with each other; are capable of accounting for the variation of the obtained data and are derived in such a way that the first few of them can often provide enough information about the data and so the dimensionality of the problem can considerably be reduced. The variables with higher component loadings in a particular principal component are considered to be the important ones and it is assumed that the principal component is the representative of these variables; hence it is interpreted only in terms of these variables. This approach of interpretation of principal components is acceptable if principal components are extracted using the correlation matrix R. The variance–covariance matrix Σ is as well used to derive principal components; since the principal component technique is scale dependent, the principal component loadings with this approach will therefore be much influenced by the unit of the measurements of the variables under consideration, hence, the interpretation of principal components just based on the magnitude of the loading may become questionable and misleading. Ali et al. (1985) suggested using the correlation between the original variables and the principal component for selection of representative variables in a particular principal component instead of using principal component loadings.

PCA is extremely useful technique when interest lies in investigating the interrelationship within a set of variables; when the relationship of two sets of variables, within and among the sets is of interest, the PCA is not a valid technique. The agricultural researchers always encounter with such types of problems where the assessment of relationship among and within the twosets is essential e.g. the interdependence of nutritional status and vegetative related characteristics with the crop yield related characteristics is pivotal. For such cases, ►Canonical Correlation Analysis (CCA) technique developed by Hotelling (1936) is of great benefit. It has certain maximal properties similar to those of PCA and in a way is an extension of the multiple regression analysis. The object of this approach is to find the linear functions of the variables for each of the sets such that the correlation between these linear functions is as high as possible. After locating such a pair of linear functions which are maximally correlated with each other, we look for other pairs of linear functions which are maximally correlated subject to the restriction that the new pair of linear functions must be uncorrelated with all other previously located functions. For the purpose of interpretation of the results, it is proposed to use correlation between the canonical variates and the original variables instead of canonical weights as has been already proposed for interpretation of PCA results. Details of PCA and CCA may be found in Mardia et al. (1979) and Jolliffe (2002).

About the Author

Asghar Ali holding M.Sc and D.Phil in Statistics from Sussex University and Post Doctorate from University of Kent at Canterbury, UK is Professor of Statistics at Bahauddin Zakariya University (BZU) Multan, Pakistan. Since 1975, he is serving BZU in various capacities: Presently, he is Chairman, Department of Statistics and also for a period of three years, he held the position of chairmanship, Department of Computer Science. Especially commendable have been his period as Principal, College of Agriculture, BZU. He has been publishing regularly related to data analysis in the field of agricultural sciences which is now being cited in related text books.

Cross References

►Agriculture, Statistics in
►Canonical Correlation Analysis
►Farmer Participatory Research Designs
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Principal Component Analysis

References and Further Reading

Ali A (1983) Interpretation of multivariate data: Comparison of several methods of interpreting multivariate data from a series of nutritional experiments, University of Sussex, Unpublished PhD thesis

Ali A, Clarke GM, Trustrum K (1985) Principal component analysis applied to some data from fruit nutrition experiments. The Statistician 34:365–370

Ali A, Clarke GM, Trustrum K (1986) Log-linear response functions and their use to model data from plant nutrition experiments. J Sci Food & Agric 37:1165–1177
Hotelling H (1936) Relation between two sets of variates. Biometrika 28:321–377
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, USA
Mardia KV, Kent JT, Bibi JM (1979) Multivariate analysis. Academic, London
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
Mead R, Pike DJ (1975) A review of response surface methodology from a biometric viewpoint. Biometrics 31(4):803–851
Nelder JA (1966) Inverse polynomials, a useful group of multifactor response functions. Biometrics 22:128–141
Nelder JA (1977) A reformation of linear models (with discussion), J R Stat Soc A140:48–76
Nelder JA, Wedderburn WM (1972) Generalized linear models. J R Stat Soc (General) A135(3):370–384
Pike DJ (1977) Inverse polynomials: A study of parameter estimation procedures and comparison of the performance of several experimental design criteria. University of Reading, Unpublished PhD thesis

Analysis of Variance

Gudmund R. Iversen
Professor Emeritus
Swarthmore College, Swarthmore, PA, USA

Analysis of variance is the name given to a collection of statistical methods originally used to analyze data obtained from experiments. The experiments make us of a quantitative dependent variable, also known as a metric variable or an interval or ratio variable, and one or more qualitative independent variables, also known as categorical or nominal variables. These analysis methods grew out of agricultural experiments in the beginning of the twentieth century, and the great English statistician Sir Ronald Fisher developed many of these methods. As an example, the dependent variable could be the yield in kilos of wheat from different plots of land and the independent variable could by types of fertilizers used on the plots of land.

Experimental Design

The way an experiment is run affects the particular analysis of variance method used for the analysis of the data. Experiments are designed according to different plans, and the choice of the design of the experiment affects which analysis of variance method being used. Without going into details about designs of experiments, an experiment could follow a factorial design, a randomized block design, a Latin square design, etc. There exist too many designs of experiments and accompanying analysis of variance methods for the analysis of the resulting data to cover all of them in this short presentation. But it is possible to present the underlying features of all analysis of variance methods.

Analysis of Variance and Multiple Regression

But first it is worth noting that analysis of variance is closely related to regression analysis. Indeed, it is possible to see both analyses as special cases of the so-called general linear model. In particular, using ▶dummy variables for the independent variables in analysis of variance, the analysis quickly turns into a regression analysis. The main difference is that when data are collected through a properly designed experiment, it is possible to conclude that there is a causal effect of the independent variable(s) on the dependent variable. When data are collected through observational, studies there may be a causal effect of the independent variable(s) or not.

Statistical Software

Much of the early work on analysis of variance consisted of finding efficient ways of making the necessary computations with the use of simple calculators. With the introduction of modern statistical software for electronic computers, this line of work is now less important. Instead, statisticians have worked on showing the similarities of the computations needed for both analysis of variance and multiple regression, and the old distinction between the two approaches to data analysis is no longer of any importance. However, statistical software packages still make a distinction between the two, and the output from the two methods often look very different.

One-Way Analysis of Variance

This name is given to the design where there is one independent nominal variable with several categories and a quantitative dependent variable with a unit of measurement and often a meaningful zero. An example of an experiment could be where students are randomly assigned to two different groups and the students in one group were taught using a new method of teaching while the students in the second group, as a control group, were taught using the old method. The random assignment to the different groups means that the effects of all other variables, for

example gender, is canceled out, and any observed difference between the two groups is causally due to the teaching method being used. In this simple case, the statistical method is the same as the *t*-test for the difference between two groups. From a regression point of view we could use a dummy variable and assign the value 0 to all the students in the control group and the value 1 to all the students in the experiment group. In this case, the intercept of the regression line would equal the mean of the dependent variable for the control group and the slope of the line would equal the difference between the means of the two groups. Thus, the *t*-test for the null hypothesis that the population regression line has zero slop becomes the same as the *t*-test for the difference between the two means.

The fundamental question in an analysis of variance is whether the population means for different groups are equal or not. But the methods for analysis of variance use variances to answer the question about means, thus the name analysis of variance. The analysis is based on identifying two factors that determine the values of the dependent variable. One such factor is the net effect of all factors except the independent variable, known as the residual variable, and the other factor is the independent variable.

The Residual Sum of Squares

If the residual variable had no effect, then all the values of the dependent variable for the control group would be equal to each other, and all the values of the dependent variable for the experimental group would be equal to each other. The best estimates of these two values would be the mean of the dependent variable for the group. To the extent that the values within each group are not equal, is due to the residual variable. Thus, the effect of the residual variable for a single observation can be seen as the difference between the observed value and the group mean. For each observation we now have such a difference. One way to summarize the values of these differences for a group is to square each difference and add all these squares. We then have a sum of squares for each of the two groups, and by adding these two sums we have a measure of the overall effect of the residual variable. If the dependent variable is known as Y and y_{ij} is the ith observation in the jth group and $\bar{y}_j$ is the mean in the jth group, then the residual sum of squares RSS can be written

$$RSS = \sum \sum (y_{ij} - \bar{y}_j)^2.$$

Note that there are many other ways we could combine these differences. For example, we could have taken the absolute value of each difference and added those differences instead of using squares. Thus, the final conclusion from the analysis should include a statement that the conclusion is based on squares and not some other mathematical operation. Even though nobody does include such a warning, it should be made clear that the analysis is based on squares.

The Treatment Sum of Squares

We also need a measure of how different the two groups are from each other. One way to do that is find how different the group means are from the overall mean. If the treatment variable has no effect, then the two group means would be equal and equal to the overall mean. One way to measure how different the group means are from the overall mean is to take each group mean and subtract the overall mean. By squaring each difference and weighing each square by the number of observations in the group n_j, then the treatment sum of squares between the groups GSS can be written

$$GSS = \sum \sum n_j (\bar{y}_j - \bar{y})^2.$$

The *F*-Test

The residual sum of squares is also known as the within group sum of squares and the group sum of squares is sometimes known as the between group squares. The final step consists of making a comparison between the two sums of squares. If the residual sum of squares is large in comparison with the group sum of squares, then it seems that the difference between the group means is not statistically significant. For this comparison we take into how many groups we have, here 2 and in general k groups, and how many observations n there are all together. A mathematical development shows that we should compute the ratio

$$F = \frac{GSS/(k-1)}{RSS/(n-k)}$$

This is known as the F-ratio and is named in honor of Ronald Fisher. It gives rise to the F-distribution, and the distribution has been extensively tabulated. The two numbers $(k-1)$ and $(n-k)$ are the so-called degrees of freedom, and they are used to take into account how many groups there are in the experiment and how many observations there are in the experiment. For example, for a 5% significance level with $k = 2$ groups and $n = 30$ observations, the critical value of F on 1 and 28 degrees of freedom equals 4.20. Thus, for any observed value of F larger than 4.20, we conclude that there is a statistically significant difference between the two groups. In this case, had we done a *t*-test

for the difference between the special case of two group means, the critical value of t becomes $\sqrt{4.20} = 2.05$.

Other Analyses

It is possible to generalize to an experiment with more than just two groups. The null hypothesis of equal group means is tested the same way as with two groups, and the computations follow the same plan as above. With two or more independent variables the analysis becomes more extensive. We can still represent the independent variables by dummy variables and do a regression analysis. But that way it is easy to overlook the possible interaction effect of the two independent variables. This means we could have an effect of the independent variables together over and beyond their separate effects. Finally, in analysis of variance we distinguish between using all values of the independent variables (Model I) and only using a sample of possible values (Model II).

About the Author

Dr. Gudmund Iversen is Professor Emeritus of Statistics at Swarthmore College, Swarthmore PA 19081. He chaired the Department of Mathematics and Statistics at three different intervals and directed the College's Center for Social and Policy Studies for several years. Prior to Swarthmore he taught statistics at the University of Michigan and also directed the Summer Training Program, Inter-university Consortium for Political and Social Research. He was a visiting lecturer in the American Statistical Association visiting lecture program. He has been a visiting professor at Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, West Germany 1986, at Department of Political Science, University of Oslo, Norway (1978–1979), at The Graduate School of Social Work and Social Research, Bryn Mawr College, spring 1990, fall 2000, and at School of Social Policy and Practice at the University of Pennsylvania 1999. He was a member of the joint committee of Mathematical Association of America and American Statistical Association on the statistics curriculum (1991–1997). He was Associate book review editor, *Journal of American Statistical Association* (1986–1989), and Associate editor, *Journal of Statistics Education* (1993–2000). He has published 22 articles and 10 books on statistics and statistical education.

Cross References

►Agriculture, Statistics in
►Analysis of Covariance
►Analysis of Multivariate Agricultural Data
►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Data Analysis
►Experimental Design: An Introduction
►F Distribution
►Farmer Participatory Research Designs
►General Linear Models
►Graphical Analysis of Variance
►Multiple Comparison
►Multiple Comparisons Testing from a Bayesian Perspective
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Data Analysis: An Overview
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Rank Transformations
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Statistical Software: An Overview
►Statistics: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Hinkelmann K, Kempthorne O (2008) Design and analysis of experiments, I and II, 2nd edn. Wiley, New York
Iversen GR, Norpoth H (1987) Analysis of variance, 2nd edn. Sage Baverly Hills, CA

Analysis of Variance Model, Effects of Departures from Assumptions Underlying

Hardeo Sahai[1], Mohammed I. Ageel[2], Anwer Khurshid[3]
[1]Professor
University of Puerto Rico, San Juan, Puerto Rico
[2]Founder and President of the Saudi Association of Statistical Sciences, Professor
Jazan University, Jazan, Saudi Arabia
[3]Professor
University of Karachi, Karachi, Pakistan

Introduction

Every statistical model has its own underlying "assumptions" that must be verified to validate the results. In some situations, violations of these assumptions will not change substantive research conclusions, while in others,

violation of assumptions can be critical to meaningful research. For a meaningful and conclusive data analysis by ►Analysis of Variance (ANOVA), the following assumptions are needed:

(a) Errors be normally distributed
(b) Errors have same variances (homogeneity of variances)
(c) Errors be independently distributed

However, the question arising is What would be the effects of any departure from the assumptions of the model on the inferences made? The answer is simple: It may either influence the probability of making Type I error (i.e., incorrectly rejecting null hypothesis) or a Type II error (i.e., failing to reject a null hypothesis when it is false). For a thorough discussion of the topic, the reader is referred to Scheffé (1959), Miller (1986), Snedecor and Cochran (1989), Sahai and Ojeda (2005), and Sahai and Ageel (2000). Some of the main findings are discussed in the following section.

Effects of Departures from Assumptions

Departures from Normality

For fixed effects model, due to the central limit theorem (see ►Central Limit Theorems) the lack of normality causes no problems in large samples, as long as the assumptions hold. In general, when true ►randomization occurs the violations of normality is acceptable. Also, heterogeneity of variances can result in nonnormality, so ensuring homogeneity of variances may also result in normality. Only highly skewed distributions would have a marked effect either on the level of significance or the power of the F test. However, it is worth mentioning that kurtosis of the error distribution (either more or less peaked than a normal distribution) is more important than skewness of the distribution in terms of the effects on inferences. Both analytical results (see, e.g., Scheffé 1959:345–351) and the empirical studies by Pearson (1931), Geary (1947), Gayen (1950), Box and Anderson (1955), Boneau (1960, 1962), Srivastava (1959), Bradley (1964), Tiku (1964, 1971), and Donaldson (1968) attest to the conclusion that lack of normality would have little effect of F test either in terms of level of significance or power. Hence, the F test is generally robust against departures from normality (in skewness and/or kurtosis) if sample sizes are large or even if moderately large. For instance, the specified level of significance might be 0.05, whereas the actual level for a nonnormal error distribution might vary from 0.044 to 0.052 depending on the sample size and the magnitude of the kurtosis. Generally, the actual level of significance in the presence of positive kurtosis (platykurtic) is slightly higher than the specified one and the real power of the test for positive kurtosis is slightly higher than the normal one. If the underlying population has negative kurtosis (leptokurtic), the actual power of the test will be slightly lower than the normal one (Glass et al. 1972). Single interval estimates of the factor level means and contrasts and some of the multiple comparison methods are also not much affected by the lack of normality provided the sample sizes are not too small. The robustness of multiple comparison tests in general has not been as thoroughly studied. Among few studies in this area is that of Brown (1974). Some other studies have investigated the robustness of several multiple comparison procedures, including Tukey and Scheffé, for exponential and chi-square distributions and found little effect on both α and power (see, e.g., Petrinovich and Hardyck 1969; Keselman and Rogan 1978). Dunnett (1982) reported that Tukey is conservative both with respect to α and power for long-tailed distributions and to ►outliers. Similarly, Ringland (1983) found that Scheffé was conservative for distributions with influence to outliers.

Lange and Ryan (1989) gave several examples that show that nonnormality of random effects is, indeed, encountered in practice. For random effects model, the lack of normality has more serious implications than fixed effects model. The estimates of the variance components are still unbiased, but the actual confidence coefficients for interval estimates of σ_e^2, σ_α^2, $\sigma_\alpha^2/\sigma_e^2$ may be substantially different from the specified one (Singhal and Sahai 1992). Moreover, when testing the null hypothesis, if the variance of a random effect is some specified value different from zero, the test is not robust to the assumption of normality. For some numerical results of this, the reader is referred to Arvesen and Schmitz (1970) and Arvesen and Layard (1975). However, if one is concerned only with a test of hypothesis $\sigma_\alpha^2 = 0$, then slight departures from normality have only minor consequences for the conclusions reached when the sample size is reasonably large (see, e.g., Tan and Wong 1980; Singhal et al. 1988).

Departures from Equal Variances

Both the analytical derivations by Box (1954) and the empirical studies indicate that if the variances are unequal, the F test for the equality of means under fixed effects model is only slightly affected provided there is no remarkable difference in sample sizes and the parent populations are approximately normally distributed. When the variances are unequal, an approximate test similar to the approximate t test when two group variances are unequal may be used (Welch 1956). Generally, unequal error variances increase the actual level of significance slightly

higher than the specified level and result in a slight elevation of the power function to a degree related to the magnitude of differences among variances. If larger variances are associated with larger sample sizes, the level of significance will be slightly less than the nominal value, and if they are associated with smaller sample sizes, the level of significance will be slightly greater than the nominal value (Horsnell 1953; Kohr and Games 1974). Similarly, the Scheffé's multiple comparison procedure based on the ►*F* distribution is not affected to any appreciable degree by unequal variances if the sample sizes are approximately equal. Thus, the *F* test and the related analyses are robust against unequal variances if the sample sizes are nearly equal.

On the other hand, when different number of cases appear in various samples, violation of the assumption of homogeneous variances can have serious effects in the validity of the final inference (see, e.g., Scheffé 1959; Welch 1956; James 1951; Box 1954; Brown and Forsythe 1974; Bishop and Dudewicz 1978; Tan and Tabatabai 1986). Krutchkoff (1988) made an extensive simulation study to determine the size and power of several analysis of variance procedures, including the *F* test, Kruskal–Wallis test, and a new procedure called the *K* test. It was found that both the *F* test and the Kruskal–Wallis test are highly sensitive whereas the *K* test is relatively insensitive to the heterogeneity of variances. Kruskal–Wallis test, however, is not as sensitive to the unequal error variances as the *F* test and was found to be more robust to nonnormality (when the error variances are equal) than either the *F* test or the *K* test. Thus, whenever possible, the experimenter should try to achieve the same number of cases in each factor level unless the assumption of equal population variances can reasonably be assured in the experimental context. The use of equal sample sizes for all factor levels not only tends to minimize the effects of unequal variances using the *F* test, but also simplifies the computational procedure.

For random effects model, however, the lack of homoscedasticity or unequal error variances can have serious effects on inferences about the variance components, even when all factor levels contain equal sample sizes. Boneau (1960) has shown that when variances are different in the various groups and sample sizes are small and different, ANOVA can produce highly misleading results.

Departures from Independence of Error Terms

Lack of independence can result from biased measurements or possibly from a poor allocation of treatments to experimental units. Nonindependence of the error terms can have important effects on inferences for both fixed and random effects models. If this assumption is not met, the *F* ratio may be strongly affected severely in serious errors in inferences (Scheffé 1959). The direction of the effect depends on the nature of the dependence of the error terms. In most cases encountered in practice, the dependence tends to make the value of the ratio too large and consequently the significance level will be smaller than it should be (although the opposite can also be true). Since the remedy of violation of this assumption is often difficult, every possible effort should be made to obtain independent random samples. The use of randomization in various stages of the study can be most important protection against independence of error terms. In general, great care should be taken to ensure that the data are based on independent observations, both between and within groups, i.e., each observation is in no way related to any of the other observations. Although, dependency among the error terms creates a special problem in any analysis of variance, it is not necessary that the observations themselves must be completely independent for applying the random effects model.

In summary, ANOVA is very robust to violations of the assumptions, as long as only one assumption is violated. If two or more assumptions are severely violated the results are not to be trusted. Further if the data are:

(a) Not normally distributed, but satisfies the homogeneity of variance and independent assumptions, the findings may still be valid.
(b) Normally distributed and are independent samples, but does not satisfy the homogeneity of variance assumption, the findings may still be valid.

The above review and discussion are restricted to the one-way analysis of variance. A similar finding for two-way classification without and with interaction can be found in Sahai and Ageel (2000).

Tests for Departures from Assumptions

As we have seen in the preceding section, the analysis of variance procedure is robust and can tolerate certain departures from the specified assumptions. It is, nevertheless, recommended that whenever a departure is suspected it should be checked out. In this section, we shall briefly state the tests for normality and homoscedasticity.

Tests for Normality

A relatively simple technique to determine the appropriateness of the assumption of normality is to graph the data points on a normal probability paper. If a straight line can be drawn through the plotted points, the assumption of normality is considered to be reasonable. Some formal tests for normality are the chi-square goodness of fit test, and the

tests for skewness and kurtosis that are often used as supplements to the chi-square test (see ▶Chi-Square Tests). For a detailed discussion of these tests refer to Sahai and Ageel (2000).

The tests mentioned above are some of the classical tests of normality. Over the years, a large number of other techniques have been developed for testing for departures from normality. Some powerful omnibus tests proposed for the problem are Shapiro–Wilk's test (Shapiro and Wilk 1965), Shapiro–Francia's test (Shapiro and Francia 1972), and D'Agostino's test (D'Agostino 1971).

For a discussion of tests especially designed for detecting outliers see Barnett and Lewis (1994). Robust estimation procedures have also been employed in detecting extreme observations. The procedures give less weight to data values that are extreme in comparison to the rest of the data. Robust estimation techniques have been reviewed by Hampel et al. (1986).

Tests for Homoscedasticity

If there are just two populations, the equality of two population variances can be tested by using the usual F test. However, more than two population, rather than making all pairwise F tests, we want a single test that can be used to verify the assumption of equality of population variances. There are several tests available for this purpose. The three most commonly used tests are the Bartlett's, Hartley's, and Cochran's tests. The ▶Bartlett's test (Bartlett 1937a, b) compares the weighted arithmetic and geometric means of the sample variances. The Hartley's test (Hartley 1950) compares the ratio of the largest to the smallest variance. The Cochran's test (Cochran 1941) compares the largest sample variance with the average of all the sample variances. For a full description of these procedures and illustration of their applications with examples see Sahai and Ageel (2000). These tests, however, have lower power than is desired for most applications and are adversely affected by nonnormality. Detailed practical comments on Bartlett's, Hartley's, and Cochran's tests are also given by Sahai and Ageel. In recent years, there have appeared a number of tests in the literature that are less sensitive to normality in the data and are found to have a good power for a variety of population distributions see Levene (1960). Following Levene (1960), a number of other robust procedures have been proposed, which are essentially based on techniques of applying ANOVA to transformed scores. For example, Brown and Forsythe (1974a) proposed applying an ANOVA to the absolute deviations from the mean. A somewhat different approach known as ▶jackknife was proposed by Miller (1968) where the original scores in each group are replaced by the contribution of that observation to the group variance. O'Brien (1979, 1981) proposed a procedure, which is a blend of Levene's squared deviation scores and the jackknife. In recent years, there have been a number of studies investigating the robustness of these procedures. For a further discussion and details, the reader is referred to Conover et al. (1981), Olejnik and Algina (1987), and Ramsey (1994).

Corrections for Departures from Assumptions of the Model

Departure from independence could arise in an experiment in which experimental units or plots are laid out in a field so that adjacent plots give similar yields. Lack of independence can also result from correlation in time rather than in space. If the data set in a given problem violates the assumptions of the analysis of variance model, a choice of possible corrective measures is available. One approach is to modify the model. However, this approach has the disadvantage that more often than not the modified model involves fairly complex analysis. Another approach may be to consider using some nonparametric tests. A third approach to be discussed in this section is to use transformations on the data. Sometimes it is possible to make an algebraic transformation of the data to make them appear more nearly normally distributed, or to make the variances of the error terms constant. Conclusions derived from the statistical analyses performed on the transformed data are also applicable to the original data. In this section, we briefly discuss some commonly used transformations to correct for the lack of normality and homoscedasticity. An extremely thorough and detailed monograph on transformation methodology has been prepared by Thöni (1967). An excellent and thorough introduction and a bibliography of the topic can be found in a review paper by Hoyle (1973). For a more recent bibliography of articles on transformations see Draper and Smith (1981:683–684).

Transformations to Correct Lack of Normality

Some transformations to correct for the departures from normality are logarithmic transformation, square-root transformation, and arcsine transformation.

Transformations to Correct Lack of Homoscedasticity

There are several types of data in which the variances of the error terms are not constant. If there is evidence of some systematic relationship between treatment mean and variance, homogeneity of the error variance may be achieved through an appropriate transformation of the data. Bartlett (1936) has given a formula for deriving such transformations provided the relationship between μ_i and σ_e^2 is known. In many cases where the nature of the relationship is not clear, the experimenter can, through trial

and error, find a transformation that will stabilize the variance. We give some commonly employed transformations to stabilize the variance. These are logarithmic transformation, square-root transformation, reciprocal transformation, arcsine transformation, and power transformation. For a detailed discussion of these transformations and their applicability refer to Sahai and Ageel (2000).

These are some of the more commonly used transformations. Still other transformations can be found applicable for various other relationships between the means and the variances. Further, the transformations to stabilize the variance also often make the population distribution nearly normal. For equal sample sizes, however, these transformations may not usually be necessary. Moreover, the use of such transformations may often result in different group means. It is possible that the means of the original scores are equal but the means of the transformed scores are not, and vice versa. Further, the means of transformed scores are often changed in ways that are not intuitively meaningful or are difficult to interpret.

Acknowledgment

We are grateful to Professor Miodrag Lovric, editor-in chief, for his nice editorial efforts and for his valuable suggestions and comments that led to the improvement in the contribution.

About the Authors

Dr. Hardeo Sahai held Professorial and visiting Professorial positions at the University of Puerto Rico, Mayaguez and San Juan, University of Ceara, Brazil, University of Granada, Spain, University of Veracruzana, Mexico, University of Nacional de Colombia, University of Nacional de Trujillo, Peru. He has received the University of Kentuky Outstanding Alumnus award, Medal of honor University of Granada (Spain), Plaque of honor University of Nacional de Trujillo (Peru). He has published over 150 papers in the statistical (bio)medical and epidemiological literature and is coauthor of the several books which include: *Statistics in Epidemiology: Methods, Techniques and Applications* (with Anwer Khurshid, CRC Press, Boca Raton, Florida, 1996), and *Analysis of Variance for Random Models, Vol. 1: Balanced Data and Vol. 2: Unbalanced Data* (with Mario M. Ojeda, Birkhäuser 2004). Professor Sahai is a Fellow of the American Statistical Association, Royal Statistical Society and Elected Member of the International Statistical Institute.

Dr. Mohammed Ibrahim Ali Ageel was a Full Professor and Chairman of Mathematics, Department of Mathematics, King Saud University (KSU), Saudi Arabia, and later the Chairman of Mathematics Department, King Khalid University (KKU), Saudi Arabia. He was the Dean of Graduate School, King Khalid University, Saudi Arabia. Professor Ageel was also a Full Professor of Mathematics and Dean of Engineering, Najran University (NU), Saudi Arabia. Ageel is currently a Full Professor of Mathematics, Jazan University, Jazan, Saudi Arabia. He is a Founder and President of Saudi Association of Statistical Sciences (SASS). He is an Elected member of the International Statistical Institute, and was elected as a Fellow of the Royal Statistical Society. He has published more than 50 research papers and articles in both theoretical and applied areas. Professor Ageel is a coauthor of the book *The Analysis of Variance: Fixed, Random and Mixed Models* (with Hardeo Sahai, Birkhäuser, Boston 2000).

Anwer Khurshid is a Professor at the Department of Statistics, University of Karachi, Pakistan. During 2004–2010 he had a faculty position at the Sultan Qaboos University, Oman and University of Nizwa, Oman. He is author or coauthor of more than 70 papers and two books (both with Professor Hardeo Sahai). In recognition of his teaching and research contributions Professor Khurshid was awarded a certificate of appreciation in 1997 by the Chancellor, University of Karachi, Pakistan.

Cross References

►Analysis of Variance
►Bartlett's Test
►Heteroscedasticity
►Normality Tests
►Robust Inference
►Robust Statistics
►Tests for Homogeneity of Variance

References and Further Reading

Arvesen JN, Layard MWJ (1975) Asymptotically robust tests in unbalanced variance component models. Ann Stat 3:1122–1134

Arvesen JN, Schmitz TH (1970) Robust procedures for variance component problems using the jackknife. Biometrics 26:677–686

Barnett VD, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, New York

Bartlett MS (1936) The square root transformation in the analysis of variance. J R Stat Soc 3:68–78

Bartlett MS (1937a) Properties of sufficiency and statistical tests. Proc R Soc Lond Ser A 160:268–282

Bartlett MS (1937b) Some examples of statistical methods of research in agriculture and applied biology. J R Stat Soc Suppl 4: 137–183

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

Boneau CA (1962) A comparison of the power of the U and t tests. Psychol Rev 59:246–256

Box GEP (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. Ann Math Stat 25:290–302

Box GEP, Anderson SL (1955) Permutation theory in the derivation of robust criteria and the study of departures from assumption. J R Stat Soc Ser B 17:1–26

Bradley JV (1964) Studies in research methodology, VI. The central limit effect for a variety of populations and the robustness of Z, t, and F. Technical report no. 7 AMRL-54-123. Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio

Brown RA (1974) Robustness of the studentized range statistic. Biometrika 61:171–175

Brown MB, Forsythe AB (1974a) Robust tests for the equality of variances. J Am Stat Assoc 69:364–367

Brown MB, Forsythe AB (1974b) The small size sample behavior of some statistics which test the equality of several means. Technometrics 16:129–132

Brown MB, Forsythe AB (1974c) The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics 30:719–724

Cochran WG (1941) The distribution of the largest of a set of estimated variances as a fraction of their total. Ann Eugen 11:47–52

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances with applications to outer continental shelf bidding data. Technometrics 23:351–361 (Corrigendum Technometrics 26:302)

D'Agostino RB (1971) An omnibus test of normality for moderate and large size samples. Biometrika 58:341–348

Donaldson TS (1968) Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. J Am Stat Assoc 63:660–676

Draper NR, Smith H (1981) Applied regression analysis, 2nd edn. Wiley, New York

Dunnett CW (1982) Robust multiple comparisons. Commun Stat Part A: Theory Methods 11:2611–2629

Gayen AK (1950) The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika 37:236–255

Geary RC (1947) Testing for normality. Biometrika 34:209–242

Glass GV, Peckham PD, Sanders JR (1972) Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Rev Educ Res 42:239–288

Hampel FR, Rochetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Horsnell G (1953) The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika 40: 128–136

Hoyle MH (1973) Transformations – an introduction and a bibliography. Int Stat Rev 41:203–223

James GS (1951) The comparison of several groups of observations when the ratio of population variances are unknown. Biometrika 38:324–329

Keselman HJ, Rogan JC (1978) A comparison of modified-Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. J Am Stat Assoc 73:47–51

Kohr RL, Games PA (1974) Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. J Exp Educ 43:61–69

Krutchkoff RG (1988) One way fixed effects analysis of variance when the error variances may be unequal. J Stat Comput Simul 30:259–271

Lange N, Ryan L (1989) Assessing normality in random effects models. Ann Stat 17:624–642

Levene H (1960) Robust tests for equality of variances. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds) Contributions to probability and statistics. Stanford University Press, Stanford, pp 278–292

Miller RG Jr (1968) Jackknifing variances. Ann Math Stat 39: 567–582

O'Brien RG (1979) An improved ANOVA method for robust tests of additive models for variances. J Am Stat Assoc 74:877–880

O'Brien RG (1981) A simple test for variance effects in experimental designs. Psychol Bull 89:570–574

Olejnik SF, Algina J (1987) Type I error rates and power estimates of selected parametric and nonparametric tests of scales. J Educ Stat 12:45–61

Pearson ES (1931) The analysis of variance in cases of non-normal variation. Biometrika 23:114–133

Petrinovich LF, Hardyck CD (1969) Error rates for multiple comparison methods. Psychol Bull 71:43–54

Ramsey PH (1994) Testing variances in psychological and educational research. J Educ Stat 19:23–42

Ringland JT (1983) Robust multiple comparisons. J Am Stat Assoc 78:145–151

Sahai H, Ageel MI (2000) The analysis of variance: fixed, random and mixed models. Birkhäuser/Springer, Boston

Sahai H, Ojeda M (2005) Analysis of variance for random models: unbalanced data. Birkhauser, USA

Scheffé H (1959) The analysis of variance. Wiley, New York

Shapiro SS, Francia RS (1972) An approximate analysis of variance test for normality. J Am Stat Assoc 67:215–216

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611

Singhal RA, Sahai H (1992) Sampling distribution of the ANOVA estimator of between variance component in samples from a non-normal universe. J Stat Comput Simul 43:19–30

Singhal RA, Tiwari CB, Sahai H (1988) A selected and annotated bibliography on the robustness studies to non-normality in variance component models. J Jpn Stat Soc 18: 195–206

Snedecor GW, Cochran WG (1989) Statistical methods, 8th edn. Iowa State University Press, Ames

Srivastava ABL (1959) Effects of non-normality on the power of the analysis of variance test. Biometrika 46:114–122

Tan WY, Tabatabai MA (1986) Some Monte Carlo studies on the comparison of several means under heteroscedasticity and robustness with respect to departure from normality. Biom J 28:801–814

Tan WY, Wong SP (1980) On approximating the null and non-null distributions of the F ratio in unbalanced random effects models from non-normal universes. J Am Stat Assoc 75: 655–662

Thöni H (1967) Transformation of variables used in the analysis of experimental and observational data: a review. Technical report no. 7. Statistical Laboratory, Iowa State University, Ames

Tiku ML (1964) Approximating the general non-normal variance-ratio sampling distributions. Biometrika 51:83–95

Tiku ML (1971) Power function of F-test under non-normal situations. J Am Stat Assoc 66:913–916

Welch BL (1956) On linear combinations of several variances. J Am Stat Assoc 51:132–148

Anderson–Darling Tests of Goodness-of-Fit

Theodore W. Anderson
Professor of Statistics and Economics, Emeritus
Stanford University, Stanford, CA, USA

Introduction

A "goodness-of-fit" test is a procedure for determining whether a sample of n observations, $x_1, \ldots, x_n$, can be considered as a sample from a given specified distribution. For example, the distribution might be a normal distribution with mean 0 and variance 1. More generally, the specified distribution is defined as

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad -\infty < x < \infty, \tag{1}$$

where $f(y)$ is a specified density. This density might be suggested by a theory, or it might be determined by a previous study of similar data.

When X is a random variable with distribution function $F(x) = \Pr\{X \le x\}$, then $U = F(X)$ is a random variable with distribution function

$$\Pr\{U \le u\} = \Pr\{F(X) \le u\} = u, \quad 0 \le u \le 1. \tag{2}$$

The model specifies $u_1 = F(x_1), \ldots, u_n = F(x_n)$ as a sample from the distribution (2), that is, the standard uniform distribution (see ▶Uniform Distribution in Statistics) on the unit interval $[0,1]$ written $U(0,1)$.

A test of the hypothesis that $x_1, \ldots, x_n$ is a sample from a specified distribution, say $F^0(x)$, is equivalent to a test that $u_1 = F^0(x_1), \ldots, u_n = F^0(x_n)$ is a sample from $U(0,1)$. Define the *empirical distribution function* as

$$F_n(x) = \frac{k}{n}, \quad -\infty < x < \infty, \tag{3}$$

if k of $(x_1, \ldots, x_n)$ are $\le x$. A goodness-of-fit test is a comparison of $F_n(x)$ with $F^0(x)$. The hypothesis $H_0 : F(x) = F^0(x)$, $-\infty < x < \infty$, is rejected if $F_n(x)$ is very different from $F^0(x)$. "Very different" is defined here as

$$W_n^2 = n\int_{-\infty}^{\infty} \left[F_n(x) - F^0(x)\right]^2 \psi\left[F^0(x)\right] dF^0(x)$$
$$= n\int_{-\infty}^{\infty} \left[F_n(x) - F^0(x)\right]^2 \psi\left[F^0(x)\right] f^0(x)dx \tag{4}$$

being large; here (1) holds and $\psi(z)$ is a weight function such that $\psi(z) \ge 0$, and $f^0(x)$ is the density of $F^0(x)$.

If $\psi(z) = 1$, the statistic W_n^2 is the Cramér-von Mises statistic, denoted by $n\omega^2$. Anderson and Darling (1952) gave a table of the limiting distribution of $n\omega^2$ as $n \to \infty$. For example, the 5% significance point is .46136 and the 1% significance point is .74346.

The Anderson–Darling Statistic

For a given x and hypothetical distribution $F^0(\cdot)$, the random variable $n F_n(x)$ has a ▶binomial distribution with probability $F^0(x)$. The expected value of $n F_n(x)$ is $nF^0(x)$ and the variance is $n F^0(x)\left[1 - F^0(x)\right]$. The definition of the goodness-of-fit statistic (4) permits the choice of weight function $\psi(\cdot)$. In particular the investigator may want to emphasize the tails of the presumed distribution $F^0(x)$. In that case the choice is

$$\psi(u) = \frac{1}{u(1-u)}. \tag{5}$$

Then for a specified x

$$\sqrt{n}\frac{F_n(x) - F^0(x)}{\sqrt{F^0(x)\left[1 - F^0(x)\right]}} \tag{6}$$

has mean 0 and variance 1 when the null hypothesis is true. The Anderson–Darling statistic is

$$A_n^2 = n\int_{-\infty}^{\infty} \frac{\left[F_n(x) - F^0(x)\right]^2}{F^0(x)\left[1 - F^0(x)\right]} dF^0(x). \tag{7}$$

It was shown in Anderson and Darling (1954) that (7) can be written as

$$A_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\log u_{(j)} + \log\left(1 - u_{(n-j+1)}\right)\right] \tag{8}$$

where $u_{(j)} = F^0\left(x_{(j)}\right)$ and $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ is the ordered sample.

Anderson and Darling found the limiting distribution of A_n^2 [for weight function (5)]. In the next section the development of this distribution is outlined. The 5% significance point of the limiting distribution is 2.492 and the 1% point is 3.880. The mean of this limiting distribution is 1 and the variance is $2(\pi^2 - 9)/3 \sim .57974$.

Outline of Derivation

Let $u = F^0(x)$, $u_i = F^{(0)}(x_i)$, $i = 1, \ldots, n$, and $u_{(i)} = F^{(0)}(x_{(i)})$, $i = 1, \ldots, n$. Let $G_n(u)$ be the empirical distribution function of $u_1, \ldots, u_n$; that is

$$G_n(u) = \frac{k}{n}, \quad 0 \le u \le 1, \tag{9}$$

if k of $u_1, \ldots, u_n$ are $\le u$. Thus

$$G_n\left[F^0(x)\right] = F_n^0(x), \tag{10}$$

and

$$W_n^2 = n \int_0^1 [G_n(u) - u]^2 \psi(u) du, \tag{11}$$

when the null hypothesis $F(x) = F^{(0)}(x)$ is true. For every u $(0 \le u \le 1)$

$$Y_n(u) = \sqrt{n}\,[G_n(u) - u] \tag{12}$$

is a random variable, and the set of these may be considered as a stochastic process with parameter u. Thus

$$\Pr\left\{W_n^2 \le z\right\} = \Pr\left\{\int_0^1 Y_n^2(u)\psi(u)du \le z\right\} = A_n(z), \tag{13}$$

say. For a fixed set $u_1, \ldots, u_k$ the k-variate distribution of $Y_n(u_1), \ldots, Y_n(u_k)$ approaches a multivariate normal distribution (see ▶Multivariate Normal Distributions) as $n \to \infty$ with mean and covariance function

$$\mathcal{E}[Y_n(u)] = 0, \quad \mathcal{E}Y_n(u)Y_n(v) = \min(u,v) - uv. \tag{14}$$

The limiting process of $\{Y_n(u)\}$ is a Gaussian process $y(u)$, $0 \le u \le 1$, and $\mathcal{E}y(u) = 0$ and $\mathcal{E}y(u)y(v) = \min(u,v) - uv$. Let

$$a(z) = \Pr\left\{\int_0^1 y^2(u)\psi(u)du \le z\right\}. \tag{15}$$

Then $A_n(z) \to a(z)$, $0 \le z < \infty$. The mathematical problem for the Anderson–Darling statistic is to find the distribution function $a(z)$ when $\psi(u) = 1/u(1-u)$.

We briefly sketch the procedure to find the distribution of $\int_0^1 z^2(u)du$, where $z(u)$ is a Gaussian stochastic process with $\mathcal{E}z(u) = 0$ and $\mathcal{E}z(u)z(v) = k(u,v)$. When the kernel is continuous and square integrable (as is the case here), it can be written as

$$k(u,v) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(u) f_j(v), \tag{16}$$

where λ_j is an eigenvalue and $f_j(u)$ is the corresponding normalized eigenfunction of the integral equation

$$\lambda \int_0^1 k(u,v) f(u) du = f(v), \tag{17}$$

$$\int_0^1 f_j^2(u)du = 1, \quad \int_0^1 f_i(u)f_j(u)du = 0, \quad i \ne j. \tag{18}$$

Then the process can be written

$$z(u) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} X_j f_j(u), \tag{19}$$

where $X_1, X_2, \ldots,$ are independent $N(0,1)$ variables. Then

$$\int_0^1 z^2(u)du = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} X_j^2, \tag{20}$$

with characteristic function

$$\mathcal{E}\exp\left[it\int_0^1 z^2(u)du\right] = \prod_{j=1}^{\infty} \mathcal{E}\left(\exp itX_j^2/\lambda_j\right) = \prod_{j=1}^{\infty}\left(1 - 2it/\lambda_j\right)^{-\frac{1}{2}}. \tag{21}$$

The process $Y_n^*(u) = \sqrt{\psi(u)}\,Y_n(u)$ has covariance function

$$k(u,v) = \sqrt{\psi(u)}\sqrt{\psi(v)}\,[\min(u,v) - uv]; \tag{22}$$

as $n \to \infty$, the process $Y_n^*(u)$ approaches $y^*(u) = \sqrt{\psi(u)}\,y(u)$ with covariance (22). The characteristic function of the limiting distribution of $n\omega^2$ is

$$\sqrt{\frac{\sqrt{2it}}{\sin\sqrt{2it}}} \tag{23}$$

for $\psi(u) = 1$, and that of the limiting distribution of A_n^2 is

$$\sqrt{\frac{-2\pi it}{\cos\left(\frac{\pi}{2}\sqrt{1+8it}\right)}}. \tag{24}$$

for $\psi(u) = 1/u(1-u)$.

The integral equation (17) can be transformed to a differential equation

$$h''(t) + \lambda\psi(t)\,h((t) = 0. \tag{25}$$

Anderson–Darling Tests with Unknown Parameters

When parameters in the tested distribution are not known, but are estimated efficiently, the covariance (14) is modified, and the subsequent limiting distribution theory for both $n\omega^2$ and A_n^2 follows the same lines as above, with this new covariance. If the parameters are location and/or scale, the limiting distributions do not depend on the true parameter values, but depend on the class of tested distributions. If the parameters are shape parameters, the limiting distribution depends on shape. Limiting distributions have been evaluated and percentage points given for a number of different tested distributions; see Stephens (1976, 1986). Tests for three parameter Weibull, and von Mises have been given by Lockhart and Stephens (1985, 1994).

The percentage points for these tests are much smaller than those given above for the case when parameters are known.

Acknowledgments

The assistance of Michael A. Stephens is gratefully acknowledged.

About the Author

Professor Anderson was born June 5, 1918 in Minneapolis, Minnesota. He is Past President, Institute of Mathematical Statistics (1963), Vice President, American Statistical Association (1971–1973), Fellow of the American Academy of Arts and Sciences (elected 1974), Member of the National Academy of Sciences (elected 1976). Professor Anderson has been awarded the R. A. Fisher Award of Committee of Presidents of Statistical Societies (1985) and Samuel S. Wilks Memorial Medal, American Statistical Association (1988). He holds four honorary doctorates. Professor Anderson has published over 170 articles in statistical, econometric, and mathematical journals, and seven books, including the internationally recognized text *An Introduction to Multivariate Statistical Analysis* (1958, Wiley; 3rd edition 2003). The 17th International Workshop in Matrices and Statistics was held in Tomar (Portugal July 2008), in honour of Professor Theodore Wilbur Anderson 90th birthday. *The Collected Papers of T. W. Anderson: 1943–1985* (edited by George P. H. Styan) were published by Wiley in 1990, comprising 109 papers and 16 commentaries, in a 2-volume set covering 1,681 pages.

Cross References

►Cramér-Von Mises Statistics for Discrete Distributions
►Jarque-Bera Test
►Kolmogorov-Smirnov Test
►Normality Tests
►Normality Tests: Power Comparison
►Omnibus Test for Departures from Normality
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Anderson TW, Darling DA (1952) Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Ann Math Stat 23:193–212
Anderson TW, Darling DA (1954) A test of goodness-of-fit. J Am Stat Assoc 49:765–769
Lockhart RA, Stephens MA (1985) Tests of fit for the von-Mises distribution. Biometrika 72:647–652
Lockhart RA, Stephens MA (1994) Estimation and tests of fit for the three-parameter Weibull distribution. J R Stat Soc B 56:491–500
Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369
Stephens MA (1986) In: D'Agostino R, Stephens MA (eds) Goodness-of-fit techniques, chap. 4. Marcel Dekker, New York

Approximations for Densities of Sufficient Estimators

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

Introduction

Durbin (1980a) proposed a simple method for obtaining asymptotic expansions for the densities of sufficient estimators. The expansion is a series which is effectively in powers of n^{-1}, where n is the sample size, as compare with the ►Edgeworth expansion which is in powers of $n^{-1/2}$. The basic approximation is just the first term of this series. This has an error of order n^{-1} compare to the error of $n^{-1/2}$ in the usual asymptotic normal approximation (see ►Asymptotic Normality). The order of magnitude of the error can generally be reduced to order $n^{-3/2}$ by renormalization.

Suppose that the real m-dimensional random vector $\mathbf{S}_n = (S_{1n}, S_{2n}, \ldots, S_{mn})'$ has a density with respect to Lebesgue measure which depends on integer $n > N$ for some positive N and on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a subset of $\mathbb{R}^q$ for q an arbitrary positive integer.

Let

$$\mathbf{D}_n(\boldsymbol{\theta}) = n^{-1} E\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\}\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\}' \qquad (1)$$

which we assume is finite and positive-definite for all n and $\boldsymbol{\theta}$, and which we assume converges to a finite positive-definite matrix $\mathbf{D}(\boldsymbol{\theta}_0)$ as $n \to \infty$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a particular value of $\boldsymbol{\theta}$, usually the true value.

Let $\phi_n(\mathbf{z}, \boldsymbol{\theta}) = E(e^{i\mathbf{z}'\mathbf{S}_n})$ be the characteristic function of $\mathbf{S}_n$ where $\mathbf{z} = (z_1, z_2, \ldots, z_m)'$. Whenever the appropriate derivatives exists, let

$$\frac{\partial^j \log \phi_n(\widetilde{\mathbf{z}}, \boldsymbol{\theta})}{\partial \mathbf{z}^j}$$

denote the set of jth order derivatives $\partial^j \log \phi_n(\mathbf{z},\boldsymbol{\theta})/\partial z_1^{j_1}\cdots\partial z_m^{j_m}$ for all integers $j_1, j_2, \ldots, j_m \geq 0$ satisfying $\sum_k j_k = j$, evaluated at $\mathbf{z} = \tilde{\mathbf{z}}$. The jth cumulant $\boldsymbol{\kappa}_{nj}(\boldsymbol{\theta})$ of $\mathbf{S}_n$, where it exists, satisfies the relation

$$i^j \boldsymbol{\kappa}_{nj}(\boldsymbol{\theta}) = \frac{\partial^j \log \phi_n(\mathbf{0},\boldsymbol{\theta})}{\partial \mathbf{z}^j}. \tag{2}$$

In what follows, let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ be points in an open subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$, and let r be a specified integer. We use the word limit in the sense of joint limit, and introduce three assumptions.

Assumption 1. If n is large enough $|\phi_n(\mathbf{z},\boldsymbol{\theta})|$ is integrable over $\mathbb{R}^m$, and if δ_1 is an arbitrary positive constant the limit of

$$n^{\frac{r}{2}-1} \int_{B_{\delta_1\sqrt{n}}} \left|\phi_n\left(\mathbf{z}/\sqrt{n},\boldsymbol{\theta}\right)\right| d\mathbf{z},$$

as $n \to \infty$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$ is zero, where $B_{\delta_1\sqrt{n}}$ is the region $\|\mathbf{z}\| \geq \delta_1\sqrt{n}$ and $\|\cdot\|$denotes the Euclidean norm.

Assumption 2. The rth derivative $\partial^r \log \phi_n(\mathbf{z},\boldsymbol{\theta})/\partial \mathbf{z}^r$ exists for $\mathbf{z}$ in a neighborhood of the origin and the limit of

$$n^{-1}\frac{\partial^r \log \phi_n(\mathbf{z},\boldsymbol{\theta})}{\partial \mathbf{z}^r}$$

as $n \to \infty$, $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$ and $\|\mathbf{z}\| \to 0$ exists.

Assumption 3. The cumulant $\boldsymbol{\kappa}_{nj}(\boldsymbol{\theta}) = O(n)$ uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ for $j = 3, \ldots, r-1$.

Now we present the Edgeworth expansion and the corresponding approximation to the density $h_n(\mathbf{x},\boldsymbol{\theta})$ of $\mathbf{X}_n = n^{-1/2}E\{\mathbf{S}_n - E(\mathbf{S}_n)\}$. Suppose that there is an integer $r \geq 3$ such that Assumptions 1–3 hold. Then there is a neighborhood $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_2$ of $\boldsymbol{\theta}_0$ such that

$$h_n(\mathbf{x},\boldsymbol{\theta}) - \widehat{h}_n(\mathbf{x},\boldsymbol{\theta}) = o\left\{n^{-(r/2)+1}\right\} \tag{3}$$

uniformly in $\mathbf{x}$ and in $\boldsymbol{\theta}$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_2$, where

$$\widehat{h}_n(\mathbf{x},\boldsymbol{\theta}) = \frac{|\mathbf{D}_n(\boldsymbol{\theta})|^{-1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{D}_n^{-1}(\boldsymbol{\theta})\mathbf{x}\right\} \left\{1 + \sum_{j=3}^{r} n^{-(j/2)+1} P_{nj}(\mathbf{x},\boldsymbol{\theta})\right\}, \tag{4}$$

and where $P_{nj}(\mathbf{x},\boldsymbol{\theta})$ is a generalized Edgeworth polynomial of order j the definition of which is given in Durbin (1980a). The practical construction of $P_{nj}(\mathbf{x},\boldsymbol{\theta})$ is described by Chambers (1967, pp. 368–369).

Approximations to the Densities of Sufficient Estimators

Suppose that $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ is a matrix of observations of n continuos or discrete random $\ell \times 1$ vectors, not necessarily independent or identically distributed, with density

$$f(\mathbf{y},\boldsymbol{\theta}) = G(\mathbf{t},\boldsymbol{\theta})H(\mathbf{y}), \qquad \mathbf{y} \in \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{5}$$

where $\mathbf{t} = (t_1, \ldots, t_m)'$ is the value computed from $\mathbf{y}$ of an estimator $\mathbf{T}_n$ of the m-dimensional parameter $\boldsymbol{\theta}$, where $\mathcal{Y}$ and $\boldsymbol{\Theta}$ are observation and parameter spaces and where $\mathcal{Y}$ and H do not depend upon $\boldsymbol{\theta}$. We assume that $f(\mathbf{y},\boldsymbol{\theta}) > 0$ for all $\mathbf{y} \in \mathcal{Y}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By the factorization theorem $\mathbf{T}_n$ is sufficient for $\boldsymbol{\theta}$.

Suppose that a transformation $\mathbf{y}_1, \ldots, \mathbf{y}_n \to t_1, \ldots, t_m, u_{m+1}, \ldots, u_{n\ell}$ exists such that on substituting for $\mathbf{y}$ on the right-hand side of (5) and integrating or summing out $u_{m+1}, \ldots, u_{n\ell}$ we obtain the marginal density $g(\mathbf{t},\boldsymbol{\theta})$ of $\mathbf{T}_n$ in the form $g(\mathbf{t},\boldsymbol{\theta}) = G(\mathbf{t},\boldsymbol{\theta})H_1(\mathbf{t})$ where H_1 does not depend upon $\boldsymbol{\theta}$. We therefore have

$$f(\mathbf{y},\boldsymbol{\theta}) = g(\mathbf{t},\boldsymbol{\theta})h(\mathbf{y}), \tag{6}$$

where $h(\mathbf{y}) = H(\mathbf{y})/H_1(\mathbf{t})$. The derivation of (6) from (5) has been given in this form to avoid measure-theoretic complications.

Suppose further that although functions $G(\mathbf{t},\boldsymbol{\theta})$ satisfying (5) can be deduced immediately from inspection of $f(\mathbf{y},\boldsymbol{\theta})$, the density $g(\mathbf{t},\boldsymbol{\theta})$ is unknown and we want to obtain an approximation to it for a particular value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. Since (6) holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have

$$f(\mathbf{y},\boldsymbol{\theta}_0) = g(\mathbf{t},\boldsymbol{\theta}_0)h(\mathbf{y}). \tag{7}$$

On dividing (7) by (6) the unknown factor $h(\mathbf{y})$ is eliminated and we obtain immediately

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\boldsymbol{\theta})}g(\mathbf{t},\boldsymbol{\theta}). \tag{8}$$

If we substitute $\mathbf{t}$ for $\boldsymbol{\theta}$ in (8), as is legitimate since we have assumed that $\mathbf{t} \in \boldsymbol{\Theta}$, we obtain

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}g(\mathbf{t},\mathbf{t}). \tag{9}$$

The basic idea is to obtain an approximation $\widehat{g}(\mathbf{t},\boldsymbol{\theta}_0)$ for $g(\mathbf{t},\boldsymbol{\theta}_0)$ by substituting a series approximation $\widehat{g}(\mathbf{t},\mathbf{t})$ for $g(\mathbf{t},\mathbf{t})$ in (9), giving

$$\widehat{g}(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\widehat{g}(\mathbf{t},\mathbf{t}). \tag{10}$$

In effect, the method rescales the approximation $\widehat{g}(\mathbf{t},\mathbf{t})$ at $\boldsymbol{\theta} = \mathbf{t}$ by the likelihood ratio $f(\mathbf{y},\boldsymbol{\theta}_0)/f(\mathbf{y},\mathbf{t})$.

A second idea is to substitute an Edgeworth series approximation $\widehat{g}(\mathbf{t},\widetilde{\boldsymbol{\theta}})$ for $g(\mathbf{t},\boldsymbol{\theta})$ in (8), where $\widetilde{\boldsymbol{\theta}}$ is chosen as the value of $\boldsymbol{\theta}$ for which the mean of the distribution of $\mathbf{T}_n$ coincides with $\mathbf{t}$. The reason for using this indirect approach instead of approximating $g(\mathbf{t},\boldsymbol{\theta})$ directly is that a straightforward Edgeworth approximation of $g(\mathbf{t},\boldsymbol{\theta})$, would normally be in powers of $n^{-1/2}$ whereas an Edgeworth approximation of $g(\mathbf{t},\mathbf{t})$ or $\widehat{g}(\mathbf{t},\widetilde{\boldsymbol{\theta}})$ is normally a series in powers of n^{-1}.

Suppose that $E(\mathbf{T}_n) = \boldsymbol{\theta} - \boldsymbol{\beta}_n(\boldsymbol{\theta})$, where $\boldsymbol{\beta}_n(\boldsymbol{\theta}) = O(n^{-1})$ uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, and that $n\mathbf{T}_n = \mathbf{S}_n$, where $\mathbf{S}_n$ satisfies the Assumptions 1–3 given above with $r = 4$. Maximum likelihood estimators often satisfy these assumptions. We make the following further assumption:

Assumption 4. Uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$,

$$|\mathbf{D}_n(\boldsymbol{\theta})| = |\mathbf{D}(\boldsymbol{\theta})|\left\{1 + O(n^{-1})\right\}.$$

The assumption is, of course, satisfied when $\mathbf{S}_n$ is a sum of independent and identically distributed vectors but it is also satisfied in other cases of interest, notably in some applications in time series analysis. We suppose that we require a single-term approximation which has an error of order n^{-1} at most.

Since $\mathbf{X}_n = n^{-1/2}E\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\} = \sqrt{n}\left\{\mathbf{T}_n - \boldsymbol{\theta} + \boldsymbol{\beta}_n(\boldsymbol{\theta})\right\}$, the value of $\mathbf{X}_n$ when $\mathbf{T}_n = \mathbf{t}$ and $\boldsymbol{\theta} = \mathbf{t}$ is $\mathbf{x} = \boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n}$. With $r = 4$, () gives

$$\widehat{h}_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}_n(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}n\boldsymbol{\beta}_n(\mathbf{t})'\mathbf{D}_n^{-1}(\mathbf{t})\boldsymbol{\beta}_n(\mathbf{t})\right\} \times\left[1 + \sum_{j=3}^{4} n^{-(j/2)+1}P_{nj}\left\{\boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n},\mathbf{t}\right\}\right]. \tag{11}$$

Now $n\boldsymbol{\beta}_n(\mathbf{t})'\mathbf{D}_n^{-1}(\mathbf{t})\boldsymbol{\beta}_n(\mathbf{t}) = O(n^{-1})$ and the constant term of P_{n4} is $O(1)$. Moreover P_{n3} contains no constant term and hence is $O(n^{-1/2})$ when $\mathbf{x} = \boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n}$. We note that these orders of magnitude are uniform for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\theta}_0$ under the Assumptions 1–3. Because of Assumption 4, we have

$$\widehat{h}_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\}$$

uniformly for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\theta}_0$.

Let $h_n(\mathbf{x},\mathbf{t})$ be the true density of $\mathbf{X}_n$, then by (3)

$$h_n(\mathbf{x},\mathbf{t}) = \widehat{h}_n(\mathbf{x},\mathbf{t}) + o(n^{-1}) = \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\} + o(n^{-1}). \tag{12}$$

Since the term $o(n^{-1})$ is uniform for $\mathbf{t}$ in a neighborhood of $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$, where $\delta_2$is a suitably chosen positive constant independent of n, and since $|\mathbf{D}(\mathbf{t})|$ is continuous at $\boldsymbol{\theta}_0$ and hence is bounded away from zero for $\mathbf{t}$ in the neighborhood, the term $o(n^{-1})$ of (12) can be absorbed inside the curly bracket. We thus have uniformly

$$h_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\}.$$

Transforming from $\mathbf{x}$ to $\mathbf{t}$ we obtain for the density of $\mathbf{T}_n$ at $\mathbf{T}_n = \boldsymbol{\theta} = \mathbf{t}$,

$$g(\mathbf{t},\mathbf{t}) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathbf{D}(\mathbf{t})|^{-1/2}\left\{1 + O(n^{-1})\right\}. \tag{13}$$

Substituting in (9) we obtain

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathbf{D}(\mathbf{t})|^{-1/2}\frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\left\{1 + O(n^{-1})\right\}, \tag{14}$$

uniformly in $\mathbf{t}$ for $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$.

Expression (14) is the basic approximation for the density of the sufficient estimator $\mathbf{T}_n$. The fact that the error is a proportional error which is uniform over the region $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$ is important since the limiting probability that $\mathbf{T}_n$ falls outside this region is zero.

Assuming appropriate regularity conditions to be satisfied, $\mathbf{D}^{-1}(\boldsymbol{\theta})$ is the limiting mean information matrix $\mathcal{I}(\boldsymbol{\theta})$, where

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{n\to\infty} E\left[-n^{-1}\frac{\partial^2 \log f(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right].$$

We then have for the basic approximation

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathcal{I}(\mathbf{t})|^{1/2}\frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\left\{1 + O(n^{-1})\right\}, \tag{15}$$

uniformly in $\mathbf{t}$ for $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$.

The simplicity of the structure of this approximation should be noted. It consists of the normal approximation to the density when $\boldsymbol{\theta} = \mathbf{t}$, namely $\{n/(2\pi)\}^{m/2}|\mathcal{I}(\mathbf{t})|^{1/2}$, multiplied by the likelihood ratio $f(\mathbf{y},\boldsymbol{\theta}_0)/f(\mathbf{y},\mathbf{t})$.

Durbin (1980a) proved that when either (14) or (15) is integrated over any subset of $\mathbb{R}^m$, the error term remains $O(n^{-1})$. This result is, in fact, of great importance in practical situations since it demonstrates that the basic approximation can be integrated for inference purposes with an error which is of order n^{-1} at most. He proved as well that when the constant term of the approximation (14), and consequently also of (15), is adjusted to make the integral over the whole space equal to unity, the order of magnitude of the error is often reduced from $O(n^{-1})$ to $O_x(n^{-3/2})$, where $O_x(n^{-q})$ denotes a quantity which is $O(n^{-q})$ for each fixed $\mathbf{x} = \sqrt{n}\{\mathbf{t} - E(\mathbf{T}_n)\}$ but which is not $O(n^{-q})$ uniformly for all $\mathbf{x}$. This process of adjusting the constant term is generally called renormalization.

About the Author

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

Cross References

►Approximations to Distributions
►Edgeworth Expansion
►Properties of Estimators
►Sufficient Statistics

References and Further Reading

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

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

Bhattacharya RN, Ghosh JK (1978) On the validity of the formal Edgeworth expansion. Ann Statist 6:434–451

Bhattacharya RN, Rao RR (1976) Normal approximation and asymptotic expansions. Wiley, New York

Chambers JM (1967) On Methods of asymptotic approximation for multivariate distributions. Biometrika 54:367–383

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

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

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

Durbin J (1980b) The approximate distribution of partial serial correlation coefficient calculated from residual from regression on Fourier series. Biometrika 67:335–349

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

Hampel FR (1973) Some small sample asymptotics. Proc Prague Symp Asymptotic Stat 2:109–126

Loève M (1977) Probability theory, vol I, 4th edn. Springer, Berlin

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

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

Approximations to Distributions

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

Introduction

The exact probability distribution of estimators for finite samples is only available in convenient form for simple functions of the data and when the likelihood function is completely specified. Frequently, these conditions are not satisfied and the inference is based on approximations to the sample distribution. Typically, large sample methods based on the central limit theorem (see ►Central Limit Theorems) are generally used. For example, if T_n is an estimator of the parameter θ based on a sample of size n, it is sometimes possible to obtain functions $\sigma(\theta)$ such that the distribution of the random variable $\sqrt{n}(T_n - \theta)/\sigma(\theta)$ converges to the standard normal distribution as n tends to infinity. In such a case, it is very common to approximate the distribution of T_n by a normal distribution with mean θ and variance $\sigma^2(\theta)/n$.

These asymptotic approximations can be good even for very small samples. The mean of independent draws from a rectangular distribution has a bell-shaped density for n as small as three. But it is easy to construct examples where the asymptotic approximation is bad even when the sample has hundreds of observations. It is therefore desirable to know the conditions under which the asymptotic approximations are reasonable and to have alternative methods available when these approximations do not work properly. Most of the material discussed here is closely related with the topic *Asymptotic, higher order* which is presented as well in this Encyclopedia.

There is a good literature treating the theory and practice of approximations to distributions, but introductory

texts are relatively few. A very brief summary can be seen in Bickel and Doksum (1977), while some discussion is given in Johnson and Kotz (1970). The extension to asymptotic expansions can be seen in the excellent paper by Wallace (1958), although it is outdated. For a good treatment of the subject, an incursion upon the advanced probability and numerical analysis textbooks is needed. For those with enough time and patience, Chaps. 15 and 16 of Feller (1971) are well worth reading.

The Central Limit Theorem

The center of a large part of the asymptotic theory is the central limit theorem, initially formulated for sums of independent random variables. Let $\{Y_n\}$ be a sequence of independent random variables. Denote by H_n the distribution function of the standardized sum

$$X_n = \frac{\sum_{j=1}^{n} \{Y_j - E(Y_j)\}}{\sqrt{\left\{\sum_{j=1}^{n} V(Y_j)\right\}}},$$

where $V(Y_j)$ is the variance of Y_j, and by $\mathcal{N}(.)$ the standard normal distribution function. The central limit theorem then states that $\lim H_n(x) = \mathcal{N}(x)$, as $n \to \infty$, for every fixed x, provided only that the means and variances are finite. If the $\{Y_j\}$ are not identically distributed, an additional condition guaranteeing that the distributions are not too unbalanced is necessary.

For time series problems, for example, where in general the variables are not independent, there have been particularized versions of this theorem guaranteeing the asymptotic behavior of statistics used in this area. Good references are the textbook by Anderson (1971), Brockwell and Davis (1991), Hannan (1970), and Priestley (1982) where one can find an excellent treatment of the asymptotic theory applied to time series problems.

Some authors have shown that the order of magnitude of the errors in the central limit theorem is $O(n^{-1/2})$.

While the central limit theorem is very useful theoretically and often in practice, it is not always satisfactory since for small or moderate n the errors of the normal approximation may be too large.

Curve Fitting

The most simplest form for obtaining an approximation to a distribution is to look for a family of curves with the correct shape and select the member that fits best. If the moments, specially those of low order, of the true distribution are known, they can be used in the fitting process. Otherwise one can use Monte Carlo simulations or any other information about the true distribution.

Durbin and Watson (1971) describe a number of different approximations to the null distribution of the statistic d used for testing serial correlation in regression analysis. One of the most accurate is the beta approximation proposed by Henshaw (1966). Since d is between zero and four and it seems to have a unimodal density, it is reasonable to think that a linear transformation from a beta distributed variable can be a good approximation to the true distribution. Suppose that Y is a random variable with beta distribution function

$$\Pr(Y \le y) = \frac{1}{B(p,q)} \int_0^y t^{p-1}(1-t)^{q-1} dt = G(y; p.q),$$

where

$$B(p,q) = \int_0^\infty t^{p-1}(1-t)^{q-1} dt.$$

Then, for a and b constant, the random variable $a + bY$ has moments depending on p, q, a and b. These moments are easy to express analytically. Moreover, the moments of the Durbin–Watson's statistic d are simple functions of the matrix of explanatory variables. Equating the first four moments of d with the corresponding moments of $a + bY$, one obtains four equations with four unknowns. For a given matrix of explanatory variables these equations give a unique solution, p^*, q^*, a^* and b^* say. So $\Pr(d \le y)$ can be approximated by $G\{(y - a^*)/b^*; p^*, q^*\}$. This approximation gives good results in many cases. Theil and Nagar (1961) proposed a similar approximation but using the approximated moments of d instead of the true moments. Since these approximated moments are independent of the matrix of explanatory variables, Theil–Nagar's approximation does not depend on the data and can be tabulated without any problem. Unfortunately the approximated moments are not always accurate and the resulting approximation to the distribution is less satisfactory than Henshaw's approximation.

If one has enough information over the true density, the curve fitting methods give simple and correct approximations. However these methods are not so attractive when the purpose is not quantitative but qualitative. The comparison of alternative procedures is difficult because the curve fitting methods does not produce, in general, parametric families of curves easily comparable. If two statistics are approximately normal, they can be compared by their means and variances. If one statistic is approximately beta and another is approximately normal, the comparison between them is not easy since the usual parameters that describe one of the distributions are

not of much interest for obtaining information about the other. The flexibility that makes the curve fitting method so accurate is, as well, an inconvenience for using it in comparisons.

Transformations

Suppose that Y is a random variable and b a monotonically increasing function such that $b(Y)$ has a distribution function H which can be approximated by $\widehat{H}$. Since $\Pr(Y \le y)$ is equal to $\Pr\{b(Y) \le b(y)\}$, the distribution function of Y can be approximated by $\widehat{H}\{b(y)\}$. A well known example of this technique is Fisher's z transformation. The sample correlation coefficient $\widehat{\rho}$ based on a random sample from a bivariate normal population is very far from symmetry when the true coefficient ρ is large in absolute value. But, $z = b(\widehat{\rho}) = 2^{-1}\log\{(1+\widehat{\rho})/(1-\widehat{\rho})\}$ is almost symmetric and can be approximated by a normally distributed random variable with mean $2^{-1}\log\{(1+\rho)/(1-\rho)\}$ and variance n^{-1}. Therefore $\Pr(\widehat{\rho} \le y)$ can be approximated by $\mathcal{N}\left\{\sqrt{n}b(y) - \sqrt{n}b(\rho)\right\}$ for moderate sample size n.

The use of transformations for approximating distributions is an art. Sometimes, as in the case of the correlation coefficient, the geometry of the problem can suggest the appropriate transformation b. Since $\widehat{\rho}$ can be interpreted as the cosine of the angle between two normally distributed random vectors, an inverse trigonometric transformation can be useful. In other cases, arguments based on approximations to the moments are helpful. Suppose that $b(Y)$ can be expanded as a power series about $\mu = E(Y)$

$$b(Y) = b(\mu) + b'(\mu)(Y-\mu) + \frac{1}{2}b''(\mu)(Y-\mu)^2 + \cdots,$$

where $Y - \mu$ is in some sense small. so we can do

$$\begin{aligned} E(b) &\approx b(\mu) + \frac{1}{2}b''(\mu)E(Y-\mu)^2, \\ V(b) &\approx \left\{b'(\mu)\right\}^2 V(Y), \\ E\left\{b - E(b)\right\}^3 &\approx \left\{b'(\mu)\right\}^3 E(Y-\mu)^3 \\ &\quad + \frac{3}{2}\left\{b'(\mu)\right\}^2 b''(\mu)E(Y-\mu)^4, \end{aligned}$$

and choose b in such a way that these approximates moments are equal to the moments of the approximated distribution. If the approximated distribution is normal, we can require that the variance $V(b)$ be a constant independent of μ; or we can require that the third order moment be zero. If the moments of Y are (almost) known and the above approximation is used, the criterion leads to differential equations in $b(\mu)$. Note that Fisher's transformation of $\widehat{\rho}$ stabilizes the approximated variance of b making it independent of ρ.

Jenkins (1954) and Quenouille (1948) apply inverse trigonometric transformations to the case of the autocorrelation coefficient in time series. The use of transformations in econometrics seems, however, to be minimum due mainly to the fact that the method is closely related with univariate distributions.

Asymptotic Expansions

Frequently it is possible to decompose the problem of finding the distribution in a sequence of similar problems. If the sequence has a limit which can easily be found, one can obtain an approximation to the solution of the original problem by a solution of the limit problem. The sequence of the problem is indexed by a parameter, which usually is the sample size n. Suppose for instance that we want an approximation to the distribution of an estimator, computed from a sample, of a parameter θ. We define an infinite sequence $\widehat{\theta}_n$ of estimators, one for each sample size $n = 1, 2, \ldots$, and we consider the problem of obtaining the distribution of each $\widehat{\theta}_n$. Of course, it is necessary to have some description of the joint distribution of the observations for each n. Given such a sequence of problems, the asymptotic approach implies three steps:

(a) To look for a simple monotonic transformation $X_n = b(\widehat{\theta}_n; \theta, n)$ such that the estimator X_n is not very sensitive to n. Since the majority of estimators are centerd upon the true value of the parameter and they have a dispersion which decreases at the same rate as $n^{-1/2}$, the transformation $X_n = \sqrt{n}(\widehat{\theta}_n - \theta)$ is frequently used.

(b) To look for an approximation $\widehat{H}_n(x)$ to the distribution function $H_n(x) = \Pr(X_n \le x)$ such that, when n tends to infinity, the error

$$\left|\widehat{H}_n(x) - H_n(x)\right|$$

tends to zero.

(c) The distribution function of $\widehat{\theta}_n$ is approximated by $\widehat{H}_n$, i.e., $\Pr(\widehat{\theta}_n \le a) = \Pr\{X_n \le b_n(a;\theta,n)\}$ is approximated by $\widehat{H}_n\{b_n(a;\theta,n)\}$.

Let $\widehat{H}_n(x)$ be an approximation to the distribution function $H_n(x)$. If, for every x,

$$\lim_{n\to\infty} n^{(r/2)-1}\left|\widehat{H}_n(x) - H_n(x)\right| = 0, \quad r = 2, 3, \ldots,$$

we write

$$H_n(x) = \widehat{H}_n(x) + o\left\{n^{(r/2)-1}\right\}, \quad r = 2, 3, \ldots,$$

and we say that $\widehat{H}_n(x)$ is an approximation $o\left\{n^{(r/2)-1}\right\}$ or an approximation of order $r-1$. These names are used as well when approximating density functions. The asymptotic distribution is an approximation $o(n^0) = o(1)$ or a first order approximation. These concepts are related with the topic *Asymptotic, higher order* which is presented as well in this Encyclopedia.

The number n measures the velocity at which the error of approximation tends to zero as n tends to infinity. If we choose the transformation b such that H_n and $\widehat{H}_n$ vary gently with n, the value of r can give an indication of the error of approximation for moderate values of n.

There are two well known methods for obtaining high order approximations to distributions, both based on the Fourier inversion of the characteristic function. Let $\phi_n(z,\theta) = E\{\exp(izX_n)\}$ be the characteristic function of X_n and let $\psi_n(z,\theta) = \log \phi_n(z,\theta)$ be the cumulant generating function. If ϕ_n is integrable, the density function h_n of X_n can be written as

$$h_n(x;\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixz}\phi_n(z,\theta)dz$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{-ixz + \psi_n(z,\theta)\}\, dz. \qquad (1)$$

Frequently it is possible to expand $\psi_n(z,\theta)$ in power series where the successive terms are increasing powers of $n^{-1/2}$. In this case the integrand can be approximated by the first few terms of this series expansion. Integrating term by term, one obtains a series approximation to h_n; afterward integration will give an approximation to the distribution function. The approximation known as *Edgeworth approximation* or ►*Edgeworth expansion* consists in expanding $\psi_n(z,\theta)$ at $z = 0$. This method is the most frequently used in practice because of its relative simplicity. It does not require a complete knowledge of $\psi_n(z,\theta)$. It is enough if one knows the first low order cumulants of X_n. More details about this method is given in this Encyclopedia under the name *Edgeworth expansion*. The approximation known as *saddlepoint approximation* is obtained by expanding $\psi_n(z,\theta)$ at the "saddlepoint" value z^* where the integrand of (1) is maximized. This method, introduced by Daniels (1954), is more complex and requires a deeper knowledge of the function $\psi_n(z,\theta)$.When this knowledge is available, the method gives accurate approximations specially in the "tail" region of the distribution. Daniels (1956) and Phillips (1978) applied this method to some autocorrelation statistics in time series analysis. More details about this method is given in this Encyclopedia under the name *Saddlepoint approximations*.

Wallace (1958) gives an excellent introduction to the approximations based on expansions of the characteristic function. An exposition with emphasis on multivariate expansions can be found in Barndorff-Nielsen and Cox (1979). Durbin (1980) proposed a simple method for obtaining a second order approximation to the density of a large class of statistics. This method is discuss in this Encyclopedia under the name *Approximations for densities of sufficient estimators*.

Attitudes and Perspectives

The theory of approximate distributions, like the theory of exact distributions, depends on the assumptions made about the stochastic process which generates the data. The quality of the approximations will not be better than the quality of the specifications sustaining them. one certainly will not rely upon a theory of distribution unless the conclusions are so robust that they do not vary significantly in front of moderate changes of basic assumptions. Since the majority of the methods of approximation use information about the first four moments at least, while the usual asymptotic theory only need information about the first two moments, some loss of robustness has to be expected. However, if some idea about the degree of skewness and kurtosis is available, this information can be helpful to obtain better approximations to the distribution of statistics.

Recently there has been an increasing interest in asymptotic theory. Great efforts have been made in order to demonstrate that some statistics are asymptotically normal and efficient. Of course, the asymptotic theory is important to have an idea of the sample properties of a given statistical procedure. Unfortunately there has been some confusion with the use of the terms "asymptotic" and "approximated."The fact that a standardized estimator has an asymptotic normal distribution is purely a mathematical proposition about the limit of the probabilities measures under a set of previously specified assumptions. The fact that a given estimator is approximately normal suggests that, for this particular problem, one believes in the possibility of treating the estimator as if it was normal.

Sometimes, under certain circumstances, asymptotic arguments lead to good approximations, bet frequently they do not. A careful analyst, with some knowledge of statistical theory, a modest computer and a great amount of common sense can find reasonable approximations for a given inferential problem.

About the Author

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

Cross References

►Approximations for Densities of Sufficient Estimators
►Asymptotic Normality
►Asymptotic, Higher Order
►Central Limit Theorems
►Cornish-Fisher Expansions
►Edgeworth Expansion
►Limit Theorems of Probability Theory
►Saddlepoint Approximations
►Strong Approximations in Probability and Statistics

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD thesis, The London School of Economics and Political Science, University of London, England
Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312
Bickel PJ, Doksum KA (1977) Mathematical statistics. Holden-Day, San Francisco
Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer, New York
Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650
Daniels HE (1956) The approximate distribution of serial correlation coefficients, Biometrika 43:169–185
Durbin J, (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333
Durbin J, Watson GS (1971) Testing for serial correlation in least squares regression, III. Biometrika 58:1–19
Feller W (1971) An Introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Hannan EJ (1970) Multiple time series. Wiley, New York
Henshaw RC (1966) Testing single-equation least-squares regression models for autocorrelated disturbances. Econometrica 34:646–660
Jenkins GM (1954) An angular transformation for the serial correlation coefficient. Biometrika 41:261–265
Johnson NI, Kotz S (1970) Continuous univariate distributions, vol 1. Wiley, New York
Phillips PCB (1978) Edgeworth and saddlepoint approximations in a first order autoregression. Biometrika 65:91–98
Priestley MB (1982) Spectral analysis and time series. Academic, London
Quenouille MH (1948) Some results in the testing of serial correlation coefficients. Biometrika 35:261–284
Theil H, Nagar AL (1961) Testing the independence of regression disturbances. J Am Stat Assoc 56:793–806
Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Association Measures for Nominal Categorical Variables

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

As a means of summarizing the potential relationship between two (or more) random categorical variables X and Y, a number of measures of association have been proposed over the years. A historical review of such measures and new proposals have been presented in a series of papers by Goodman and Kruskal (1979) [see also Kendall and Stuart (1979), Ch. 33 and Liebetrau (1983)]. Such summary measures depend on whether X and Y are nominal or ordinal as well as on whether X and Y are to be treated symmetrically or asymmetrically. In the symmetric case, X and Y are treated equivalently and no causal relationship is assumed to exist between them. In the asymmetric case, a causal relationship between X and Y is considered to exist so that one variable is treated as the explanatory variable (X) and the other variable treated as the response variable (Y).

The focus here will be on the case when both X and Y are nominal categorical variables, i.e., no natural ordering exists for the variables. Association measures for both the symmetric and asymmetric case will be considered.

Symmetric Measures

For the variable X with I categories and the variable Y with J categories, their joint and marginal probabilities are defined as $\Pr(X = i, Y = j) = p_{ij}$, $\Pr(X = i) = p_{i+}$, and $\Pr(Y = j) = p_{+j}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ where $\sum_1^2 i = 1I \sum_1^2 j = 1J p_{ij} = \sum_1^2 i = 1I p_{i+} = \sum_1^2 j = 1J p_{+j} = 1$. In terms of a two-way contingency table with I rows and J columns, the cell entries come from the joint distribution $\{p_{ij}\}$, with p_{ij} being the entry in cell (i, j), and $\{p_{i+}\}$ and $\{p_{+j}\}$ are the marginal distributions (totals) for the rows and columns, respectively. The conditional distribution of Y given X is defined in terms of $p_{j|i} = p_{ij}/p_{i+}$ for all i and j.

Several early suggested association measures were based on the (Pearson) coefficient of mean square contingency defined by

$$\Phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i+}p_{+j}} - 1. \quad (1)$$

If the p_{ij} represent sample estimates (of population probabilities π_{ij}) $p_{ij} = n_{ij}/N$ based on the multinominal frequencies n_{ij} for all i, j and sample size $N = \sum_{1}^{2} i = 1 J \sum_{1}^{2} j = 1 J n_{ij}$, then it is recognized that $\Phi^2 = X^2/N$ where X^2 is the familiar Pearson chi-square goodness-of-fit statistic for testing the null hypothesis or independence between X and Y, i.e.,

$$X^2 = N\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i+}n_{+j}} - 1\right). \quad (2)$$

The most popular such association measure based on X^2 appears to be Cramér's (1946) V defined as

$$V = \sqrt{\frac{X^2}{N(M-1)}}, \quad M = \min\{I, J\}. \quad (3)$$

This V ranges in value between 0 and 1, inclusive, for any given I and J, with $V = 0$ if, and only if, X and Y are independent and $V = 1$ when there is no more than one non-zero entry in either each row or in each column. The V is invariant with any permutations of the rows or the columns. The estimated standard error of V is given in Bishop et al. (1975, p. 386), but its expression is rather messy.

Kendall and Stuart (1979, p. 606), have shown that V^2 is the mean squared canonical correlation. However, it has been argued that values of V are difficult to interpret since V has no obvious probabilistic meaning or interpretation. Nevertheless, V does reflect the divergence (or "distance") of the distribution $\{p_{ij}\}$ from the independence distribution $\{p_{i+}p_{+j}\}$ relative to the maximum divergence.

There is some uncertainty in the literature as to whether V or V^2 is the proper measure to use. This issue will be addressed in a section below. It may also be pointed out that a similar association measure can be formulated in terms of the likelihood-ratio statistic G^2, which has the same asymptotic chi-square distribution as χ^2 under the null hypothesis and is often used instead of χ^2. For the G^2 under independence, i.e., for

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{Nn_{ij}}{n_{i+}n_{+j}}\right) \quad (4)$$

and since $n_{ij} \le n_{i+}$ and $n_{ij} \le n_{+j}$ for all i and j, it follows that

$$\begin{aligned} G^2 \le G_X^2 &= 2\sum_{i=1}^{I} n_{i+}\log\left(\frac{N}{n_{i+}}\right) \text{ and } G^2 \le G_Y^2 \\ &= 2\sum_{j=1}^{J} n_{+j}\log\left(\frac{N}{n_{+j}}\right). \end{aligned} \quad (5)$$

Thus, analogously to V in (3), one could define the association measure

$$W = \sqrt{\frac{G^2}{\min\{G_X^2, G_Y^2\}}} \quad (6)$$

with G_X^2 and G_Y^2 as given in (5).

This new measure W can also be interpreted as the divergence ("distance") of the distribution $\{p_{ij}\}$ from the independence distribution $\{p_{i+}p_{+j}\}$ relative to its maximum [see also Kvålseth (1987)]. The W has the same type of properties as Cramér's V in (3) and can be expected to take on values quite similar to those of V. For instance, for the data

$$n_{11} = 20 \quad n_{12} = 15 \quad n_{13} = 25$$

$$n_{21} = 5 \quad n_{22} = 25 \quad n_{23} = 10$$

it is found from (3) and (6) that $V = .38$ and $W = .33$.

Asymmetric Measures

Goodman and Kruskal 1979 have discussed two different asymmetric association measures $(\lambda_{Y|X})$ and $(\tau_{Y|X})$ for the case when X can be considered to be the explanatory variable and Y the response variable. Such measures are frequently referred to as proportional reduction in error (PRE) measures since they can be interpreted in terms of the relative difference between two error probabilities P_Y and $P_{Y|X}$, i.e.,

$$PRE_{Y|X} = \frac{P_Y - P_{Y|X}}{P_Y} \quad (7)$$

where P_Y is the probability of error when predicting the Y – category of a randomly selected observation or item without knowing its X – category and $P_{Y|X}$ is the corresponding expected (weighted mean) error probability given its X – category.

The optimal prediction strategy would clearly be to predict that a randomly selected observation (item) would belong to a maximum-probability (modal) category, so that with

$$p_{+m} = \max\{p_{+1}, \ldots, p_{+J}\}; \text{ and } p_{im} = \max\{p_{i1}, \ldots, p_{iJ}\}, \quad i = 1, \ldots, I$$

the error probabilities P_Y and $P_{Y|X}$ become

$$P_Y = 1-p_{+m},\ P_{Y|X} = \sum_{i=1}^{I} p_{i+}\,(1-p_{im}/p_{i+}) = 1-\sum_{i=1}^{I} p_{im}. \quad (8)$$

From (7)–(8), the so-called Goodman–Kruskal *lambda* becomes

$$\lambda_{Y|X} = \frac{\sum_{1}^{2} i = 1 I p_{im} - p_{+m}}{1-p_{+m}} \quad (9)$$

which is the relative decrease in the error probability when predicting the Y-category as between not knowing and knowing the X-category.

Another asymmetric measure is based on a different prediction rule: Predictions are made according to the given probabilities. Thus, a randomly chosen observation (item) is predicted to fall in the jth category of Y with probability p_{+j} $(j = 1,\ldots,J)$ if its X-category is unknown. If, however, the observation is known to belong to the ith category of X, it is predicted to belong to the jth category of Y with the (conditional) probability p_{ij}/p_{i+} $(j = 1,\ldots,J)$. The error probabilities are then given by

$$P_Y = 1-\sum_{j=1}^{J} p_{+j}^2, \quad P_{Y|X} = \sum_{i=1}^{I} p_{i+}\left[1-\sum_{j=1}^{J}(p_{ij}/p_{i+})^2\right] \quad (10)$$

so that, from (7) and (10), the following so-called Goodman–Kruskal *tau* results:

$$\tau_{Y|X} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}^2/p_{i+} - \sum_{j=1}^{J} p_{+j}^2}{1-\sum_{j=1}^{J} p_{+j}^2} \quad (11)$$

which gives the relative reduction in the error probability when predicting an observation's Y-category as between its X-category not given and given.

Both measures in (9) and (11), and whose estimated standard errors are given elsewhere [e.g., Bishop et al. (1975, pp. 388–391), Goodman and Kruskal (1979), and Liebetrau (1983)], can assume values between 0 and 1, inclusive. Both equal 1 if, and only if, each row of the contingency table contains no more than one non-zero cell entry. Both are invariant under permutations of rows or of columns. However, their zero-value conditions differ. The $\tau_{Y|X} = 0$ if, and only if, X and Y are independent, whereas $\lambda_{Y|X} = 0$ if (1) X and Y are independent or (2) the modal probabilities p_{im} in all rows fall in the same column. This second condition is most likely to occur when the marginal distribution $\{p_{i+}\}$ is highly uneven (non-uniform). Thus, in cases of highly uneven $\{p_{i+}\}$, $\lambda_{Y|X}$ may be 0 or very small, while other measures such as $\tau_{Y|X}$ may be substantially larger. The high sensitivity of $\lambda_{Y|X}$ to $\{p_{i+}\}$ is one limitation of this measure that may lead to misleadingly low association values.

Symmetric version of *lambda* and *tau* can also be formulated in terms of weighted averages (Goodman and Kruskal 1979). Thus, in terms of the general expression in (7), a symmetric *PRE* could be formulated as the following weighted mean of $PRE_{Y|X}$ and $PRE_{X|Y}$:

$$PRE = \frac{P_Y - P_{Y|X} + P_X - P_{X|Y}}{P_Y + P_X}.$$

However, there would seem to be no strong reason for preferring such symmetricized measures over the V or W in (3) and (6).

It should be pointed out that asymmetric association measures can also be formulated in terms of relative reduction in variation, somewhat analogously to the coefficient of determination (R^2) used in regression analysis. This can be done by basically replacing the prediction error probabilities in (7) with appropriate measures of categorical variation (Agresti 2002, pp. 56–69).

Concluding Comment

For Cramér's V in (3), there is inconsistency in the literature concerning the use of V versus V^2 (and Cramér himself proposed V^2 (Cramér 1946, p. 443)). Also, concern has been expressed that different measures such as those in (9) and (11) can produce quite different results. Such issues are indeed important and are often overlooked.

As with any summary measure, and so it is with association measures, it is essential that a measure takes on values throughout its range that are reasonable in that they provide true or valid representations of the attribute being measured. In order to make such an assessment for the above association measures, consider the simple case of a 2×2 table with all the marginal probabilities equal to .5 and with the following cell entries:

$$p_{11} = (1-w)/4 p_{12} = (1+w)/4$$

$$p_{21} = (1+w)/4 p_{22} = (1-w)/4$$

with $0 \le w \le 1$. Each of these probabilities are seen to be the weighted mean of the corresponding probabilities for the case of perfect association and zero association (independence) for the given marginal probabilities, i.e.,

$$p_{11} = p_{22} = w(0)+(1-w)(.25), \quad p_{12} = p_{21} = w(.5)+(1-w)(.25)$$

In order for some association measure $A \in [0,1]$ to take on reasonable values in this ease, the only logical requirement is clearly that

$$A = w(A=1) + (1-w)(A=0) = w,\ w \in [0,1]$$

It is readily seen that the measures in (3) and (9) meet this requirement for all w, i.e., $V = w$ (and not V^2) and $\lambda_{Y|X} = w$ for the above $\{p_{ij}\}$ – distribution. However, it is seen that, for (11), $\lambda_{Y|X} = w^2$. This shows that $\tau'_{Y|X} = \sqrt{\tau_{Y|X}}$ should be used as an association measure rather than $\tau_{Y|X}$. In the case of W in (6), it is apparent that W is only approximately equal to w, but the approximation appears to be sufficiently close for W to be a competitive association measure.

About the Author

For biography *see* the entry ►Entropy.

Cross References

►Categorical Data Analysis
►Scales of Measurement
►Variables
►Variation for Categorical Variables

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken, NJ

Bishop YMM, Fienberg SE, Holland PW(1975) Discrete multivariate analysis. MIT, Cambridge, MA

Cramér H (1946) Mathematical methods for statistics. Princeton University Press, Princeton, NJ

Goodman LA, Kruskal WH (1979) Measures of association for cross-classifications. Springer, New York

Kendall M, Stuart A (1979) The advanced theory of statistics, vol.2, 4th edn. Charles Griffin, London

Kvålseth TO (1987) Entropy and correlation: some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-17, 517–519

Liebetrau AM (1983) Measures of Association. Beverly Hills, CA: Sage Publications.

Astrostatistics

Eric D. Feigelson
Professor, Associate Director
Penn State Center for Astrostatistics Pennsylvania State University, University Park, PA, USA

Introduction

The term "astronomy" is best understood as short-hand for "astronomy and astrophysics." Astronomy is the observational study of matter beyond Earth: planets and other bodies in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations of mass. The perspective is rooted in our viewpoint on or near Earth, typically using telescopes or robotic satellites. Astrophysics is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an inferential intellectual effort based on the well-confirmed assumption that physical processes established to rule terrestrial phenomena – gravity, thermodynamics, electromagnetism, quantum mechanics, plasma physics, chemistry, and so forth – also apply to distant cosmic phenomena.

Statistical techniques play an important role in analyzing astronomical data and at the interface between astronomy and astrophysics. Astronomy encounters a huge range of statistical problems: samples selected with truncation; variables subject to censoring and heteroscedastic measurement errors; parameter estimation of complex models derived from astrophysical theory; anisotropic spatial clustering of galaxies; time series of periodic, stochastic, and explosive phenomena; image processing of both gray-scale and Poissonian images; ►data mining of terabyte-petabyte datasets; and much more. Thus, astrostatistics is not focused on a narrow suite of methods, but rather brings the insights from many fields of statistics to bear on problems arising in astronomical research.

History

As the oldest observational science, astronomy was the driver for statistical innovations over many centuries (Stigler 1986; Hald 1998). Hipparchus, Ptolemy, al-Biruni, and Galileo Galilei were among those who discussed methods for averaging discrepant astronomical measurements. The least squares method (see ►Least Squares) and its understanding in the context of the normal error distribution were developed to address problems in Newtonian celestial mechanics during the early nineteenth century by Pierre-Simon Laplace, Adrian Legendre, and Carl Friedrich Gauss. The links between astronomy and statistics considerably weakened during the first decades of the twentieth century as statistics turned its attention to social and biological sciences while astronomy focused on astrophysics. Maximum likelihood methods emerged slowly starting in the 1970s, and Bayesian methods are now gaining considerably popularity.

Modern astrostatistics has grown rapidly since the 1990s. Several cross-disciplinary research groups emerged to develop advanced methods and critique common practices (http://hea-www.harvard.edu/AstroStat; http://www.incagroup.org; http://astrostatistics.psu.edu). Monographs were written on astrostatistics (Babu and Feigelson 1996), galaxy clustering (Martinez and Saar 2002), image processing (Starck and Murtagh 2006), Bayesian analysis

(Gregory 2005), and Bayesian cosmology (Hobson et al. 2010). The *Statistical Challenges in Modern Astronomy* (Babu and Feigelson 2007) conferences bring astronomers and statisticians together to discuss methodological issues.

The astronomical community is devoting considerable resources to the construction and promulgation of large archival datasets, often based on well-designed surveys of large areas of the sky. These surveys can generate petabytes of images, spectra and time series. Reduced data products include tabular data with approximately ten variables measured for billions of astronomical objects. Major projects include the Sloan Digital Sky Survey, International Virtual Observatory, and planned Large Synoptic Survey Telescope (http://www.sdss.org, http://www.ivoa.net, http://www.lsst.org). Too large for traditional treatments, these datasets are spawning increased interest in computationally efficient data visualization, data mining, and statistical analysis. A nascent field of astroinformatics allied to astrostatistics is emerging.

Topics in Contemporary Astrostatistics

Given the vast range of astrostatistics, only a small portion of relevant issues can be outlined here. We outline three topics of contemporary interest (The astronomical research literature can be accessed online through the SAO/NASA Astrophysics Data System, http://adsabs.harvard.edu.).

Heteroscedastic Measurement Errors

Astronomical measurements at telescopes are made with carefully designed and calibrated instruments, and "background" levels in dark areas of the sky are examined to quantitatively determine the noise levels. Thus, unlike in social and biological science studies, heteroscedastic measurement error are directly obtained for each astronomical measurement. This produces unusual data structures. For example, a multivariate table of brightness of quasars in six photometric bands will have 12 columns of numbers giving the measured brightness and the associated measurement error in each band.

Unfortunately, few statistical techniques are available for this class of non-identically distributed data. Most errors-in-variables methods are designed to treat situations where the heteroscedasticity is not measured, and instead becomes part of the statistical model (Carroll et al. 2006). Methods are needed for density estimation, regression, multivariate analysis and classification, spatial processes, and time series analysis. Common estimation procedures in the astronomical literature weight each measurement by its associated error. For instance, in a functional regression model, the parameters $\hat{\boldsymbol{\theta}}$ in model M are estimated by minimizing the weighted sum of squared residuals $\sum_i (O_i - M_i(\hat{\boldsymbol{\theta}})^2/\sigma_i^2$ of the observed data O_i where σ_i^2 are the known variances of the measurement errors.

More sophisticated methods are being developed, but have not yet entered into common usage. Kelly (2007) treats structural regression as an extension of a normal mixture model, constructing a likelihood which can either be maximized with the EM Algorithm or used in ►Bayes' theorem. The Bayesian approach is more powerful, as it also can simultaneous incorporate censoring and truncation into the measurement error model. Delaigle and Meister (2008) describe a nonparametric kernel density estimator that takes into account the heteroscedastic errors. More methods (e.g., for multivariate clustering and time series modeling) are needed.

Censoring and Truncation

In the telescopic measurement of quasar brightnesses outlined above, some targeted quasars may be too faint to be seen above the background noise level in some photometric bands. These nondetections lead to censored data points. The situation is similar in some ways to censoring treated by standard survival analysis, but differs in other ways: the data are left-censored rather than right-censored; censoring can occur in any variable, not just a single response variable; and censoring levels are linked to measurement error levels. Survival techniques have come into common usage in astronomy since their introduction (Isobe et al. 1986). They treat some problems such as density estimation (with the Kaplan-Meier product-limit estimator), two-sample tests (with the Gehan, logrank and Peto-Prentice tests), correlation (using a generalization of Kendall's τ), and linear regression (using the Buckley-James line).

Consider a survey of quasars at a telescope with limited sensitivity where the quasar sample is not provided in advance, but is derived from the photometric colors of objects in the survey. Now quasars which are too faint for detection are missing entirely from the dataset. Recovery from this form of truncation is more difficult than recovery from censoring with a previously established sample. A major advance was the derivation of the nonparametric estimator for a randomly truncated dataset, analogous to the Kaplan-Meier estimator for censored data, by astrophysicist Lynden-Bell (1971). This solution was later recovered by statistician Woodroofe (1985), and bivariate extensions were developed by Efron and Petrosian (1992).

Periodicity Detection in Difficult Data

Stars exhibit a variety of periodic behaviors: binary star or planetary orbits; stellar rotation; and stellar oscillations. While Fourier analysis is often used to find and characterize such periodicities, the data often present problems such as non-sinusoidal repeating patterns, observations of limited duration, and unevenly-spaced observations. Non-sinusoidal periodicities occur in elliptical orbits, eclipses, and rotational modulation of surface features. Unevenly-spaced data arise from bad weather at the telescope, diurnal cycles for ground-based telescopes, Earth orbit cycles for satellite observatories, and inadequate observing time provided by telescope allocation committees.

Astronomers have developed a number of statistics to locate periodicities under these conditions. The Lomb-Scargle periodogram (Scargle 1982) generalizes the Schuster periodogram to treat unevenly-spaced data. Stellingwerf (1978) presents a widely used least-squared technique where the data are folded modulo trial periods, grouped into phase bins, and intra-bin variance is compared to inter-bin variance using χ^2. The method treats unevenly spaced data, measurement errors, and non-sinusoidal shapes. Dworetsky (1983) gives a similar method without binning suitable for sparse datasets. Gregory and Loredo (1992) develop a Bayesian approach for locating non-sinusoidal periodic behaviors from Poisson distributed event data. Research is now concentrating on methods for computationally efficient discovery of planets orbiting stars as they eclipse a small fraction during repeated transits across the stellar surface. These methods involve matched filters, Bayesian estimation, least-squares box-fitting, maximum likelihood, ►analysis of variance, and other approaches (e.g., Pontopappas et al. 2005).

About the Author

Dr. Eric Feigelson was trained in astronomy at Haverford College and Harvard University during the 1970s, and entered the Astronomy and Astrophysics faculty at Pennsylvania State University in 1983 where he received an NSF Presidential Young Investigator Award and is now Professor. In addition to X-ray astronomical studies of star and planet formation, he has a long-standing collaboration with statisticians. Working with G. Jogesh Babu at Penn State's Center for Astrostatistics, he organizes summer schools, conferences and other resources for advancing statistical methodology in astronomy. He serves as an Associate Editor of the *Astrophysical Journal*, on the Virtual Astronomical Observatory Science Council, and other organizations relating statistics to astronomy.

Cross References

►Chaotic Modelling
►False Discovery Rate
►Heteroscedasticity
►Linear Regression Models
►Statistics, History of

References and Further Reading

Babu GJ, Feigelson ED (1996) Astrostatistics. Chapman and Hall, London

Babu GJ, Feigelson ED (2007) Statistical challenges in modern astronomy IV. Astronomical Society of the Pacific, San Fransisco, California

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement errors in nonlinear models. Chapman and Hall/CRC, Boca Raton, FL

Delaigle A, Meister A (2008) Density estimation with heteroscedastic error. Bernoulli 14:562–579

Dworetsky MM (1983) A period-finding method for sparse randomly spaced observations of 'How long is a piece of string?'. Mon Not Royal Astro Soc 203:917–924

Efron B, Petrosian V (1992) A simple test of independence for truncated data with applications to redshift surveys. Astrophys J 399:345–352

Gregory PC (2005) Bayesian logical data analysis for the physical sciences. Cambridge University Press, Cambridge, UK

Gregory PC, Loredo TJ (1992) A new method for the detection of a periodic signal of unknown shape and period. Astrophys J 398:146–168

Hald A (1998) A history of mathematical statistics from 1750 to 1930. Wiley, New York

Hobson MP et al (eds) (2010) Bayesian methods in cosmology. Cambridge University Press, Cambridge

Isobe T, Feigelson ED, Nelson PI (1986) Statistical methods for astronomical data with upper limits. II–correlation and regression. Astrophys J 306:490–507

Kelly BC (2007) Some Aspects of Measurement Error in Linear Regression of Astronomical Data. Astrophys J 665:1489–1506

Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon Not R Astro Soc 155:95–118

Martinez VJ, Saar E (2002) Statistics of the galaxy distribution. CRC, Boca Raton, USA

Protopapas P, Jimenez R, Alcock C (2005) Fast identification of transits from light-curves. Mon Not R Astro Soc 362: 460–468

Scargle JD (1982) Studies in astronomical time series analysis. II Statistical aspects of spectral analysis of unevenly spaced data. Astrophys J 263:835–853

Starck J-L, Murtagh F (2006) Astronomical image and data analysis. Springer, New York

Stellingwerf RF (1978) Period determination using phase dispersion minimization. Astrophys J 224:953–960

Stigler SM (1986) The history of Statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge, MA

Woodroofe MB (1985) Estimating a distribution function with truncated data. Ann Statist 13:163–177

A

Asymptotic Normality

John E. Kolassa
Professor
Rutgers University, Newark, NJ, USA

Consider a sequence of random variables T_n, whose distribution depends on a parameter n that generally represents sample size. The sequence is said to be asymptotically normal if there exists a sequences μ_n and σ_n such that $\lim_{n\to\infty} P\left[(T_n - \mu_n)/\sigma_n \le x\right] = \Phi(x)$ for all x, where $\Phi(x)$ is the standard Gaussian distribution function

$$\int_{-\infty}^{x} \exp(-y^2/2)(2\pi)^{-1/2}\, dy. \tag{1}$$

One often writes

$$T_n \sim AN(\mu_n, \sigma_n^2) \tag{2}$$

to express asymptotic normality. Note that μ_n generally depend on n, and furthermore may be data-dependent. Furthermore, in some cases T_n might be a sequence of random vectors; in this case, μ_n is a sequence of vectors, σ_n^2 is a sequence of matrices, and Φ the vector valued counterpart of (1). In the scalar case, for fixed n, the quantity σ_n is called the standard error of T_n.

Many frequentist statistical inferential procedures are performed by constructing a T_n so that (2) holds under a null hypothesis, with a dissimilar distribution under interesting alternative hypotheses, and reporting

$$2(1 - \Phi(|(T_n - \mu_n)/\sigma_n|)) \tag{3}$$

as a two-sided p-value; the application for one-sided p-values is similar, and there are also Bayesian applications of a similar flavor. Serfling (1980) provides further information.

Consider the following examples of quantities that are asymptotically normal:

- If T_n is the mean of n independent and identically distributed random variables, each with expectation μ and standard deviation σ, then

  $$T_n \sim AN(\mu, \sigma^2/n). \tag{4}$$

  Furthermore, if s_n is the traditional standard deviation of the contributors the the mean,

  $$T_n \sim AN(\mu, s_n^2/n); \tag{5}$$

  note that the standard error here is data-dependent, and it is incorrect to call $s_n/\sqrt{n}$ a standard deviation of T_n, even approximately. In the present case square root of the second argument to the AN operator estimates the standard deviation of T_n, but a further example shows that even this need not be true. In this case, the standard Z-test for a single sample mean follows from using (4) when σ is known, and when the components of T_n are binary, the standard standard Z-test for a single sample mean follows from using (4) with σ^2 the standard Bernoulli variance. When σ is unknown, (5) is often used instead, and for $n \le 30$, the t distribution function is generally used in place of Φ in (3) for greater accuracy.
- Many rank-based statistics are asymptotically normal; for example, if T_n is the Wilcoxon signed-rank statistic (see ▶Wilcoxon-Signed-Rank Test) for testing whether the expectation of n independent and identically distributed random variables takes on a particular null value, assuming symmetry and continuity of the underlying distribution. Without loss of generality, take this null mean to be zero. Then T_n is obtained by ranking the absolute values of the observations, and summing the ranks of those observations that are positive. Hettmansperger (1984) notes that (2) holds with $\mu_n = n(n+1)/2$ and $\sigma_n = \sqrt{n(n+1)(2n+1)/24}$, and the test against the two-sided alternative reports the p-value (3). In this case, T_n may be written as the sum of independent but not identically-distributed random variables, or as the sum of identically-distributed but not independent random variables.
- Many parameter estimates resulting from fitting models with independent observations are asymptotically normal. For example, consider independent Bernoulli observations Y_i with $P\left[Y_i = 1\right] = \exp(\beta_1 + \beta_2 x_i)/(1 + \exp(\beta_1 + \beta_2 x_i))$. Let

  $$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i\beta_1 + x_i Y_i \beta_2 - \log(1 + \exp(\beta_1 + \beta_2 x_i))\right], \tag{6}$$

  and let $\hat{\boldsymbol{\beta}}$ maximize ℓ; here $\hat{\boldsymbol{\beta}}$ implicitly depends on n. Then

  $$\hat{\boldsymbol{\beta}} \sim AN(\boldsymbol{\beta}, [-\ell''(\boldsymbol{\beta})]^{-1}), \tag{7}$$

  as one can see by heuristically expressing $\ell'(\boldsymbol{\beta}) + \ell''(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \ell'(\hat{\boldsymbol{\beta}}) = 0$, and solving for $\hat{\boldsymbol{\beta}}$ to obtain $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta} - [\ell''(\boldsymbol{\beta})]^{-1}\ell'(\boldsymbol{\beta})$, noting that $\ell''(\boldsymbol{\beta})$ is nonrandom, and noting that a variant of the Central Limit Theorem proves the asymptotic normality of $\ell'(\boldsymbol{\beta})$, and hence of $\hat{\boldsymbol{\beta}}$. This heuristic argument is easily made rigorous once one notes $\hat{\boldsymbol{\beta}}$ is consistent (i.e., for any $\epsilon > 0$, $\lim_{n\to\infty} P\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > \epsilon\right] = 0$; see Cox and Hinkley 1974). In this example, the outcome $Y_i = 1\ \forall i$

has positive probability, and for such $\{Y_1, \ldots, Y_n\}$, $\hat{\beta}_1$ is infinite. A similar result holds for $Y_i = 0\ \forall i$. Hence the variance of $\hat{\boldsymbol{\beta}}$ does not exist.

About the Author

John Kolassa is Professor of Statistics and Biostatistics, and Director of the Graduate Programs in Statistics and Biostatistics at Rutgers University. John Kolassa was previously Assistant and Associate Professor in the Department of Biostatistics, and Graduate Student Advisor, at the University of Rochester. John Kolassa is an Elected Fellow of the ASA and IMS, and has received grants from and served on grant review panels for the National Science Foundation and the National Institutes of Health.

Cross References

►Approximations to Distributions
►Asymptotic Relative Efficiency in Testing
►Central Limit Theorems
►Chernoff-Savage Theorem
►Limit Theorems of Probability Theory
►Martingale Central Limit Theorem
►Properties of Estimators
►Sampling Problems for Stochastic Processes
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

References and Further Reading

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, New York

Hettmansperger TP (1984) Statistical inference based on ranks. Krieger, Melbourne, FL

Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York

Asymptotic Relative Efficiency in Estimation

Robert Serfling
Professor
University of Texas at Dallas, Richardson, TX, USA

Asymptotic Relative Efficiency of Two Estimators

For statistical estimation problems, it is typical and even desirable that several reasonable estimators can arise for consideration. For example, the mean and median parameters of a symmetric distribution coincide, and so the *sample mean* and the *sample median* become competing estimators of the point of symmetry. *Which is preferred? By what criteria shall we make a choice?*

One natural and time-honored approach is simply to compare the sample sizes at which two competing estimators meet a given standard of performance. This depends upon the chosen measure of performance and upon the particular population distribution F.

To make the discussion of sample mean versus sample median more precise, consider a distribution function F with density function f symmetric about an unknown point θ to be estimated. For $\{X_1, \ldots, X_n\}$ a sample from F, put $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\mathrm{Med}_n = \mathrm{median}\{X_1, \ldots, X_n\}$. Each of $\overline{X}_n$ and Med_n is a consistent estimator of θ in the sense of convergence in probability to θ as the sample size $n \to \infty$. To choose between these estimators we need to use further information about their performance. In this regard, one key aspect is *efficiency*, which answers: *How spread out about θ is the sampling distribution of the estimator?* The smaller the variance in its sampling distribution, the more "efficient" is that estimator.

Here we consider "large-sample" sampling distributions. For $\overline{X}_n$, the classical central limit theorem (see ►Central Limit Theorems) tells us: if F has finite variance σ_F^2, then the sampling distribution of $\overline{X}_n$ is approximately $N\left(\theta, \sigma_F^2/n\right)$, i.e., Normal with mean θ and variance σ_F^2/n. For Med_n, a similar classical result (Serfling 1980) tells us: if the density f is continuous and positive at θ, then the sampling distribution of Med_n is approximately $N(\theta, 1/4[f(\theta)]^2 n)$. On this basis, we consider $\overline{X}_n$ and Med_n to perform equivalently at respective sample sizes n_1 and n_2 if

$$\frac{\sigma_F^2}{n_1} = \frac{1}{4[f(\theta)]^2 n_2}.$$

Keeping in mind that these sampling distributions are only approximations assuming that n_1 and n_2 are "large," we define the *asymptotic relative efficiency (ARE)* of Med to $\overline{X}$ as the *large-sample limit* of the ratio n_1/n_2, i.e.,

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, F\right) = 4[f(\theta)]^2 \sigma_F^2. \tag{1}$$

Definition in the General Case

For any parameter η of a distribution F, and for estimators $\widehat{\eta}^{(1)}$ and $\widehat{\eta}^{(2)}$ approximately $N(\eta, V_1(F)/n)$ and $N(\eta, V_2(F)/n)$, respectively, the *ARE of $\widehat{\eta}^{(2)}$ to $\widehat{\eta}^{(1)}$* is given by

$$\mathrm{ARE}\left(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F\right) = \frac{V_1(F)}{V_2(F)}. \tag{2}$$

Interpretation. If $\widehat{\eta}^{(2)}$ is used with a sample of size n, the number of observations needed for $\widehat{\eta}^{(1)}$ to perform equivalently is $\mathrm{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) \times n$.

Extension to the case of multidimensional parameter. For a parameter $\boldsymbol{\eta}$ taking values in $\mathbb{R}^k$, and two estimators $\widehat{\boldsymbol{\eta}}^{(i)}$ which are k-variate Normal with mean $\boldsymbol{\eta}$ and nonsingular covariance matrices $\boldsymbol{\Sigma}_i(F)/n$, $i = 1, 2$, we use [see Serfling (1980)]

$$\mathrm{ARE}\left(\widehat{\boldsymbol{\eta}}^{(2)}, \widehat{\boldsymbol{\eta}}^{(1)}, F\right) = \left(\frac{|\boldsymbol{\Sigma}_1(F)|}{|\boldsymbol{\Sigma}_2(F)|}\right)^{1/k}, \tag{3}$$

the ratio of *generalized variances* (determinants of the covariance matrices), raised to the power $1/k$.

Connection with the Maximum Likelihood Estimator

Let F have density $f(x|\eta)$ parameterized by $\eta \in \mathbb{R}$ and satisfying some differentiability conditions with respect to η. Suppose also that $I(F) = E_\eta\left\{\left[\frac{\partial}{\partial\eta}\log f(x|\eta)\right]^2\right\}$ (the *Fisher information*) is positive and finite. Then (Lehmann and Casella 1988) it follows that (a) the *maximum likelihood estimator* $\widehat{\eta}^{(\mathrm{ML})}$ of η is approximately $N(\eta, 1/I(F)n)$, and (b) for a wide class of estimators $\widehat{\eta}$ that are approximately $N(\eta, V(\widehat{\eta}, F)/n)$, a *lower bound* to $V(\widehat{\eta}, F)$ is $1/I(F)$. In this situation, (2) yields

$$\mathrm{ARE}\left(\widehat{\eta}, \widehat{\eta}^{(\mathrm{ML})}, F\right) = \frac{1}{I(F)V(\widehat{\eta}, F)} \le 1, \tag{4}$$

making $\widehat{\eta}^{(\mathrm{ML})}$ (asymptotically) the most efficient among the given class of estimators $\widehat{\eta}$. We note, however, as will be discussed later, that (4) does not necessarily make $\widehat{\eta}^{(\mathrm{ML})}$ the estimator of choice, when certain other considerations are taken into account.

Detailed Discussion of Estimation of Point of Symmetry

Let us now discuss in detail the example treated above, with F a distribution with density f symmetric about an unknown point θ and $\{X_1, \ldots, X_n\}$ a sample from F. For estimation of θ, we will consider not only $\overline{X}_n$ and Med_n but also a third important estimator.

Mean versus Median

Let us now formally compare $\overline{X}_n$ and Med_n and see how the ARE differs with choice of F. Using (1) with $F = N\left(\theta, \sigma_F^2\right)$, it is seen that

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, N\left(\theta, \sigma_F^2\right)\right) = 2/\pi = 0.64.$$

Thus, for sampling from a *Normal* distribution, the sample mean performs as efficiently as the sample median using only 64% as many observations. (Since θ and σ_F are location and scale parameters of F, and since the estimators $\overline{X}_n$ and Med_n are location and scale equivariant, their ARE does not depend upon these parameters.) The superiority of $\overline{X}_n$ here is no surprise since it is the MLE of θ in the model $N\left(\theta, \sigma_F^2\right)$.

As noted above, *asymptotic* relative efficiencies pertain to large sample comparisons and need not reliably indicate small sample performance. In particular, for F *Normal*, the *exact* relative efficiency of Med to $\overline{X}$ for sample size $n = 5$ is a very high 95%, although this decreases quickly, to 80% for $n = 10$, to 70% for $n = 20$, and to 64% in the limit.

For sampling from a *double exponential* (or *Laplace*) distribution with density $f(x) = \lambda e^{-\lambda|x-\theta|}/2$, $-\infty < x < \infty$ (and thus variance $2/\lambda^2$), the above result favoring $\overline{X}_n$ over Med_n is reversed: (1) yields

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, \mathrm{Laplace}\right) = 2,$$

so that the sample mean requires 200% as many observations to perform equivalently to the sample median. Again, this is no surprise because for this model the MLE of θ is Med_n.

A Compromise: The Hodges–Lehmann Location Estimator

We see from the above that the ARE depends dramatically upon the shape of the density f and thus must be used cautiously as a benchmark. For Normal versus Laplace, $\overline{X}_n$ is either greatly superior or greatly inferior to Med_n. This is a rather unsatisfactory situation, since in practice we might not be quite sure whether F is Normal or Laplace or some other type. A very interesting solution to this dilemma is given by an estimator that has excellent *overall performance*, the so-called *Hodges–Lehmann location estimator* (Hodges and Lehmann 1963; see ▶Hodges-Lehmann Estimators):

$$\mathrm{HL}_n = \mathrm{Median}\left\{\frac{X_i + X_j}{2}\right\},$$

the median of all pairwise averages of the sample observations. (Some authors include the cases $i = j$, some not.) We have (Lehmann 1998a) that HL_n is asymptotically $N(\theta, 1/12[\int f^2(x)dx]^2 n)$, which yields that $\mathrm{ARE}\left(\mathrm{HL}, \overline{X}, N\left(\theta, \sigma_F^2\right)\right) = 3/\pi = 0.955$ and $\mathrm{ARE}(\mathrm{HL}, \overline{X}, \mathrm{Laplace}) = 1.5$. Also, for the ▶*Logistic* distribution with density $f(x) = \sigma^{-1}e^{(x-\theta)/\sigma}/[1 + e^{(x-\theta)/\sigma}]^2$, $-\infty < x < \infty$, for which HL_n is the MLE of θ and thus optimal, we have $\mathrm{ARE}(\mathrm{HL}, \overline{X}, \mathrm{Logistic}) = \pi^2/9 = 1.097$ [see Lehmann (1998b)]. Further, for $\mathcal{F}$ the class of all distributions symmetric about θ and having finite variance, we have $\inf_{\mathcal{F}} \mathrm{ARE}(\mathrm{HL}, \overline{X}, F) = 108/125 = 0.864$ [see Lehmann (1998a)]. The estimator HL_n is highly competitive with $\overline{X}$ at Normal distributions, can be infinitely more efficient at some other symmetric distributions F, and is never much

less efficient at any distribution F in $\mathcal{F}$. The computation of HL_n appears at first glance to require $O(n^2)$ steps, but a much more efficient $O(n \log n)$ algorithm is available [see Monohan (1984)].

Efficiency versus Robustness Trade-Off

Although the asymptotically most efficient estimator is given by the MLE, the particular MLE depends upon the shape of F and can be drastically inefficient when the actual F departs even a little bit from the nominal F. For example, if the assumed F is $N(\mu, 1)$ but the actual model differs by a small amount ε of "contamination," i.e., $F = (1-\varepsilon)N(\mu, 1) + \varepsilon N(\mu, \sigma^2)$, then

$$\mathrm{ARE}(\mathrm{Med}, \overline{X}, F) = \frac{2}{\pi}\left(1 - \varepsilon + \varepsilon\sigma^{-1}\right)^2\left(1 - \varepsilon + \varepsilon\sigma^2\right),$$

which equals $2/\pi$ in the "ideal" case $\varepsilon = 0$ but otherwise $\to \infty$ as $\sigma \to \infty$. A small perturbation of the assumed model thus can destroy the superiority of the MLE.

One way around this issue is to take a *nonparametric* approach and seek an estimator with ARE satisfying a favorable lower bound. Above we saw how the estimator HL_n meets this need.

Another criterion by which to evaluate and compare estimators is *robustness*. Here let us use finite-sample *breakdown point (BP)*: the minimal fraction of sample points which may be taken to a limit L (e.g., $\pm\infty$) without the estimator also being taken to L. A *robust* estimator remains stable and effective when in fact the sample is only partly from the nominal distribution F and contains some non-F observations which might be relatively extreme contaminants.

A single observation taken to ∞ (with n fixed) takes $\overline{X}_n$ with it, so $\overline{X}_n$ has BP = 0. Its optimality at Normal distributions comes at the price of a complete sacrifice of robustness. In comparison, Med_n has extremely favorable BP = 0.5 but at the price of a considerable loss of efficiency at Normal models.

On the other hand, the estimator HL_n appeals broadly, possessing *both* quite high ARE over a wide class of F and relatively high BP $= 1 - 2^{-1/2} = 0.29$.

As another example, consider the problem of estimation of scale. Two classical scale estimators are the *sample standard deviation* s_n and the *sample MAD* (median absolute deviation about the median) MAD_n. They estimate scale in different ways but can be regarded as competitors in the problem of estimation of σ in the model $F = N(\mu, \sigma^2)$, as follows. With both μ and σ unknown, the estimator s_n is (essentially) the MLE of σ and is asymptotically most efficient. Also, for this F, the population MAD is equal to $\Phi^{-1}(3/4)\sigma$, so that the estimator $\widehat{\sigma}_n = \mathrm{MAD}_n/\Phi^{-1}(3/4) = 1.4826\,\mathrm{MAD}_n$ competes with s_n for estimation of σ. (Here Φ denotes the standard normal distribution function, and, for any F, $F^{-1}(p)$ denotes the pth quantile, $\inf\{x : F(x) \geq p\}$, for $0 < p < 1$.) To compare with respect to robustness, we note that a single observation taken to ∞ (with n fixed) takes s_n with it, s_n has BP = 0. On the other hand, MAD_n and thus $\widehat{\sigma}_n$ have BP = 0.5, like Med_n. However, $\mathrm{ARE}(\widehat{\sigma}_n, s_n, N(\mu, \sigma^2)) = 0.37$, even worse than the ARE of Med_n relative to $\overline{X}$. Clearly desired is a more balanced trade-off between efficiency and robustness than provided by either of s_n and $\widehat{\sigma}_n$. Alternative scale estimators having the same 0.5 BP as $\widehat{\sigma}_n$ but much higher ARE of 0.82 relative to s_n are developed in Rousseeuw and Croux (1993). Also, further competitors offering a range of trade-offs given by (BP, ARE) $= (0.29, 0.86)$ or $(0.13, 0.91)$ or $(0.07, 0.96)$, for example, are developed in Serfling (2002).

In general, efficiency and robustness trade off against each other. Thus ARE should be considered in conjunction with robustness, choosing the balance appropriate to the particular application context. This theme is prominent in the many examples treated in Staudte and Sheather (1990).

A Few Additional Aspects of ARE

Connections with Confidence Intervals

In view of the asymptotic normal distribution underlying the above formulation of ARE in estimation, we may also characterize the ARE given by (2) as the limiting ratio of sample sizes at which the *lengths of associated confidence intervals at approximate level* $100(1-\alpha)\%$,

$$\widehat{\eta}^{(i)} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{V_i(F)}{n_i}},\ i = 1, 2,$$

converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$. (In practice, of course, consistent estimates of $V_i(F)$, $i = 1, 2$, are used in forming the CI.)

Fixed Width Confidence Intervals and ARE

One may alternatively consider confidence intervals of *fixed length*, in which case (under typical conditions) the noncoverage probability depends on n and tends to 0 at an exponential rate, i.e., $n^{-1}\log\alpha_n \to c > 0$, as $n \to \infty$. For fixed width confidence intervals of the form

$$\widehat{\eta}^{(i)} \pm d\,\sigma_F,\ i = 1, 2,$$

we thus define the *fixed width asymptotic relative efficiency (FWARE)* of two estimators as the limiting ratio of sample sizes at which the respective *noncoverage probabilities* $\alpha_n^{(i)}$, $i = 1, 2$, of the associated fixed width confidence intervals

converge to zero at the same exponential rate. In particular, for Med versus $\overline{X}$, and letting $\eta = 0$ and $\sigma_F = 1$ without loss of generality, we obtain (Serfling and Wackerly 1976)

$$\mathrm{FWARE}(\mathrm{Med}, \overline{X}, F) = \frac{\log m(-d)}{\log[2(F(d) - F^2(d))^{1/2}]}, \quad (5)$$

where $m(-d)$ is a certain parameter of the ▶moment generating function of F. The FWARE is derived using *large deviation theory* instead of the central limit theorem. As $d \to 0$, the FWARE converges to the ARE. Indeed, for F a Normal distribution, this convergence (to $2/\pi = 0.64$) is quite rapid: the expression in (5) rounds to 0.60 for $d = 2$, to 0.63 for $d = 1$, and to 0.64 for $d \le 0.1$.

Confidence Ellipsoids and ARE

For an estimator $\widehat{\boldsymbol{\eta}}$ which is asymptotically k-variate Normal with mean $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma}/n$, as the sample size $n \to \infty$, we may form (see Serfling 1980) an *associated ellipsoidal confidence region of approximate level* $100(1-\alpha)\%$ for the parameter $\boldsymbol{\eta}$,

$$E_{n,\alpha} = \{\boldsymbol{\eta} : n\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})' \boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \le c_\alpha\},$$

with $P\left(\chi^2_k > c_\alpha\right) = \alpha$ and in practice using a consistent estimate of $\boldsymbol{\Sigma}$. The *volume* of the region $E_{n,\alpha}$ is

$$\frac{\pi^{k/2}(c_\alpha/n)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}{\Gamma((k+1)/2)}.$$

Therefore, for two such estimators $\widehat{\boldsymbol{\eta}}^{(i)}$, $i = 1, 2$, the ARE given by (3) may be characterized as the limiting ratio of sample sizes at which the *volumes of associated ellipsoidal confidence regions at approximate level* $100(1-\alpha)\%$ converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$.

Under regularity conditions on the model, the maximum likelihood estimator $\widehat{\boldsymbol{\eta}}^{(\mathrm{ML})}$ has a confidence ellipsoid $E_{n,\alpha}$ attaining the *smallest possible volume* and, moreover, lying wholly within that for any other estimator $\widehat{\boldsymbol{\eta}}$.

Connections with Testing

Parallel to ARE in estimation as developed here is the notion of *Pitman ARE* for comparison of two hypothesis test procedures. Based on a different formulation, although the central limit theorem is used in common, the Pitman ARE agrees with (2) when the estimator and the hypothesis test statistic are linked, as for example $\overline{X}$ paired with the *t-test*, or Med_n paired with the ▶*sign test*, or HL_n paired with the ▶*Wilcoxon-signed-rank test*. See Lehmann 1998b, Nikitin 1995, Nikitin 2010, and Serfling 1980.

Other Notions of ARE

As illustrated above with FWARE, several other important approaches to ARE have been developed, typically using either moderate or large deviation theory. For example, instead of asymptotic variance parameters as the criterion, one may compare *probability concentrations* of the estimators in an ε-neighborhood of the target parameter η: $P\left(\left|\widehat{\eta}^{(i)} - \eta\right| > \varepsilon\right)$, $i = 1, 2$. When we have

$$\frac{\log P(|\widehat{\eta}^{(i)}_n - \eta| > \varepsilon)}{n} \to \gamma^{(i)}(\varepsilon, \eta), \; i = 1, 2,$$

as is typical, then the ratio of sample sizes n_1/n_2 at which these concentration probabilities converge to 0 at the same rate is given by $\gamma^{(1)}(\varepsilon, \eta)/\gamma^{(2)}(\varepsilon, \eta)$, which then represents another ARE measure for the efficiency of estimator $\widehat{\eta}^{(2)}_n$ relative to $\widehat{\eta}^{(1)}_n$. See Serfling (1980, 1.15.4) for discussion and Basu (1956) for illustration that the variance-based and concentration-based measures need not agree on which estimator is better. For general treatments, see Nikitin (1995), Puhalskii and Spokoiny (1998), Nikitin (2010), and Serfling (1980, Chap. 10), as well as the other references cited below. A comprehensive bibliography is beyond the present scope. However, very productive is *ad hoc* exploration of the literature using a modern search engine.

Acknowledgments

Support by NSF Grant DMS-0805786 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

About the Author

Robert Serfling is author of the classic textbook *Approximation Theorems of Mathematical Statistics*, Wiley, 1980, and has published extensively in statistical journals. He received a Humboldt-Preis, awarded by the Alexander von Humboldt Stiftung, Germany, "in recognition of accomplishments in research and teaching" (1985). He is a Fellow of the American Statistical Association and of Institute of Mathematical Statistics, and an Elected Member of International Statistical Institute. Professor Serfling was Editor of the IMS Lecture Notes Monograph Series (1988–1993) and currently is an Associate Editor for *Journal of Multivariate Analysis* (2007–) and for *Journal of Nonparametric Statistics* (2007–).

Cross References

▶Asymptotic Relative Efficiency in Testing
▶Estimation
▶Estimation: An Overview
▶Mean Median and Mode
▶Normality Tests

►Properties of Estimators
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Basu D (1956) On the concept of asymptotic relative efficiency. Sankhyā 17:193–196
Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. Ann Math Stat 34:598–611
Lehmann EL (1998a) Elements of large-sample theory. Springer, New York
Lehmann EL (1998b) Nonparametrics: statistical methods based on ranks. Prentice-Hall, Upper Saddle River, NJ
Lehmann EL, Casella G (1988) Theory of point estimation, 2nd edn. Springer, New York
Monohan JF (1984) Algorithm 616: fast computation of the Hodges–Lehmann location estimator. ACM T Math Software 10:265–270
Nikitin Y (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge
Nikitin Y (2010) Asymptotic relative efficiency in testing. International Encyclopedia of Statistical Sciences. Springer, New York
Puhalskii A, Spokoiny V (1998) On large-deviation efficiency in statistical inference. Bernoulli 4:203–272
Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88:1273–1283
Serfling R (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York
Serfling R (2002) Efficient and robust fitting of lognormal distributions. N Am Actuarial J 4:95–109
Serfling R, Wackerly DD (1976) Asymptotic theory of sequential fixed-width confidence interval procedures. J Am Stat Assoc 71:949–955
Staudte RG, Sheather SJ (1990) Robust estimation and testing. Wiley, New York

Asymptotic Relative Efficiency in Testing

Yakov Nikitin
Professor, Chair of Probability and Statistics
St. Petersburg University, St. Petersburg, Russia

Asymptotic Relative Efficiency of Two Tests

Making a substantiated choice of the most efficient statistical test of several ones being at the disposal of the statistician is regarded as one of the basic problems of Statistics. This problem became especially important in the middle of XX century when appeared computationally simple but "inefficient" rank tests.

Asymptotic relative efficiency (ARE) is a notion which enables to implement in large samples the quantitative comparison of two different tests used for testing of the same statistical hypothesis. The notion of the asymptotic efficiency of *tests* is more complicated than that of asymptotic efficiency of *estimates*. Various approaches to this notion were identified only in late forties and early fifties, hence, 20–25 years later than in the estimation theory. We proceed now to their description.

Let $\{T_n\}$ and $\{V_n\}$ be two sequences of statistics based on n observations and assigned for testing the null-hypothesis H against the alternative A. We assume that the alternative is characterized by real parameter θ and for $\theta = \theta_0$ turns into H. Denote by $N_T(\alpha, \beta, \theta)$ the sample size necessary for the sequence $\{T_n\}$ in order to attain the power β under the level α and the alternative value of parameter θ. The number $N_V(\alpha, \beta, \theta)$ is defined in the same way.

It is natural to prefer the sequence with smaller N. Therefore the relative efficiency of the sequence $\{T_n\}$ with respect to the sequence $\{V_n\}$ is specified as the quantity

$$e_{T,V}(\alpha, \beta, \theta) = N_V(\alpha, \beta, \theta) \big/ N_T(\alpha, \beta, \theta),$$

so that it is the reciprocal ratio of sample sizes N_T and N_V.

The merits of the relative efficiency as means for comparing the tests are universally acknowledged. Unfortunately it is extremely difficult to explicitly compute $N_T(\alpha, \beta, \theta)$ even for the simplest sequences of statistics $\{T_n\}$. At present it is recognized that there is a possibility to avoid this difficulty by calculating the limiting values $e_{T,V}(\alpha, \beta, \theta)$ as $\theta \to \theta_0$, as $\alpha \to 0$ and as $\beta \to 1$ keeping two other parameters fixed. These limiting values $e^{P}_{T,V}$, $e^{B}_{T,V}$ and $e^{HL}_{T,V}$ are called respectively the Pitman, Bahadur and Hodges–Lehmann asymptotic relative efficiency (ARE), they were proposed correspondingly in Pitman (1949), Bahadur (1960) and Hodges and Lehmann (1956).

Only close alternatives, high powers and small levels are of the most interest from the practical point of view. It keeps one assured that the knowledge of these ARE types will facilitate comparing concurrent tests, thus producing well-founded application recommendations.

The calculation of the mentioned three basic types of efficiency is not easy, see the description of theory and many examples in Serfling (1980), Nikitin (1995) and Van der Vaart (1998). We only mention here, that Pitman efficiency is based on the central limit theorem (see ►Central Limit Theorems) for test statistics. On the contrary, Bahadur efficiency requires the large deviation asymptotics of test statistics under the null-hypothesis, while Hodges–Lehmann efficiency is connected with large deviation asymptotics under the alternative. Each type of efficiency has its own merits and drawbacks.

Pitman Efficiency

Pitman efficiency is the classical notion used most often for the asymptotic comparison of various tests. Under some regularity conditions assuming ▶asymptotic normality of test statistics under H and A, it is a number which has been gradually calculated for numerous pairs of tests.

We quote now as an example one of the first Pitman's results that stimulated the development of nonparametric statistics. Consider the two-sample problem when under the null-hypothesis both samples have the same continuous distribution and under the alternative differ only in location. Let $e^P_{W,t}$ be the Pitman ARE of the two-sample Wilcoxon rank sum test (see ▶Wilcoxon–Mann–Whitney Test) with respect to the corresponding Student test (see ▶Student's t-Tests). Pitman proved that for Gaussian samples $e^P_{W,t} = 3/\pi \approx 0.955$, and it shows that the ARE of the Wilcoxon test in the comparison with the Student test (being optimal in this problem) is unexpectedly high. Later Hodges and Lehmann (1956) proved that

$$0.864 \le e^P_{W,t} \le +\infty,$$

if one rejects the assumption of normality and, moreover, the lower bound is attained at the density

$$f(x) = \begin{cases} 3\left(5 - x^2\right)/\left(20\sqrt{5}\right) & \text{if } |x| \le \sqrt{5}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence the Wilcoxon rank test can be infinitely better than the parametric test of Student but their ARE never falls below 0.864. See analogous results in Serfling (2010) where the calculation of ARE of related estimators is discussed.

Another example is the comparison of independence tests based on Spearman and Pearson correlation coefficients in bivariate normal samples. Then the value of Pitman efficiency is $9/\pi^2 \approx 0.912$.

In numerical comparisons, the Pitman efficiency appear to be more relevant for moderate sample sizes than other efficiencies Groeneboom and Oosterhoff (1981). On the other hand, Pitman ARE can be insufficient for the comparison of tests. Suppose, for instance, that we have a normally distributed sample with the mean θ and variance 1 and we are testing $H : \theta = 0$ against $A : \theta > 0$. Let compare two significance tests based on the sample mean $\bar{X}$ and the Student ratio t. As the t-test does not use the information on the known variance, it should be inferior to the optimal test using the sample mean. However, from the point of view of Pitman efficiency, these two tests are equivalent. On the contrary, Bahadur efficiency $e^B_{t,\bar{X}}(\theta)$ is strictly less than 1 for any $\theta > 0$.

If the condition of asymptotic normality fails, considerable difficulties arise when calculating the Pitman ARE as the latter may not at all exist or may depend on α and β. Usually one considers limiting Pitman ARE as $\alpha \to 0$. Wieand (1976) has established the correspondence between this kind of ARE and the limiting approximate Bahadur efficiency which is easy to calculate.

Bahadur Efficiency

The Bahadur approach proposed in Bahadur (1960; 1967) to measuring the ARE prescribes one to fix the power of tests and to compare the exponential rate of decrease of their sizes for the increasing number of observations and fixed alternative. This exponential rate for a sequence of statistics $\{T_n\}$ is usually proportional to some non-random function $c_T(\theta)$ depending on the alternative parameter θ which is called the *exact slope* of the sequence $\{T_n\}$. The Bahadur ARE $e^B_{V,T}(\theta)$ of two sequences of statistics $\{V_n\}$ and $\{T_n\}$ is defined by means of the formula

$$e^B_{V,T}(\theta) = c_V(\theta) \,/\, c_T(\theta).$$

It is known that for the calculation of exact slopes it is necessary to determine the large deviation asymptotics of a sequence $\{T_n\}$ under the null-hypothesis. This problem is always nontrivial, and the calculation of Bahadur efficiency heavily depends on advancements in large deviation theory, see Dembo and Zeitouni (1998) and Deuschel and Stroock (1989).

It is important to note that there exists an upper bound for exact slopes

$$c_T(\theta) \le 2K(\theta)$$

in terms of Kullback–Leibler information number $K(\theta)$ which measures the "statistical distance" between the alternative and the null-hypothesis. It is sometimes compared in the literature with the ▶Cramér–Rao inequality in the estimation theory. Therefore the absolute (nonrelative) Bahadur efficiency of the sequence $\{T_n\}$ can be defined as $e^B_T(\theta) = c_T(\theta)/2K(\theta)$.

It is proved that under some regularity conditions the likelihood ratio statistic is asymptotically optimal in Bahadur sense (Bahadur 1967; Van der Vaart 1998, Sect. 16.6; Arcones 2005).

Often the exact Bahadur ARE is uncomputable for any alternative θ but it is possible to calculate the limit of Bahadur ARE as θ approaches the null-hypothesis. Then one speaks about the *local* Bahadur efficiency.

The indisputable merit of Bahadur efficiency consists in that it can be calculated for statistics with non-normal asymptotic distribution such as Kolmogorov-Smirnov, omega-square, Watson and many other statistics.

Asymptotic Relative Efficiency in Testing. Table 1 Some local Bahadur efficiencies

Statistic	Distribution					
	Gauss	Logistic	Laplace	Hyperbolic cosine	Cauchy	Gumbel
D_n	0.637	0.750	1	0.811	0.811	0.541
ω_n^2	0.907	0.987	0.822	1	0.750	0.731

Consider, for instance, the sample with the distribution function (df) F and suppose we are testing the goodness-of-fit hypothesis $H_0 : F = F_0$ for some known continuous df F_0 against the alternative of location. Well-known distribution-free statistics for this hypothesis are the Kolmogorov statistic D_n and omega-square statistic ω_n^2. The following table presents their local absolute efficiency in case of six standard underlying distributions:

We see from Table 1 that the integral statistic ω_n^2 is in most cases preferable with respect to the supremum-type statistic D_n. However, in the case of Laplace distribution the Kolmogorov statistic is locally optimal, the same happens for the Cramér-von Mises statistic in the case of hyperbolic cosine distribution. This observation can be explained in the framework of Bahadur local optimality, see Nikitin (1995 Chap. 6).

See also Nikitin (1995) for the calculation of local Bahadur efficiencies in case of many other statistics.

Hodges–Lehmann efficiency

This type of the ARE proposed in Hodges and Lehmann (1956) is in the conformity with the classical Neyman-Pearson approach. In contrast with Bahadur efficiency, let us fix the level of tests and let compare the exponential rate of decrease of their second-kind errors for the increasing number of observations and fixed alternative. This exponential rate for a sequence of statistics $\{T_n\}$ is measured by some non-random function $d_T(\theta)$ which is called the Hodges–Lehmann index of the sequence $\{T_n\}$. For two such sequences the Hodges–Lehmann ARE is equal to the ratio of corresponding indices.

The computation of Hodges–Lehmann indices is difficult as requires large deviation asymptotics of test statistics under the alternative.

There exists an upper bound for the Hodges–Lehmann indices analogous to the upper bound for Bahadur exact slopes. As in the Bahadur theory the sequence of statistics $\{T_n\}$ is said to be *asymptotically optimal in the Hodges–Lehmann sense* if this upper bound is attained.

The drawback of Hodges–Lehmann efficiency is that most *two-sided* tests like Kolmogorov and Cramér-von Mises tests are asymptotically optimal, and hence this kind of efficiency cannot discriminate between them. On the other hand, under some regularity conditions the one-sided tests like linear rank tests can be compared on the basis of their indices, and their Hodges–Lehmann efficiency coincides locally with Bahadur efficiency, see details in Nikitin (1995).

Coupled with three "basic" approaches to the ARE calculation described above, intermediate approaches are also possible if the transition to the limit occurs simultaneously for two parameters at a controlled way. Thus emerged the Chernoff ARE introduced by Chernoff (1952), see also Kallenberg (1982); the intermediate, or the Kallenberg ARE introduced by Kallenberg (1983), and the Borovkov–Mogulskii ARE, proposed in Borovkov and Mogulskii (1993).

Large deviation approach to asymptotic efficiency of tests was applied in recent years to more general problems. For instance, the change-point, "signal plus white noise" and regression problems were treated in Puhalskii and Spokoiny (1998), the tests for spectral density of a stationary process were discussed in Kakizawa (2005), while Taniguchi (2001) deals with the time series problems, and Otsu (2010) studies the empirical likelihood for testing moment condition models.

About the Author

Professor Nikitin is Chair of Probability and Statistics of St. Petersburg University. He is an Associate editor of *Statistics and Probability Letters*, and member of the editorial Board, *Mathematical Methods of Statistics* and *Metron*. He is a Fellow of the Institute of Mathematical Statistics. Professor Nikitin is the author of the text *Asymptotic efficiency of nonparametric tests*, Cambridge University Press, NY, 1995, and has authored more than 100 papers, in many international journals, in the field of Asymptotic efficiency of statistical tests, large deviations of test statistics and nonparametric Statistics.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Chernoff-Savage Theorem
►Nonparametric Statistical Inference
►Robust Inference

References and Further Reading

Arcones M (2005) Bahadur efficiency of the likelihood ratio test. Math Method Stat 14:163–179

Bahadur RR (1960) Stochastic comparison of tests. Ann Math Stat 31:276–295

Bahadur RR (1967) Rates of convergence of estimates and test statistics. Ann Math Stat 38:303–324

Borovkov A, Mogulskii A (1993) Large deviations and testing of statistical hypotheses. Siberian Adv Math 2(3, 4); 3(1, 2)
Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations. Ann Math Stat 23:493–507
Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York
Deuschel J-D, Stroock D (1989) Large deviations. Academic, Boston
Groeneboom P, Oosterhoff J (1981) Bahadur efficiency and small sample efficiency. Int Stat Rev 49:127–141
Hodges J, Lehmann EL (1956) The efficiency of some nonparametric competitors of the t-test. Ann Math Stat 26:324–335
Kakizawa Y (2005) Bahadur exact slopes of some tests for spectral densities. J Nonparametric Stat 17:745–764
Kallenberg WCM (1983) Intermediate efficiency, theory and examples. Ann Stat 11:170–182
Kallenberg WCM (1982) Chernoff efficiency and deficiency. Ann Stat 10:583–594
Nikitin Y (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge
Otsu T (2010) On Bahadur efficiency of empirical likelihood. J Econ 157:248–256
Pitman EJG (1949) Lecture notes on nonparametric statistical inference. Columbia University, Mimeographed
Puhalskii A, Spokoiny V (1998) On large-deviation efficiency in statistical inference. Bernoulli 4:203–272
Serfling R (1980) Approximation theorems of mathematical statistics. Wiley, New York
Serfling R (2010) Asymptotic relative efficiency in estimation. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer
Taniguchi M (2001) On large deviation asymptotics of some tests in time series. J Stat Plann Inf 97:191–200
Van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wieand HS (1976) A condition under which the Pitman and Bahadur approaches to efficiency coincide. Ann Statist 4:1003–1011

Asymptotic, Higher Order

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

Higher order asymptotic deals with two sorts of closely related things. First, there are questions of approximation. One is concerned with expansions or inequalities for a distribution function. Second, there are inferential issues. These involve, among other things, the application of the ideas of the study of higher order efficiency, admissibility and minimaxity. In the matter of expansions, it is as important to have usable, explicit formulas as a rigorous proof that the expansions are valid in the sense of truly approximating a target quantity up to the claimed degree of accuracy.

Classical asymptotics is based on the notion of asymptotic distribution, often derived from the central limit theorem (see ►Central Limit Theorems), and usually the approximations are correct up to $O(n^{-1/2})$, where n is the sample size. Higher order asymptotics provides refinements based on asymptotic expansions of the distribution or density function of an estimator of a parameter. They are rooted in the Edgeworth theory, which is itself a refinement of the central limit theorem. The theory of higher order asymptotic is very much related with the corresponding to *Approximations to distributions* treated as well in this Encyclopedia.

When higher order asymptotic is correct up to $o(n^{-1/2})$, it is second order asymptotic. When further terms are picked up, so that the asymptotic is correct up to $o(n^{-1})$, it is third order asymptotic. In his pioneering papers, C. R. Rao coined the term second order efficiency for a concept that would now is called third order efficiency. The new terminology is essentially owing to Pfanzagl and Takeuchi.

About the Author

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

Cross References

►Approximations to Distributions
►Edgeworth Expansion

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD thesis. The London School of Economics and Political Science, University of London, England
Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312
Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650
Durbin J (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333
Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Ghosh JK (1994) Higher order Asymptotic. NSF-CBMS Regional Conference Series in Probability and Statistics, 4. Hayward and Alexandria: Institute of Mathematical Statistics and American Statistical Association
Pfanzagl J (1979) Asymptotic expansions in parametric statistical theory. In: Krishnaiah PR (ed) Developments in statistics, vol. 3. Academic, New York, pp 1–97
Rao CR (1961) Asymptotic efficiency and limiting information. In Proceedings of Fourth Berkeley Symposium on Mathematical

Statistics and Probability. Berkeley: University of California Press, pp 531–546
Rao CR (1962) Efficient estimates and optimum inference procedure in large samples. J R Stat Soc B 24:46–63
Rao CR (1963) Criteria of estimation in large samples. Sankhya B 25:189–206
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Autocorrelation in Regression

Badi H. Baltagi
Distinguished Professor of Economics
Syracuse University, Syracuse, NY, USA

Linear regressions are a useful empirical tool for economists and social scientists and the standard ►least squares estimates are popular because they are the best linear unbiased estimators (BLUE) under some albeit strict assumptions. These assumptions require the regression disturbances not to be correlated with the regressors, also homoskedastic, i.e., with constant variance, and not autocorrelated. Violation of the no autocorrelation assumption on the disturbances, will lead to inefficiency of the least squares estimates, i.e., no longer having the smallest variance among all linear unbiased estimators. It also leads to wrong standard errors for the regression coefficient estimates. This in turn leads to wrong t-statistics on the significance of these regression coefficients and misleading statistical inference based on a wrong estimate of the variance–covariance matrix computed under the assumption of no autocorrelation. This is why standard regression packages have a robust heteroskedasticity and autocorrelation-consistent covariance matrix (HAC) option for these regressions which at least robustifies the standard errors of least squares and shows how sensitive they would be to such violations, see Newey and West (1987).

Autocorrelation is more likely to occur in time-series than in cross-section studies. Consider estimating the consumption function of a random sample of households. An unexpected event, like a visit of family members will increase the consumption of this household. However, this positive disturbance need not be correlated with the disturbances affecting consumption of other randomly drawn households. However, if we were estimating this consumption function using aggregate time-series data for the U.S., then it is very likely that a recession year affecting consumption negatively that year, may have a carry over effect to the next few years. A shock to the economy like an oil embargo in 1973 is likely to affect the economy for several years. A labor strike this year may affect production for the next few years. The simplest work horse for illustrating this autocorrelation in time series on the regression disturbances, say u_t is the first-order autoregressive process denoted by AR(1):

$$u_t = \rho u_{t-1} + \epsilon_t \qquad t = 1, 2, \ldots, T$$

where ϵ_t is independent and identically distributed with mean 0 and variance σ_ϵ^2. It is autoregressive because u_t is related to its lagged value u_{t-1}. One can show, see for example Baltagi (2008), that the correlation coefficient between u_t and u_{t-1} is ρ. Also, that the correlation coefficient between u_t and u_{t-r} is ρ^r. When $\rho = 0$, there is no autocorrelation and one test for this null hypothesis is the Durbin and Watson (1951) test discussed as a separate entry in this encyclopedia by Krämer. This AR(1) process is also stationary as long as $|\rho| < 1$. If $\rho = 1$, then this process has a unit root and it is called a ►random walk. See the entry by Dickey on testing for this unit root using the ►Dickey-Fuller tests. Note that if the process is stationary, then ρ is a fraction and the correlation for two disturbances r periods apart is ρ^r, i.e., a fraction raised to an integer power. This means that the correlation is decaying between the disturbances the further apart they are. This is reasonable in economics and may be the reason why this AR(1) form is so popular. One should note that this is not the only form that would correlate the disturbances across time. Other forms like the Moving Average (MA) process, and higher order Autoregressive Moving Average (ARMA) processes are popular, see Box and Jenkins (1970), but these are beyond the scope of this entry.

Since least squares is no longer BLUE under autocorrelation of the disturbances, Cochrane and Orcutt (1949) suggested a simple estimator that corrects for autocorrelation of the AR(1) type. This method starts with an initial estimate of ρ, the most convenient is 0, and minimizes the residual sum of squares of the regression. This gives us the least squares estimates of the regression coefficients and the corresponding least squares residuals which we denote by e_t. In the next step, one regresses e_t on e_{t-1} to get an estimate of ρ, say $\widehat{\rho}$. The second step of the Cochrane–Orcutt procedure (2SCO) is to perform the regression of $(Y_t - \widehat{\rho} Y_{t-1})$ on $(X_t - \widehat{\rho} X_{t-1})$ to get estimates of the regression coefficients. One can iterate this procedure (ITCO) until convergence. Both the 2SCO and the ITCO are asymptotically efficient as the sample size gets large. The argument for iterating must be justified in terms of small sample gains. Other methods of correcting for serial correlation include Prais and Winsten (1954), Durbin (1960), as well as maximum likelihood methods, all studied more extensively in Chap. 5 of Baltagi (2008). The Prais and

Winsten method recaptures the initial observation lost in the Cochrane–Orcutt method. Monte Carlo studies using an autoregressive regressor, and various values of ρ, found that least squares is still a viable estimator as long as $|\rho| < 0.3$, but if $|\rho| > 0.3$, then it pays to perform procedures that correct for serial correlation based on an estimator of ρ. For trended regressors, which is usually the case with economic data, least squares outperforms 2SCO, but not the Prais-Winsten procedure that recaptures the initial observation. In fact, Park and Mitchell (1980) who performed an extensive Monte Carlo using trended and untrended regressors recommend that one should not use regressions based on $(T - 1)$ observations as in the Cochrane and Orcutt procedure. They also found that test of hypotheses regarding the regression coefficients performed miserably for all estimators based on an estimator of ρ.

Correcting for serial correlation is not without its critics. Mizon (1995) argues this point forcefully in his article entitled "A simple message for autocorrelation correctors: Don't." The main point being that serial correlation is a symptom of dynamic misspecification which is better represented using a general unrestricted dynamic specification.

About the Author

Badi H. Baltagi is distinguished Professor of Economics, and Senior Research Associate at the Center for Policy Research, Syracuse University. He received his Ph.D. in Economics at the University of Pennsylvania in 1979. He served on the faculty at the University of Houston and Texas A&M University. He was a visiting Professor at the University of Arizona and the University of California, San Diego. He is the author of *Econometric Analysis of Panel Data* (Wiley, 4*th* edn. 2008); *Econometrics* (Springer, 4*th* edn. 2008), and editor of *A Companion to Theoretical Econometrics* (Wiley–Blackwell); *Recent Developments in the Econometrics of Panel Data*, Volumes I and II (Edward-Elgar); *Nonstationary Panels*, *Panel Cointegration*, and *Dynamic Panels* (Elsevier); *Panel Data Econometrics: Theoretical Contributions and Empirical Applications* (Physica-Verlag, 2004); *Spatial Econometrics: Methods and Applications*, (with Giuseppe Arbia), Physica-Verlag, 2009. He is author or co-author of over 100 publications, all in leading economics and statistics journals. Professor Baltagi was the holder of the George Summey, Jr. Professor Chair in Liberal Arts and was awarded the Distinguished Achievement Award in Research at Texas A&M University. He is co-editor of *Empirical Economics*, and Associate editor of *Journal of Econometrics* and *Econometric Reviews*. He is the replication editor of the *Journal of Applied Econometrics* and the series editor for *Contributions to Economic Analysis*. He is a fellow of the Journal of Econometrics and a recipient of the Multa and Plura Scripsit Awards from Econometric Theory. He is a founding fellow and member of the Board of Directors of the Spatial Econometrics Association.

Cross References

►Approximations to Distributions
►Box–Jenkins Time Series Models
►Correlation Coefficient
►Dickey-Fuller Tests
►Durbin–Watson Test
►Linear Regression Models
►Structural Time Series Models
►Tests of Independence
►Time Series
►Time Series Regression

References and Further Reading

Baltagi BH (2008) Econometrics, 4th edn., Springer, Berlin

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden Day, San Francisco

Cochrane D, Orcutt G (1949) Application of least squares regression to relationships containing autocorrelated error terms. J Am Stat Assoc 44:32–61

Durbin J (1960) Estimation of parameters in time-series regression model. J R Stat Soc, B, 22:139–153

Durbin J, Watson G (1951) Testing for serial correlation in least squares regression-II. Biometrika 38:159–178

Mizon GE (1995) A simple message for autocorrelation correctors: don't. J Econometrics 69:267–288

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55:703–708

Park RE, Mitchell BM (1980) Estimating the autocorrelated error model with trended data. J Econometrics 13:185–201

Prais S, Winsten C (1954) Trend estimation and serial correlation. Discussion Paper 383, Cowles Commission, Chicago

Axioms of Probability

Vincenzo Capasso
Professor of Probability and Mathematical Statistics
University of Milan, Milan, Italy

Ingredients of Probability Spaces

Definition 1 A collection $\mathcal{F}$ of subsets of a set Ω is called a *ring* on Ω if it satisfies the following conditions:

1. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$,
2. $A, B \in \mathcal{F} \Rightarrow A \setminus B \in \mathcal{F}$.

A ring $\mathcal{F}$ is called an *algebra* if $\Omega \in \mathcal{F}$.

Definition 2 A ring $\mathcal{F}$ on Ω is called a *σ-ring* if it satisfies the following additional condition:

3. For every countable family $(A_n)_{n\in\mathbb{N}}$ of subsets of $\mathcal{F}$: $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

A σ-ring $\mathcal{F}$ on Ω is called a *σ-algebra* (or *σ-field*) if $\Omega \in \mathcal{F}$.

Proposition 1 The following properties hold:

1. If $\mathcal{F}$ is a σ-algebra of subsets of a set Ω, then it is an algebra.
2. If $\mathcal{F}$ is a σ-algebra of subsets of Ω, then
   - For any countable family $(E_n)_{n\in\mathbb{N}\setminus\{0\}}$ of elements of $\mathcal{F}$: $\bigcap_{n=1}^{\infty} E_n \in \mathcal{F}$,
   - For any finite family $(E_i)_{1\le i\le n}$ of elements of $\mathcal{F}$: $\bigcap_{i=1}^{n} E_i \in \mathcal{F}$,
   - $B \in \mathcal{F} \Rightarrow \Omega \setminus B \in \mathcal{F}$.

Definition 3 Every pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-ring $\mathcal{F}$ of subsets of Ω is a *measurable space*. Furthermore, if $\mathcal{F}$ is a σ-algebra, then $(\Omega, \mathcal{F})$ is a *measurable space on which a probability measure can be built*.

Example 1

1. *Generated σ-algebra*. If $\mathcal{A}$ is a family of subsets of a set Ω, then there exists the smallest σ-algebra of subsets of Ω that contains $\mathcal{A}$. This is the σ-algebra *generated* by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$. If, now, $\mathcal{G}$ is the set of all σ-algebras of subsets of Ω containing $\mathcal{A}$, then it is not empty because it has the set $\mathcal{P}(\Omega)$ of all subsets of Ω, among its elements, so that $\sigma(\mathcal{A}) = \bigcap_{\mathcal{C}\in\mathcal{G}} \mathcal{C}$.
2. *Borel σ-algebra*. Let Ω be a topological space. Then the *Borel σ-algebra* on Ω, denoted by $\mathcal{B}_\Omega$, is the σ-algebra generated by the set of all open subsets of Ω. Its elements are called Borel sets.
3. *Product σ-algebra*. Let $(\Omega_i, \mathcal{F}_i)_{1\le i\le n}$ be a family of measurable spaces, with all $\mathcal{F}_i, 1 \le i \le n$, σ-algebras, and let $\Omega = \prod_{i=1}^{n} \Omega_i$. Defining

$$\mathcal{R} = \left\{ E \subset \Omega \,|\, \forall i = 1, \ldots, n \; \exists E_i \in \mathcal{F}_i \text{ such that } E = \prod_{i=1}^{n} E_i \right\},$$

the σ-algebra on Ω generated by $\mathcal{R}$ is called the *product σ-algebra* of the σ-algebras $(\mathcal{F}_i)_{1\le i\le n}$.

Proposition 2 Let $(\Omega_i)_{1\le i\le n}$ be a family of topological spaces with a countable base and let $\Omega = \prod_{i=1}^{n} \Omega_i$. Then the Borel σ-algebra $\mathcal{B}_\Omega$ is identical to the product σ-algebra of the family of Borel σ-algebras $(\mathcal{B}_{\Omega_i})_{1\le i\le n}$.

Axioms of Probability

We assume that the reader is already familiar with the basic motivations and notions of probability theory. We present the axioms of probability according to the Kolmogorov approach [see Kolmogorov (1956)].

Definition 4 Given a set Ω, and a σ-algebra $\mathcal{F}$ of subsets of Ω, a probability measure on $\mathcal{F}$ is any function $P : \mathcal{F} \to [0,1]$ such that

P_1. $P(\Omega) = 1$,

P_2. for any countable family $A_1, \ldots, A_n, \ldots$ of elements of $\mathcal{F}$ such that $A_i \cap A_j = \emptyset$, whenever $i \neq j$:

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n).$$

Definition 5 A *probability space* is an ordered triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $P : \mathcal{F} \to [0,1]$ is a probability measure on $\mathcal{F}$. The set Ω is called the *sample space*, the elements of $\mathcal{F}$ are called *events*.

Definition 6 A probability space $(\Omega, \mathcal{F}, P)$ is *finite* if Ω has finitely many elements.

Remark 1 If Ω is at most countable, then it is usual to assume that $\mathcal{F} = \mathcal{P}(\Omega)$, the σ-algebra of all subsets of Ω. In this case all sets $\{\omega\}$ reduced to sample points $\omega \in \Omega$ are events; they are called *elementary events*.

Remark 2 If the σ−algebra of events $\mathcal{F}$ is finite, then the requirement of countable additivity in the definition of the probability measure P can be reduced to finite additivity.

Remark 3 It is worth mentioning that an important alternative approach to probability theory is the so called *subjective probability*; this approach does not insist on Axiom P_2., and rather uses the finite version of it (De Finetti 1974–1975).

Definition 7 A finite probability space $(\Omega, \mathcal{F}, P)$ with $\mathcal{F} = \mathcal{P}(\Omega)$ is an *equiprobable* or *uniform* space, if

$$\forall\, \omega \in \Omega : \qquad P(\{\omega\}) = k \text{ (constant)};$$

i.e., its elementary events are equiprobable.

Remark 4 Following the axioms of a probability space and the definition of a uniform space, if $(\Omega, \mathcal{F}, P)$ is equiprobable, then

$$\forall \omega \in \Omega : \qquad P(\{\omega\}) = \frac{1}{|\Omega|},$$

where $|\cdot|$ denotes the cardinal number of elementary events in Ω, and

$$\forall A \in \mathcal{F} \equiv \mathcal{P}(\Omega) : \qquad P(A) = \frac{|A|}{|\Omega|}.$$

Intuitively, in this case we may say that $P(A)$ is the ratio of the number of favorable outcomes, divided by the number of all possible outcomes.

Example 2 Consider an urn that contains 100 balls, of which 80 are red and 20 are black but that are otherwise identical, from which a player draws a ball. Define the event

R: The first drawn ball is red.

Then

$$P(R) = \frac{|R|}{|\Omega|} = \frac{80}{100} = 0.8.$$

Definition 8 We shall call any event $F \in \mathcal{F}$ such that $P(F) = 0$, a *null event*.

Elementary consequences of the above definitions are the following ones.

Proposition 3 Let $(\Omega, \mathcal{F}, P)$ be is a probability space.

1. $P(A^c) = 1 - P(A)$, for any $A \in \mathcal{F}$;
2. $P(\emptyset) = 0$;
3. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(B) = P(A) + P(B \setminus A)$;
4. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(A) \leq P(B)$ (monotonicity);
5. If $A, B \in \mathcal{F}$, then

$$P(B \setminus A) = P(B) - P(B \cap A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B);$$

6. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$;
7. (Principle of inclusion-exclusion) Let $A_1, \ldots, A_n \in \mathcal{F}$, then

$$P\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n);$$

8. If $A_1, \ldots, A_n \in \mathcal{F}$, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

About the Author

Fellow of the European Academy of Sciences (2003–), President of ECMI (the European Consortium for Mathematics in Industry) (1999–2001); Founder (1991) and President of the European Society for Mathematical and Theoretical Biology (2000–2002); Founder (1985) and Director of SASIAM: "School for Advanced Studies in Industrial and Applied Mathematics", Tecnopolis, Bari, Italy (1985–1991), Founder and Director of MIRIAM (Milan Research Centre for Industrial and Applied Mathematics) (1999–2005) and later of ADAMSS (Research Centre for Advanced Applied Mathematical and Statistical Sciences) of the University of Milano (2005–2007), Director of CIMAB (InterUniversity Centre for Mathematics Applied to Biology, Medicine, Environment, etc.) (2008–).

Cross References

►Foundations of Probability
►Imprecise Probability
►Measure Theory in Probability
►Philosophy of Probability
►Probability Theory: An Outline

References and Further Reading

Ash RB (1972) Real analysis and probability. Academic, London
Bauer H (1981) Probability theory and elements of measure theory. Academic, London
Billingsley P (1995) Probability and measure. Wiley, New York
Breiman L (1968) Probability. Addison–Wesley, Reading, MA
Chung KL (1974) A course in probability theory, 2nd edn. Academic, New York
De Finetti B (1974–1975) Theory of probability, vols 1 and 2. Wiley, London
Dudley RM (2002) Real analysis and probability. Cambridge Studies in Advanced Mathematics 74, Cambridge University Press, Cambridge
Fristedt B, Gray L (1997) A modern approach to probability theory. Birkhäuser, Boston
Kolmogorov AN (1956) Foundations of the theory of probability. Chelsea, New York
Métivier M (1968) Notions fondamentales de la théorie des probabilités., Dunod, Paris