F

F Distribution

Enrique M. Cabaña
Professor
Universidad de la República, Monterideo, Uruguay

George W. Snedecor (1882–1974) promoted the development of statistics in the USA by contributing to the foundation of a department of statistics at Iowa State University, reputed to be the first one in the country, and helping with his writings the diffusion and application of Sir Ronald A. Fisher's (1890–1962) work on the ►analysis of variance and covariance (Fisher 1950, 1971).

Snedecor named "F" the distribution of the ratio of independent estimates of the variance in a normal setting as a tribute to Fisher, and now that distribution is known as the *Snedecor F*. It is a continuous skew probability distribution with range $[0, +\infty)$, depending on two parameters denoted ν_1, ν_2 in the sequel. In statistical applications, ν_1 and ν_2 are positive integers.

Definition of the F Distribution

Let Y_1 and Y_2 be two independent random variables distributed as chi-square, with ν_1 and ν_2 degrees of freedom, respectively (abbreviated $Y_i \sim \chi^2_{\nu_i}$). The distribution of the ratio $Z = \dfrac{Y_1/\nu_1}{Y_2/\nu_2}$ is called the F distribution with ν_1 and ν_2 degrees of freedom.

The notation $Z \sim \mathrm{F}_{\nu_1,\nu_2}$ expresses that Z has the F distribution with ν_1 and ν_2 degrees of freedom. The role of ν_1 and ν_2 in this definition is often emphasized by saying that ν_1 are the degrees of freedom of the numerator, and ν_2 are the degrees of freedom of the denominator.

Remark 1 Let $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ denote the usual estimator of the variance σ^2 obtained from $X_1, X_2, \ldots, X_n$ i.i.d. Normal(μ, σ^2), $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Since s^2 is distributed as $\sigma^2\chi^2_{n-1}/(n-1)$ (this means that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$), then the ratio of two such independent estimators of the same variance has the F distribution that, for this reason, is often referred to as *the distribution of the variance ratio*.

This leads to an immediate application of the F distribution: Assume that s_1^2 and s_2^2 are the estimators of the variances of two normal populations with variances σ_1^2 and σ_2^2 respectively, computed from independent samples of sizes n_1 and n_2 respectively. Then the ratio $F = s_1^2/s_2^2$ is distributed as $\frac{\sigma_1^2}{\sigma_2^2} F_{n_1-1,n_2-1}$.

When $\sigma_1^2 = \sigma_2^2$, $F \sim F_{n_1-1,n_2-1}$. On the other hand, when $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$, F is expected to be respectively larger or smaller than a random variable with distribution F_{n_1-1,n_2-1}, and this suggests the use of F to test the null hypothesis $\sigma_1^2 = \sigma_2^2$: the null hypothesis is rejected when F is significantly large or small.

Remark 2 The joint probability density of the i.i.d. Normal$(0,1)$ variables $X_1, X_2, \ldots, X_n$ in $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|\boldsymbol{x}\|^2/2}$, with $\|\boldsymbol{x}\|^2 = \sum_{i=1}^n x_i^2$ equal to the Euclidean norm of $\boldsymbol{x}$. Since it depends on $\boldsymbol{x}$ only through its norm, it follows that the new coordinates $X_1^*, X_2^*, \ldots, X_n^*$ of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ in any orthonormal basis of $\mathbf{R}^n$ have the same joint density and therefore are i.i.d. Normal$(0,1)$.

If $\mathcal{R}$ denotes the subspace generated by the first p vectors of the new basis, and $\mathcal{R}^\perp$ is its orthogonal complement, then the angle Ψ of $\boldsymbol{X}$ with $\mathcal{R}^\perp$ has tangent

$$\tan\Psi = \sqrt{\frac{Y_1^*}{Y_2^*}}, \text{ with } Y_1^* = \sum_{i=1}^p \left(X_i^*\right)^2 \sim \chi^2_p, Y_2^* = \sum_{i=p+1}^n \left(X_i^*\right)^2 \sim \chi^2_{n-p},$$

and hence

$$Z := \frac{n-p}{p}\tan^2\Psi \sim \mathrm{F}_{p,n-p}.$$

This geometrical interpretation of the F distribution is closely related to the F test in the analysis of variance (Scheffe 1959).

Important applications of F distribution include: F test for testing equality of two population variances, F test for fit of regression models, and Scheffe's method of multiple comparison.

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

The Density and Distribution Function of F_{ν_1,ν_2}

Let $Y_i \sim \chi^2_{\nu_i}, i = 1, 2$. In order to compute the probability density of $Z = \frac{Y_1/\nu_1}{Y_2/\nu_2} \sim F_{\nu_1,\nu_2}$, introduce the random angle Ψ by $Z = \frac{\nu_2}{\nu_1}\tan^2\Psi$, and start by computing the distribution of $C = \cos^2\Psi = (1+\tan^2\Psi)^{-1} = \frac{\nu_2}{\nu_1 Z+\nu_2} = \frac{Y_2}{Y_1+Y_2}$:

$$\begin{aligned}\mathbf{P}\{C \le c\} &= \mathbf{P}\left\{Y_1 \ge \left(\tfrac{1}{c}-1\right)Y_2\right\}\\ &= \int_0^\infty f_2(y_2) \int_{(\frac{1}{c}-1)y_2}^\infty f_1(y_1)dy_1dy_2,\end{aligned}$$

where $f_i(t) = \frac{e^{-t/2}t^{\nu_i/2-1}}{2^{\nu_i/2}\Gamma(\nu_i/2)}$ is the density of the χ^2 distribution with ν_i degrees of freedom.

By replacing the analytical expressions of the χ^2 densities in

$$f_C(c) := \frac{d}{dc}\mathbf{P}\{C \le c\} = \int_0^\infty f_2(t)\frac{t}{c^2}f_1\left(\left(\tfrac{1}{c}-1\right)t\right)dt$$

one gets

$$\begin{aligned}f_C(c) &= \int_0^\infty \frac{e^{-t/2}t^{\nu_2/2-1}}{2^{\nu_2/2}\Gamma(\nu_2/2)}\\ &\quad\times\frac{t}{c^2}\frac{e^{-(1/c-1)t/2}(1/c-1)^{\nu_1/2-1}t^{\nu_1/2-1}}{2^{\nu_1/2}\Gamma(\nu_1/2)}dt\\ &= \frac{(1/c-1)^{\nu_1/2-1}}{c^2 2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\\ &\quad\times\int_0^\infty e^{-t/2c}t^{(\nu_1+\nu_2)/2-1}dt\\ &= c^{\nu_2/2-1}(1-c)^{\nu_1/2-1}\frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\\ &= \frac{c^{\nu_2/2-1}(1-c)^{\nu_1/2-1}}{B(\nu_1/2,\nu_2/2)}.\end{aligned}$$

This last expression, obtained by using the well-known relation $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x,y)}$ between Euler's Beta and Gamma functions, shows that $\cos^2\Psi$ has the ▶Beta distribution with parameters $(\nu_2/2, \nu_1/2)$ and consequently $\sin^2\Psi = 1-\cos^2\Psi$ has the Beta distribution with parameters $(\nu_1/2, \nu_2/2)$ and density $f_S(s) = \frac{s^{\nu_1/2-1}(1-s)^{\nu_2/2-1}}{B(\nu_1/2,\nu_2/2)}$.

The distribution function of $Z \sim F_{\nu_1,\nu_2}$ is

$$\begin{aligned}F_{\nu_1,\nu_2}(z) &= \mathbf{P}\{Z \le z\} = \mathbf{P}\left\{\tan^2\Psi \le \frac{\nu_1}{\nu_2}z\right\}\\ &= \mathbf{P}\left\{\cos^2\Psi \ge \frac{\nu_2}{\nu_1 z+\nu_2}\right\}\\ &= \mathbf{P}\left\{\sin^2\Psi \le \frac{\nu_1 z}{\nu_1 z+\nu_2}\right\} = \int_0^{\frac{\nu_1 z}{\nu_1 z+\nu_2}} f_S(s)ds\\ &= \frac{B\left(\frac{\nu_1 z}{\nu_1 z+\nu_2};\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)},\end{aligned}$$

where $B(t;a,b) = \int_0^t s^{a-1}(1-s)^{b-1}ds$ denotes the incomplete Beta function with parameters a, b evaluated in t. It may be noticed that the distribution function of F_{ν_1,ν_2} evaluated at z is the same as the distribution function of a $\text{Beta}(\nu_1/2, \nu_2/2)$ random variable evaluated at $\frac{\nu_1 z}{\nu_1 z+\nu_2}$.

By differentiating the c.d.f. the density of the F distribution is obtained:

$$\begin{aligned}f_{\nu_1,\nu_2}(z) &= \frac{\nu_1\nu_2}{(\nu_1 z+\nu_2)^2}f_S\left(\frac{\nu_1 z}{\nu_1 z+\nu_2}\right)\\ &= \frac{\sqrt{\nu_1^{\nu_1}\nu_2^{\nu_2}}}{zB(\nu_1/2,\nu_2/2)}\sqrt{\frac{z^{\nu_1}}{(\nu_1 z+\nu_2)^{\nu_1+\nu_2}}}.\end{aligned}$$

Figures 1 and 2 show graphs of f_{ν_1,ν_2} for several values of the parameters.

Some Properties of F Distribution

The moments of $Y \sim \chi^2_\nu$ are $\mathbf{E}Y^k = 2^k\frac{\Gamma(\frac{\nu}{2}+k)}{\Gamma(\frac{\nu}{2})}$ for $k > -\frac{\nu}{2}$, and infinite otherwise. Therefore, from the expression of $Z = \frac{\nu_2 Y_1}{\nu_1 Y_2}$ as the ratio of independent random variables $Y_i \sim \chi^2_{\nu_i}$ we get $\mathbf{E}Z^k = \left(\frac{\nu_2}{\nu_1}\right)^k \mathbf{E}Y_1^k\mathbf{E}Y_2^{-k} = \left(\frac{\nu_2^k}{\nu_1^k}\right)2^k\frac{\Gamma(\frac{\nu_1}{2}+k)}{\Gamma(\frac{\nu_1}{2})} \times 2^{-k}\frac{\Gamma(\frac{\nu_2}{2}-k)}{\Gamma(\frac{\nu_2}{2})} = \left(\frac{\nu_2}{\nu_1}\right)^k\frac{\Gamma(\frac{\nu_1}{2}+k)\Gamma(\frac{\nu_2}{2}-k)}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}$, provided $k < \nu_2/2$. If this last restriction does not hold, the moment is infinite. In particular,

$$\mathbf{E}Z = \frac{\nu_2}{\nu_2-2} \text{ for } \nu_2 > 2,$$

$$\mathbf{Var}Z = \frac{2(\nu_1+\nu_2-2)\nu_2^2}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2 > 4.$$

Other descriptive parameters are the mode $\frac{(\nu_1-2)\nu_2}{\nu_1(\nu_2+2)}$ for $\nu_1 > 2$, the skewness coefficient

$$\frac{2\sqrt{2}(2\nu_1+\nu_2-2)\sqrt{\nu_2-4}}{\sqrt{\nu_1(\nu_1+\nu_2-2)}(\nu_2-6)}$$

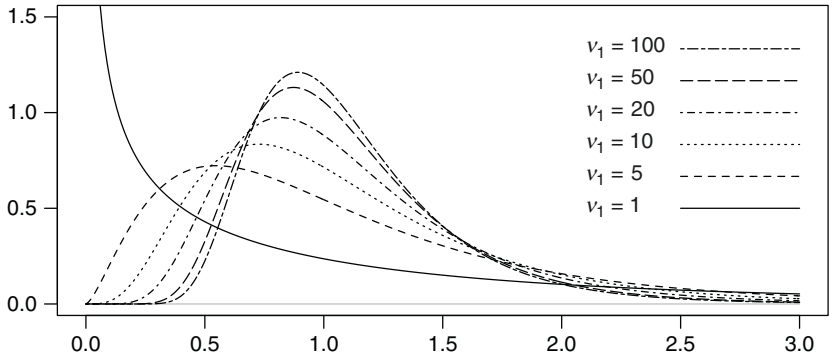
for $\nu_2 > 6$ and the kurtosis

$$\frac{3(\nu_2-4)\left(4(\nu_2-2)^2+\nu_1^2(\nu_2+10)+\nu_1(\nu_2-2)(\nu_2+10)\right)}{\nu_1(\nu_2-8)(\nu_2-6)(\nu_1+\nu_2-2)} - 3$$

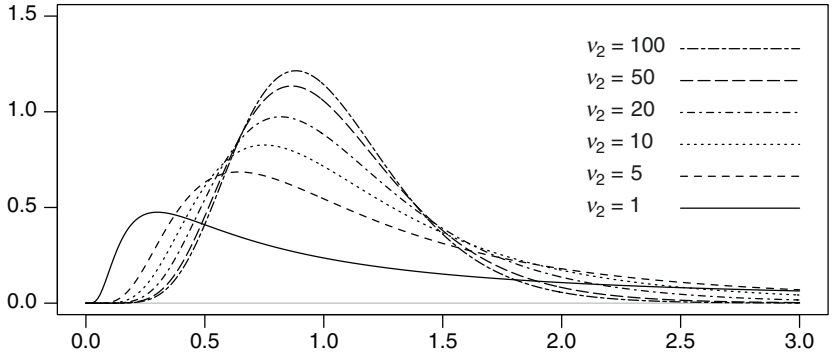
for $\nu_2 > 8$.

On Numerical Computations

There exist many tables of the F distribution, but the simpler way to obtain the numerical values of the density, the distribution function or its inverse, is to use the facilities provided by statistical software. The pictures here included and the numerical computations required by them were made by using the free software "R" (R Development Core Team 2008).



F Distribution. Fig. 1 Densities of F distribution with $\nu_1 = 1, 5, 10, 20, 50, 100$ and $\nu_2 = 20$



F Distribution. Fig. 2 Densities of F distribution with $\nu_1 = 20$ and $\nu_2 = 1, 5, 10, 20, 50, 100$

About the Author

Professor Enrique Cabaña is a member of the International Statistical Institute (elected in 1994) and founding President of the Latin American Regional Committee of The Bernoulli Society (1981–1983). He was Head of the Mathematical Centre of the Universidad de la República (1987–1990), Pro-Vice-Chancellor for Research of the same University (1999–2006) and Director of the Programme for the Development of Basic Sciences (PEDECIBA) of Uruguay (1997–2000). He has been teaching probability and statistics since 1958, mainly in Uruguay but also in Chile (1975–1977) and Venezuela (1978–1986) and his recent papers coauthored with Alejandra Cabaña develop several applications of L^2-techniques for the assessment of models based on transformations of stochastic processes.

Cross References

►Analysis of Variance
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Fisher RA (1950) Contributions to mathematical statistics. Wiley, New York
Fisher RA (1971) Collected Papers of Fisher RA. In Bennet JH (ed) The University of Adelaide
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
Scheffe H (1959) The analysis of variance. Wiley, New York

Factor Analysis and Latent Variable Modelling

David J. Bartholomew
Professor Emeritus of Statistics
London School of Economics and Political Science, London, UK

Background

Factor analysis was invented in 1904 by Professor Charles Spearman at University College London. Spearman was a psychologist and, for half century, factor analysis largely

remained the preserve of psychologists. Latent class analysis was developed by Paul Lazarsfeld at Columbia University in New York in the nineteen fifties and was designed for use by sociologists. Yet both of these techniques, and others like them, are essentially statistical and are now recognized as sharing a common conceptual framework which can be used in a wide variety of fields. It was not until the late nineteen thirties that statisticians, such as M. S. Bartlett, made serious contributions to the field.

Both factor analysis and latent class analysis are examples of the application of what would now be called latent variable models. Statistics deals with things that vary and in statistical theory such quantities are represented by random variables. In most fields these variables are observable and statistical analysis works with the observed values of such variables. But there are some important applications where we are interested in variables which cannot be observed. Such variables are called *latent* variables. In practice these often arise in the social sciences and include such things as human intelligence and political attitudes.

A latent variable model provides the link between the latent variables, which cannot be observed, and the manifest variables which can be observed. The purpose of the analysis is to determine how many latent variables are needed to explain the correlations between the manifest variable, to interpret them and, sometimes, to predict the values of the latent variables which have given rise to the manifest variables.

The Linear Factor Model

The basic idea behind factor analysis and other latent variable models is that of regression, or conditional expectation. We may regress each of the manifest (observed) variables on the set of latent variables (or factors). Thus, if we have p manifest variables, denoted by $x_1, x_2, \ldots x_p$ and q factors, denoted by $f_1, f_2, \ldots f_q$, the model may be written

$$x_i = \alpha_0 + \alpha_1 f_1 + \alpha_2 f_2 + \ldots \alpha_q f_q + e_i \qquad (i = 1, 2, \ldots, p) \quad (1)$$

where, without loss of generality, the fs are assumed to have zero means and unit standard deviations. The error term e_i is also assumed to have zero mean and standard deviation, σ_i. We often assume that all distributions are normal, in which case we refer to this as the *normal linear factor model*.

There are p linear equations here but the model cannot be fitted like the standard linear regression model (see ►Linear Regression Models) because the number of factors is unknown and the values of the f_s are not known, by definition. We, therefore, have to use indirect methods which depend on the fact that the correlation coefficients between the xs depend only on the αs and the σs. In practice, efficient computer programs are available which take care of the fitting.

The Latent Class Model

In a latent class model the manifest and latent variables are both categorical, often binary, instead of continuous. Thus the xs may consist of binary answers to a series of questions of the YES/NO variety. These are often coded 0 and 1 so that the manifest variable, x_i takes one of the two values 0 and 1. On the basis of these data we may wish to place individuals into one of several categories. In such cases the model is usually expressed in terms of probabilities. For example, for the ith manifest variable we may specify that

$$\Pr[x_i = 1] = \frac{\exp - \alpha_{i0} - \alpha_{i1} f}{1 + \exp - \alpha_{i0} - \alpha_{i1} f} \quad (2)$$

Because x_i is binary, the left hand side of the equation may also be written, $E(x_i)$ The reason for this somewhat strange expression is that probabilities necessarily lie between 0 and 1. The link with the linear expression of the previous section is made clearer if we write it in terms of the logit function. In that case we have

$$\text{logitE}[x_i] = \alpha_{i0} + \alpha_{i1} f. \quad (3)$$

This becomes a latent class model if we let f be a binary variable; this is a way of letting the probability on the left hand side of the equation take just two values.

Other Latent Variable Models

Prior to the last step, we actually had a latent profile model with one continuous factor, f. Further latent variables could have been added to the right hand side in exactly the same way as with the general linear factor model. Similarly, we could have had a continuous variable on the left hand side with discrete variables on the right hand side. Beyond this, in principle, there could be mixtures of continuous and/or categorical variables on both sides of the equation.

Much recent work is on what are called linear structural relations models where the interest is in the assumed (linear) relationships among the latent variables.

The Literature

There is an enormous literature on factor analysis and latent variable models, much of it very old and difficult to follow. This is not helped by the fact that much of the work has been published in books or journals appropriate to the disciplinary origins of the material and the level of mathematical expertise expected of the readers. One of the very few broad treatments from a statistical angle is given in:

Bartholomew, D.J. and Knott, M. (2011) *Latent Variable Models and Factor Analysis*, 2nd edition, Kendall's Library of Statistics 7, Arnold.

The references given there will lead on to many other aspects of the field, some of which have been touched on above.

About the Author

Past President of the Royal Statistical Society (1993–1995), David Bartholomew, was born in England in 1931. After undergraduate and postgraduate study at University College London, specializing in statistics, he worked for two years in the operational research branch of the National Coal Board. In 1957 he began his academic career at the University of Keele and then moved to the University College of Wales, Aberystwyth as lecturer, then senior lecturer in statistics. This was followed by appointment to a chair in statistics at the University of Kent in 1967. Six years later he moved to the London School of Economics as Professor of Statistics where he stayed until his retirement in 1996. During this time at the LSE he also served as Pro-Director for three years. He is a Fellow of the British Academy, a Member of the International Statistical Institute, and a Fellow of the Institute of Mathematical Statistics, He has authored, co-authored or edited about 20 books and about 120 research papers and articles, including the text *Latent Variable Models and Factor Analysis* (1987, Griffin; 2nd edition (with Martin Knott), Edward Arnold, 1999).

Cross References

►Correspondence Analysis
►Data Analysis
►Mixed Membership Models
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Principal Component Analysis
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Psychiatry, Statistics in
►Psychology, Statistics in
►Statistical Inference in Ecology
►Statistics: An Overview
►Structural Equation Models

References and Further Reading

Bartholomew DJ, Knott M (2011) Latent variable models and factor ananlysis: A unified approach, 3rd edn. Wiley Black well (in press)

Factorial Experiments

Kalina Trenevska Blagoeva
Associate Professor, Faculty of Economics
University "Ss. Cyril and Methodius", Skopje, Macedonia

Statistically designed experiments are an important tool in data analysis. The objective of such experimentation is to estimate the effect of each experimental factor on a response variable and to determine how the effect of one factor varies over the levels of other factors. Each measurement or observation is made on an item denoted as an *experimental unit*. Although some ideas of the several varying factors simultaneously appeared in England in the nineteenth century, the first major systematic discussion on factorial designs was given by Sir Ronald Fisher in his seminal book The Design of Experiments (Chap. 6) in 1935.

A *factorial experiment* is an experiment in which several factors (such as fertilizers or antibiotics) are applied to each experimental unit and each factor is applied at two, or more, levels. The levels may be quantitative (as with amounts of some ingredient) or qualitative (where the level refers to different varieties of wheat) but in either case are represented by elements of a finite set, usually by $0, 1, 2, \ldots, k_i - 1$ where the ith factor occurs at k_i levels. A factorial experiment in which t independent factors are tested, and in which the ith factor has k_i levels is labeled a $k_1 \times k_2 \times \cdots \times k_t$ factorial experiment. If $k_1 = k_2 = \cdots = k_t = k$, then the experiment is designated as a k^t *symmetrical factorial experiment*. An important feature of a complete factorial experiment is that all possible factor-level combinations are included in the design of the experiment.

Each controllable experimental variable, such as temperature or diet, in a factorial experiment is termed a *factor*. The *effect* of a factor on the response variable is the change in the average response between two experimental conditions. When the effect is computed as the difference between the average response at a given level of one factor and the overall average based on all of its levels after averaging over the levels of all the other factors, it is labeled the *main effect* of that factor. The difference in the effects of factors at different levels of other factors represents the *interaction* between factors. We can estimate the effect of each factor, independently of the others (the main effect), and the effect of the interaction of two (or more) factors (the interaction effect).

A factorial experiment allows for estimation of experimental error in two ways. The experiment can be replicated, or the sparsity-of-effects principle can often be

exploited. Replication is more common for small experiments and is a very reliable way of assessing experimental error. When the number of factors is large, replication of the design can become operationally difficult. In these cases, it is common to only run a single replicate of the design and to assume that factor interactions of more than a certain order (say, between three or more factors) are negligible. A *single replicate factorial design* is a factorial experiment in which every treatment combination appears precisely once in the design. As with any statistical experiment, the experimental runs in a factorial experiment should be randomized to reduce the impact that bias could have on the experimental result. In practice, this can be a large operational challenge.

Factorial experiments also can be run in block designs, where blocks refer to groups of experimental units or test runs (such as batches of raw material) that are more homogeneous within a block than between blocks. Combinations of the levels of two or more factors are defined as *treatment combinations*. If the number of treatment combinations is not too large, it is often possible to run the experiment in block designs in which some information is available within blocks on all factorial effects (i.e., main effects and interactions). Such effects are said to be partially confounded with blocks. However, factorial experiments with many factors, or with factors at many levels, involve large numbers of treatment combinations. The use of designs that require a number of replicates of each treatment combination then becomes impractical. To overcome this problem, designs using a single replicate are frequently used. Information on all or part of some of the factorial effects will consequently no longer be available from comparisons within blocks; these effects, or some components of them, will be said to be totally confounded with blocks.

Any number of factor levels can be used in a factorial experiment provided there is an adequate number of experimental units. However, the number of experimental runs required for three-level (or more) factorial designs will be considerably greater than for their two-level counterparts. Factorial designs are therefore less attractive if a researcher wishes to consider more than two levels. When the number of test runs required by a complete factorial experiment cannot be run due to time or cost constraints, a good alternative is to use fractional factorial experiments. These types of designs reduce the number of test runs.

Cross References

►Design of Experiments: A Pattern of Progress
►Interaction
►Research Designs
►Statistical Design of Experiments (DOE)

References and Further Reading

Box GEP, Hunter WG, Hunter JS (2005) Statistics for experimenters: design, innovation, and discovery, 2nd edn. Wiley, New York
Cox DR, Reid N (2000) The theory of the design of experiments. Chapman & Hall/CRC, London
Fisher R (1935) The design of experiments. Collier Macmillan, London
Mukherjee R, Wu CFJ (2006) A modern theory of factorial design. Springer Series in Statistics, Springer, New York
John A, Williams ER (1995) Cyclic and computer generated designs. Chapman & Hall, New York
Raktoe BL (1992) Factorial designs. Krieger Pub Co, Reprinted edition, Malabar, Florida

False Discovery Rate

John D. Storey
Associate Professor
Princeton University, Princeton, NJ, USA

Multiple Hypothesis Testing

In hypothesis testing, *statistical significance* is typically based on calculations involving ►p-values and Type I error rates. A p-value calculated from a single statistical hypothesis test can be used to determine whether there is statistically significant evidence against the null hypothesis. The upper threshold applied to the p-value in making this determination (often 5% in the scientific literature) determines the Type I error rate; i.e., the probability of making a Type I error when the null hypothesis is true. *Multiple hypothesis testing* is concerned with testing several statistical hypotheses simultaneously. Defining statistical significance is a more complex problem in this setting.

A longstanding definition of statistical significance for multiple hypothesis tests involves the probability of making one or more Type I errors among the family of hypothesis tests, called the *family-wise error rate*. However, there exist other well established formulations of statistical significance for multiple hypothesis tests. The Bayesian framework for classification naturally allows one to calculate the probability that each null hypothesis is true given the observed data (Efron et al. 2001; Storey 2003), and several frequentist definitions of multiple hypothesis testing significance are also well established (Shaffer 1995).

Soric (1989) proposed a framework for quantifying the statistical significance of multiple hypothesis tests based on

the proportion of Type I errors among all hypothesis tests called statistically significant. He called statistically significant hypothesis tests *discoveries* and proposed that one be concerned about the rate of false discoveries when testing multiple hypotheses. (A false discovery, Type I error, and false positive are all equivalent. Whereas the false positive rate and Type I error rate are equal, the false discovery rate is an entirely different quantity.) This false discovery rate is robust to the false positive paradox and is particularly useful in exploratory analyses, where one is more concerned with having mostly true findings among a set of statistically significant discoveries rather than guarding against one or more false positives. Benjamini and Hochberg (1995) provided the first implementation of false discovery rates with known operating characteristics. The idea of quantifying the rate of false discoveries is directly related to several preexisting ideas, such as Bayesian misclassification rates and the positive predictive value (Storey 2003).

Applications

In recent years, there has been a substantial increase in the size of data sets collected in a number of scientific fields, including genomics, astrophysics, neurobiology, and epidemiology. This has been due in part to an increase in computational abilities and the invention of various technologies, such as high-throughput biological devices. The analysis of high-dimensional data sets often involves performing simultaneous hypothesis tests on each of thousands or millions of measured variables. Classical multiple hypothesis testing methods utilizing the family-wise error rate were developed for performing just a few tests, where the goal is to guard against any single false positive occurring. However, in the high-dimensional setting, a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Hypothesis testing in high-dimensional genomics data sets has been particularly influential in increasing the popularity of false discovery rates (Storey and Tibshirani 2003). For example, DNA microarrays measure the expression levels of thousands of genes from a single biological sample. It is often the case that microarrays are applied to samples collected from two or more biological conditions, such as from multiple treatments or over a time course. A common goal in these studies is to identify genes that are differentially expressed among the biological conditions, which involves performing a hypothesis tests on each gene. In addition to incurring false positives, failing to identify truly differentially expressed genes is a major concern, leading to the false discovery rate being in widespread use in this area. In a notably different area of application, the false discovery rate was utilized in an astrophysics study to detect acoustic oscillations on the distribution of matter in present time, which had implications towards confirming the Big Bang theory of the creation of the universe (Lindsay et al. 2004). The body of scientific problems to which the false discovery rate is applied continues to grow.

Mathematical Definitions

Although multiple hypothesis testing with false discovery rates can be formulated in a very general sense (Storey 2007; Storey et al. 2007), it is useful to consider the simplified case where m hypothesis tests are performed with corresponding p-values $p_1, p_2, \ldots, p_m$. The typical procedure is to call hypotheses statistically significant whenever their corresponding p-values are less than or equal to some threshold t, where $0 < t \leq 1$. This threshold can be fixed or data-dependent, and the procedure for determining the threshold involves quantifying a desired error rate.

Table 1 describes the various outcomes that occur when applying this approach to determining which of the m hypothesis tests are statistically significant. Specifically, V is the number of Type I errors (equivalently false positives or false discoveries) and R is the total number of hypothesis tests called significant (equivalently total discoveries). The *family-wise error rate* (FWER) is defined to be

$$\text{FWER} = \mathbf{Pr}(V \geq 1),$$

and the *false discovery rate* (FDR) is usually defined to be (Benjamini and Hochberg 1995):

$$\text{FDR} = \mathbf{E}\left[\frac{V}{R \vee 1}\right] = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]\mathbf{Pr}(R > 0).$$

The effect of "$R \vee 1$" in the denominator of the first expectation is to set $V/R = 0$ when $R = 0$. As demonstrated by Benjamini and Hochberg (1995), the FDR offers a less strict

False Discovery Rate. Table 1 Possible outcomes from m hypothesis tests based on applying a significance threshold $t \in (0,1]$ to their corresponding p-values

	Not significant (p-value $> t$)	Significant (p-value $\leq t$)	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

multiple testing criterion than the FWER, allowing it to be more appropriate for some applications.

Two other false discovery rate definitions have been proposed in the literature, where the main difference is in how the $R = 0$ event is handled. These quantities are called the *positive false discovery rate* (pFDR) and the *marginal false discovery rate* (mFDR), and they are defined as follows (Storey 2003, 2007):

$$\text{pFDR} = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right],$$
$$\text{mFDR} = \frac{\mathbf{E}[V]}{\mathbf{E}[R]}.$$

Note that pFDR = mFDR = 1 whenever all null hypotheses are true, whereas FDR can always be made arbitrarily small because of the extra term $\mathbf{Pr}(R > 0)$. Some have pointed out that this extra term in the FDR definition may lead to misinterpreted results, and pFDR or mFDR offer more scientifically relevant values (Storey 2003; Zaykin et al. 1998), while others have argued that FDR is preferable because it allows for the traditional "strong control" criterion to be met (Benjamini and Hochberg 1995). All three quantities can be utilized in practice, and they are all similar when the number of hypothesis tests is particularly large.

Control and Estimation

There are two approaches to utilizing false discovery rates in a conservative manner when determining multiple testing significance. One approach is to fix the acceptable FDR level beforehand, and find a data-dependent thresholding rule so that the expected FDR of this rule over repeated studies is less than or equal to the pre-chosen level. This property is called *FDR control* (Benjamini and Hochberg 1995; Shaffer 1995). Another approach is to fix the p-value threshold at a particular value and then form a point estimate of the FDR whose expectation is greater than or equal to the true FDR at that particular threshold (Storey 2002). The latter approach has been useful in that it places multiple testing in the more standard context of point estimation, whereas the derivation of algorithms in the former approach may be less tractable. Indeed, it has been shown that the point estimation approach provides a comprehensive and unified framework (Storey et al. 2004).

For the first approach, (Benjamini and Hochberg 1995) proved that the algorithm below for determining a data based p-value threshold controls the FDR at level α when the p-values corresponding to true null hypotheses are independent and identically distributed (i.i.d.) Uniform(0,1). Other p-value threshold determining algorithms for FDR control have been subsequently studied (e.g., Benjamini and Liu 1999). This algorithm was originally introduced by Simes (1986) to control the FWER when all p-values are independent and all null hypotheses are true, although it also provides control of the FDR for any configuration of true and false null hypotheses.

FDR Controlling Algorithm (Simes, 1986; Benjamini and Hochberg, 1995)
1. Let $p_{(1)} \le \ldots \le p_{(m)}$ be the ordered, observed p-values.
2. Calculate $\widehat{k} = \max\{1 \le k \le m : p_{(k)} \le \alpha \cdot k/m\}$.
3. If $\widehat{k}$ exists, then reject null hypotheses corresponding to $p_{(1)} \le \ldots \le p_{(\widehat{k})}$. Otherwise, reject nothing.

To formulate the point estimation approach, let $\text{FDR}(t)$ denote the FDR when calling null hypotheses significant whenever $p_i \le t$, for $i = 1, 2, \ldots, m$. For $t \in (0, 1]$, we define the following ▶stochastic processes based on the notation in Table 1:

$$V(t) = \#\{\text{true null } p_i : p_i \le t\},$$
$$R(t) = \#\{p_i : p_i \le t\}.$$

In terms of these, we have

$$\text{FDR}(t) = \mathbf{E}\left[\frac{V(t)}{R(t) \vee 1}\right].$$

For fixed t, Storey (2002) provided a family of conservatively biased point estimates of $\text{FDR}(t)$:

$$\widehat{\text{FDR}}(t) = \frac{\widehat{m}_0(\lambda) \cdot t}{[R(t) \vee 1]}.$$

The term $\widehat{m}_0(\lambda)$ is an estimate of m_0, the number of true null hypotheses. This estimate depends on the tuning parameter λ, and it is defined as

$$\widehat{m}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda)}.$$

It can be shown that $\mathbf{E}[\widehat{m}_0(\lambda)] \ge m_0$ when the p-values corresponding to the true null hypotheses are Uniform(0,1) distributed (or stochastically greater). There is an inherent bias/variance trade-off in the choice of λ. In most cases, when λ gets smaller, the bias of $\widehat{m}_0(\lambda)$ gets larger, but the variance gets smaller. Therefore, λ can be chosen to try to balance this trade-off. Storey and Tibshirani (2003) provide an intuitive motivation for the $\widehat{m}_0(\lambda)$ estimator, as well as a method for smoothing over the $\widehat{m}_0(\lambda)$ to obtain a tuning parameter free $\widehat{m}_0$ estimator. Sometimes instead

of m_0, the quantity $\pi_0 = m_0/m$ is estimated, where simply $\widehat{\pi}_0(\lambda) = \widehat{m}_0(\lambda)/m$.

To motivate the overall estimator $\widehat{\text{FDR}}(t) = \widehat{m}_0(\lambda) \cdot t/[R(t) \vee 1]$, it may be noted that $\widehat{m}_0(\lambda) \cdot t \approx V(t)$ and $[R(t) \vee 1] \approx R(t)$. It has been shown under a variety of assumptions, including those of Benjamini and Hochberg (1995), that the desired property $\mathbf{E}\left[\widehat{\text{FDR}}(t)\right] \geq \text{FDR}(t)$ holds.

Storey et al. (2004) have shown that the two major approaches to false discovery rates can be unified through the estimator $\widehat{\text{FDR}}(t)$. Essentially, the original FDR controlling algorithm can be obtained by setting $\widehat{m}_0 = m$ and utilizing the p-value threshold $t^*_\alpha = \max\left\{t : \widehat{\text{FDR}}(t) \leq \alpha\right\}$. By allowing for the different estimators $\widehat{m}_0(\lambda)$, a family of FDR controlling procedures can be derived in this manner. In the asymptotic setting where the number of hypothesis tests m is large, it has also been shown that the two approaches are essentially equivalent.

Q-Values

In single hypothesis testing, it is common to report the p-value as a measure of significance. The "q-value" is the FDR based measure of significance that can be calculated simultaneously for multiple hypothesis tests. Initially it seems that the q-value should capture the FDR incurred when the significance threshold is set at the p-value itself, $\text{FDR}(p_i)$. However, unlike Type I error rates, the FDR is not necessarily strictly increasing with an increasing significance threshold. To accommodate this property, the q-value is defined to be the minimum FDR (or pFDR) at which the test is called significant (Storey 2002, 2003):

$$\text{q-}value(p_i) = \min_{t \geq p_i} \text{FDR}(t) \quad \text{or}$$
$$\text{q-}value(p_i) = \min_{t \geq p_i} \text{pFDR}(t).$$

To estimate this in practice, a simple plug-in estimate is formed, for example:

$$\widehat{\text{q}}\text{-}value(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t).$$

Various theoretical properties have been shown for these estimates under certain conditions, notably that the estimated q-values of the entire set of tests are simultaneously conservative as the number of hypothesis tests grows large (Storey et al. 2004).

Bayesian Derivation

The pFDR has been shown to be exactly equal to a Bayesian derived quantity measuring the probability that a significant test is a true null hypothesis. Suppose that (a) $H_i = 0$ or 1 according to whether the ith null hypothesis is true or not, (b) $H_i \overset{i.i.d.}{\sim} \text{Bernoulli}(1-\pi_0)$ so that $\mathbf{Pr}(H_i = 0) = \pi_0$ and $\mathbf{Pr}(H_i = 1) = 1-\pi_0$, and (c) $P_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot G_0 + H_i \cdot G_1$, where G_0 is the null distribution and G_1 is the alternative distribution. Storey (2001, 2003) showed that in this scenario

$$\begin{aligned}\text{pFDR}(t) &= \mathbf{E}\left[\frac{V(t)}{R(t)}\middle| R(t) > 0\right] \\ &= \mathbf{Pr}(H_i = 0|P_i \leq t),\end{aligned}$$

where $\mathbf{Pr}(H_i = 0|P_i \leq t)$ is the same for each i because of the i.i.d. assumptions. Under these modeling assumptions, it follows that $\text{q-value}(p_i) = \min_{t \geq p_i} \mathbf{Pr}(H_i = 0|P_i \leq t)$, which is a Bayesian analogue of the p-value – or rather a "Bayesian posterior Type I error rate." Related concepts were suggested as early as 1955 (Morton 1955). In this scenario, it also follows that $\text{pFDR}(t) = \int \mathbf{Pr}(H_i = 0|P_i = p_i) dG(p_i|p_i \leq t)$, where $G = \pi_0 G_0 + (1-\pi_0)G_1$. This connects the pFDR to the posterior error probability $\mathbf{Pr}(H_i = 0|P_i = p_i)$, making this latter quantity sometimes interpreted as a *local false discovery rate* (Efron et al. 2001; Storey 2001).

Dependence

Most of the existing procedures for utilizing false discovery rates in practice involve assumptions about the p-values being independent or weakly dependent. An area of current research is aimed at performing multiple hypothesis tests when there is dependence among the hypothesis tests, specifically at the level of the data collected for each test or the p-values calculated for each test. Recent proposals suggest modifying FDR controlling algorithms or extending their theoretical characterizations (Benjamini and Yekutieli 2001), modifying the null distribution utilized in calculating p-values (Devlin and Roeder 1999; Efron 2004), or accounting for dependence at the level of the originally observed data in the model fitting (Leek and Storey 2007, 2008).

About the Author

Dr. John D. Storey is Associate Professor of Genomics and Molecular Biology at Princeton University, with associated appointments in the Program in Applied and Computational Mathematics and the Department of Operations Research and Financial Engineering. He is currently an Associate editor of the *Annals of Applied Statistics* and *PLoS Genetics*, and has previously served on the editorial boards of *Biometrics*, *Biostatistics*, and *PLoS ONE*. He has published over 40 articles, including several highly cited articles on multiple hypothesis testing. He was recently recognized by Thomson Reuters as one of the top ten most cited mathematicians in the last decade.

Cross References

►Multiple Comparison
►Simes' Test in Multiple Testing

References and Further Reading

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 85:289–300
Benjamini Y, Liu W (1999) A step-down multiple hypothesis procedure that controls the false discovery rate under independence. J Stat Plann Infer 82:163–170
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188
Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104
Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96: 1151–1160
Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3:e161
Leek JT, Storey JD (2008) A general framework for multiple testing dependence. Proc Natl Acad Sci 105:18718–18723
Lindsay BG, Kettenring J, Siegmund DO (2004) A report on the future of statistics. Stat Sci 19:387–407
Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318
Shaffer J (1995) Multiple hypothesis testing. Ann Rev Psychol 46:561–584
Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754
Soric B (1989) Statistical discoveries and effect-size estimation. J Am Stat Assoc 84:608–610
Storey JD (2001) The positive false discovery rate: a Bayesian interpretation and the q-value. Technical Report 2001–2012, Department of Statistics, Stanford University
Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498
Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31:2013–2035
Storey JD (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Ser B 69: 347–368
Storey JD, Dai JY, Leek JT (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics 8:414–432
Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66:187–205
Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci 100:9440–9445
Zaykin DV, Young SS, Westfall PH (1998) Using the false discovery approach in the genetic dissection of complex traits: a response to weller et al. Genetics 150:1917–1918

Farmer Participatory Research Designs

Kaku Sagary Nokoe
Professor
University for Development Studies, Navrongo, Ghana

Introduction

Multilocation trials often follow classical on-station agronomic and breeding trials to test developed varieties under varying local conditions. Often these trials are imposed on farmer fields or set up as demonstration trials only to be viewed and ultimately to be adopted by rural poor farmers. On-farm trials involving the participation or use of farmers' fields have been applied in various studies, including: taungya and intercropping trials; mother–baby breeding trials aimed at selecting for specific traits in breed; augmented block designs (ABD) with emphasis on technology or selection of best or adaptable variety of crop under conditions of urgency and insufficient quantities of planting materials; crop livestock systems involving farmer management practices and animal preferences; and in the evaluation of adaptation and adoption of technologies. These trials are characterized by a high degree of variability within and between farmer fields (Mutsaers et al. 1997; Nokoe 1999; Odong 2002). Statistical issues of primary concern embrace the need for trial locations under farmer conditions, and why and how farmers may be involved to ensure acceptability and analyzability of selected designs.

From an intuitive but nonstatistical point of view, the involvement of all stakeholders (end user, researcher, community, donor) in the design of a trial, and the testing of trials under real-farm conditions utilizing options including maximum farmer management, is a sure way of enhancing adaptability and adoption. For breeders, there is a considerable advantage in time as duration from on-station to on-farm and then release is considerably shortened and results made more certain and output acceptable. On-farm research (OFR) has been variously classified, but generally could be grouped according to the level of farmer involvement. The class of interest in this entry is that involving the active participation of the farmer right from the design to the execution phases.

As an example, a participatory on-farm trial involving a crop–livestock system involved the following steps:

- Farmers and research institutions established formal collaborative linkages.
- Farmers discussed needs and prevailing cultural practices with researchers.

- Researchers and farmers evaluated intervention strategies.
- Statisticians guided selection of farmers and treatment allocations.
- Farmers randomly assigned inferior treatments allowed to change over time.

Arising from the above that is relevant from the point of view of designs is the fact that farmers are involved in the choice of treatments, blocks, and consequently the sampling or experimental designs. The implication is that block sizes are rarely of the same size or homogeneous, while considerable variability in some factors (such as variation in planting times) is common. In addition, it is common practice to have several standard controls (farmer practices), while in crop yield assessments the entire (not net or inner) plots are observed. Since, farmer differences are confounded in treatment, comparison of on-farm trials extend beyond differences in treatment effects. Mutsaers et al. (1997) point out that testing under farmer-field conditions and with their involvement provides a realistic assessment of the technologies or innovations under evaluation. Furthermore, the large number of farmers required is essential for capturing the expectedly high variation among farmer practices and sites. This large number should not be seen as a disadvantage, as the trade-off is the potentially high rate of adaptation and adoption of promising technologies (Nokoe 1999, 2000).

We shall consider general approaches aimed at the effective construction and analysis of farmer participatory designs.

The Design

Basic experimental principles (►randomization, replication, blocking, scope, and experimental error minimization) hold for participatory designs. The enforcement of these principles enables objectivity, estimation of standard errors, and effective comparison of treatment effects.

Identifying the Blocks

On-farm trials expectedly involve the use of block designs. The usual practice is to assume as blocks villages/communities (singly or cluster) and farm sites. This practice of assigning or identifying blocks is not very appropriate, though convenient. A preferred procedure would involve the use of statistical methods such as principal component and cluster analysis for constructing clusters. Classification variables must be relevant to the principal objectives of the study and would include socioeconomic, demographic, agronomic, and historical variables among others. It is emphasized that clusters are based on nontreatment characteristics that have the potential of influencing yields. As expected, the resulting blocks would not necessarily be contiguous, and that several farm sites from different villages, possibly distant-apart, may belong to the same block. Examples include classification of farm sites in an agroforestry and socioeconomic study and classification of farmers on the basis of soil type and cultural practices adopted. An alternative procedure, the post-model-based approach, would involve fitting a model (including discriminant functions) and then creating groups on the basis of limits of expected values, to which individuals may then be assigned.

Standard Block Structures

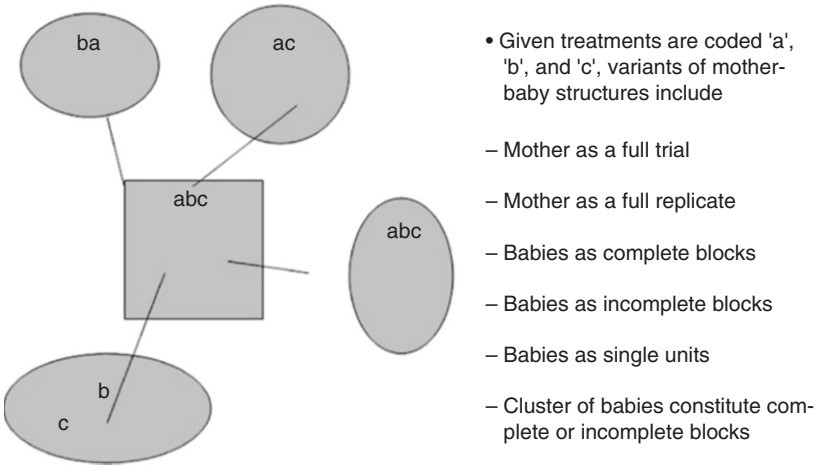
The block sizes may be equal (with each block receiving the same number of treatments) or balanced/unbalanced incomplete. Balanced complete block structure would generally imply treatments and pairs appear the same number of times in the experiment, and are present in all blocks. For incomplete block structures, several variants are available. These include alpha and cyclic incomplete block structures, and may be balanced with pairs of treatments appearing the same number of times in the experiment. Discussions on such designs are well documented (see, e.g., Cox and Reid 2000; and basic texts on designs). When block sizes are unequal, a fully balanced structure is not imaginable. Expectedly, in participatory designs, natural block structures are usually of the unequal and unbalanced type.

Augmented Block Structures

In recent times, block structures augmented with additional treatments, which are usually not replicated, are common in use. Augmented block designs (ABD) involve enlargement of blocks of a design (with treatments already assigned) to accommodate new treatments which appear usually once in the entire experiment (Federer 1955). The design allows for a wide range of technologies to be tested without necessarily straining resources or stifling farmers' interest [e.g., "1000 lines" in an on-farm situation is feasible]. Pinney (1991) provides an illustrated example for participatory agroforestry research.

Mother–Baby Structures

These block structures involve several blocks of varying sizes (Figure 1), with usually the larger sized one (the mother) having all or accommodating more treatments than the other blocks (baby/babies). The mother could also constitute a full trial with necessary replicates, and



Farmer Participatory Research Designs. Fig. 1 Treatment and Allocation to Blocks in Mother–Baby Trial

could represent an on-station or researcher-managed component of the trial. Babies may also be complete or incomplete blocks or as single experimental units with clusters of babies constituting blocks. The final block structure is arrived at after determining the number and sizes of the clusters.

Choice, Number of Treatments and Treatment Structures

The choice and number of treatments are made by consensus and on factors involved. The number need not be small as popularized in earlier works on OFR. However, the guiding principles are wide coverage (in the allocation of the treatments to the blocks), the willingness of the participating farmers, and availability of resources for the trial at the farmer/site level. The structure could be a single factor (say already packaged technologies or crop variety), factorial, or nested involving two or more factors. For factorial treatment structures, an alternative is the sequential or stepwise (step-up or step-down) arrangements. Stepwise allocation of treatments enable fewer number of treatments constituted from a number of factors, but the order of factor levels is crucial and needs to be well determined with all stakeholders. An example of factorial and corresponding stepwise is given in Table 1, where only four out of eight treatments are required for the stepwise (step-up and step-down options).

It is recommended that the decision to include a level of a factor, and at what step in the stepwise structure, must be determined jointly with all stakeholders. The sequence of factor levels affects and restricts the type of contrasts or comparisons that could reasonably be made (see, e.g., Mutsaers et al. 1991). It is also important the inability to estimate interactions is a major drawback of stepwise structures.

Observations and Measurements

Each farmer site represents different environments. Observations or measurements should therefore cover all other variables likely to account for the expectedly high variability. In crop trials with the response variable of interest being yield per hectare, there may be a need to include as many covariates (e.g., stand at establishment at plot level; soil depth, slope at field level; rainfall and labor cost at village level) and regressors (e.g., shade, labor size) as possible. In addition it is the gross and not the net plot that is observed. One basic advantage of such measurements from gross instead of usually uniform micro-net plot (as in on-station trials) is that yields are more likely to be realistic. Studies have shown that conversion of yield from micro to farmer level on the basis of small uniform on-station plot sizes considerably overestimates the real or farmer yield (Odulaja and Nokoe 1997).

Analytical Options

Several analytical options may be adapted from conventional ▶analysis of variance and regression modeling. The particular option to use will be influenced by the desired objective, the hypotheses of interests, and the nature of the response variables. The response variables may be continous (normally or non-normally distributed) or discrete

Farmer Participatory Research Designs. Table 1 Breakdown of treatment structure using factorial and stepwise procedures

Treatment code	Maize variety	Fertilizer use	Planting density	Step
Factorial 2^3				
1	Local	Local	Farmers own	1
2	Local	Local	Recommended	1
3	Local	Recommended	Farmers own	1
4	Local	Recommended	Recommended	1
5	Recommended	Local	Farmers own	1
6	Recommended	Local	Recommended	1
7	Recommended	Recommended	Farmers own	1
8	Recommended	Recommended	Recommended	1
Stepwise (step-up)				
1	Local	Local	Farmers own	1
2	Recommended	Local	Farmers own	2
3	Recommended	Recommended	Farmers own	3
4	Recommended	Recommended	Recommended	4
Stepwise (step-down)				
1	Recommended	Recommended	Recommended	1
2	Recommended	Local	Recommended	2
3	Recommended	Local	Farmers own	3
4	Local	Local	Farmers own	4

(counts, ordinal, binary outcomes) or nominal outcomes. The options are briefly outlined.

Simple Analysis (Adjusted by Local Controls)

Treatments in farmer/village blocks are adjusted by farmer/village's own treatment (control) either directly (e.g., difference in response) or used as covariate (especially in situations where the farmer site is not a block).

Stability Analysis

Adaptability (stability) analysis made popular by Hildebrand in the 1980s for genotype by environment interaction can readily be adapted. This is a regression-based method used initially in genotype by environment interaction studies, where the treatment response is fitted to site mean and the estimated slope examined (see Example of output in Table 2). In the example, Variety differences were small; while high yields were associated with Fertilizer input. In particular, tropical zea streak resistance (TZSR) with Fertilizer input should be recommended – stable yields (with slope close to 1) imply same yield may be expected across all sites for this treatment combination. It may also be noted that Local/300 associated with certain sites (i.e., high yields of local with 300L fertilizer expected at some locations).

An alternative and enhanced procedure is through the use of *biplot and AMMI (additive main effect multiplicative analysis) models* (See, e.g., text of Milliken and Johnson 1987 or notes on Matmodel Software by Gauch). AMMI involves the partitioning of variance. For data presented in a contingency table format, biplot may also be obtained via the method of ▶correspondence analysis which involves the partitioning of the chi-square.

Analysis of variance (ANOVA) with mixed models, where the error structure is adequately catered for and Contrasts, can be effectively used. In the mixed model scenario blocks, farmers, sites, etc., are usually treated as random effects being respectively a random sample from a large bulk. In particular, mixed modeling is most appropriate for augmented block designs (ABDs) and mother baby trials (MBTs) (Nokoe 1999), as it enables recovery of both inter-block and inter-treatment variation (Wolfinger et al. 1997). In ABDs, replicated lines are treated as fixed while non-replicated lines are considered as random.

Regression modeling is recommended for several situations where the design is not balanced, and/or when several auxiliary variables are available. These covariates or regressors, when included in the model, lead to considerable improvement in fit (Mutsaers et al. 1997; Carsky et al. 1998) and reduction of experimental error.

Categorical response variables, which constitute a substantial percentage of response variables, have not been modeled appropriately in several studies. These are better fitted by appropriate categorical modeling procedures, including the logistic and loglinear models (Bellon and Reeves 2002; Agresti 1990). An example of a trial with categorical responses is given in Table 3 (experimental setting and results) and Table 4 (partial data). In Table 4, the reader is to note the different types of response variables in the same data set – nominal (site history, indicating previous crop on farm site), ordinal (size of finger of plantains coded 1–3), and binary (size of plantain bunch and acceptability of product). It is important to indicate that a mixture of categorical and continuous variables is common in participatory design analysis.

Farmer Participatory Research Designs. Table 2 A simple regression based stability analysis for on-farm trial

Extracted output			
Fit Yield for variety, fertilizer, etc., as function of site mean (index). Comment on regression slopes.			
Variety	**Mean**	**slope, *b***	***P* > *t***
Local	2.464	1.02	0.0009
TZSR	2.831	0.97	0.0007
Fertilizer			
0	2.189	0.758	0.0003
300	3.107	1.240	< .0001
Treatment			
Local/ 0	1.984	0.697	0.0076
Local/300	2.945	1.356	0.0011
TZSR/0	2.393	0.818	0.0110
TZSR/300	3.269	1.124	0.0014
Conclusions: Variety differences low; high yields associated with Fertilizer input. TZSR with Fertilizer recommended			

Farmer Participatory Research Designs. Table 3 Partial data for plantain trial with categorical input and output variables

Zone	Farm Site	Treatment	Site Crop History	Weevil History	Bunch Size	Finger Size	Market Acceptability
East	1	A	Fallow	High	Small	1	Acceptable
	1	B	Fallow	High	Normal	1	Not
	1	C	Fallow	High	Normal	2	Not
	1	D	Plantain	Moderate	Normal	1	Acceptable
	2	A	Fallow	High	Small	2	Acceptable
	2	E	Fallow	Moderate	Small	1	Not
	2	D	Maize	High	Normal	2	Not
	2	C	Plantain	Low	Normal	3	Acceptable
West	8	C	Maize	Moderate	Small	2	Not
	8	B	Plantain	High	Normal	2	Not
	8	E	Plantain	Low	Small	3	Acceptable

Farmer Participatory Research Designs. Table 4 A summary of the plantain on-farm trial with categorical response variables

The experimental setting and variables	Models fitted and conclusions
Four Agro-Ecological zones across West/ Central/Eastern Africa involved Six plantain cultivars A, B, C, D, E, and F evaluated against local cultivar Only four cultivars allocated to farmers; each farmer provided four sites, one for each of the four allocated cultivars. Farmers local planted along with each cultivar or separately (as they wished) Previous field history recorded as: Fal – Fallow prior to trial Maz – Maize planted previous year Pla – Field already with plantain Weevil history (previous year when plantain had been cultivated): H – High Weevil population M – Medium Weevil population L – Low Medium population Response variables (assessed against local variety by farmers and chief) Bunch size: Binary - normal, small Finger size: Ordinal - 1 (least) to 4 (highest) Marketabilty: Binary - accept(able), not	*Logistic model* for bunch size, marketability as function of field history, weevil history, and other covariates, and regressors that may be available *Cumulative logit* model fitted to finger size Main conclusions: 1. Weevil history is the most important determinant of marketability 2. For bunch size, only finger size was significant at $p = 0.0434$. At 10% level of significance, study identified Weevil history and finger size as significant determinants of bunch size.

Conclusion

Participatory research designs can be planned and executed with scientifically verifiable results as outcome. It is emphasized that the understanding and active involvement of the farmer or end user are nontrivial considerations that need to be strictly adhered to for a successful research design aimed at addressing the needs of the usually poor rural farmer in developing countries. Adoption and adaptation of improved technologies are enhanced if all stakeholders are involved in the entire process. Blocking is a key ingredient in participatory designs, while the use of several regressors and covariates facilitate proper handling of the expectedly large variation between and within farm sites and farmer practices. The cocktail of analytical options requires adequate knowledge of statistical designs and the use of appropriate statistical software.

Acknowledgments

An earlier version of this entry had been presented at the regional sub-Saharan Africa Network of IBS meeting, and had benefited from inputs from workshop participants.

About the Author

Professor K S Nokoe, a Fellow of the Ghana Academy of Arts and Sciences, obtained his PhD from the University of British Columbia, and has over 30 years of teaching and research experience. He was Head of Department of Mathematical Sciences at the University of Agriculture in Nigeria (2002–2004), and the acting Vice-Chancellor of the University for Development Studies in Ghana (2007–2010). He had also served as biometrician in national and international research centres, published extensively in several peer-reviewed journals and supervised over 20 PhD and MSc students in Statistics, Statistical Ecology, Quantitative Entomology, Mensuration, Biometrics, and Statistical Computing among others. He is an Elected Member of the International Statistical Institute, Member of the International Association of Statistical Computing, Member of the International Biometric Society (IBS), and Founder of Sub-Saharan Africa Network (SUSAN) of IBS. He is the first recipient of the Rob Kempton Award of the International Biometric Society for "outstanding contribution to the development of biometry in developing countries."

Cross References

►Agriculture, Statistics in
►Research Designs
►Statistical Design of Experiments (DOE)

References and Further Reading

Agresti A (1990) Categorical data analysis. Wiley, New York
Carsky R, Nokoe S, Lagoke STO, Kim SK (1998) Maize yield determinants in farmer-managed trials in the Nigerian Northern Guinea Savanna. Experiment Agric 34:407–422

Cox DR, Reid N (2000) The theory of the design of experiments. Chapman & Hall/CRC, London
Federer WT (1955) Experimental design. MacMillan, New York
FCNS (2001) Food consumption and nutrition survey (Nigeria). Preliminary Report, IITA/USAID/UNICEF/FGN
Milliken GA, Dallas EJ (1987) Analysis of messy data vol. 2 nonreplicated experiments. Von Nostrand Reinhold, New York
Mutsaers HJW, Weber GK, Walker P (eds) (1991) On farm research in theory and practice. IITA Ibadan, Nigeria
Mutsaers HJW, Weber GK, Walker P, Fischer NM (1997) A field guide for on-farm experimentation. IITA/CTA/ISNAR, Ibadan
Nokoe S (1999) On farm trials: preventive and surgical approaches. J Trop For Resources (Special edition) 15(2):93–103
Nokoe S (2000) Biometric issues in agronomy: further on-station and on-farm designs and analyses. In: Akoroda MO (ed) Agronomy in Nigeria. Department of Agronomy, University of Ibadan, Nigeria, pp 35–42
Odong TL (2002) Assessment of variability in on-farm trial: a Uganda case. Unpublished dissertation, University of Natal MSc (Biometry), 103 pp
Odulaja A, Nokoe S (1997) Conversion of yield data from experimental plot to larger plot units. Discov Innovat 9: 137–141
Pinney A (1991) Farmer augmented designs for participatory agroforestry research. In: Patel MS, Nokoe S (eds) Biometry for development. ICIPE Science Press, Nairobi, pp 39–50
Wolfinger RD, Federer WT, Cordero-Brana O (1997) Recovering information in augmented designs using SAS PROC GLM and PROC MIXED. Agron J 89:856–859

Federal Statistics in the United States, Some Challenges

Edward J. Spar
Executive Director
Council of Professional Associations on Federal Statistics, Alexandria, VA, USA

The Current Status

Are the federal statistical agencies in the United States meeting their mandates? Overall, the answer is yes. Surveys that are required for policy purposes in health, education, labor, and other areas are being conducted with well-tested statistical designs that so far have reasonable margins of error. The decennial census, even with an under and over count meets the needs of the Constitution and thousands of federal, state and local data users. Measures, including labor force data, gross domestic product, the system of national accounts, health, education, and income estimates are excellently covered by the federal statistical agencies. Estimates of the population are reasonable even in situations where high immigration and/or internal migration, that have disproportionate influence, take place. The agencies are very sensitive of the need to maintain the confidentiality of respondents. Based on the above, it sounds as if the federal statistical system in the United States is healthy and on track; yet what about the future?

Ongoing and Upcoming Issues

Many new problems are facing the statistical agencies in the United States, and it will take enormous effort to solve them. Indeed, the agencies are fully aware of them and understand that there is a need for innovative thinking. An example of the type of innovation that has already taken place is the U.S. Census Bureau's American Community Survey. This is a replacement for the decennial census long form, and at the same time as an ongoing annual survey of about three million housing units, is unique. The ability to have data available every year for national, state, and local geographies is an important step for a dynamic country such as the United States. Another innovative set of data is the U.S. Census Bureau's Longitudinal Employer–Household Employer Dynamic. Using a mathematical model to insure non-disclosure, data are available for detailed employment statistics at very local geographic levels.

An issue that is becoming critical and is being looked at closely is the declining response rates in key federal surveys that measure employment, income, consumer expenditures, health and education, for example. Surveys that were achieving rates in the middle to high 90% range are now attaining response rates well below that. Clearly, the continuing decline in non-response will have serious effects on the usefulness of data collected. Either the statistical error will become so high so as to make the estimates of limited value, or, perhaps even worse, with biases due to non-response, the data may lose most of its value. Clearly the statistical agencies are aware of the problem and much research is being conducted to determine if address-listing techniques, for example, can be of use in conjunction with telephone interviewing. Some work has been accomplished in the areas of non-response and statistical and non-statistical biases but much more is required. The issue of conducting telephone surveys (see ►Telephone Sampling: Frames and Selection Techniques), given the elimination of land lines on the part of households and their turning to the increasing use of cell phones, must be addressed.

The data retrieval world has been transformed by the world-wide-web. The concept of charging for governmental data is no longer realistic given the assumption on the part of users that all data should be free on-line. Also, search engines such as Google have enabled users to retrieve diverse information as an integrated "package."

However, data integration across federal statistical agencies is for the most part limited. For example, there is no way to analyze and reconcile the many different measures of income between and sometimes even within an agency. Each agency creates its own web site and its own data dissemination system with little or no regard to the fact that the user has to go to a over a dozen sites and learn a dozen approaches to data retrieval to get a complete review of the socio-economic data of the United States. Indeed, if the user wants to integrate the data, it's much easier, but more expensive to go to a private sector vendor to do the work for you. At a time when the web is there for the specific purpose to retrieve information easily, freely, and comprehensively, this approach is outdated. The time has come for an integration of data processing and retrieval systems. This should be accomplished even though the structure of the federal statistical system in the United States is highly decentralized. The concept of a single system in the case of the United States, and probably most countries, is misleading. In reality what you have is a confederation of agencies for the most part reporting to different jurisdictions and quite independent of each other. In the United States, there is very limited administrative record data sharing and with separate Internet sites mentioned above, little integration of tabulated data sets. Each agency has its own budget and except for the purchasing of surveys from the U.S. Census Bureau, little in the way of financial interaction. Unfortunately, because of this lack of centralization, the agencies don't have great influence with the Congress. (This is not the case during the decennial census cycle where the apportionment of Congressional seats can impact a member of the House of Representatives. Other data series such as employment and inflation are also closely looked at.) This lack of influence can be a problem for an agency that each year must request funding for its programs. Would a centralized single agency help solve this? An agency large enough to be noticed by Congress and the Administration as being critical to the overall health of the nation would have a better opportunity of receiving the needed resources to implement innovative statistical techniques.

To perhaps overstate the case, the days of taking censuses and surveys may soon be coming to an end. We may be at the crossroads of having to rely, for the very most part, on administrative records. The use of administrative record data brings up issues of confidentiality on the part of agencies and the sensitivity to the privacy needs of the public. Yet these data may have to become the basis for measuring health, education, employment, expenditure, transportation, energy use and many more statistical needs on the part of the federal government. Using administrative data will call for public/private sector coordinated analyses and the allocation of talent and research dollars. If the use of administrative data becomes the norm, it is not too outré to see a time when no data will be real – put another way, they will be modeled estimates based on the original data series. As previously mentioned, we already see such a transformation in the U.S. Census Bureau's Longitudinal Employer-Household Employer Dynamic program produced at the local level. Indeed, once the block group level data from the American Community Survey are analyzed, we may also see some move in the same direction.

Over the next few years, much of the senior staffs of statistical agencies will be of retirement age. At the same time, it's difficult for agencies to hire new personnel and hold on to talented statisticians and economists that have entered the federal statistical system. The private sector offers both higher salaries and the opportunity to diversify. Indeed, the problem of "stove-piping" within statistical agencies, where talented people are expected to stay in one place for an overly extended period of time, is counter-productive. There is a need to develop a system whereby people can move not only within an agency, but also across agencies. Such a system of diverse training will be required so that personnel can develop the skills needed to address the concerns that have been mentioned in this review.

The challenges reviewed above are only the beginning. In order to properly measure the effects of the current and probably future economic crises in the United States, timely and relevant data are needed for those who have to make informed decisions affecting all Americans.

Cross References

►Census
►Nonresponse in Surveys
►Telephone Sampling: Frames and Selection Techniques

Fiducial Inference

Jan Hannig[1], Hari Iyer[2], Thomas C. M. Lee[3]
[1]Associate Professor
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Professor
Colorado State University, Fort Collins, CO, USA
[3]Professor
The University of California at Davis, Davis, CA, USA

Introduction

The origin of Generalized Fiducial Inference can be traced back to R. A. Fisher (Fisher 1930, 1933, 1935) who introduced the concept of a fiducial distribution for a

parameter, and proposed the use of this fiducial distribution, in place of the Bayesian posterior distribution, for interval estimation of this parameter. In the case of a one-parameter family of distributions, Fisher gave the following definition for a fiducial density $f(\theta|x)$ of the parameter based on a single observation x for the case where the cdf $F(x|\theta)$ is a monotonic decreasing function of θ:

$$f(\theta/x) = -\frac{\partial F(x|\theta)}{\partial \theta}. \tag{1}$$

In simple situations, especially in one parameter families of distributions, Fisher's fiducial intervals turned out to coincide with classical confidence intervals. For multiparameter families of distributions, the fiducial approach led to confidence sets whose frequentist coverage probabilities were close to the claimed confidence levels but they were not exact in the frequentist sense. Fisher's proposal led to major discussions among the prominent statisticians of the 1930's, 40's and 50's (e.g., Dempster 1966, 1968; Fraser 1961a, b, 1966, 1968; Jeffreys 1940; Lindley 1958; Stevens 1950). Many of these discussions focused on the nonexactness of the confidence sets and also nonuniqueness of fiducial distributions. The latter part of the 20th century has seen only a handful of publications Barnard (1995); Dawid and Stone (1982); Dawid et al. (1973); Salome (1998); Wilkinson (1977) as the fiducial approach fell into disfavor and became a topic of historical interest only.

Recently, the work of Tsui and Weerahandi (1989, 1991) and Weerahandi (1993, 1994, 1995) on generalized confidence intervals and the work of Chiang (2001) on the *surrogate variable method* for obtaining confidence intervals for variance components, led to the realization that there was a connection between these new procedures and fiducial inference. This realization evolved through a series of works (Hannig 2009b; Hannig et al. 2006b; Iyer and Patterson 2002; Iyer et al. 2004; Patterson et al. 2004). The strengths and limitations of the fiducial approach is becoming to be better understood, see, especially, Hannig (2009b). In particular, the asymptotic exactness of fiducial confidence sets, under fairly general conditions, was established in Hannig et al. (2006b); Hannig (2009a,b).

Subsequently Hannig et al. (2003); Iyer et al. (2004); McNally et al. (2003); Wang and Iyer (2005, 2006a,b) applied this fiducial approach to derive confidence procedures in many important practical problems. Hannig (2009b) extended the initial ideas and proposed a Generalized Fiducial Inference procedure that could be applied to arbitrary classes of models, both parametric and nonparametric, both continuous and discrete. These applications include Bioequivalence Hannig et al. (2006a), Variance Components Lidong et al. (2008), Problems of Metrology Hannig et al. (2007, 2003); Wang and Iyer (2005, 2006a, b), Interlaboratory Experiments and International Key Comparison Experiments Iyer et al. (2004), Maximum Mean of a Multivariate Normal Distribution Wandler and Hannig (2009), Mixture of a Normal and Cauchy Glagovskiy (2006), Wavelet Regression Hannig and Lee (2009), ▶Logistic Regression and LD_{50} Lidong et al. (2009). Recently, other authors have also contributed to research on fiducial methods and related topics (e.g., Berger and Sun 2008; Wang 2000; Xu and Li 2006).

Generalized Fiducial Distribution

The idea underlying Generalized Fiducial Inference comes from an extended application of Fisher's fiducial argument, which is briefly described as follows. Generalized Fiducial Inference begins with expressing the relationship between the data, $\mathbf{X}$, and the parameters, $\boldsymbol{\theta}$, as

$$\mathbf{X} = G(\boldsymbol{\theta}, \mathbf{U}), \tag{2}$$

where $G(\cdot,\cdot)$ is termed structural equation, and $\mathbf{U}$ is the random component of the structural equation whose distribution is completely known. The data $\mathbf{X}$ are assumed to be created by generating a random variable $\mathbf{U}$ and plugging it into the structural equation (2).

For simplicity, this section only considers the case where the structural relation (2) can be inverted and the inverse $G^{-1}(\cdot,\cdot)$ always exists. Thus, for any observed $\mathbf{x}$ and for any arbitrary $\mathbf{u}$, $\boldsymbol{\theta}$ is obtained as $\boldsymbol{\theta} = G^{-1}(\mathbf{x}, \mathbf{u})$. Fisher's *Fiducial Argument* leads one to define the fiducial distribution for $\boldsymbol{\theta}$ as the distribution of $G^{-1}(\mathbf{x}, \mathbf{U}^\star)$ where $\mathbf{U}^\star$ is an independent copy of $\mathbf{U}$. Equivalently, a sample from the fiducial distribution of $\boldsymbol{\theta}$ can be obtained by generating $\mathbf{U}_i^\star$, $i = 1, \ldots, N$ and using $\boldsymbol{\theta}_i = G^{-1}\left(\mathbf{x}, \mathbf{U}_i^\star\right)$. Estimates and confidence intervals for $\boldsymbol{\theta}$ can be obtained based on this sample.

Hannig (2009b) has generalized this to situations where G is not invertible. The resulting fiducial distribution is called a Generalized Fiducial Distribution. To explain the idea we begin with Eq. 2 but do not assume that G is invertible with respect to $\boldsymbol{\theta}$. The inverse $G^{-1}(\cdot,\cdot)$ may not exist for one of the following two reasons: for any particular $\mathbf{u}$, either there is no $\boldsymbol{\theta}$ satisfying (2), or there is more than one $\boldsymbol{\theta}$ satisfying (2).

For the first situation, Hannig (2009b) suggests removing the offending values of $\mathbf{u}$ from the sample space and then re-normalizing the probabilities. Such an approach has also been used by Fraser (1968) in his work on structural inference. Specifically, we generate $\mathbf{u}$ conditional on the event that the inverse $G^{-1}(\cdot,\cdot)$ exists. The rationale

for this choice is that we know our data $\mathbf{x}$ were generated with some $\boldsymbol{\theta}_0$ and $\mathbf{u}_0$, which implies there is at least one solution $\boldsymbol{\theta}_0$ satisfying (2) when the "true" $\mathbf{u}_0$ is considered. Therefore, we restrict our attention to only those values of $\mathbf{u}$ for which $G^{-1}(\cdot,\cdot)$ exists. However, this set has probability zero in many practical situations leading to non-uniqueness due to the Borel paradox (Casella and Berger 2002, Section 4.9.3). The Borel paradox is the fact that when conditioning on an event of probability zero, one can obtain any answer.

The second situation can be dealt with either by selecting one of the solutions or by the use of the mechanics underlying Dempster-Shafer calculus Dempster (2008). In any case, Hannig (2009a) proved that this non-uniqueness disappears asymptotically under very general assumptions.

Hannig (2009b) proposes the following formal definition of the generalized fiducial recipe. Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with a distribution indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Recall that the data generating mechanism for $\mathbb{X}$ is expressed by (2) where G is a jointly measurable function and $\mathbf{U}$ is a random variable or vector with a completely known distribution independent of any parameters. We define for any measurable set $A \in \mathbb{R}^n$ a set-valued function

$$Q(A,\mathbf{u}) = \{\boldsymbol{\theta} : G(\boldsymbol{\theta},\mathbf{u}) \in A\}. \tag{3}$$

The function $Q(A,\mathbf{u})$ is the generalized inverse of the function G. Assume $Q(A,\mathbf{u})$ is a measurable function of $\mathbf{u}$.

Suppose that a data set was generated using (2) and it has been observed that the sample value $\mathbf{x} \in A$. Clearly the values of $\boldsymbol{\theta}$ and $\mathbf{u}$ used to generate the observed data will satisfy $G(\boldsymbol{\theta},\mathbf{u}) \in A$. This leads to the following definition of a generalized fiducial distribution for $\boldsymbol{\theta}$:

$$Q(A,\mathbf{U}^\star) \mid \{Q(A,\mathbf{U}^\star) \neq \emptyset\}, \tag{4}$$

where $\mathbf{U}^\star$ is an independent copy of $\mathbf{U}$.

The object defined in (4) is a random set of parameters (such as an interval or a polygon) with distribution conditioned on the set being nonempty. It is well-defined provided that $P(Q(A,\mathbf{U}^\star) \neq \emptyset) > 0$. Otherwise additional care needs to be taken to interpret this distribution (c.f., Hannig 2009b). In applications, one can define a distribution on the parameter space by selecting one point out of $Q(A,\mathbf{U}^\star)$.

Examples

The following examples provide simple illustrations of the definition of a generalized fiducial distribution.

Example 1 Suppose $\mathbf{U} = (U_1,U_2)$ where U_i are i.i.d. $N(0,1)$ and $\mathbb{X} = (X_1,X_2) = G(\mu,\mathbf{U}) = (\mu+U_1,\mu+U_2)$ for some $\mu \in \mathbb{R}$. So X_i are iid $N(\mu,1)$. Given a realization $\mathbf{x} = (x_1,x_2)$ of $\mathbb{X}$, the set-valued function Q maps $\mathbf{u} = (u_1,u_2) \in R^2$ to a subset of $\mathbb{R}$ and is given by

$$Q(\mathbf{x},\mathbf{u}) = \begin{cases} \{x_1 - e_1\} & \text{if} \quad x_1 - x_2 = u_1 - u_2, \\ \emptyset & \text{if} \quad x_1 - x_2 \neq u_1 - u_2. \end{cases}$$

By definition, a generalized fiducial distribution for μ is the distribution of $x_1 - U_1^\star$ conditional on $U_1^\star - U_2^\star = x_1 - x_2$ where $\mathbf{U}^\star = \left(U_1^\star, U_2^\star\right)$ is an independent copy of $\mathbf{U}$. Hence a generalized fiducial distribution for μ is $N(\bar{x},1/2)$ where $\bar{x} = (x_1+x_2)/2$.

Example 2 Suppose $\mathbf{U} = (U_1,\ldots,U_n)$ is a vector of i.i.d. uniform $(0,1)$ random variables U_i. Let $p \in [0,1]$. Let $X = (X_1,\ldots,X_n)$ be defined by $X_i = I(U_i < p)$. So X_i are iid Bernoulli random variables with success probability p. Suppose $x = (x_1,\ldots,x_n)$ is a realization of X. Let $s = \sum_{i=1}^n x_i$ be the observed number of 1's. The mapping $Q : [0,1]^n \to [0,1]$ is given by

$$Q(x,\mathbf{u}) = \begin{cases} [0,u_{1:n}] & \text{if} \quad s = 0, \\ (u_{1:n},1] & \text{if} \quad s = n, \\ (u_{s:n},u_{s+1:n}] & \text{if} \quad s = 1,\ldots,n-1 \text{ and} \\ & \sum_{i=1}^n I(x_i = 1)I(u_i \leq u_{s:n}) = s, \\ \emptyset & \text{otherwise.} \end{cases}$$

Here $u_{r:n}$ denotes the rth order statistic among $u_1,\ldots,u_n$. So a generalized fiducial distribution for p is given by the distribution of $Q(x,\mathbf{U}^\star)$ conditional on the event $Q(x,\mathbf{U}^\star) \neq \emptyset$. By the exchangeability of $U_1^\star,\ldots,U_n^\star$ it follows that the stated conditional distribution of $Q(x,\mathbf{U}^\star)$ is the same as the distribution of $\left[0,U_{1:n}^\star\right]$ when $s = 0$, $\left(U_{s:n}^\star,U_{s+1:n}^\star\right]$ for $0 < s < n$, and $\left(U_{n:n}^\star,1\right]$ for $s = n$.

Next, we present a general recipe that is useful in many practical situations.

Example 3 Let us assume that the observations $X_1,\ldots,X_n$ are i.i.d. univariate with distribution function $F(x,\xi)$ and density $f(x,\xi)$, where ξ is a p-dimensional parameter. Denote the generalized inverse of the distribution function by $F^{-1}(\xi,u)$ and use the structural equation

$$X_i = F^{-1}(\xi,U_i) \qquad \text{for} \quad i = 1,\ldots,n. \tag{5}$$

If all the partial derivatives of $F(x,\xi)$ with respect to ξ are continuous and the Jacobian

$$\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(x_{i_1},\xi),\ldots,F(x_{i_p},\xi)\right)\right) \neq 0$$

for al distinct $x_1, \ldots, x_p$, then Hannig (2009b) shows that the generalized fiducial distribution (4) is

$$r(\xi) = \frac{f_{\mathbf{X}}(\mathbf{x}|\xi)J(\mathbf{x},\xi)}{\int_{\Xi} f_{\mathbf{X}}(\mathbf{x}|\xi')J(\mathbf{x},\xi')\,d\xi'}, \tag{6}$$

where

$$J(\mathbf{x},\xi) = \sum_{\mathbf{i}=(i_1,\ldots,i_p)} \left| \frac{\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(x_{i_1},\xi),\ldots,F(x_{i_p},\xi)\right)\right)}{f(x_{i_1},\xi)\cdots f(x_{i,p},\xi)} \right|. \tag{7}$$

This provides a form of generalized fiducial distribution that is usable in many practical applications, see many of the papers mentioned in introduction. Moreover, if $n = p = 1$ (6) and (7) simplify to the Fisher's original definition (1).

Equation 6 is visually similar to Bayes posterior. However, the role of the prior is taken by the function $J(\mathbf{x}, \xi)$. Thus unless $J(\mathbf{x}, \xi) = k(\mathbf{x})l(\xi)$ where k and l are measurable functions, the generalized fiducial distribution is not a posterior distribution with respect to any prior. A classical example of such a situation is in Grundy (1956).

Moreover, $\binom{n}{p}^{-1} J(\mathbf{x}, \xi)$ is a ►U-statistic and therefore it often converges a.s. to

$$\pi_{\xi_0}(\xi) = E_{\xi_0}\left| \frac{\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(X_1,\xi),\ldots,F(X_p,\xi)\right)\right)}{f(X_1,\xi)\cdots f(X_p,\xi)} \right|$$

At first glance $\pi_{\xi_0}(\xi)$ could be viewed as an interesting non-subjective prior. Unfortunately, this prior is not usable in practice, because the expectation in the definition of $\pi(\xi)$ is taken with respect to the true parameter ξ_0 which is unknown. However, since $\binom{n}{p}^{-1} J(\mathbf{x}, \xi)$ is an estimator of $\pi_{\xi_0}(\xi)$, the generalized fiducial distribution (6) could be interpreted as an empirical Bayes posterior.

Acknowledgments

The authors' research was supported in part by the National Science Foundation under Grant No. 0707037.

About the Authors

Jan Hannig is the Associate Professor of Statistics and Operations Research at The University of North Carolina at Chapel Hill. He is an Associate Editor of *Electronic Journal of Statistics* and an elected member of International Statistical Institute (ISA).

Hari Iyer is a Research Fellow at Caterpillar Inc. He is also a Professor of Statistics at Colorado State University.

Thomas C. M. Lee is a Professor of Statistics at the University of California, Davis. Before joining UC Davis, he had held regular and visiting faculty positions at University of Chicago, Chinese University of Hong Kong, Colorado State University and Harvard University. He is an elected Fellow of the American Statistical Association, and an elected Senior Member of the IEEE. He has published more than 50 papers in refereed journals and conference proceedings. Currently he is an Associate editor for *Bernoulli*, *Journal of Computational and Graphical Statistics*, and *Statistica Sinica*.

Cross References

►Behrens–Fisher Problem
►Confidence Distributions
►Statistical Inference: An Overview

References and Further Reading

Barnard GA (1995) Pivotal models and the fiducial argument. Int Stat Rev 63:309–323
Berger JO, Sun D (2008) Objective priors for the bivariate normal model. Ann Stat 36:963–982
Casella G, Berger RL (2002) Statistical inference, 2nd edn. Wadsworth and Brooks/Cole, Pacific Grove, CA
Chiang A (2001) A simple general method for constructing confidence intervals for functions of variance components. Technometrics 43:356–367
Dawid AP, Stone M (1982) The functional-model basis of fiducial inference (with discussion). Ann Stat 10:1054–1074
Dawid AP, Stone M, Zidek JV (1973) Marginalization paradoxes in Bayesian and structural inference (with discussion). J R Stat Soc Ser B 35:189–233
Dempster AP (1966) New methods for reasoning towards posterior distributions based on sample data. Ann Math Stat 37:355–374
Dempster AP (1968) A generalization of Bayesian inference (with discussion). J R Stat Soc Ser B 30:205–247
Dempster AP (2008) The Dempster-Shafer calculus for statisticians. Int J Approx Reason 48:365–377
Fisher RA (1930) Inverse probability. Proc Cambridge Philos Soc 26:528–535
Fisher RA (1933) The concepts of inverse probability and fiducial probability referring to unknown parameters. Proc R Soc Lond A 139:343–348
Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugenics 6:91–98
Fraser DAS (1961a) The fiducial method and invariance. Biometrika 48:261–280
Fraser DAS (1961b) On fiducial inference. Ann Math Stat 32:661–676
Fraser DAS (1966) Structural probability and a generalization. Biometrika 53:1–9
Fraser DAS (1968) The structure of inference. Wiley, New York
Glagovskiy YS (2006) Construction of fiducial confidence intervals for the mixture of cauchy and normal distributions. Master's thesis, Department of Statistics, Colorado State University
Grundy PM (1956) Fiducial distributions and prior distributions: an example in which the former cannot be associated with the latter. J R Stat Soc Ser B 18:217–221
Hannig J (2009a) On asymptotic properties of generalized fiducial inference for discretized data. Tech. Rep. UNC/STOR/09/02, Department of Statistics and Operations Research, The University of North Carolina

Hannig J (2009b) On generalized fiducial inference. Stat Sinica 19:491–544
Hannig J, Abdel-Karim LEA, Iyer HK (2006a) Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. Aust J Stat 35:261–269
Hannig J, Iyer HK, Patterson P (2006b) Fiducial generalized confidence intervals. J Am Stat Assoc 101:254–269
Hannig J, Iyer HK, Wang JC-M (2007) Fiducial approach to uncertainty assessment accounting for error due to instrument resolution. Metrologia 44:476–483
Hannig J, Lee TCM (2009) Generalized fiducial inference for wavelet regression. Biometrika 96(4):847–860
Hannig J, Wang CM, Iyer HK (2003) Uncertainty calculation for the ratio of dependent measurements. Metrologia, 4: 177–186
Iyer HK, Patterson P (2002) A recipe for constructing generalized pivotal quantities and generalized confidence intervals. Tech. Rep. 2002/10, Department of Statistics, Colorado State University
Iyer HK, Wang JC-M, Mathew T (2004) Models and confidence intervals for true values in interlaboratory trials. J Am Stat Assoc 99:1060–1071
Jeffreys H (1940) Note on the Behrens-Fisher formula. Ann Eugenics 10:48–51
Lidong E, Hannig J, Iyer HK (2008) Fiducial Intervals for variance components in an unbalanced two-component normal mixed linear model. J Am Stat Assoc 103:854–865
Lidong E, Hannig J, Iyer HK (2009) Fiducial generalized confidence interval for median lethal dose (LD50). (Preprint)
Lindley DV (1958) Fiducial distributions and Bayes' theorem. J R Stat Soc Ser B 20:102–107
McNally RJ, Iyer HK, Mathew T (2003) Tests for individual and population bioequivalence based on generalized p-values. Stat Med 22:31–53
Patterson P, Hannig J, Iyer HK (2004) Fiducial generalized confidence intervals for proportion of conformance. Tech. Rep. 2004/11, Colorado State University
Salome D (1998) Staristical inference via fiducial methods. Ph.D. thesis, University of Groningen
Stevens WL (1950) Fiducial limits of the parameter of a discontinuous distribution. Biometrika 37:117–129
Tsui K-W, Weerahandi S (1989) Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J Am Stat Assoc 84:602–607
Tsui K-W, Weerahandi S (1991) Corrections: generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. [J Am Stat Assoc 84 (1989), no. 406, 602–607; MR1010352 (90g:62047)]. J Am Stat Assoc 86:256
Wandler DV, Hannig J (2009) Fiducial inference on the maximum mean of a multivariate normal distribution (Preprint)
Wang JC-M, Iyer HK (2005) Propagation of uncertainties in measurements using generalized inference. Metrologia 42: 145–153
Wang JC-M, Iyer HK (2006a) A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. Measurement 39:856–863
Wang JC-M, Iyer HK (2006b) Uncertainty analysis for vector measurands using fiducial inference. Metrologia 43: 486–494
Wang YH (2000) Fiducial intervals: what are they? Am Stat 54: 105–111
Weerahandi S (1993) Generalized confidence intervals. J Am Stat Assoc 88:899–905
Weerahandi S (1994) Correction: generalized confidence intervals [J Am Stat Assoc 88 (1993), no. 423, 899–905; MR1242940 (94e:62031)]. J Am Stat Assoc 89:726
Weerahandi S (1995) Exact statistical methods for data analysis. Springer series in statistics. Springer-Verlag. New York
Wilkinson GN (1977) On resolving the controversy in statistical inference (with discussion). J R Stat Soc Ser B 39: 119–171
Xu X, Li G (2006) Fiducial inference in the pivotal family of distributions. Sci China Ser A Math 49:410–432

Financial Return Distributions

Matthias Fischer
University of Erlangen-Nürnberg, Erlangen, Germany

Describing past and forecasting future asset prices has been attracting the attention of several generations of researchers. Rather than analyzing the asset prices P_t at times $t = 1, \ldots, T$ themselves, one usually focusses on the corresponding log-returns defined by $R_t^c = \log(P_t) - \log(P_{t-1})$ for $t = 2, \ldots, T$. Considering prices (and consequently log-returns) as realizations of random variables, it seems natural to identify the underlying data-generating probability distribution. The search for an adequate model for the distribution of stock market returns dates back to the beginning of the twentieth century: Following Courtault et al. (2000), "*The date March 29, 1900, should be considered as the birthdate of mathematical finance. On that day, a French postgraduate student, Louis Bachélier, successfully defended at the Sorbonne his thesis Théorie de la Spéculation. […] This pioneering analysis of the stock and option markets contains several ideas of enormous value in both finance and probability. In particular, the theory of Brownian motion (see* ►*Brownian Motion and Diffusions*), *was initiated and used for the mathematical modelling of price movements and the evaluation of contingent claims in financial markets*".

Whereas Bachélier (1900) rests upon normally distributed return distributions, the history of heavy tails in finance began in 1963: Assuming independence of successive increments and the validity of the principle of scaling invariance, Mandelbrot (1963) advocates the (Lévy) stable distributions for price changes, supported by Fama (1965) and Fielitz (1976). In fact, tails of stable distributions are very heavy, following a power-law distribution with an exponent $\alpha < 2$. In contrast, empirical studies indicate that tails of most financial time series have to be modeled with $\alpha > 2$ (see, e.g., Lau et al. 1990 or Pagan 1996). In particular,

Akgiray et al. (1989) support this conjecture for German stock market returns. Rejecting the stable hypothesis, several proposals came up in the subsequent years: Praetz (1972), Kon (1984) or Akgiray and Booth (1987) favour finite mixtures of normal distributions, whereas, e.g., Ball and Torous (1983) propose an infinite number of normal distributions mixtures with Poisson probabilities.

Since the early seventies of the last century, the Student-*t* distribution increases in popularity (see, e.g., Blattberg and Gonedes 1974). Depending on the shape and tail parameter ν, moments of the Student-*t* distribution exist only up to a certain order depending on ν, whereas the ►moment-generating function doesn't exist. In order to increase its flexibility regarding ►skewness, peakedness and tail behavior, several generalized Student-*t* versions followed up within the past years (see, e.g., McDonald and Newey 1988; Theodossiou 1998; Hansen et al. 2003 or Adcock and Meade 2003). Finally, if both (semi-)heavy tails and existence of the corresponding moment-generating function are required, different multi-parametric distribution families with exponential tail behavior were successfully applied to financial returns: Among them, the generalized logistic distribution family (see, e.g., Bookstaber and McDonald 1987; McDonald 1991 or McDonald and Bookstaber 1991), the generalized hyperbolic secant distribution families (see, e.g., Fischer 2004, 2006) and the generalized hyperbolic distribution family (see, e.g., Eberlein and Keller 1995; Barndorff-Nielsen 1995; Küchler et al. 1999 and Prause 1999) which in turn includes a subfamily (in the limit) where one tail has polynomial and the other exponential tails, see Aas and Haff (2006).

Selecting a suitable probability distribution for a given return data sample is by far not an easy task. In general, there is no information about the tail behavior of the unknown distribution or, conversely, about the order up to which the corresponding moments exit. Discussions as to whether moments, in particular variances, do exist or not, have a long tradition in financial literature (see, for instance, Tucker 1992). In order to check whether certain moments do exist or not, Granger and Orr (1972) introduced the so-called running-variance plot. Alternatively, the test statistic of Yu (2000) – which is determined by the range of the sample interquartile and the sample standard deviation – may come to application. Recently, Pisarenko and Sornette (2006) came up with a test statistic to discriminate between exponential and polynomial tail behavior.

Cross References

►Hyperbolic Secant Distributions and Generalizations
►Statistical Modeling of Financial Markets

References and Further Reading

Aas K, Haff IH (2006) The generalized hyperbolic skew Student's t-distribution. J Financ Econom 4(2):275–309

Adcock CJ, Meade N (2003) An extension of the generalised skew Student distribution with applications to modelling returns on financial assets. Working paper, Department of Economics, University of Sheffield

Akgiray V, Booth GG (1987) Compound distribution models of stock returns: an empirical comparison. J Financ Res 10(3): 269–280

Akgiray V, Booth GG, Loistl O (1989) Statistical models of German stock returns. J Econ 50(1):17–33

Bachélier L (1900) Théorie de la spéculation. Annales Scientiéques de l'Ecole Nor-male Supérieure 17(3):21–81

Ball CA, Torous W (1983) A simplified jump process for common stock returns. J Financ Quant Anal 18:53–65

Barndorff-Nielsen OE (1995) Normal inverse Gaussian processes and the modelling of stock returns. Research Report No. 300, Department of Theoretical Statistics, University of Aarhus

Blattberg R, Gonedes N (1974) Stable and student distributions for stock prices. J Bus 47:244–280

Bookstaber RM, McDonald JB (1987) A general distribution for describing security price returns. J Bus 60(3):401–424

Courtault J-M, Kabanov J, Bru B, Crépel P (2000) Louis Bachelier – On the centenary of théorie de la spéculation. Math Financ 10(3):341–353

Eberlein E, Keller U (1995) Hyperbolic distributions in finance. Bernoulli 1(3):281–299

Fama E (1965) The behaviour of stock market prices. J Bus 38:34–105

Fielitz B (1976) Further results on asymmetric stable distributions of stock prices changes. J Financ Quant Anal 11:39–55

Fischer M (2004) Skew generalized secant hyperbolic distributions: unconditional and conditional fit to asset returns. Austrian J Stat 33(3):293–304

Fischer M (2006) A skew generalized secant hyperbolic family. Austrian J Stat 35(4):437–444

Granger CWJ, Orr R (1972) Infinite variance and research strategy in time series analysis. J Am Stat Assoc 67:275–285

Hansen CB, McDonald JB, Theodossiou P (2003) Some exible parametri models for skewed and leptokurtic data. Working paper, Department of Economics, Brigham Young University

Kon SJ (1984) Models of stock returns – A comparison. J Financ 39(1):147–165

Küchler E, Neumann K, Sørensen M, Streller A (1999) Stock returns and hyperbolic distributions. Math Comp Model 29:1–15

Lau AH, Lau KC, Wingender JR (1990) The distribution of stock returns: new evidence against the stable model. J Bus Econ Stat 8(2):217–223

Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–519

McDonald JB (1991) Parametric models for partially adaptive estimation with skewed and leptokurtic residuals. Econ Lett 37: 273–278

McDonald JB, Bookstaber RM (1991) Option pricing for generalized distributions. Commun Stat Theory Meth 20(12):4053–4068

McDonald JB, Newey WK (1988) Partially adaptive estimation of regression models via the generalized-t-distribution. Econometr Theory 1(4):428–457

Pagan A (1996) The econometrics of financial markets. J Empirical Financ 3:15–102

Pisarenko V, Sornette D (2006) New statistic for financial return distributions: power-law or exponential? Physika A 366: 387–400
Praetz PD (1972) The distribution of stock price changes. J Bus 45:49–55
Prause K (1999) The generalized hyperbolic model: estimation, financial derivatives and risk measures. PhD thesis, University of Freiburg, Freiburg
Theodossiou P (1998) Financial data and the skewed generalized t distribution. Manag Sci 44:1650–1661
Tucker A (1992) A reexamination of finite- and infinite variance distributions as models of daily stock returns. J Bus Econ Stat 10:73–81
Yu J (2000) Testing for a finite variance in stock return distributions. In: Dunis CL (ed) Advances in quantitative asset management, studies in computational finance vol 1, Chap 6. Kluwer Academic, Amsterdam, pp 143–164

First Exit Time Problem

Christos H. Skiadas[1], Charilaos Skiadas[2]
[1]Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece
[2]Hanover College, Hanover, IN, USA

The first exit time distribution for a stochastic process is the distribution of the times at which particles following this process cross a certain (often linear) barrier. It is often referred to also as hitting time. It is closely related to the probability density function $p(x_t, t)$ of a stochastic process x_t over time t.

For a linear horizontal barrier located at a, the first exit time density function relation is given by: $g(t) = \frac{|a|}{t} p(a, t)$.

For other types barriers (e.g., quadratic), a tangent approximation may be used to obtain a satisfactory estimate as is presented below.

The probability density function may be computed in some cases using the Fokker-Planck equation. In particular in the one-dimensional diffusion problem expressed by a stochastic differential equation of the form:

$$dx_t = \sigma dw_t,$$

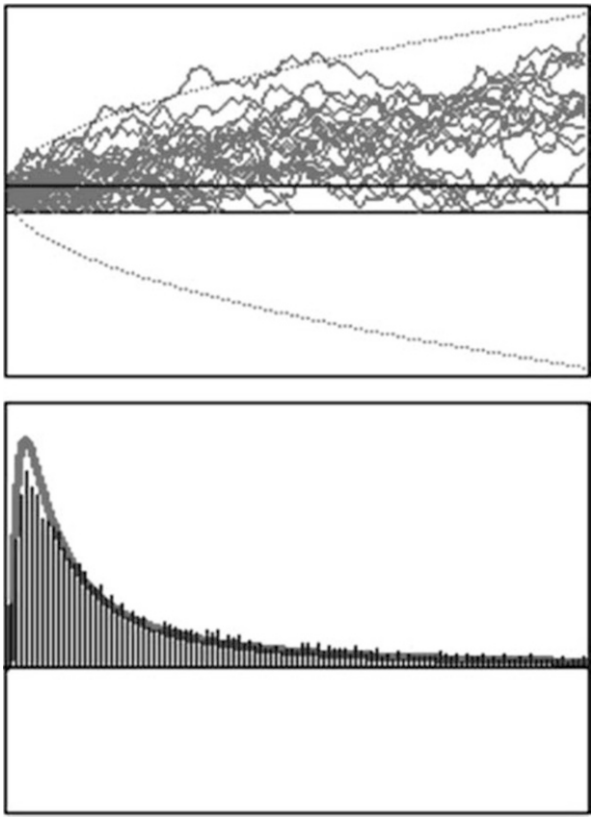
where σ is the variance and w_t is the standard Wiener process, the corresponding Fokker-Planck equation for the probability density function $p(x_t, t)$ associated to the above stochastic differential equation has the form:

$$\frac{\partial p(x_t, t)}{\partial t} = \frac{\sigma^2}{2} \frac{\partial^2 p(x_t, t)}{\partial x_t^2}$$

This partial differential equation form is also known as the one-dimensional heat equation first solved by Joseph Fourier (1822). Later on Fick (1855a; 1855b) applied this equation to express one-dimensional diffusion in solids. Albert Einstein (1905) proposed the same form for the one-dimensional diffusion for solving the Brownian motion process (see ►Brownian Motion and Diffusions). It was the first derivation and application of a probabilistic-stochastic theory to the classical Brownian motion problem that is the movement of a particle or a molecule into a liquid. He resulted in giving the development over space and time of this particle. One year later Smoluchowski (1906) proposed also a theory for solving the Brownian motion problem.

Solving this partial differential equation with the boundary conditions, $p(x_t, 0 : 0, 0) = \delta(x_t, 0)$ and $\frac{\partial p(x_t, t : 0, t)}{\partial x} = 0$ as $x_t \to \infty$ the probability density function p_t for the stochastic process results:

$$p(x_t, t) = \frac{1}{\sigma\sqrt{2\pi t}} e^{-\frac{x_t^2}{2\sigma^2 t}}.$$



First Exit Time Problem. Fig. 1 Linear barrier

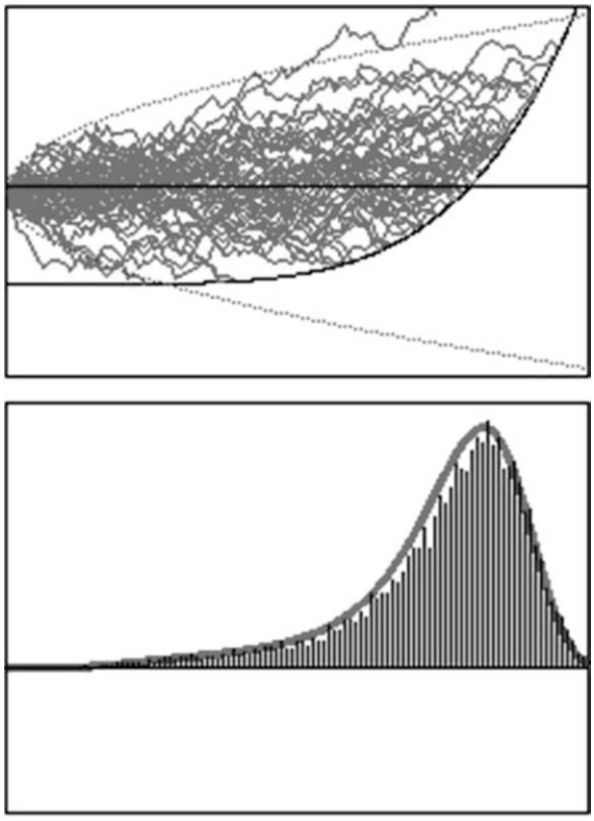
The First Exit Time Density Function

The finding of a density function expressing the distribution of the first exit time of particles escaping from a boundary is due to Schrödinger (1915) and Smoluchowski (1915) in two papers published in the same journal issue. Later on Siegert (1951) gave an interpretation closer to our modern notation whereas Jennen (1985), Lerche (1986) and Jennen and Lerche (1981) gave the most interesting first exit density function form. For the simple case presented earlier the proposed form is:

$$g(t) = \frac{|a|}{t} p(a,t) = \frac{|a|}{\sigma\sqrt{2\pi t^3}} e^{-\frac{a^2}{2\sigma^2 t}}.$$

Jennen (1985) proposed a more general form using a tangent approximation of the first exit density. Application of this theory to the mortality modeling leads to the following form (earlier work can be found in Janssen and Skiadas (1995) and Skiadas and Skiadas (2007):

$$g(t) = \frac{|H_t - tH_t'|}{t} p(t) = \frac{|H_t - tH_t'|}{\sigma\sqrt{2\pi t^3}} e^{-\frac{(H_t)^2}{2\sigma^2 t}}.$$



First Exit Time Problem. Fig. 2 Curved barrier

The last form is associated to the following stochastic process Skiadas (2010):

$$dx_t = \mu_t dt + \sigma dw_t$$

where μ_t is a function of time and there exists a function H_t related to μ_t with the differential equation: $\mu_t = dH_t/dt$.

The associated Fokker–Planck equation is:

$$\frac{\partial p(x_t,t)}{\partial t} = -\mu_t \frac{\partial p(x_t,t)}{\partial x_t} + \frac{\sigma^2}{2} \frac{\partial^2 p(x_t,t)}{\partial x_t^2}$$

and the solution is given by:

$$p(t) = \frac{1}{\sigma\sqrt{2\pi t^3}} e^{-\frac{(H_t)^2}{2\sigma^2 t}}.$$

Two realizations are provided in Figs. 1 and 2. In the first case the first exit time probability density function is provided and stochastic simulations are done for a linear barrier located at α. Figure 2 illustrates the case when a curved barrier is present.

About the Author

For biography *see* the entry ►"Chaotic modelling."

Cross References

►Brownian Motion and Diffusions
►First-Hitting-Time Based Threshold Regression
►Random Walk
►Stochastic Processes

References and Further Reading

Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik 17:549–560
Fick A (1855) Über Diffusion. Poggendorff's Annalen. 94:59–86
Fick A (1855) On liquid diffusion. Philos Mag J Sci 10:31–39
Fourier J (1822) Theorie Analytique de la Chaleur. Firmin Didot, Paris
Fourier J (1878) The analytical theory of heat. Cambridge University Press, New York
Janssen J, Skiadas CH (1995) Dynamic modelling of life-table data. Appl Stoch Model Data Anal 11(1):35–49
Jennen C (1985) Second-order approximation for Brownian first exit distributions. Ann Probab 13:126–144
Jennen C, Lerche HR (1981) First exit densities of Brownian motion through one-sided moving boundaries. Z Wahrsch uerw Gebiete 55:133–148
Lerche HR (1986) Boundary crossing of Brownian motion. Springer-Verlag, Berlin
Schrödinger E (1915) Zur theorie der fall - und steigversuche an teilchenn mit Brownsche bewegung. Phys Zeit 16:289–295
Siegert AJF (1951) On the first passage time probability problem. Phys Rev 81:617–623
Skiadas CH (2010) Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential. Meth Comput Appl Probab 12(2):261–270

Skiadas CH, Skiadas C (2007) A modeling approach to life table data. In Skiadas CH (ed) Recent advances in stochastic modeling and data analysis. World Scientific, Singapore, pp 350–359

Skiadas C, Skiadas CH (2010) Development, simulation and application of first exit time densities to life table data. Comm Stat Theor Meth 39(3):444–451

Smoluchowski M (1906) Zur kinetischen theorie der Brownschen molekularbewegung und der suspensionen. Ann D Phys 21:756–780

Smoluchowski M (1915) Notizüber die berechning der Brownschen molekular-bewegung bei der ehrenhaft-millikanchen versuchsanordnung. Phys Zeit 16:318–321

First-Hitting-Time Based Threshold Regression

Xin He[1], Mei-Ling Ting Lee[2]
[1]Assistant Professor
University of Maryland, College Park, MD, USA
[2]Professor, Director, Biostatistics and Risk Assessment Center (BRAC)
University of Maryland, College Park, MD, USA

First-hitting-time (FHT) based threshold regression (TR) model is a relatively new methodology for analyzing ►survival data where the time-to-event is modeled as the first time the stochastic process of interest hits a boundary threshold. FHT models have been applied in analyzing the failure time of engineering systems, the length of hospital stay, the survival time of AIDS patients, and the duration of industrial strikes, etc.

First-Hitting-Time (FHT) Model

A first-hitting-time (FHT) model has two basic components, namely a stochastic process $\{Y(t), t \in \mathcal{T}, y \in \mathcal{Y}\}$ with initial value $Y(0) = y_0$, where $\mathcal{T}$ is the time space and $\mathcal{Y}$ is the state space of the process; and a boundary set $\mathcal{B}$, where $\mathcal{B} \subset \mathcal{Y}$. Assume that the initial value of the process y_0 lies outside the boundary set $\mathcal{B}$, then the first hitting time can be defined by the random variable

$$S = \inf\{t : Y(t) \in \mathcal{B}\},$$

where S is the first time the sample path of the stochastic process reaches the boundary set $\mathcal{B}$. In a medical context, the stochastic process $\{Y(t)\}$ may describe a subject's latent health condition or disease status over time t. The boundary set $\mathcal{B}$ represents a medical end point, such as death, or disease onset. Although the boundary set $\mathcal{B}$ is set to be fixed in time in basic FHT models, it may vary with time in some applications. The stochastic process $\{Y(t)\}$ in the FHT model may take many forms. The most commonly used process is a Wiener diffusion process with a positive initial value and a negative drift parameter. Alternative processes including the gamma process, the Ornstein-Uhlenbeck (OU) process, and the semi-Markov process have also been investigated. For a review, see Lee and Whitmore (2006) and Aalen et al. (2008).

Threshold Regression

Threshold regression (TR) is an extension of the first-hitting-time model by adding regression structures to it so as to accommodate important covariates. The threshold regression model does not required the proportional hazards assumption and hence it provides an alternative model for analyzing time-to-event data. The unknown parameters in the stochastic process $\{Y(t)\}$ and the boundary set $\mathcal{B}$ are connected to covariates using suitable regression link functions. For example, the initial state y_0 and the drift parameter μ of a Wiener diffusion process $\{Y(t)\}$ can be linked to covariates using general link functions of the form

$$y_0 = g_1(\mathbf{x})$$

and

$$\mu = g_2(\mathbf{x}),$$

where $\mathbf{x}$ is the vector of covariates (Lee et al. 2000; Lee and Whitmore 2006). Pennell et al. (2010) proposed a TR model with Bayesian random effects to account for unmeasured covariates in both the initial state and the drift. Yu et al. (2009) incorporated penalized regression and regression splines to TR models to accommodate semi-parametric nonlinear covariate effects.

Analytical Time Scale

In stead of calendar time, in many applications involving time-dependent cumulative effects, an alternative time scale can be better used in describing the stochastic process. Let $r(t|\mathbf{x})$ denote a monotonic transformation of calendar time t to analytical time r (or referred to as process time) with $r(0|\mathbf{x}) = 0$. In a medical context, the analytical time may be some time-dependent measure to describe cumulative toxic exposure or the progression of disease. The process $\{Y(r)\}$ defined in terms of analytical time r can be expressed as a subordinated process $\{Y[r(t)]\}$ in terms of calendar time t. Lee and Whitmore (1993, 2004) examined the connection between subordinated stochastic processes and analytical time.

Whitmore et al. (1998) proposed a bivariate Wiener model in which failure is governed by a latent process while auxiliary readings are available from a correlated marker process. Lee et al. (2000) extended this model to bivariate threshold regression by including CD4 cell counts as a marker process in the context of AIDS clinical trials.

Tong et al. (2008) generalized the bivariate TR model to current status data. Using Markov decomposition methods, Lee et al. (2010) generalized threshold regression to include time-dependent covariates. Lee and Whitmore (2010) discussed the connections between TR and proportional hazard regressions and demonstrated that proportional hazard functions can be generated by TR models.

About the Author

Professor Lee was named the Mosteller Statistician of the Year in 2005 by the American Statistical Association, Boston Chapter. She is Elected member of the International Statistical Institute (1995), and Elected Fellow of: Royal Statistical Society (1998), American Statistical Association (1999) and the Institute of Mathematical Statistics (2005). Professor Lee is the Founding Editor and Editor-in-Chief of the international journal *Lifetime Data Analysis*.

Cross References

►First Exit Time Problem
►Survival Data

References and Further Reading

Aalen OO, Borgan Ø, Gjessing HK (2008) Survival and event history analysis: a process point of view. Springer, New York

Lee M-LT, Whitmore GA (1993) Stochastic processes directed by randomized time. J Appl Probab 30:302–314

Lee M-LT, Whitmore GA (2004) First hitting time models for liefetime data. In: Rao CR, Balakrishnan N (eds) Advances in survival ananlysis. North Holland, Amsterdam pp 537–543

Lee M-LT, Whitmore GA (2006) Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat Sci 21:501–513

Lee M-LT, Whitmore GA (2010) Proportional hazards and threshold regression: their theoretical and practical connections. Lifetime Data Anal 16:196–214

Lee M-LT, DeGruttola V, Schoenfeld D (2000) A model for markers and latent health status. J R Stat Soc B 62:747–762

Lee M-LT, Whitmore GA, Rosner B (2010) Threshold regression for survival data with time-varying covariates. Stat Med 29: 896–905

Pennell ML, Whitmore GA, Lee M-LT (2010) Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. Biostat 11:111–126

Tong X, He X, Sun J, Lee M-LT (2008) Joint analysis of current status and marker data: an extension of a bivariate threshold model. Int J Biostat 4, Article 21

Whitmore GA, Crowder MJ, Lawless JF (1998) Failure inference from a marker process based on a bivariate Wiener model. Lifetime Data Anal 4:229–251

Yu Z, Tu W, Lee M-LT (2009) A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women. Stat Med 28:3029–3042

Fisher Exact Test

Peter Sprent
Emeritus Professor
University of Dundee, Dundee, UK

The Fisher Exact test was proposed by Fisher (1934) in the fifth edition of *Statistical Methods for Research Workers*. It is a test for independence as opposed to association in 2×2 contingency tables.

A typical situation where such tables arise is where we have counts of individuals categorized by each of two dichotomous attributes, e.g., one attribute may be religious affiliation dichotomized into Christian and non-Christian and the other marital status recorded as married or single.

Another example is that where one of the attributes that are dichotomized corresponds to treatments, e.g., Drug A prescribed, or Drug B prescribed, and the other attribute is the responses to those treatments, e.g., patient condition improves or patient condition does not improve.

In the latter situation if 9 patients are given Drug A and 12 patients are given drug B we might observe the following counts in cells of a 2 × 2 table:

	Improvement	No improvement	Row total
Drug A	8	1	9
Drug B	3	9	12
Column totals	11	10	21

Fisher pointed out that if we assume row and column totals are fixed then once we know the entry in any cell of the table (e.g., here 8 in the top left cell) then the entries in the remaining three cells are all fixed by the constraint that the marginal totals are fixed. This is usually expressed by saying the table has one degree of freedom. What Fisher noted is that under the hypothesis of independence, if we assume the marginal totals fixed then the distribution of the numbers in the first cell (or any other cell) has a hypergeometric distribution under independence for any of the common models associated with such a table as described, for example in Agresti (2002) or Sprent and Smeeton (2007). These common models are (1) that responses to each drug, for example, are binomially distributed with a common value for the binomial parameter p or (2) have a common Poisson distribution, or (3)

the four cell counts are a sample from a ►multinomial distribution.

If a general a 2×2 contingency table has cell entries n_{ij} $(i, j = 1, 2)$ and row totals n_{i+} and column totals n_{+j} and grand total of all 4 cell entries is n, then the hypergeometric distribution for the observed cell values has an associated probability

$$\frac{(n_{1+})!(n_{2+})!(n_{+1})!(n_{+2})!}{(n_{11})!(n_{12})!(n_{21})!(n_{22})!n!}$$

To perform the test one calculates these probabilities for all possible n_{11} consistent with the fixed marginal totals and computes the P-value as the sum of all such probabilities that are less than or equal to that associated with the observed configuration. For the numerical example given above n_{11} may take any integral value between 0 and 9 and the following table gives the corresponding hypergeometric probabilities:

n_{11}	Hypergeometric probability
0	0.00003
1	0.00168
2	0.02245
3	0.11788
4	0.28295
5	0.33008
6	0.18862
7	0.05052
8	0.00561
9	0.00019

From this table we see that the observed $n_{11} = 8$ has associated probability $p = 0.00561$ and the only other outcomes with this or a lower probability correspond to $n_{11} = 0, 1$ or 9. Thus the test P-value is $P = 0.00561 + 0.00168 + 0.00019 + 0.00003 = 0.00751$.

This low P-value provides very strong evidence of association, i.e., that the drugs differ in effectiveness.

In practice, except for very small samples appropriate statistical software is required to compute P. When the expected numbers in each cell assuming independence are not too small the standard chi-squared test for contingency tables gives a close approximation to the exact test P-value especially if Yates's correction (*see* Sprent and Smeeton 2007) is used.

The exact test procedure was extended by Freeman and Halton (1951) to tables with any specified numbers of rows and columns.

Some statisticians have argued that it is inappropriate to condition the test statistics on fixed marginal totals, but it is now widely accepted that in most, though not perhaps in all, situations arising in practice such conditioning is appropriate.

About the Author

For biography *see* the entry ►Sign Test.

Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Exact Inference for Categorical Data
►Hypergeometric Distribution and Its Application in Statistics
►Nonparametric Statistical Inference
►Proportions, Inferences, and Comparisons

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Fisher RA (1934) Statistical methods for research workers, 5th edn. Oliver & Boyd, Edinburgh
Freeman GH, Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38:141–149
Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods, 4th edn. Chapman & Hall/CRC, Boca Raton

Fisher-Tippett Theorem

Bojan Basrak
University of Zagreb, Zagreb, Croatia

In 1928, Fisher and Tippett presented a theorem which can be considered as a founding stone of the *extreme value theory*. They identified all ►*extreme value distributions*, which means all possible nondegenerate limit laws for properly centered and scaled partial maxima $M_n = \max\{X_1, \ldots, X_n\}$, where (X_n) is a sequence of independent and identically distributed random variables. More precisely, if there exist real sequences (a_n) and (b_n) where $a_n > 0$ for all n, such that the random variables

$$\frac{M_n - b_n}{a_n} \quad \text{as } n \to \infty,$$

converge in distribution to a nondegenerate random variable with a distribution function G, then G is called an extreme value distribution. The theorem states that G (permitting centering and scaling) necessarily belongs to one of the following three classes: *Fréchet, Gumbel*, and *Weibull distributions*.

Rigorous proofs of the theorem appearing in contemporary literature are due to Gnedenko in 1943, and works of de Haan and Weissman in 1970s. The class of ►extreme value distributions coincides with the class of *max–stable distributions*, i.e. those distributions of the random variable X_1, for which there exist real constants $c_n > 0$ and d_n for each $n \geq 1$, such that $(M_n - d_n)/c_n$ has the same distribution as X_1.

To determine whether partial maxima of a given family of random variables, after scaling and centering, has asymptotically one of those distributions is one of the main tasks of extreme value analysis. Many of such questions are answered using the notion of regular variation which was introduced in mathematical analysis by Karamata, a couple of years after the publication of Fisher–Tippett theorem.

Cross References

►Extreme Value Distributions
►Statistics of Extremes

References and Further Reading

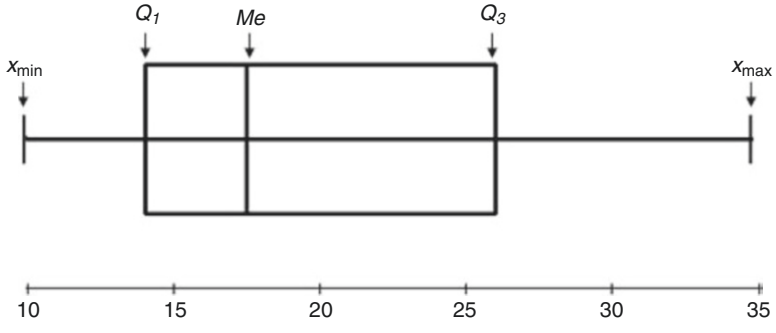
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc Camb Philos Soc 24:180–190
Gnedenko B (1943) Sur la distribuion limite du terme maximum d'une série aléatoire. Ann Math 44:423–453
Resnick SI (1987) Extreme values, regular variation, and point processes. Springer, New York

Five-Number Summaries

Mirjana Čižmešija
Professor, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

The five-number summary $(5'S)$ is a technique of exploratory data analyses developed with the aim of investigating one or more data sets. It consists of five descriptive measures (Anderson 2007): minimum value $(x_{\min})$, first quartile (Q_1), median (Me), third quartile (Q_3), and maximum value $(x_{\max})$. The graphical presentation of the five-number summary is the box-and-whisker plot (box-plot) developed by John Tukey (Levine 2008). In determining the $5'S$, the data set of observations on a single variable must be arranged from the smallest to the largest value, and therefore the $5'S$ are arranged as follows: $x_{\min} \leq Q_1 \leq Me \leq Q_3 \leq x_{\max}$. Each of these five parameters is important in descriptive statistics for providing information about the dispersion and skew of data sets. ►Outliers in the data set may be detected in the box-plot. In measuring dispersion, the distance between the minimum and maximum value is important (particularly in financial analyses).

The difference between the first and third quartile is the range of the middle 50% of the data in the data set (interquartile range). These differences and the differences between quartiles and the median are important in detecting the shape of the data set. In a symmetrical distribution, the difference between the first quartile and the minimum value is the same as the difference between the maximum value and the third quartile, and the difference between the median and the first quartile is the same as the difference between the third quartile and the median. In a right-skewed distribution, the difference between the first quartile and the minimum value is smaller than the



Five-Number Summaries. Fig. 1 Box-plot

difference between the maximum value and the third quartile, and the difference between the median and the first quartile is smaller than the difference between the third quartile and the median. In a left-skewed distribution, the difference between the first quartile and the minimum value is greater than the difference between the maximum value and the third quartile and the difference between the median and the first quartile is greater than the difference between the third quartile and the median. The five-number summary is a useful tool in comparing the dispersion of two or more data sets.

For example, the following data set

$$10, 11, 14, 15, 17, 18, 20, 26, 26, 35$$

can be described as $5'S$:

$$x_{\min} = 10, \quad Q_1 = 14, \quad Me = 17,5, \quad Q_3 = 26, \quad x_{\max} = 35$$

and graphically displayed by the box-plot in the Fig. 1.

Cross References

►Data Analysis
►Summarizing Data with Boxplots

References and Further Reading

Anderson DR, Sweeney DJ, Williams TA, Freeman J, Shoesmith E (2007) Statistics for business and economics. Thomson, London

Levine DM, Stephan DF, Krehbiel FC, Berenson ML (2008) Statistics for managers using Microsoft Excel, 5th edn. Personal Education International Upper Saddle River

Forecasting Principles

Kesten C. Green[1], Andreas Graefe[2], J. Scott Armstrong[3]
[1]Associate Professor
University of South Australia, Adelaide, SA, Australia
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Professor of Marketing
University of Pennsylvania, Philadelphia, PA, USA

Introduction

Forecasting is concerned with making statements about the as yet unknown. There are many ways that people go about deriving forecasts. This entry is concerned primarily with procedures that have performed well in empirical studies that contrast the accuracy of alternative methods.

Evidence about forecasting procedures has been codified as condition-action statements, rules, guidelines or, as we refer to them, *principles*. At the time of writing there are 140 principles. Think of them as being like a safety checklist for a commercial airliner – if the forecast is important, it is important to check all relevant items on the list. Most of these principles were derived as generalized findings from empirical comparisons of alternative forecasting methods. Interestingly, the empirical evidence sometimes conflicts with common beliefs about how to forecast.

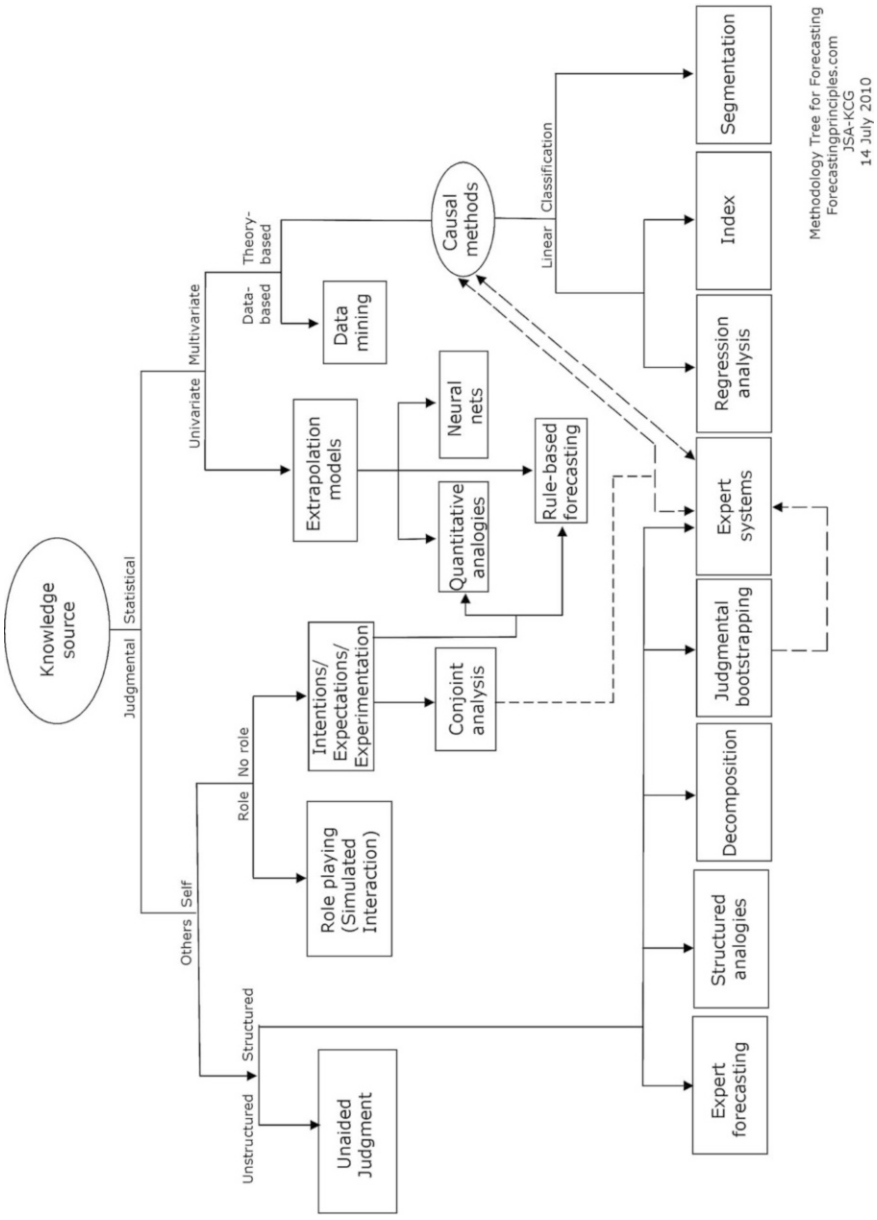
Primarily due to the strong emphasis placed on empirical comparisons of alternative methods, researchers have made many advances in forecasting since 1980. The most influential paper in this regard is the M-competition paper (Makridakis et al. 1982). This was based on a study in which different forecasters were invited to use what they thought to be the best method to forecast many time series. Entry into the competition required that methods were fully disclosed. Entrants submitted their forecasts to an umpire who calculated the errors for each method. This was only one in a series of M-competition studies, the most recent being Makridakis and Hibon (2000). For a summary of the progress that has been made in forecasting since 1980, see Armstrong (2006).

We briefly describe valid forecasting methods, provide guidelines for the selection of methods, and present the *Forecasting Canon* of nine overarching principles. The *Forecasting Canon* provides a gentle introduction for those who do not need to become forecasting experts but who nevertheless rightly believe that proper knowledge about forecasting would help them to improve their decision making. Those who wish to know more can find what they seek in *Principles of Forecasting: A Handbook for Practitioners and Researchers*, and at the Principles of Forecasting Internet site (ForPrin.com).

Forecasting Methods

As shown in Fig. 1, the *Forecasting Methodology Tree*, forecasting methods can be classified into those that are based primarily on judgmental sources of information and those that use statistical data. There is overlap between some judgmental and statistical approaches.

If available data are inadequate for quantitative analysis or qualitative information is likely to increase the accuracy, relevance, or acceptability of forecasts, one way to make forecasts is to ask experts to think about a situation and predict what will happen. If experts' forecasts are not derived using structured forecasting methods, their forecasting method is referred to as *unaided judgment*. This is the most commonly used method. It is fast, inexpensive



Forecasting Principles. Fig. 1 Methodology tree

when few forecasts are needed, and may be appropriate when small changes are expected. It is most likely to be useful when the forecaster knows the situation well and gets good feedback about the accuracy of his forecasts (e.g., weather forecasting, betting on sports, and bidding in bridge games).

Expert forecasting refers to forecasts obtained in a structured way from two or more experts. The most appropriate method depends on the conditions (e.g., time constraints, dispersal of knowledge, access to experts, expert motivation, need for confidentiality). In general, diverse experts should be recruited, questions should be chosen carefully and tested, and procedures for combining across experts (e.g., the use of medians) should be specified in advance.

The *nominal group technique* (NGT) tries to account for some of the drawbacks of traditional meetings by imposing a structure on the interactions of the experts. This process consists of three steps: First, group members work independently and generate individual forecasts. The group then conducts an unstructured discussion to deliberate on the problem. Finally, group members work independently and provide their final individual forecasts. The NGT forecast is the mean or median of the final individual estimates.

Where group pressures are a concern or physical proximity is not feasible, the *Delphi method,* which involves at least two rounds of anonymous interaction, may be useful. Instead of direct interaction, individual forecasts and arguments are summarized and reported as feedback to participants after each round. Taking into account this information, participants provide a revised forecast for the next round. The Delphi forecast is the mean or median of the individual forecasts in the final round. Rowe and Wright (2001) found that Delphi improved accuracy over unstructured groups in five studies, harmed accuracy in one, and the comparison was inconclusive in two. Delphi is most suitable if experts are expected to possess different information, but it can be conducted as a simple one-round survey for situations in which experts possess similar information. A free version of the Delphi software is available at ForPrin.com.

In situations where dispersed information frequently becomes available, *prediction markets* can be useful for providing continuously updated numerical or probability forecasts. In a prediction market, mutually anonymous participants reveal information by trading contracts whose prices reflect the aggregated group opinion. Incentives to participate in a market may be monetary or nonmonetary. Although prediction markets seem promising, to date there has been no published ▶meta-analysis of the method's accuracy. For a discussion of the relative advantages of prediction markets and Delphi, see Green et al. (2007).

With *structured analogies*, experts identify situations that are analogous to a target situation, identify similarities and differences to the target situation, and determine an overall similarity rating. The outcome or decision implied by each expert's top-rated analogy is used as the structured analogies forecast. Green and Armstrong (2007) analyzed structured analogies for the difficult problem of forecasting decisions people will make in conflict situations. When experts were able to identify two or more analogies and their closest analogy was from direct experience, 60% of structured analogies forecasts were accurate compared to 32% of experts' unaided judgment forecasts, the latter being little better than guessing.

Decomposition involves breaking down a forecasting problem into components that are easier to forecast. The components may either be multiplicative (e.g., to forecast a brand's sales, one could estimate total market sales and market share) or additive (estimates could be made for each type of product when forecasting new product sales for a division). Decomposition is most likely to be useful in situations involving high uncertainty, such as when predicting large numbers. MacGregor (2001) summarized results from three studies involving 15 tests and found that judgmental decomposition led to a 42% reduction in error under high levels of uncertainty.

Judgmental bootstrapping derives a model from knowledge of experts' forecasts and the information experts used to make their forecasts. This is typically done by regression analysis. It is useful when expert judgments have validity but data are scarce (e.g., forecasting new products) and outcomes are difficult to observe (e.g., predicting performance of executives). Once developed, judgmental bootstrapping models are a low-cost forecasting method. Armstrong (2001a) found judgmental bootstrapping to be more accurate than unaided judgment in 8 of 11 comparisons. Two tests found no difference, and one found a small loss in accuracy.

Expert systems are based on rules for forecasting that are derived from the reasoning experts use when making forecasts. They can be developed using knowledge from diverse sources such as surveys, interviews of experts, protocol analysis in which the expert explains what he is doing as he makes forecasts, and research papers. Collopy et al. (2001) summarized evidence from 15 comparisons that included expert systems. Expert systems were more accurate than unaided judgment in six comparisons, similar in one, and less accurate in another. Expert systems were less accurate than judgmental bootstrapping in two

comparisons and similar in two. Expert systems were more accurate than econometric models in one comparison and as accurate in two.

It may be possible to ask people directly to predict how they would behave in various situations. However, this requires that people have valid *intentions* or *expectations* about how they would behave. Both are most useful when (1) responses can be obtained from a representative sample, (2) responses are based on good knowledge, (3) people have no reason to lie, and (4) new information is unlikely to change behavior. Intentions are more limited than expectations in that they are most useful when (5) the event is important, (6) the behavior is planned, and (7) the respondent can fulfill the plan (e.g., their behavior is not dependent on the agreement of others). Better yet, in situations in which it is feasible to do so, conduct an experiment by changing key causal variables in a systematic way, such that the independent variables are not correlated with one another. Estimate relationships from responses to the changes and use these estimates to derive forecasts.

Role playing involves asking people to think and behave in ways that are consistent with a role and situation described to them. Role playing for the purpose of predicting the behavior of people with different roles who are interacting with each other is called *simulated interaction*. Role players are assigned roles and asked to act out prospective interactions in a realistic manner. The decisions are used as forecasts of the actual decision. Green (2005) found that 62% of simulated interaction forecasts were accurate for eight diverse conflict situations. By comparison, 31% of forecasts from the traditional approach – expert judgments unaided by structured techniques – were accurate. Game theory experts' forecasts were no better, also 31%, and both unaided judgment and game theory forecasts were little better than chance at 28% accurate.

Conjoint analysis is a method for eliciting people's preferences for different possible offerings (e.g., for alternative mobile phone designs or for different political platforms) by using combinations of features (e.g., size, camera, and screen of a mobile phone.) The possibilities can be set up as experiments where each variable is unrelated to the other variable. Regression-like analyses are then used to predict the most desirable design.

Extrapolation models use time-series data on the situation of interest (e.g., data on automobile sales from 1940–2009) or relevant cross-sectional data. For example, exponential smoothing, which relies on the principle that more recent data is weighted more heavily, can be used to extrapolate over time. Quantitative extrapolation methods do not harness people's knowledge about the data but assume that the causal forces that have shaped history will continue. If this assumption turns out to be wrong, forecast errors can be large. As a consequence, one should only extrapolate trends when they correspond to the prior expectations of domain experts. Armstrong (2001b) provides guidance on the use of extrapolation.

Quantitative analogies are similar to structured analogies. Experts identify analogous situations for which time-series or cross-sectional data are available, and rate the similarity of each analogy to the data-poor target situation. These inputs are used to derive a forecast. This method is useful in situations with little historical data. For example, one could average data from cinemas in suburbs identified by experts as similar to a new (target) suburb in order to forecast demand for cinema seats in the target suburb.

Rule-based forecasting is an expert system for combining expert domain knowledge and statistical techniques for extrapolating time series. Most series features can be identified automatically, but experts are needed to identify some features, particularly causal forces acting on trends. Collopy and Armstrong (1992) found rule-based forecasting to be more accurate than extrapolation methods.

If data are available on variables that might affect the situation of interest, causal models are possible. Theory, prior research, and expert domain knowledge provide information about relationships between the variable to be forecasted and explanatory variables. Since causal models can relate planning and decision-making to forecasts, they are useful if one wants to create forecasts that are conditional upon different states of the environment. More important, causal models can be used to forecast the effects of different policies.

Regression analysis involves estimating causal model coefficients from historical data. Models consist of one or more regression equations used to represent the relationship between a dependent variable and explanatory variables. Regression models are useful in situations with few variables and many reliable observations where the causal factors vary independently of one another. Important principles for developing regression (econometric) models are to (1) use prior knowledge and theory, not statistical fit, for selecting variables and for specifying the directions of effects (2) use simple models, and (3) discard variables if the estimated relationship conflicts with theory or prior evidence.

Real-world forecasting problems are, however, more likely to involve few observations and many relevant variables. In such situations, the *index method* can be used. Index scores are calculated by adding the values of the explanatory variables, which may be assessed subjectively,

for example as zero or one, or may be normalized quantitative data. If there is good prior domain knowledge, explanatory variables may be weighted relative to their importance. Index scores can be used as forecasts of the relative likelihood of an event. They can also be used to predict numerical outcomes, for example by regressing index scores against historical data.

Segmentation is useful when a heterogeneous whole can be divided into homogenous parts that act in different ways in response to changes, and that can be forecasted more accurately than the whole. For example, in the airline industry, price has different effects on business and personal travelers. Appropriate forecasting methods can be used to forecast individual segments. For example, separate regression models can be estimated for each segment. Armstrong (1985:287) reported on three comparative studies on segmentation. Segments were forecasted either by extrapolation or regression analysis. Segmentation improved accuracy for all three studies.

Selection of Methods

The Forecasting Method Selection Tree, shown in Fig .2, provides guidance on selecting the best forecasting method for a given problem. The Tree has been derived from evidence-based principles. Guidance is provided in response to the user's answers to questions about the availability of data and state of knowledge about the situation for which forecasts are required. The first question is whether sufficient objective data are available to perform statistical analyses. If not, the forecaster should use judgmental methods.

In deciding among judgmental procedures, one must assess whether the future is likely to be substantially different from the past, whether the situation involves decision makers who have conflicting interests, and whether policy analysis is required. Other considerations affecting the selection process are whether forecasts are made for recurrent and well-known problems, whether domain knowledge is available, and whether information about similar types of problems is available.

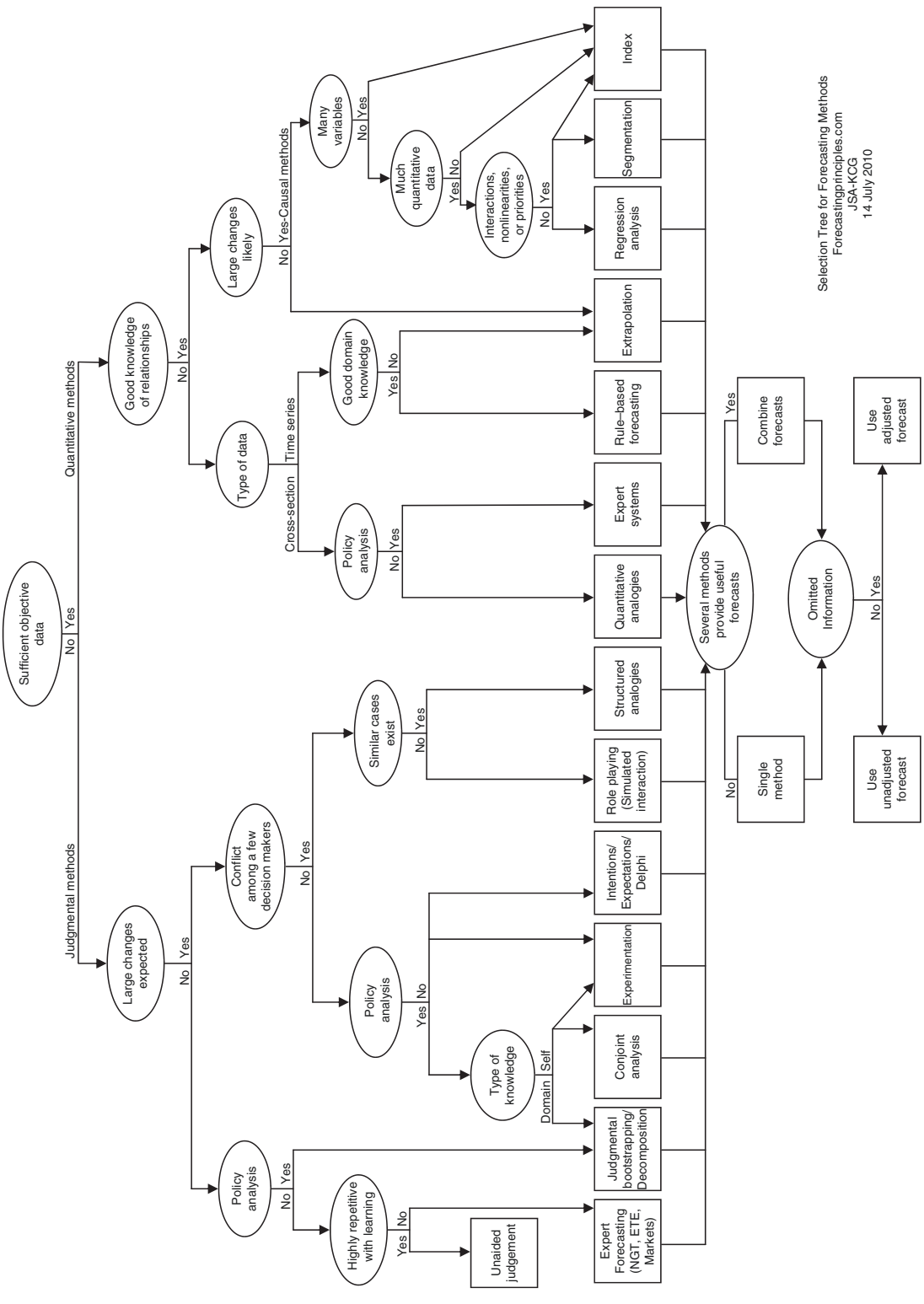
If, on the other hand, much objective data are available and it is possible to use quantitative methods, the forecaster first has to assess whether there is useful knowledge about causal relationships, whether cross-sectional or time-series data are available, and whether large changes are involved. In situations with little knowledge about empirical relationships, the next issues are to assess whether policy analysis is involved and whether there is expert domain knowledge about the situation. If there is good prior knowledge about empirical relationships and the future can be expected to substantially differ from the past, the number of variables and presence or absence of intercorrelation between them, and the number of observations determine which causal method to use. For example, regression models that rely on non-experimental data can typically use no more than three or four variables – even with massive sample sizes. For problems involving many causal variables, variable weights should not be estimated from the dataset. Instead it is useful to draw on independent sources of evidence (such as empirical studies and prior expert knowledge) for assessing the impact of each variable on the situation.

The Forecasting Method Selection Tree provides guidance but on its own, the guidance is not comprehensive. Forecasters may have difficulty identifying the conditions that apply. In such situations, one should use different methods that draw on different information and combine their forecasts according to pre-specified rules. Armstrong (2001c) conducted a meta-analysis of 30 studies and estimated that the combined forecast yielded a 12% reduction in error compared to the average error of the components; the reductions of forecast error ranged from 3% to 24%. In addition, the combined forecasts were often more accurate than the most accurate component. Studies since that meta-analysis suggest that under favorable conditions (many forecasts available for a number of different valid methods and data sources when forecasting for an uncertain situation), the error reductions from combining are much larger. Simple averages are a good starting point but differential weights may be used if there is strong evidence about the relative accuracy of the method. Combining forecasts is especially useful if the forecaster wants to avoid large errors and if there is uncertainty which method will be most accurate.

The final issue is whether there is important information that has not been incorporated in the forecasting methods. This includes situations in which recent events are not reflected in the data, experts possess good domain knowledge about future events or changes, or key variables could not be included in the model. In the absence of these conditions, one should not adjust the forecast. If important information has been omitted and adjustments are needed, one should use a structured approach. That is, provide written instructions, solicit written adjustments, request adjustments from a group of experts, ask for adjustments to be made prior to seeing the forecast, record reasons for the revisions, and examine prior forecast errors.

Forecasting Canon

The Forecasting Canon provides a summary of evidence-based forecasting knowledge, in this case in the form of



Forecasting Principles. Fig. 2 Selection tree

nine overarching principles that can help to improve forecast accuracy. The principles are often ignored by organizations, so attention to them offers substantial opportunities for gain.

Match the Forecasting Method to the Situation

Conditions for forecasting problems vary. No single best method works for all situations. The Forecasting Method Selection Tree (Fig. 2) can help identify appropriate forecasting methods for a given problem. The recommendations in the Selection Tree are based on expert judgment grounded in research studies. Interestingly, generalizations based on empirical evidence sometimes conflict with common beliefs about which forecasting method is best.

Use Domain Knowledge

Managers and analysts typically have useful knowledge about situations. While this domain knowledge can be important for forecasting, it is often ignored. Methods that are not well designed to incorporate domain knowledge include exponential smoothing, stepwise regression, ►data mining and ►neural networks.

Managers' expectations are particularly important when their knowledge about the direction of the trend in a time series conflicts with historical trends in the data (called "contrary series"). If one ignores domain knowledge about contrary series, large errors are likely.

A simple rule can be used to obtain much of the benefit of domain knowledge: when one encounters a contrary series, do not extrapolate a trend. Instead, extrapolate the latest value – this approach is known as the naive or no-change model.

Structure the Problem

One of the basic strategies in management research is to break a problem into manageable pieces, solve each piece, then put them back together. This decomposition strategy is effective for forecasting, especially when there is more knowledge about the pieces than about the whole. Decomposition is particularly useful when the forecasting task involves extreme (very large or very small) numbers.

When contrary series are involved and the components of the series can be forecasted more accurately than the global series, using causal forces to decompose the problem increases forecasting accuracy. For example, to forecast the number of people who die on the highways each year, forecast the number of passenger miles driven (a series that is expected to grow) and the death rate per million passenger miles (a series that is expected to decrease) and then multiply these forecasts.

Model the Experts' Forecasts

Expert systems represent forecasts made by experts and can reduce the costs of repetitive forecasts while improving accuracy. However, expert systems are expensive to develop.

An inexpensive alternative to expert systems is judgmental bootstrapping. The general proposition borders on the preposterous; it is that a simple model of the man will be more accurate than the man. The reasoning is that the model applies the man's rules more consistently than the man can.

Represent the Problem Realistically

Start with the situation and develop a realistic representation. This generalization conflicts with common practice, in which one starts with a model and attempt to generalize to the situation. Realistic representations are especially important when forecasts based on unaided judgment fail. Simulated interaction is especially useful for developing a realistic representation of a problem.

Use Causal Models When You Have Good Information

Good information means that the forecaster (1) understands the factors that have an influence on the variable to forecast and (2) possesses enough data to estimate a regression model. To satisfy the first condition, the analyst can obtain knowledge about the situation from domain knowledge and from prior research. Thus, for example, an analyst can draw upon quantitative summaries of research (meta-analyses) on price or advertising elasticities when developing a sales-forecasting model. An important advantage of causal models is that they reveal the effects of alternative decisions on the outcome, such as the effects of different prices on sales. Index models are a good alternative when there are many variables and insufficient data for regression analysis.

Use Simple Quantitative Methods

Complex models are often misled by noise in the data, especially in uncertain situations. Thus, using simple methods is important when there is much uncertainty about the situation. Simple models are easier to understand than complex models, and are less prone to mistakes. They are also more accurate than complex models when forecasting for complex and uncertain situations – which is the typical situation for the social sciences.

Be Conservative When Uncertain

One should make conservative forecasts for uncertain situations. For cross-sectional data, this means staying close to

the typical behavior (often called the "base rate"). In time series, one should stay close to the historical average. If the historical trend is subject to variations, discontinuities, and reversals, one should be cautious with extrapolating the historical trend. Only when a historical time series show a long steady trend with little variation should one extrapolate the trend into the future.

Combine Forecasts

Combining is especially effective when different forecasting methods are available. Ideally, one should use as many as five different methods, and combine their forecasts using a predetermined mechanical rule. Lacking strong evidence that some methods are more accurate than others, one should use a simple average of forecasts.

Conclusion

This entry gives an overview of methods and principles that are known to reduce forecast error. The Forecasting Method Selection Tree provides guidance for which method to use under given conditions. The Forecasting Canon can be used as a simple checklist to improve forecast accuracy. Further information and support for evidence-based forecasting is available from the *Principles of Forecasting* handbook and from the ForecastingPrinciples.com Internet site.

About the Authors

Dr. Green is a developer of the simulated interactions and structured analogies methods, which have been shown to provide more accurate forecasts of decisions in conflicts than does expert judgment, including the judgments of game theorists. He has been consulted by the U.S. Department of Defense and National Security Agency on forecasting matters. He is co-director of the Forecasting Principles site (ForPrin.com).

Dr. Graefe is the prediction markets editor of *Foresight - The International Journal of Applied Forecasting*. He is currently developing and testing the index method as an alternative to regression analysis, with applications to election forecasting.

Dr. Armstrong has been involved in forecasting since 1960. He has published *Long-Range Forecasting* (1978, 1985) and *Principles of Forecasting* (2001). He is a co-founder of the *Journal of Forecasting* (1982), the *International Journal of Forecasting* (1985), and the *International Symposium on Forecasting* (1981). He is a developer of new forecasting methods including: rule-based forecasting and causal forces for extrapolation. His book, *Persuasive Advertising* was published in 2010.

Cross References

►Business Forecasting Methods
►Forecasting: An Overview
►Time Series

References and Further Reading

Armstrong JS (1985) Long-range forecasting. Wiley, New York
Armstrong JS (2006) Findings from evidence-based forecasting: methods for reducing forecast error. Int J Forecasting 22: 583–598
Armstrong JS (2001a) Judgmental bootstrapping: inferring experts' rules for forecasting. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 171–192
Armstrong JS (2001b) Extrapolation for time-series and cross-sectional data. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 217–243
Armstrong JS (2001c) Combining forecasts. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 417–440
Collopy F, Armstrong JS (1992) Rule-based forecasting: development and validation of an expert systems approach to combining time-series extrapolations. Manage Sci 38:1394–1414
Collopy F, Adya M, Armstrong JS (2001) Expert systems for forecasting. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 285–300
Green KC (2005) Game theory, simulated interaction, and unaided judgement for forecasting decisions in conicts: further evidence. Int J Forecasting 21:463–472
Green KC, Armstrong JS (2007) Structured analogies for forecasting. Int J Forecasting 23:365–376
Green KC, Armstrong JS, Graefe A (2007) Methods to elicit forecasts from groups: Delphi and prediction markets compared. Foresight Int J Appl Forecasting 8:17–20
MacGregor DG (2001) Decomposition in judgmental forecasting and estimation. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 107–124
Makridakis S, Hibon M (2000) The M-3 competition: results, conclusions and implications. Int J Forecasting 16:451–476
Makridakis S, Andersen S, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition. J Forecasting 1:111–153
Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 125–144

Forecasting with ARIMA Processes

William W. S. Wei
Professor
Temple University, Philadelphia, PA, USA

One of the most important objectives in the analysis of a time series is to forecast its future values. Let us consider

the time series Z_t from the general ARIMA(p,d,q) process

$$\phi(B)(1-B)^d \dot{Z}_t = \theta(B)a_t, \tag{1}$$

where $\dot{Z}_t = (Z_t - \mu)$ if $d = 0$ and $\dot{Z}_t = Z_t$ when $d \neq 0$, $\phi(B) = (1-\phi_1 B - \cdots - \phi_p B^p)$, $\theta(B) = (1-\theta_1 B - \cdots - \theta_q B^q)$, $\phi(B) = 0$ and $\theta(B) = 0$ share no common roots that lie outside of the unit circle, and the series a_t is a Gaussian $N\left(0, \sigma_a^2\right)$ white noise process.

Minimum Mean Square Error Forecasts and Forecast Limits

Our objective is to derive a forecast with as small an error as possible. Thus, our optimum forecast will be the forecast that has the minimum mean square forecast error. Let us consider the case when $d = 0$ in Eq. (1), and express the process in the moving average representation

$$\dot{Z}_t = \psi(b)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \tag{2}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \theta(B)/\phi(B)$, and $\psi_0 = 1$. More specifically, the ψ_j can be obtained from equating the coefficients of B^j on the both sides of

$$\begin{aligned}(1-\phi_1 B - \cdots - \phi_p B^p)(1+\psi_1 B + \psi_2 B^2 + \cdots)\\ = (1-\theta_1 B - \cdots - \theta_q B^q).\end{aligned} \tag{3}$$

For $t = n + \ell$, we have $\dot{Z}_{n+\ell} = \sum_{j=0}^{\infty} \psi_j a_{n+\ell-j}$. Suppose that at time $t = n$ we have the observations $\dot{Z}_n, \dot{Z}_{n-1}, \dot{Z}_{n-2}, \ldots$ and wish to forecast ℓ-step ahead future values of $\dot{Z}_{n+\ell}$ as a linear combination of the observations $\dot{Z}_n, \dot{Z}_{n-1}, \dot{Z}_{n-2}, \ldots$. Since $\dot{Z}_t$ for $t \leq n$ can all be written in the form of (2), we can let the minimum mean square error forecast $\hat{\dot{Z}}_n(\ell)$ of $\dot{Z}_{n+\ell}$ be

$$\hat{\dot{Z}}_n(\ell) = \psi^*_\ell a_n + \psi^*_{\ell+1} a_{n-1} + \psi^*_{\ell+2} a_{n-2} + \cdots \tag{4}$$

where the ψ^*_j are to be determined. The mean square error of the forecast is

$$E\left[\dot{Z}_{n+\ell} - \hat{\dot{Z}}_n(\ell)\right]^2 = \sigma_a^2 \sum_{j=0}^{\ell-1} \psi_j^2 + \sigma_a^2 \sum_{j=0}^{\infty} \left[\psi_{\ell+j} - \psi^*_{\ell+j}\right]^2,$$

which is easily seen to be minimized when $\psi^*_{\ell+j} = \psi_{\ell+j}$. Hence,

$$\hat{\dot{Z}}_n(\ell) = \psi_\ell a_n + \psi_{\ell+1} a_{n-1} + \psi_{\ell+2} a_{n-2} + \cdots = E(\dot{Z}_{n+\ell} | \dot{Z}_t, t \leq n). \tag{5}$$

$\hat{\dot{Z}}_n(\ell)$ is usually read as the ℓ-*step* ahead forecast of $\dot{Z}_{n+\ell}$ at the forecast origin n.

The forecast error is

$$e_n(\ell) = \dot{Z}_{n+\ell} - \hat{\dot{Z}}_n(\ell) = \sum_{j=0}^{\ell-1} \psi_j a_{n+\ell-j}. \tag{6}$$

Because $E(e_n(\ell)) = 0$ the forecast is unbiased with the error variance

$$Var(e_n(\ell)) = \sigma_a^2 \sum_{j=0}^{\ell-1} \psi_j^2. \tag{7}$$

For a normal process, the $100(1-\alpha)\%$ forecast limits are

$$\hat{\dot{Z}}_n(\ell) \pm N_{\alpha/2} \left[1 + \sum_{j=0}^{\ell-1} \psi_j^2\right]^{1/2} \sigma_a, \tag{8}$$

where $N_{\alpha/2}$ is the standard normal deviate such that $P(N > N_{\alpha/2}) = \alpha/2$.

For a general ARIMA model in (1) with $d \neq 0$ the moving average representation does not exist because when we obtain the ψ_j from equating the coefficients of B^j on the both sides of

$$\begin{aligned}(1-\phi_1 B - \cdots - \phi_p B^p)(1-B)^d (1+\psi_1 B + \psi_2 B^2 + \cdots)\\ = (1-\theta_1 B - \cdots - \theta_q B^q),\end{aligned} \tag{9}$$

the resulting series of ψ_j coefficients is not convergent. However, for practical purposes, one can use Eq. (9) to find a finite number of the ψ_j coefficients. The minimum mean square error forecast is also given by $E(\dot{Z}_{n+\ell} | \dot{Z}_t, t \leq n)$ directly through the use of Eq. (1), and Eqs. (6), (7), and (8) hold also for the general ARIMA process. The main difference between the ARMA and ARIMA processes is that $\lim_{\ell \to \infty} \sum_{j=0}^{\ell-1} \psi_j^2$ exists for a stationary ARMA process but does not exist for a nonstationary ARIMA process. Hence, the eventual forecast limits for a stationary process approach two horizontal lines. For a nonstationary process since $\sum_{j=0}^{\ell-1} \psi_j^2$ increases as ℓ increases, the forecast limits become wider and wider. It implies that the forecaster becomes less certain about the result as the forecast lead time gets larger.

Computation of Forecasts

The general ARIMA process in Eq. (1) can be written as

$$(1-\Psi_1 B - \cdots - \Psi_{p+d} B^{p+d}) \dot{Z}_t = (1-\theta_1 B - \cdots - \theta_q B^q) a_t, \tag{10}$$

where $(1-\Psi_1 B - \cdots - \Psi_{p+d} B^{p+d}) = \phi(B)(1-B)^d$. For $t = n+\ell$ we have

$$\begin{aligned}\dot{Z}_{n+\ell} = \Psi_1 Z_{n+\ell-1} + \cdots + \Psi_{p+d} Z_{n+\ell-p-d} + a_{n+\ell}\\ - \theta_1 a_{n+\ell-1} - \cdots - \theta_q a_{n+\ell-q}.\end{aligned}$$

Taking the conditional expectation at time origin n, we get

$$\begin{aligned}\hat{\dot{Z}}_n(\ell) = \Psi_1 \hat{\dot{Z}}_n(\ell-1) + \cdots + \Psi_{p+d} \hat{\dot{Z}}_n(\ell-p-d) + \hat{a}_n(\ell)\\ - \theta_1 \hat{a}_n(\ell-1) - \cdots - \theta_q \hat{a}_n(\ell-q),\end{aligned} \tag{11}$$

where

$$\hat{\dot{Z}}_n(j) = E(\dot{Z}_{n+j}|\dot{Z}_t, t \le n), \qquad j \ge 1,$$
$$\hat{\dot{Z}}_n(j) = \dot{Z}_{n+j}, \qquad j \le 0,$$
$$\hat{a}_n(j) = 0, \qquad j \ge 1,$$

and

$$\hat{a}_n(j) = \dot{Z}_{n+j} - \hat{\dot{Z}}_{n+j-1}(1) = a_{n+j}, \quad j \le 0.$$

Updating Forecasts

Note that from Eq. (6), we have

$$\begin{aligned} e_n(\ell+1) &= \dot{Z}_{n+\ell+1} - \hat{\dot{Z}}_n(\ell+1) \\ &= \sum_{j=0}^{\ell} \psi_j a_{n+\ell+1-j} = e_{n+1}(\ell) + \psi_\ell a_{n+1} \\ &= \dot{Z}_{n+\ell+1} - \hat{\dot{Z}}_{n+1}(\ell) + \psi_\ell a_{n+1}. \end{aligned}$$

Hence, we obtain the equation for updating forecasts,

$$\hat{\dot{Z}}_{n+1}(\ell) = \hat{\dot{Z}}_n(\ell+1) + \psi_\ell[\dot{Z}_{n+1} - \hat{\dot{Z}}_n(1)]. \tag{12}$$

Eventual Forecast Functions

When $\ell > q, \hat{\dot{Z}}_n(\ell)$ in Eq. (11) becomes

$$\Psi(B)\hat{\dot{Z}}_n(\ell) = 0, \tag{13}$$

where $\Psi(B) = \phi(B)(1-B)^d = 1 - \Psi_1 B - \cdots - \Psi_{p+d}B^{p+d}$, and $B\hat{\dot{Z}}_n(\ell) = \hat{\dot{Z}}_n(\ell-1)$. Thus, we can use the difference equation result to obtain the eventual forecast function. That is, if $\Psi(B) = \prod_{i=1}^{K}(1-R_iB)^{m_i}$ with $\sum_{i=1}^{K} m_i = (p+d)$, then

$$\hat{\dot{Z}}_n(\ell) = \sum_{i=1}^{K}\Big(\sum_{j=0}^{m_i-1} c_{ij}\ell^j\Big)R_i^\ell, \tag{14}$$

for $\ell \ge (q-p-d+1)$ where c_{ij} are constants that are functions of time origin n and known data.

An illustrative example: $(1-\phi_1B)(Z_t - \mu) = (1-\theta_1B)a_t$.

a. Computation of $\hat{Z}_n(\ell)$

For $t = n+\ell, Z_{n+\ell} = \mu + \phi_1(Z_{n+\ell-1} - \mu) + a_{n+\ell} - \theta_1 a_{n+\ell-1}$.

Hence

$$\hat{Z}_n(1) = E(Z_{n+1}|Z_t, t \le n) = \mu + \phi_1(Z_n - \mu) - \theta_1 a_t,$$

and

$$\begin{aligned} \hat{Z}_n(\ell) &= E(Z_{n+\ell}|Z_t, t \le n) \\ &= \mu + \phi_1[\hat{Z}_n(\ell-1) - \mu] \\ &= \mu + \phi_1^\ell(Z_n - \mu) - \phi_1^{\ell-1}\theta_1 a_n, \ell \ge 2. \end{aligned}$$

b. The Forecast Error Variance and Forecast Limits

From Eq. (3), $(1-\phi_1B)(1+\psi_1B+\psi_2B^2+\cdots) = (1-\theta_1B)$, and equating the coefficients of B^j on both sides, we get $\psi_j = \phi_1^{j-1}(\phi_1 - \theta_1), j \ge 1$. So

$$\hat{Z}_n(\ell) \pm N_{\alpha/2}\left[1 + \sum_{j=0}^{\ell-1}\left[\phi_1^{j-1}(\phi_1-\theta_1)\right]^2\right]^{1/2}\sigma_a.$$

defines the forecast limits.

c. The Eventual Forecast Function

Since $(1-\phi_1B)(\hat{Z}_n(\ell) - \mu) = 0, \ell \ge 1$, and $|\phi_1| < 1$ we have $\hat{Z}_n(\ell) = \mu + c_1\phi_1^\ell \to \mu$ as $\ell \to \infty$. For more detailed discussions and illustrative examples on time series forecasting, we refer readers to Box, Jenkins, and Reinsel (2008), and Wei (2006).

About the Author

For biography *see* the entry ▶Time Series Regression.

Cross References

▶Box–Jenkins Time Series Models
▶Forecasting: An Overview
▶Structural Time Series Models
▶Time Series

References and Further Reading

Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis: forecasting and control, 4th edn. Wiley, New York
Wei WWS (2006) Time Series Analysis–Univariate and Multivariate Methods, 2nd edn. Pearson Addison-Wesley, Boston

Forecasting: An Overview

Rob J. Hyndman
Professor of Statistics
Monash University, Melbourne, VIC, Australia

What Can Be Forecast?

Forecasting is required in many situations: deciding whether to build another power generation plant in the next 5 years requires forecasts of future demand; scheduling staff in a call centre next week requires forecasts of call volume; stocking an inventory requires forecasts of stock requirements. Forecasts can be required several years in advance (for the case of capital investments), or only a few minutes beforehand (for telecommunication routing). Whatever the circumstances or time horizons involved, forecasting is an important aid in effective and efficient planning.

Some things are easier to forecast than others. The time of the sunrise tomorrow morning can be forecast very precisely. On the other hand, currency exchange rates are very difficult to forecast with any accuracy. The predictability of an event or a quantity depends on how well we understand the factors that contribute to it, and how much unexplained variability is involved.

Forecasting situations vary widely in their time horizons, factors determining actual outcomes, types of data patterns, and many other aspects. Forecasting methods can be very simple such as using the most recent observation as a forecast (which is called the "naïve method"), or highly complex such as ►neural networks and econometric systems of simultaneous equations. The choice of method depends on what data are available and the predictability of the quantity to be forecast.

Forecasting Methods

Forecasting methods fall into two major categories: quantitative and qualitative methods.

Quantitative forecasting can be applied when two conditions are satisfied:

1. numerical information about the past is available;
2. it is reasonable to assume that some aspects of the past patterns will continue into the future.

There is a wide range of quantitative forecasting methods, often developed within specific disciplines for specific purposes. Each method has its own properties, accuracies, and costs that must be considered when choosing a specific method.

Qualitative forecasting methods are used when one or both of the above conditions does not hold. They are also used to adjust quantitative forecasts, taking account of information that was not able to be incorporated into the formal statistical model. These are not purely guesswork – there are well-developed structured approaches to obtaining good judgmental forecasts. However, as qualitative methods are non-statistical, they will not be considered further in this article.

Explanatory Versus Time Series Forecasting

Quantitative forecasts can be largely divided into two classes: time series and explanatory models. Explanatory models assume that the variable to be forecasted exhibits an explanatory relationship with one or more other variables. For example, we may model the electricity demand (ED) of a hot region during the summer period as

$$\text{ED} = f(\text{current temperature, strength of economy, population, time of day, day of week, error}). \quad (1)$$

The relationship is not exact – there will always be changes in electricity demand that can not be accounted for by the variables in the model. The "error" term on the right allows for random variation and the effects of relevant variables not included in the model. Models in this class include regression models, additive models, and some kinds of neural networks.

The purpose of the explanatory model is to describe the form of the relationship and use it to forecast future values of the forecast variable. Under this model, any change in inputs will affect the output of the system in a predictable way, assuming that the explanatory relationship does not change.

In contrast, time series forecasting uses only information on the variable to be forecast, and makes no attempt to discover the factors affecting its behavior. For example,

$$\text{ED}_{t+1} = f(\text{ED}_t, \text{ED}_{t-1}, \text{ED}_{t-2}, \text{ED}_{t-3}, \ldots, \text{error}), \quad (2)$$

where t is the present hour, $t + 1$ is the next hour, $t - 1$ is the previous hour, $t - 2$ is two hours ago, and so on. Here, prediction of the future is based on past values of a variable and/or past errors, but not on explanatory variables which may affect the system. Time series models used for forecasting include ARIMA models, exponential smoothing and structural models.

There are several reasons for using a time series forecast rather than an explanatory model for forecasting. First, the system may not be understood, and even if it was understood it may be extremely difficult to measure the relationships assumed to govern its behavior. Second, it is necessary to predict the various explanatory variables in order to be able to forecast the variable of interest, and this may be too difficult. Third, the main concern may be only to predict what will happen and not to know why it happens.

A third type of forecasting model uses both time series and explanatory variables. For example,

$$\text{ED}_{t+1} = f(\text{ED}_t, \text{current temperature, time of day, day of week, error}). \quad (3)$$

These types of models have been given various names in different disciplines. They are known as dynamic regression models, panel data models, longitudinal models, transfer function models, and linear system models (assuming f is linear).

The Basic Steps in a Forecasting Task

There are usually five basic steps in any forecasting task.

Step 1: Problem definition. Often this is most difficult part of forecasting. Defining the problem carefully requires

an understanding of how the forecasts will be used, who requires the forecasts, and how the forecasting function fits within the organization requiring the forecasts. A forecaster needs to spend time talking to everyone who will be involved in collecting data, maintaining databases, and using the forecasts for future planning.

Step 2: Gathering information. There are always at least two kinds of information required: (a) statistical data, and (b) the accumulated expertise of the people who collect the data and use the forecasts. Often, a difficulty will be obtaining enough historical data to be able to fit a good statistical model. However, occasionally, very old data will not be so useful due to changes in the system being forecast.

Step 3: Preliminary (exploratory) analysis. Always start by graphing the data. Are there consistent patterns? Is there a significant trend? Is seasonality important? Is there evidence of the presence of business cycles? Are there any ▶outliers in the data that need to be explained by those with expert knowledge? How strong are the relationships among the variables available for analysis?

Step 4: Choosing and fitting models. Which model to use depends on the availability of historical data, the strength of relationships between the forecast variable and any explanatory variables, and the way the forecasts are to be used. It is common to compare two or three potential models.

Step 5: Using and evaluating a forecasting model. Once a model has been selected and its parameters estimated, the model is to be used to make forecasts. The performance of the model can only be properly evaluated after the data for the forecast period have become available. A number of methods have been developed to help in assessing the accuracy of forecasts as discussed in the next section.

Forecast Distributions

All forecasting is about estimating some aspects of the conditional distribution of a random variable. For example, if we are interested in monthly sales denoted by y_t for month t, then forecasting concerns the distribution of y_{t+h} conditional on the values of $y_1, \ldots, y_t$ along with any other information available. Let $\mathcal{I}_t$ denote all other information available at time t. Then we call the distribution of $(y_{t+h} \mid y_1, \ldots, y_t, \mathcal{I}_t)$ the *forecast distribution*.

Typically, a forecast consists of a single number (known as a "point forecast"). This can be considered an estimate of the mean or median of the forecast distribution. It is often useful to provide information about forecast uncertainty as well in the form of a prediction interval. For example, if the forecast distribution is normal with mean $\hat{y}_{t+h}$ and variance σ^2_{t+h}, then a 95% prediction interval for y_{t+h} is $\hat{y}_{t+h} \pm 1.96\sigma^2_{t+h}$. Prediction intervals in forecasting are sometimes called "interval forecasts."

For some problems, it is also useful to estimate the forecast distribution rather than assume normality or some other parametric form. This is called "density forecasting."

Evaluating Forecast Accuracy

It is important to evaluate forecast accuracy using genuine forecasts. That is, it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model. When choosing models, it is common to use a portion of the available data for testing, and use the rest of the data for fitting the model. Then the testing data can be used to measure how well the model is likely to forecast on new data.

The issue of measuring the accuracy of forecasts from different methods has been the subject of much attention. We summarize some of the approaches here. A more thorough discussion is given by Hyndman and Koehler (2006). In the following discussion, $\hat{y}_t$ denotes a forecast of y_t. We only consider the evaluation of point forecasts. There are also methods available for evaluating interval forecasts and density forecasts (Corradi and Swanson 2006).

Scale-Dependent Errors

The forecast error is simply $e_t = y_t - \hat{y}_t$ which is on the same scale as the data. Accuracy measures that are based on e_t are therefore scale-dependent and cannot be used to make comparisons between series that are on different scales.

The two most commonly used scale-dependent measures are based on the absolute error or squared errors:

$$\text{Mean absolute error (MAE)} = \text{mean}(|e_t|),$$

$$\text{Mean squared error (MSE)} = \text{mean}\left(e_t^2\right).$$

When comparing forecast methods on a single series, the MAE is popular as it is easy to understand and compute.

Percentage Errors

The percentage error is given by $p_t = 100e_t/y_t$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance between different data sets. The most commonly used measure is:

$$\text{Mean absolute percentage error (MAPE)} = \text{mean}(|p_t|)$$

Measures based on percentage errors have the disadvantage of being infinite or undefined if $y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any y_t is close to zero. Another problem with percentage errors that is often overlooked is that they assume a meaningful zero. For example, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales.

They also have the disadvantage that they put a heavier penalty on positive errors than on negative errors. This observation led to the use of the so-called "symmetric" MAPE proposed by Armstrong (1985, p. 348), which was used in the M3 forecasting competition (Makridakis and Hibon 2000). It is defined by

$$\text{Symmetric mean absolute percentage error (sMAPE)} = \text{mean}\left(200\,|y_t - \hat{y}_t|/(y_t + \hat{y}_t)\right).$$

However, if y_t is zero, $\hat{y}_t$ is also likely to be close to zero. Thus, the measure still involves division by a number close to zero. Also, the value of sMAPE can be negative, so it is not really a measure of "absolute percentage errors" at all. Hyndman and Koehler (2006) recommend that the sMAPE not be used.

Scaled Errors

The MASE was proposed by Hyndman and Koehler (2006) as an alternative to the MAPE or sMAPE when comparing forecast accuracy across series on different scales. They proposed scaling the errors based on the *in-sample* MAE from the naïve forecast method. Thus, a scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|},$$

which is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast computed in-sample. Conversely, it is greater than one if the forecast is worse than the average one-step naïve forecast computed in-sample. The *mean absolute scaled error* is simply

$$\text{MASE} = \text{mean}(|q_t|).$$

About the Author

Rob Hyndman is Professor of Statistics and Director of the Business and Economic Forecasting Unit at Monash University, Australia. He has published extensively in leading statistical and forecasting journals. He is co-author of the highly regarded international text on business forecasting, *Forecasting: methods and applications* (Wiley, 3rd edition 1998), and more recently *Forecasting with exponential smoothing: a state space approach* (Springer, 2008). Professor Hyndman is Editor-in-Chief of the *International Journal of Forecasting* and was previously Theory and Methods Editor of the *Australian and New Zealand Journal of Statistics* (2001–2004). He was elected to the International Statistical Institute in 2005. In 2007 he was awarded the prestigious Moran Medal from the Australian Academy of Science, for his contributions to statistical research.

Cross References

►Business Forecasting Methods
►Business Statistics
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Forecasting with ARIMA Processes
►Fuzzy Logic in Statistical Data Analysis
►Optimality and Robustness in Statistical Forecasting
►Singular Spectrum Analysis for Time Series
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistics: An Overview
►Time Series

References and Further Reading

Armstrong JS (1985) Long-range forecasting: from crystal ball to computer. Wiley, New York

Corradi V, Swanson NR (2006) Predictive density evaluation. In: Handbook of economic forecasting, North-Holland, Amsterdam

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Int. J. Forecasting* 22(4):679–688

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential Smoothing: the state space approach. Springer - Verlag, Berlin

Makridakis S, Hibon M (2000) The M3-competition: results, conclusions and implications. *Int. J. Forecasting* 16:451–476

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications, 3rd edn. Wiley, New York

Forensic DNA: Statistics in

Wing Kam Fung[1], Yuk Ka Chung[2]
[1]Chair Professor
University of Hong Kong, Hong Kong, China
[2]University of Hong Kong, Hong Kong, China

Introduction

Since its introduction by Sir Alec Jeffreys in 1985, deoxyribonucleic acid (DNA) profiling, or DNA fingerprinting,

has become one of the most important tools in forensic human identification. DNA contains unique genetic information of each organism and can be found in blood, semen, hair/hair root, bone, and body fluids such as saliva and sweat. Theoretically, every individual except for identical twins can be identified by one's unique DNA sequence. However, due to technical limitations, current human identification is not based on fully sequencing the whole genome. Instead, only a number of genetic markers are used, and so the identification cannot be established without any doubt. Statistics thereby plays an important role in assessing the uncertainty in forensic identification and evaluating the weight of DNA evidence.

Random Match Probability

Suppose a crime was committed and a blood stain has been found in the crime scene and a suspect has been identified. The DNA profiles obtained from the crime stain and the blood specimen of the suspect will be compared. A DNA profile is a set of numbers representing the genetic characteristics of the forensic sample, often at nine or more DNA regions called loci. If a perfect match is found between the two DNA profiles, the suspect would not be excluded as a possible contributor to the crime stain. To evaluate the weight of the DNA evidence, the probability that another person would have the same DNA profile will be computed and reported in the courtroom. The smaller the random match probability, the stronger is the evidence to convict the suspect.

A common assumption adopted in the evaluation of the random match probability is that the population is in Hardy–Weinberg equilibrium (HWE), which means the two alleles of a genotype at a particular locus are statistically independent of each other. For example, suppose at a particular locus the alleles found in the profiles of the crime stain and the suspect are in common, say, A_iA_j. The random match probability at this particular locus can be obtained using the product rule as $2p_ip_j$ for $i \neq j$ and p_i^2 for $i = j$, where p_i and p_j are the allele frequencies of A_i and A_j, respectively. Under the assumption of linkage equilibrium, i.e., independence of alleles across all loci, multiplying the individual probabilities over all loci will give the overall random match probability, which is often found as small as one in a million or one in a billion in practice.

In some cases, the suspect is unavailable for typing and a close relative of the suspect is typed instead. In some other cases, the suspect is typed, but the prosecution about who left the crime stain involves a close relative of the suspect. Extensions of the formulas to handle these situations as well as the violation of Hardy–Weinberg and linkage equilibrium are extensively discussed in the literature.

Paternity Determination

Another application of DNA profiling is in kinship determination, which refers to the confirmation of a specific biological relationship between two individuals. In particular, a paternity test determines whether a man is the biological father of an individual. For a standard trio case in which the mother, her child, and the alleged father are typed with DNA profiles denoted by M, C, and AF respectively, the weight of evidence that the alleged father is the biological father of the child is often expressed as a likelihood ratio (LR) of the following hypotheses:

H_p : Alleged father is the biological father of the child.

H_d : The biological father is a random unrelated man.

The LR, also termed as the paternity index (PI) in the context of paternity testing, takes the form

$$LR = PI = \frac{P(Evidence|H_p)}{P(Evidence|H_d)} = \frac{P(M, C, AF|H_p)}{P(M, C, AF|H_d)}.$$

Using some results on conditional probability and the fact that the genotypes of the mother and the alleged father are not affected by the hypotheses, the index can be simplified to

$$PI = \frac{P(C|M, AF, H_p)}{P(C|M, H_d)}.$$

Suppose the genotypes at a particular locus are obtained as $C = A_1A_2$, $M = A_2A_3$, and $AF = A_1A_4$. Since the mother has half chance to pass the allele A_2 to the child and the alleged father also has half chance to pass the allele A_1 to the child under H_p, the numerator of the PI is given by $P(C|M, AF, H_p) = (1/2)(1/2) = 1/4$. Similarly, the denominator can be obtained as $P(C|M, H_d) = p_1(1/2) = p_1/2$ and as a result, $PI = 1/(2p_1)$. The overall paternity index can then be obtained by multiplying the individual PIs over all loci.

It may sometimes be argued that the alleged father is not the biological father of the child, but his relative (say, brother) is, thereby resulting in the following hypotheses:

H_p : Alleged father is the biological father of the child.

H_d : A relative (brother) of the alleged father is the biological father of the child.

The PI can still be computed by using the formula $PI = 1/[2F + 2(1 - 2F)p_1]$, where F is the kinship coefficient between the alleged father and his relative. The kinship coefficient is a measure of the relatedness between two individuals, representing the probability that two randomly sampled genes from each of them are identical. In particular, $F = 1/4$ for full siblings and, therefore, in this case, $PI = 2/(1 + 2p_1)$, which is substantially smaller

than $1/(2p_1)$, indicating that DNA profiling performs less effective in distinguishing paternity among relatives.

DNA Mixture

In practical crime cases, it is not uncommon that the biological traces collected from the crime scene are obtained as mixed stains, especially in rape cases. In general, the evaluation and interpretation of the mixed DNA sample can be very complicated due to many factors including unknown number of contributors and missing alleles. Here, we consider a simple two-person mixture problem in which the DNA mixture is assumed to be contributed by the victim and only one perpetrator. Suppose that, at a particular locus, the mixture sample contains alleles $M = \{A_1, A_2, A_3\}$, the victim has genotype $V = A_1A_2$, and the suspect has genotype $S = A_3A_3$. The following two competing hypotheses, the prosecution and defense hypotheses, about who contributes to the crime stain are considered:

H_p : The victim and the suspect are the contributors.

H_d : The victim and an unknown person are the contributors.

The weight of the evidence can be evaluated by

$$LR = \frac{P(Evidence|H_p)}{P(Evidence|H_d)} = \frac{P(M,V,S|H_p)}{P(M,V,S|H_d)} = \frac{P(M|V,S,H_p)}{P(M|V,H_d)}$$

where the last expression is obtained after some simplifications. Obviously in this case, $P(M|V,S,H_p) = 1$ as the mixture M is contributed by the victim and the suspect under H_p. Under H_d, the unknown person must have at least one A_3 allele but cannot have alleles not present in the mixture $M = \{A_1, A_2, A_3\}$. Therefore, there are only three possible genotypes for the unknown person at this locus: A_1A_3, A_2A_3, and A_3A_3. Under HWE, $P(M|V,H_d) = 2p_1p_3 + 2p_2p_3 + p_3^2$ and therefore the LR is obtained as

$$LR = \frac{1}{2p_1p_3 + 2p_2p_3 + p_3^2}$$

In the above example, the LR can be easily computed because there are only three possible genotypes for the only unknown person. In general for multiple perpetrator cases, the following general defense hypothesis may be considered:

H_d : The contributors are the victim and x unknown individuals.

For $x = 2$, there are 27 possible genotype configurations of the two unknown individuals and it is cumbersome to list them all. Over the years, general method and formulas for evaluating the LR have been developed in the literature to deal with complicated mixture problems, including situations with the presence of relatives or population substructures.

About the Author

Professor Fung is Past President, Hong Kong Statistical Society, (2003–2004), (2004–2005), (2005–2006), (2006–2007), Past Vice President, International Association for Statistical Computing, (2007–2009). He has received the Outstanding Achievement Award, Ministry of Education, China (2009) and Outstanding Researcher award, The University of Hong Kong (2001). He has been elected a Fellow of the Institute of Mathematical Statistics and a Fellow of the American Statistical Association "for significant contributions to robust statistics and forensic statistics, and for leadership in Asia for statistical research and education."

Cross References

►Bioinformatics
►Data Mining
►Medical Research, Statistics in
►Statistical Genetics
►Statistics and the Law

References and Further Reading

Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64:125–140

Brenner C (1997) Symbolic kinship program. Genetics 145:535–542

Evett IW (1992) Evaluating DNA profiles in case where the defense "It is my brother". J Forensic Sci Soc 32:5–14

Evett IW, Weir BS (1998) Interpreting DNA evidence. Sinauer, Sunderland

Fukshansky N, Bär W (2000) Biostatistics for mixed stain: the case of tested relatives of a non-tested suspect. Int J Legal Med 114: 78–82

Fung WK (2003) User-friedly programs for easy calculations in paternity testing and kinship determinations. Forensic Sci Int 136:22–34

Fung WK, Chung YK, Wong DM (2002) Power of exclusion revisited: probability of excluding relatives of the true father from paternity. Int J Legal Med 116:64–67

Fung WK, Hu YQ (2008) Statistical DNA forensics: theory, methods and computation. Wiley, Chichester

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316:76–79

Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J (1997) Interpreting DNA mixtures. J Forensic Sci 42:213–222

Foundations of Probability

Thomas Augustin, Marco E. G. V. Cattaneo
Ludwig Maximilian University, Munich, Germany

Introduction

Probability theory is that part of mathematics that is concerned with the description and modeling of random phenomena, or in a more general – but not unanimously accepted – sense, of any kind of uncertainty. Probability is assigned to random events, expressing their tendency to occur in a random experiment, or more generally to propositions, characterizing the degree of belief in their truth.

Probability is the fundamental concept underlying most statistical analyses that go beyond a mere description of the observed data. In statistical inference, where conclusions from a random sample have to be drawn about the properties of the underlying population, arguments based on probability allow to cope with the sampling error and therefore control the inference error, which is necessarily present in any generalization from a part to the whole. Statistical modeling aims at separating regularities (structure explainable by a model) from randomness. There, the sampling error and all the variation that is not explained by the chosen optimal model are comprised in an error probability as a residual category.

Different Interpretations and Their Consequences for Statistical Inference

The concept of probability has a very long history (see, e.g., Vallverdú 2010). Originally, the term had a more philosophical meaning, describing the degree of certainty or the persuasive power of an argument. The beginnings of a more mathematical treatment of probability are related to considerations of symmetry in games of chance (see, e.g., Hald 2003). The scope of the theory was extended by Bernoulli (1713), who applied similar symmetry considerations in the study of epistemic probability in civil, moral, and economic problems. In this connection he proved his "law of large numbers," which can be seen as the first theorem of mathematical statistics, and as a cornerstone of the *frequentist* interpretation of probability, which understands the probability of an event as the limit of its relative frequency in an infinite sequence of independent repetitions of a random experiment. Typically, the frequentist (or aleatoric) point of view is *objectivist* in the sense that it relates probability to random phenomena only and perceives probability as a property of the random experiment (e.g., rolling a dice) under consideration.

In contrast, the second of the two most common interpretations (see, e.g., Peterson (2010), for more details), the *subjective*, personalistic, or epistemic viewpoint, perceives probability as a property of the subject confronted with uncertainty. Consequently, here probability can be assigned to anything the very subject is uncertain about, and the question of whether or not there is an underlying random process vanishes. For the interpretation, in the tradition of Savage (1954) a fictive scenario is used where preferences between actions are described. In particular, the probability of an event is understood as the price at which the subject is indifferent between buying and selling a security paying 1 when the event occurs (and 0 otherwise).

The interpretation of probability predetermines to a considerable extent the choice of the statistical inference methods to learn the unknown parameters ϑ of a statistical model from the data. The frequentist perceives ϑ as an unknown but fixed quantity and seeks methods that are optimal under fictive infinite repetitions of the statistical experiment, while for the subjectivist it is straightforward to express his or her uncertainty about ϑ by a (*prior*) probability distribution, which is then, in the light of new data, updated by the so-called Bayes' rule to obtain the (*posterior*) probability distribution describing all her/his knowledge about ϑ (*Bayesian inference*).

Kolmorgorov's Axioms

While the interpretation of probability is quite important for statistical applications, the mathematical theory of probability can be developed almost independently of the interpretation of probability. The foundations of the modern theory of probability were laid by Kolmogorov (1933) in measure theory: Probability is axiomatized as a normalized measure.

More specifically (see, e.g., Merkle (2010) and Rudas (2010) for more details), let Ω be the set of elementary events under consideration (Ω is usually called *sample space*). The events of interest are described as sets of elementary events: it is assumed that they build a σ-algebra $\mathcal{A}$ of subsets of Ω (i.e., $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is nonempty and closed under complementation and countable union). A probability measure on $(\Omega, \mathcal{A})$ is a function $P : \mathcal{A} \to [0,1]$ such that $P(\Omega) = 1$ and

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n) \tag{1}$$

for all sequences of pairwise disjoint events $E_1, E_2, \ldots \in \mathcal{A}$. When Ω is uncountable, a Borel σ-algebra is usually selected as the set $\mathcal{A}$ of events of interest, because the natural choice $\mathcal{A} = \mathcal{P}(\Omega)$ would place too strong limitations

on the probability measure P, at least under the axiom of choice (see, e.g., Solovay (1970)).

Kolmogorov supplemented his axioms by two further basic definitions: the definition of *independence* of events and the definition of *conditional probability* $P(A|B)$ (i.e., the probability of event A given an event B).

From the axioms, fundamental theorems with a strong impact on statistics have been derived on the behavior of independent repetitions of a random experiment (see, e.g., Billingsley (1995) and Schervish (1995) for more details). They include different ►*laws of large numbers* (see above), the *central limit theorem* (see ►Central Limit Theorems), distinguishing the Gaussian distribution as a standard distribution for analyzing large samples, and the *Glivenko–Cantelli theorem* (see ►Glivenko-Cantelli Theorems), formulating convergence of the so-called empirical distribution function to its theoretical counterpart, which means, loosely speaking, that the true probability distribution can be rediscovered in a large sample and thus can be learned from data.

Current Discussion and Challenges

In statistical methodology, for a long time Kolmogorov's measure-theoretic axiomatization of probability theory remained almost undisputed: only countable additivity (1) was criticized by some proponents of the subjective interpretation of probability, such as De Finetti (1974–1975). If countable additivity is replaced by the weaker assumption of finite additivity (i.e., $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ for all pairs of disjoint events $E_1, E_2 \in \mathcal{A}$), then it is always possible to assign a probability to any set of elementary events (i.e., the natural choice $\mathcal{A} = \mathcal{P}(\Omega)$ does not pose problems anymore). However, without countable additivity many mathematical results of measure theory are not valid anymore.

In recent years, the traditional concept of probability has been questioned in a more fundamental way, especially from the subjectivist point of view. On the basis of severe problems encountered when trying to model uncertain expert knowledge in artificial intelligence, the role of probability as the exclusive methodology for handling uncertainty has been rejected (see, e.g., the introduction of Klir and Wierman (1999)). It is argued that traditional probability is only a one-dimensional, too reductionistic view on the multidimensional phenomenon of uncertainty. Similar conclusions (see, e.g., Hsu et al. (2005)) have been drawn in economic decision theory following Ellsberg's seminal experiments (Ellsberg 1961), where the extent of ambiguity (or non-stochastic uncertainty) has been distinguished as a constitutive component of decision making.

Such insights have been the driving force for the development of the theory of *imprecise probability* (see, e.g., Coolen et al. (2010) for a brief survey), comprising approaches that formalize the probability of an event A as an interval $[\underline{P}(A), \overline{P}(A)]$, with the difference between $\overline{P}(A)$ and $\underline{P}(A)$ expressing the extent of ambiguity. Here $\underline{P}$ and $\overline{P}$ are non-additive set-functions, often called *lower* and *upper probabilities*. In particular, Walley (1991) has partially extended De Finetti's framework (De Finetti 1974–1975) to a behavioral theory of imprecise probability, based on an interpretation of probability as possibly differing buying and selling prices, while Weichselberger (2001) has developed a theory of *interval-probability* by generalizing Kolmogorov's axioms.

About the Authors

Dr Thomas Augustin is Professor of Statistics at the Ludwig Maximilian University (LMU), Munich. He is Head of the group "Methodological Foundations of Statistics and their Applications." Dr Marco Cattaneo is Assistant Professor at the Ludwig Maximilian University (LMU), Munich.

Cross References

►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Measure Theory in Probability
►Philosophical Foundations of Statistics
►Philosophy of Probability
►Probability Theory: An Outline
►Probability, History of
►Statistics, History of

References and Further Reading

Bernoulli J (1713) Ars conjectandi. Thurneysen Brothers, Basel
Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York
Coolen FPA, Troffaes M, Augustin T (2010) Imprecise probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
De Finetti B (1974–1975) Theory of probability. Wiley, New York
Ellsberg D (1961) Risk, ambiguity, and the Savage axioms. Quart J Econ 75:643–669
Hald A (2003) A history of probability and statistics and their applications before 1750. Wiley, New York
Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310:1680–1683
Klir GJ, Wierman MJ (1999) Uncertainty-based information. Physica, Heidelberg
Kolmogorov A (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin
Merkle M (2010) Measure theory in probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin

Peterson M (2010) Philosophy of probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Rudas T (2010) Probability theory: an outline. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Savage LJ (1954) The foundations of statistics. Wiley, New York
Schervish MJ (1995) Theory of statistics. Springer, Berlin
Solovay RM (1970) A model of set-theory in which every set of reals is Lebesgue measurable. Ann Math (2nd series) 92:1–56
Vallverdú J (2010) History of probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman & Hall, London
Weichselberger K (2001) Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Physica, Heidelberg

Frailty Model

Paul Janssen[1], Luc Duchateau[2]
[1]Professor, President of the Belgian Statistical Society (2008–2010), Vice-rector of research at UHasselt (2008–2012)
Hasselt University, Diepenbeek, Belgium
[2]Professor and Head, President of the Quetelet Society (Belgian branch of IBS) (2010–2012)
Ghent University, Ghent, Belgium

►Survival data are often clustered; it follows that the independence assumption between event times does not hold. Such survival data occur, for instance, in cancer clinical trials, where patients share the same hospital environment. The shared frailty model can take such clustering in the data into account and provides information on the within cluster dependence. In such a model, the frailty is a measure for the relative risk shared by all observations in the same cluster. The model, a conditional hazard model, is given by

$$\begin{aligned} h_{ij}(t) &= h_0(t) u_i \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right) \\ &= h_0(t) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + w_i\right) \end{aligned}$$

where $h_{ij}(t)$ is the conditional (on u_i or w_i) hazard function for the jth observation ($j = 1, \ldots, n_i$) in the ith cluster ($i = 1, \ldots, s$): $h_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ is the fixed effects vector of dimension p, $\boldsymbol{x}_{ij}$ is the vector of covariates and w_i (u_i) is the random effect (frailty) for the ith cluster. The w_i's (u_i's) are the actual values of a sample from a density $f_W(.)$ ($f_U(.)$). Clustered survival data will be denoted by the observed (event or censoring) times $\boldsymbol{y} = (y_{11}, \ldots, y_{sn_s})^t$ and the censoring indicators $(\delta_{11}, \ldots, \delta_{sn_s})^t$. Textbooks references dealing with shared frailty models include Hougaard (2000) and Duchateau and Janssen (2008).

The one-parameter gamma density function $f_U(u) = \frac{u^{1/\theta-1}\exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}$ (with mean one and variance θ) is often used as frailty density as it simplifies model fitting, especially if a parametric baseline hazard (parameterized by $\boldsymbol{\xi}$) is assumed. The marginal likelihood for the ith cluster of the gamma frailty model can easily be obtained by first writing the conditional likelihood for the ith cluster and by then integrating out the gamma distributed frailty. With $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, we have

$$\begin{aligned} L_{marg,i}(\boldsymbol{\zeta}) = \int_0^\infty \prod_{j=1}^{n_i} \left(h_0(y_{ij}) u_i \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)\right)^{\delta_{ij}} \\ \exp\left(-H_0(y_{ij}) u_i \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)\right) \times \frac{u_i^{1/\theta-1}}{\theta^{1/\theta}\Gamma(1/\theta)} \\ \exp\left(-u_i/\theta\right) du_i \end{aligned}$$

There exists a closed form for this expression. Taking the logarithm and summing over the s clusters we obtain (Klein 1992; Duchateau and Janssen 2008, Chap. 2)

$$\begin{aligned} l_{marg}(\boldsymbol{\zeta}) = \sum_{i=1}^{s} \Bigg[& d_i \log\theta - \log\Gamma(1/\theta) + \log\Gamma(1/\theta + d_i) \\ & - (1/\theta + d_i) \log\left(1 + \theta \sum_{j=1}^{n_i} H_{ij,c}(y_{ij})\right) \\ & + \sum_{j=1}^{n_i} \delta_{ij} \left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta} + \log h_0(y_{ij})\right) \Bigg] \end{aligned} \quad (1)$$

where $H_{ij,c}(y_{ij}) = H_0(y_{ij}) \exp\left(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}\right)$ and $d_i = \sum_{j=1}^{n_i} \delta_{ij}$, the number of events in the ith cluster. The marginal loglikelihood does no longer contain the frailties and can therefore be maximized to obtain parameters estimates $\hat{\boldsymbol{\zeta}}$. The asymptotic variance-covariance matrix can also be obtained using the marginal loglikelihood expression. The preferred model in survival analysis is a (conditional) hazards model with unspecified baseline hazard (a semiparametric model, a Cox model). Leaving $h_0(.)$ and $H_0(.)$ in (1) unspecified we obtain a semiparametric gamma frailty model. For such model direct maximization of the marginal likelihood is not possible. Both the EM-algorithm (Klein 1992) and penalized likelihood maximization (Therneau et al. 2003) have been proposed to fit such models; both approaches use the fact that closed form expressions can be obtained for the expected values of the frailties.

An alternative representation of the marginal likelihood (1) for the parametric gamma frailty model is based on the Laplace transform of the gamma frailty density

$\mathcal{L}(s) = \mathrm{E}(\exp(-Us)) = (1+\theta s)^{-1/\theta}$. With $\boldsymbol{t}_{n_i} = (t_1, \ldots, t_{n_i})$ and $H_{i,c}(\boldsymbol{t}_{n_i}) = \sum_{j=1}^{n_i} H_{ij,c}(t_i)$, the joint survival function for the ith cluster is given by

$$S_{i,f}(\boldsymbol{t}_{n_i}) = \int_0^\infty \exp(-u_i H_{i,c}(\boldsymbol{t}_{n_i})) f_{U_i}(u_i) du_i$$
$$= \mathcal{L}(H_{i,c}(\boldsymbol{t}_{n_i})) = (1+\theta H_{i,c}(\boldsymbol{t}_{n_i}))^{-1/\theta}$$

The likelihood contribution of the ith cluster, with $\boldsymbol{y}_{n_i} = (y_{i1}, \ldots, y_{in_i})$ and the first l observations uncensored, is then

$$(-1)^l \frac{\partial^l}{\partial t_1 \ldots \partial t_l} S_{i,f}(\boldsymbol{y}_{n_i}) = (-1)^l \mathcal{L}^{(l)}(H_{i,c}(\boldsymbol{y}_{n_i}))$$
$$\prod_{j=1}^{l} h_0(y_{ij}) \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}) \qquad (2)$$

For the gamma frailty model the explicit form of (2) is

$$\prod_{j=1}^{n_i} h_{\boldsymbol{x}_{ij},c}^{\delta_{ij}}(y_{ij}) (1+\theta H_{i,c}(\boldsymbol{y}_{n_i}))^{-1/\theta - d_i} \prod_{l=0}^{d_i-1} (1+l\theta)$$

with $\prod_{l=0}^{d_i-1}(1+l\theta) = 1$ for $d_i = 0$.

For frailty densities different from the gamma frailty density, for which the Laplace transform exists, expression (2) is the key to obtain the appropriate marginal loglikelihood expression. Frequently used frailty densities, such as the inverse Gaussian and the positive stable densities (Hougaard 1986a), have indeed simple Laplace transforms. More complex two-parameter frailty densities are the power variance function densities (Hougaard 1986b) and the compound Poisson densities (Aalen 1992). Although the lognormal density is also used as frailty density, it does not have a simple Laplace transform; its use mainly stems from mixed models ideas (McGilchrist and Aisbett 1991), and different techniques, such as ▶numerical integration, have to be used to fit this model (Bellamy et al. 2004).

The choice of the frailty density determines the type of dependence between the observations within a cluster. A global dependence measure is Kendall's τ (Kendall 1938). For two randomly chosen clusters i and k of size two with event times (T_{i1}, T_{i2}) and (T_{k1}, T_{k2}) and no covariates, τ is defined as $\mathrm{E}[\mathrm{sign}((T_{i1}-T_{k1})(T_{i2}-T_{k2}))]$ where $\mathrm{sign}(x) = -1, 0, 1$ for $x < 0, x = 0, x > 0$. Kendall's τ can be expressed as a function of the Laplace transform. Global dependence measures do not allow us to investigate how dependence changes over time. An important local dependence measure is the cross ratio function (Clayton 1978). An interesting feature of this function is its relation with a local version of Kendall's τ (see Duchateau and Janssen 2008, Chap. 4). The positive stable distribution and the ▶gamma distribution characterize early and late dependence respectively, with the ▶inverse Gaussian distribution taking a position in between the two.

So far we discussed the shared frailty model, which is the most simple model to handle within cluster dependence. The shared frailty model can be extended in different ways. First, a frailty term can be assigned to each subject, resulting in a univariate frailty model which can be used to model overdispersion (Aalen 1994). Another extension is the correlated frailty model in which the subjects in a cluster do not share the same frailty term although their respective frailties are correlated (Yashin and Iachine 1995). Finally the model can be extended to multifrailty and multilevel frailty models. In a multifrailty model, two different frailties occur in one and the same cluster. A good example is the study of the heterogeneity of a prognostic index over hospitals in cancer clinical trials, with each hospital (cluster) containing a frailty term for the hospital effect and a frailty term for the prognostic index effect (Legrand et al. 2007). Multilevel frailty models have two or more nesting levels, with a set of smaller clusters contained in a large cluster. Fitting such models is discussed in Ripatti and Palmgren (2000) and Rondeau et al. (2006).

Cross References

▶Demographic Analysis: A Stochastic Approach
▶Hazard Ratio Estimator
▶Hazard Regression Models
▶Modeling Survival Data
▶Survival Data

References and Further Reading

Aalen OO (1992) Modelling heterogeneity in survival analysis by the compound Poisson distribution. Ann Appl Probab 2:951–972

Aalen OO (1994) Effects of frailty in survival analysis. Stat Methods Med Res 3:227–243

Bellamy SL, Li Y, Ryan LM, Lipsitz S, Canner MJ, Wright R (2004) Analysis of clustered and interval-censored data from a community-based study in asthma. Stat Med 23:3607–3621

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. Biometrika 65:141–151

Duchateau L, Janssen P (2008) The frailty model. Springer, New York

Hougaard P (1986a) A class of multivariate failure time distributions. Biometrika 73:671–678

Hougaard P (1986b) Survival models for heterogeneous populations derived from stable distributions. Biometrika 73:387–396

Hougaard P (2000) Analysis of multivariate survival data. Springer, New York

Kendall MG (1938) A new measure of rank correlation. Biometrika 30:81–93

Klein JP (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 48: 795–806

Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R (2007) Validation of prognostic indices using the frailty model. Lifetime Data Anal 15:59–78
McGilchrist CA, Aisbett CW (1991) Regression with frailty in survival analysis. Biometrics 47:461–466
Ripatti S, Palmgren J (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56:1016–1022
Rondeau V, Filleul L, Joly P (2006) Nested frailty models using maximum penalized likelihood estimation. Stat Med 25:4036–4052
Therneau TM, Grambsch PM, Pankratz VS (2003) Penalized survival models and frailty. J Comput Graph Stat 12:156–175
Yashin AI, Iachine IA (1995) Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. Genet Epidemiol 12:529–538

Fraud in Statistics

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

Fraud is an intentional distortion of the truth, whether it is a deliberate omission or false elimination, or an exaggeration or fabrication.

The aim of one who commits fraud is always self-benefit and self-interest. The main reasons for fraud are immorality, impunity, and anarchy, and the methods are deception and betrayal. From these it gives rise to the gravest of crimes – violation of law, murder, mutinies and wars. The tools to overcome fraud are law and order, auditing and control, morals, science, prosecution, and adequate punishment.

Fraud is a man-made phenomenon. The substance of fraud is unknown in the natural world. A lack of knowledge or limited knowledge, unpremeditated actions as well as unobserved phenomena (including deliberate, but legal and justified by jury's verdict, but still obviously criminal actions and perjuries) on the modern level of social mentality do not belong to the substance of fraud, and they are definitely not a subject for a scientific research.

Science, as opposed to jurisprudence, is much more liberal (despite some exceptions for genius of Galileo Galilei, Giordano Bruno, Jan Hus and Nicolaus Copernicus).

The most efficient tool for not only revealing but also overcoming fraud in economy (and further in socio-economic activity) is statistics. Its accurate methods of observation and auditing, powerful databases and knowledge bases, advanced software, and technical provision, as well as the intellectual culture verified by hundreds of years of qualitative data collection and data processing, allow the guarantee of controlled completeness, credibility, and accessibility for a wide range of people.

Being the world's most powerful information system with regulated branches in center and local areas controlled by hundred of quality criteria including those provided by IMF, the modern statistics is by its nature, as any other meter device is free from necessity to lie but at the same time it is surrounded by various kinds of lies and in turn reflects them, and as any other domain of empirical knowledge cannot be free of it.

Fraud in statistics is distortion of data, resulting from two different types of causes: 1) distortion of random errors, caused by poor observation and calculation, the characteristics of which are analyzed in another chapter in this text 2) deliberate (premeditated) distortion of data, resulting from different kinds of systematic causes and giving rise to effects beyond the statistics domain; these cannot be eliminated by methods or techniques.

The main sources of the data distortions (or in simple words – improper data reflections) are unknown, unobservable, and immeasurable phenomena. Such phenomena are not and cannot be discussed due to the objective reasons by observed phenomena; they are actually published and reflected in an incomplete and distorted form and thus they rather characterize themselves but do not reflect the real situation.

The most widespread sources of distortion in modern statistics are:

- evasion from participation in the preparation and submission of the obligatory current statistical reports;
- failure of respondents to answer ▶questionnaire for periodical and random statistical samplings;
- use of obsolete registers of individuals as well as legal entities, omissions, including deliberate omissions of the observed units and the reports units;
- underestimation (or overestimation) of the statistical data registration and reporting.

The particular type of data distortion in modern statistics, statistical estimates, connected with substitution of concepts or estimates obtained with use of inadequate techniques and algorithms that cannot be verified by the existing criteria of their credibility or with the application of other control methods, which are suitable for solving similar class of problems.

The biggest domains of fraud in statistics today are activities that cannot be prohibited and there is little means to prevent these activities. They are as follows:

- all types of illegal activities, including terrorism, counterfeiting, money laundering, corruption, smuggling, drug dealing, arms trafficking, illegal mining of rare metals, trafficking of toxic agents illegal organ transplants, and child abduction;
- illegal activity of individuals and legal entities;
- illegal business of unregistered organizations, institutions, and individuals;
- production and rendering services for self-consumption by households and individuals;
- unrecorded and omitted by statistical observations types of activities, statistical errors, rounding, errors and discrepancies, underestimated or overestimated estimates;
- second-hand activity, tolling, venture enterprises, intellectual activity; intermediate consumption, subsidies for production; offshore activity;
- transactions charges, fees, tips, VAT return, barter, payment in goods for labour, single payments, taxes, losses, including anthropogenic impacts;
- doctoring, imputed value, royalty, goodwill, repeat count of raw materials, commodities, services and capitals;
- other reasons, their identification is imposed and acceptable within the limits of standards and methods of effective statistical reporting and accounts.

The demonstrative example of fraud is fictitious estimates of capitalization of the world markets which, against the background of their real assets estimates (2008) would not exceed $70 trillion USD, and today account for over $700 trillion of USD.

The phantom of fraud in the modern world is also represented by estimates of banks assets. According to these estimates by US Statistics on Banking, 2007 (table 1136) which is considered as the most reliable one, the aggregate assets of all 730,000 of American banks in 2007 were estimated for $14.0 trillion USD (the precise sum is $13,792.5 billion of USD), whilst according to public information the allied assets of JP Morgan Bank solely at the same year were estimated for $97.5 trillion USD, Goldman and Sachs – $50 trillion USD, and HSBC – $108 trillion USD, which exceeded their accounted real equity capital by 30–40 times or more.

Another example of fraud is tax evasion, in particular VAT, the size of which reaches one-third of its total volume in the world, including over $20 billion USD per year in England or $50 billion USD in the United States.

However, there are no ideal measurements or absolutely precise estimations in science and life. Even those which seem to be absolutely accurate values obtained from the variables such as lengths, speed, weight, and temperature (degrees), are just conventional but not the absolute truth itself.

On the other hand, not all inaccurate values (estimates) are distorted ones and hence not all distorted values are false. In accordance with existing criteria in statistics, the inaccurate estimates are such and only distorted estimates that deteriorate the true core of measured phenomena and turn it into its opposite, that is to say, a lie. Inaccurate and reasonably distorted estimates, which by the way prevail in modern statistics (actually they prevailed in the past, too), are called approximated and they are widely used with reserve of some errors as acceptable asymptotic or approximations estimates.

About the Author

For biography *see* the entry ►Actuarial Methods.

Cross References

►Banking, Statistics in
►Misuse of Statistics
►Pyramid Schemes
►Statistical Fallacies
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Keen M, Smith S (2007) VAT fraud and evasion: what do we know, and what can be done. IMF Working Paper. 07/31

Mauro P (2002) The persistence of corruption and slow economic growth. IMF Working Paper /02/213

OECD (2002) Measuring the non-observed economy: a handbook. OECD Publications, Paris

Simchera VM (2003) Statistical information and economic disinformation. Federalism Magazine, Russia, 3:91–116

Simchera VM (2006) Moral economy. TENS Publishing House, Russia

Yehoue EB, Ruhashyankiko JF (2006) Corruption and technology-induced private sector development. IMFWorking Paper /06/198

Frequentist Hypothesis Testing: A Defense

Shlomo Sawilowsky
Professor and Assistant Dean
Wayne State University, Detroit, MI, USA

John Graunt, William Petty, René Déscartes, Blaise Pascal, Pierre Fermat, James Gregory, Christiaan Huygens,

Isaac Newton, Gottfried Leibniz, Jakob Bernoulli, Johann Bernoulli, Abraham DeMoivre, Daniel Bernoulli, Leonhard Euler, Joseph Lagrange, Pierre Laplace, Siméon Poisson, Jean Fourier, Friedrich Bessel, Carl Jacobi, Carl Gauss, Augustin Cauchy, Gustav Dirichlet, Georg Riemann, Michel Chasles, Augustus DeMorgan, Lambert Quetelet, Joseph Liouville, Pafnuty Chebyshev, Charles Hermite, and Andrei Markov. It was the work of these men, among others, that led to the development of the grand theorems, the mathematical inerrancies.

True, they were initially prompted by real world problems, such as winning games of chance or determining actuarial odds; and esoteric problems, such as proving the existence of a Divine plan by confirming a slightly greater proportion of male births to ensure the survival of the species. However, the grand theorems ascended to their elevated place in history because they are elegant, not because they were particularly useful in solving the problems for which they were created.

Moreover, they dictated the types of problems that are considered worthy, relegating those not subsumed under the cleverness of what mankind can solve as being intractable and designated as an eternal mystery of the universe. Their importance was buttressed by their utility for the few problems that they could solve, not for the problems that needed to be solved. Nunnally (1978) wrote mathematics "is purely an abstract enterprise that need have nothing to do with the real world… Thus the statement *iggle wug drang flous* could be a legitimate mathematical statement in a set of rules stating that when any *iggle* is *wugged* it *drang* a *flous*…Of course … [this] might not be of any practical use" (p. 9–10).

Woodward (1906) observed "since the beginning of the eighteenth century almost every mathematician of note has been a contributor to or an expositor of the theory of probability" (p. 8). But the focus on probability eventually moved away from populations and the grand theorems, and settled on just very large samples, such as, e.g., the work of Charles Darwin and Francis Galton.

Darwin collected his data between December 26, 1831 and February 27, 1832 while on the Cherokee class ten gun brig-sloop H. M. S. Beagle, sailing under the command of Captain Robert Fitzroy. Most of Darwin's data were obtained in St. Jago (Santiago) in the Cape Verde Islands from January 16 – February 7. Galton (1885) collected 17 discreet data points on 9,337 people. They were measured in a cubicle 6 feet wide and 36 feet long with the assistance of Serjeant Williams, Mr. Gammage the optician, and a doorkeeper who made himself useful. The data were obtained in the anthropometric laboratory at the International Health Exhibition and subsequently deposited at the South Kensington Museum.

Darwin's and Galton's lack of mathematical training limited their ability to quantify and analyze their trophies, but that limitation was resolved with the brilliance of Karl (née Carl) Pearson. With their data in hand, and the more immediate problem of huge data sets from the biologist/zoologist Walter Weldon, Pearson set to work. By 1900, he provided the rigor that had eluded his colleagues with the discovery of both r and χ^2, and the world was at peace. Well, at least scholars, the intelligencia, and their paparazzi were comforted.

K. Pearson (1978) assuredly knew the limitations of the grand theorems. After all, he quipped

> "As I understand De Moivre the 'Original Design' is the mean occurrence on an indefinite number of trials…The Deity fixed the 'means' and 'chance' provided the fluctuations…There is much value in the idea of the ultimate laws of the universe being statistical laws… [but] it is not an exactly dignified conception of the Deity to suppose him occupied solely with first moments and neglecting second and higher moments!" (p. 160)

But, alas and alack, as the first champion of statistics, K. Pearson was the inheritor of the grand theorems. As a co-founding editor of *Biometrika* he strove to stay above controversy by minimizing, if not ignoring, ordinary problems. And indeed there are those who still pine for the days of yore with its grand theorems, as Tukey (1954) nostalgically noted "Once upon a time the calculation of the first four moments was an honorable art in statistics" (p. 717).

But the ordinary person readily intuited that real world problems are not asymptotic in nature. William Gosset's Monte Carlo study published in 1908 with numbers written on pieces of poster board was conducted because he wasn't sure mathematicians could help him with real, small samples problems.

How could the recipe of his employer, Arthur Guinness, be improved? How many barrels of malt or hops are needed to approximate a population, or at least a large number? 2? 3? Are 4 barrels close to infinity? This chemist (whose sole credential was his undergraduate dual major in chemistry and mathematics from New College, Oxford) sounded all the great mathematical minds of his day, who assured him that he could rely on the grand theorems, and that he need not trouble himself with matters above his pay grade. Daydreams, it seems he was told, were more profitable than the time spent fretting on how large is large.

Gosset ("Student") neither wrote the t formula in the form that it appears in undergraduate textbooks today, nor

did he create the experimental design in which it would be applied. Ronald Fisher provided both the design and the null hypothesis that brought Gosset's Monte Carlo experiments and intuitive mathematics to fruition. Then, Pearson, Gosset, and Fisher became a quintet with the addition of Jerzy Neyman and Egon Pearson, who cemented the frequentist approach to hypothesis testing with the creation of an alternative hypothesis. Not satisfied, Neyman later re-expressed frequentist hypothesis testing into confidence intervals based on the same theory of probability.

Doubts were immediately raised, such as Bowley (1934), who asked and answered, "I am not at all sure that the 'confidence' is not a 'confidence trick'… Does it really take us any further?... I think it does not" (p. 609). Many scholars have adopted the shortcut to notoriety by rushing to follow in Bowley's footsteps, proclaiming the sky is falling on hypothesis testing.

But K. Pearson's development of the χ^2 test is surely listed among the greatest practical achievements of three millennia of mathematical inquiry. He captured the ability, regardless of the direct object, to quantify the difference between human observation and expectation. Remarkable! Was it wrong? Of course. Fisher had to modify the degrees of freedom. Again, remarkable! Was it still wrong? Of course. Frank Yates had to modify the method for small values of expectation. Once again, remarkable! Was it nevertheless wrong? Of course. The correction was found to sap statistical power. Where does the χ^2 test stand today? Statistical software can produce exact p values regardless of how small the expectation per cell. Remarkable!

Has society, therefore, improved with advent of the evolution of the χ^2 test? Most assuredly not:

- We live in a χ^2 society due to political correctness that dictates equality of outcome instead of equality of opportunity. The test of independence version of this statistic is accepted *sans voire dire* by many legal systems as the single most important arbiter of truth, justice, and salvation. It has been asserted that any statistical difference between (often even nonrandomly selected) samples of ethnicity, gender, or other demographic as compared with (often even inaccurate, incomplete, and outdated) census data is *primae faciea* evidence of institutional racism, sexism, or other ism. A plaintiff allegation that is supportable by a significant χ^2 is often accepted by the court (judges and juries) *praesumptio iuris et de iure*. (Sawilowsky 2010).

But is this really the fault of the χ^2 test? Any device can be lethal in the hands of a lunatic, as Mosteller (1968) warned, "I fear that the first act of most social scientists upon seeing a contingency table is to compute a chi-square for it" (p. 1).

What discipline has not followed an evolutionary path? Has agriculture, archeology, architecture, anthropology, biology, botany, chemistry, computer science, education, engineering, genetics, medicine, nursing, pharmacology, physics, psychology, sociology, and zoology always been as they exist today? Do we blame statistics for its ignoble development more so because the content disciplines were dependent on it?

There have been antagonists of Fisher–Neyman/Pearson hypothesis testing for three quarters of a century since Bowley. And it is understandable with the following analogy:

I had a summer job in 1972, working in Florida for a major manufacturer of fiberglass yachts. The hulls of the larger boats were made by laminating two 1/2 hull pieces together. The exterior paint was sprayed inside the two molds and set to dry. Next, fiberglass chop mixed with resin and methyl ethyl ketone peroxide was sprayed into the mold, laminators worked out the air bubbles, and it was set to harden.

The two 1/2 hull shells were then aligned, and held in place with many C clamps with the aid of a powerful air compressor. This was necessary because over time the molds changed in shape and they no longer matched. A small, temporary seam was laminated inside the hull to keep the two parts together. When the clamps were released, one could almost see the two 1/2 hulls trembling, working against the thin fiberglass seam to separate and go their separate ways.

My job was to be lowered inside the hull, and lay down a successively wider series of fiberglass mats and resign/catalyst, to strengthen the seam. On one boat I had laminated perhaps five or six of the required ten mats when it was quitting time. The crew chief told me I could continue the next day where I had left off.

To my chagrin, when I arrive early the next day I discovered that the night shift personnel had taken my hull down the production line, and the boat by now had floors, carpet, sink, and other amenities already installed, obviating the ability to bond the final fiberglass mats to strengthen the hull's seam. I protested to my crew chief, who nonchalantly replied not to worry myself about such things. "After all," he said, "the naval architects who designed the boat allowed considerable tolerance that should handle situations such as this." I made that my last day on the job at that company, and since then I've often wondered how that yacht fared in the middle of the Gulf of Mexico.

So too, the juxtaposition of Fisher's null with Neyman and E. Pearson's alternative leads to trembling, each part of the statistical hypothesis seemingly working against each other to go their separate ways. But this was not the end of the development of the frequentist theory. It surpassed E. Pearson (1962), who admitted "through the lack of close contact with my partner during the last 20 years, it would be a little difficult to say where precisely the Neyman and Pearson theory stands today" (p. 53). The same sentiment was also expressed by those who followed the age of the pioneers, such as Savage (1962) who echoed, "What I, and many other statisticians, call the Neyman-Pearson view may, for all I know, never have been held by Professor Neyman or by Professor Pearson" (p. 62). Wilks (1948) concluded that by now the "modern statistical method is a science in itself" (p. 1).

In truth, many of the foibles in hypothesis testing, since being admitted to the country club of mature disciplines, are traceable back to the statistician, not to the statistics. Fallacies, misconceptions, and myths abound. Which of the disciplines listed above are immune to this, and why is there an expectation that statistics should fare any better?

Yes, even under the best of circumstances there are those who have no use for hypothesis testing. Ernst Rutherford (cited in N T J Bailey 1967) said, "If your experiment needs statistics, you ought to have done a better experiment" (p. 23). But, Albert Einstein (cited in Shankland 1973) countered, "I thank you very much for sending me your careful study about the [Dayton] Miller experiments. Those experiments, conducted with so much care, merit, of course, *a very careful statistical investigation*," (p. 2283, italics added for emphasis).

Much of the criticism against hypothesis testing would presumably vanish if workers heeded the advice of Finney (1953), who advised "when you are experienced enough to make your own statistical analyses, be sure you choose the right technique and not merely any one that you can remember!" (p. 174). The sciences, physical and social, should be placated with McNemar's (1949) advice that "the student should be warned that he cannot expect miracles to be wrought by the use of statistical tools" (p. 3).

Proper selection of statistical tests based on their small samples properties, along with an understanding of their respective null and alternative hypotheses, research design, random sampling, nominal α, Type I and II errors, statistical power, and effect size would eliminate attacks against hypothesis testing from all save perhaps those who, as Bross (1969) characterized it, base their science on "a Bayesian *t*-test using an informationless prior" (p. 52). Has the world benefitted from frequentist hypothesis testing?

> The question is silly. No reputable quantitative physical, behavioral, or social scientist would overlook the breadth and depth of scholarly knowledge and its impact on society that has accrued from over a century of hypothesis testing. The definitive evidence: William Sealy Gosset created the *t* test to make better beer. (Sawilowsky 2003, p. 469)

About the Author

Shlomo S. Sawilowsky is Professor and Assistant Dean in the College of Education, and Wayne State University Distinguished Faculty Fellow. He is the author of *Statistics Through Monte Carlo Simulation With Fortran* (2003) and *Real Data Analysis* (2008), and over 100 peer-reviewed articles on applied data analysis. He is the founding editor of the *Journal of Modern Applied Statistical Methods* (http://tbf.coe.wayne.edu/jmasm). He has served as major professor on 52 doctoral dissertations, Co-advisor on 18 dissertations, 2*nd* advisor on 37 doctoral dissertations, Cognate advisor on 2 doctoral dissertations, and advisor on 23 Master's theses in applied data analysis. Approximately 1/2 of his graduates are female and 1/4 are African American. Professor Sawilowsky has won many teaching and research awards. He was the recipient of the 1998 Wayne State University Outstanding Graduate Mentor Award, and the College of Education's Excellence in Teaching Award. "Professor Sawilowsky's exceptional record as an academician is reflected in the excellence with which he mentors graduate students" (AMSTAT News, October 1998). Professor Sawilowsky was the 2008 President of the American Educational Research Association/SIG Educational Statisticians.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Confidence Interval
►Effect Size
►Full Bayesian Significant Test (FBST)
►Null-Hypothesis Significance Testing: Misconceptions
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Role of Statistics
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference: An Overview

►Statistical Significance
►Statistics: Controversies in Practice

References and Further Reading

Bailey NTJ (1967) The mathematical approach to biology and medicine. Wiley, New York
Bowley A (1934) Discussion on Dr. Neyman's paper. J Roy Stat Soc 97:607–610
Bross I (1969) Applications of probability: science versus pseudoscience. J Am Stat Assoc 64:51–57
Finney D (1953) An introduction to statistical science in agriculture. Ejnar Munksgaard, Copenhagen
Galton F (1885) On the anthropometric laboratory at the late international health exhibition. Journal of the Anthropological Institute of Grand Britain and Ireland, 14:205–221
McNemar Q (1949) Psychological statistics. Wiley, New York
Mosteller F (1968) Association and estimation in contingency tables. J Am Stat Assoc 63:1–28
Nunnally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York
Pearson ES (1962) The foundations of statistical inference. Methuen, London
Pearson K (ed. Pearson ES) (1978) The history of statistics in the 17th and 18th centuries against the changing background of intellectual, scientific and religious thought: Lectures given at University College London during the academic sessions 1921–1923. Macmillan, New York
Savage LJ (1962) The foundations of statistical inference. Methuen, London
Sawilowsky S (2010) Statistical fallacies, misconceptions, and myths, this encyclopedia
Sawilowsky S (2003) Deconstructing arguments from the case against hypothesis testing. J Mod Appl Stat Meth 2(2):467–474
Shankland R (1973) Michelson's role in the development of relativity. Appl Optics 12(10):2280–2287
Tukey JW (1954) Unsolved problems of experimental statistics. J Am Stat Assoc 49:706–731
Wilks SS (1948) Elementary statistical analysis. Princeton University Press, Princeton
Woodward R (1906) Probability and theory of errors. Wiley, New York

Full Bayesian Significant Test (FBST)

Carlos Alberto de Bragança Pereira
Professor, Head, Instituto de Matemática e Estatística
Universidade de São Paulo, São Paulo, Brazil

Introduction

Significance testing of precise (or sharp) hypotheses is an old and controversial problem: it has been central in statistical inference. Both frequentist and Bayesian schools of inference have presented solutions to this problem, not always prioritizing the consideration of fundamental issues such as the meaning of precise hypotheses or the inferential rationale for testing them. The Full Bayesian Significance Test, FBST, is an alternative solution to the problem, which attempts to ease some of the questions met by frequentist and standard Bayes tests based on Bayes factors. FBST was introduced by Pereira and Stern (1999) and reviewed by Pereira et al. (2008).

The discussion here is restricted to univariate parameter and (sufficient statistic) sample spaces;

$$\Theta \subset \mathcal{R} \text{ and } X \subset \mathcal{R}$$

A sharp hypothesis H is then a statement of the form $H : \theta = \theta_0$ where $\theta_0 \in \Theta$. The posterior probability (density) for θ is obtained after the observation of $x \in X$. While a frequentist looks for the set, C, of sample points at least as inconsistent with θ_0 as x is, a Bayesian could look for the tangential set T of parameter points that are more consistent with x than θ_0 is. This understanding can be interpreted as a partial duality between sampling and Bayesian theories. The evidence in favor of H is for frequentists the usual p-value, while for Bayesian it should be $ev = 1 - \underline{ev}$:

$$pv = Pr\{x \in C|\theta_0\} \text{ and } ev = 1 - \underline{ev} = 1 - Pr\{\theta \in T|x\}.$$

The larger pv and ev, the stronger the evidence favoring H.

In the general case, the posterior distribution is sufficient for ev to be calculated, without any complication due to dimensionality of neither the parameter nor of the sample space. This feature ceases the need for nuisance parameters elimination, a problem that disturbs some statisticians (Basu 1977). If one feels that the goal of measuring consistency between data and a null hypothesis should not involve prior opinion about the parameter, the normalized likelihood, if available, may replace the posterior distribution. The computation of ev needs no asymptotic methods, although numerical optimization and integration may be needed.

The fact that the frequentist and Bayesian measures of evidence, pv and ev, are probability values – therefore defined in a zero to one scale – does not easily help to answer the question "How small is *significant*?". For ►p-values, the NP lemma settles the question by means of subjective arbitration of critical values. For Bayesian assessment of significance through evaluation of ev, decision theory again clears the picture. Madruga et al. (2001) show that there exist loss functions the minimization of which render a test of significance based on ev into a formal Bayes test.

The FBST has successfully solved several relevant problems of statistical inference: see Pereira et al. (2008) for a list of publications.

FBST Definition

Significance FBST was created under the assumption that a significance test of a sharp hypothesis had to be performed. At this point, a formal definition of a sharp hypothesis is presented.

Consider general statistical spaces, where $\Theta \subset \mathcal{R}^m$ is the parameter space and $X \subset \mathcal{R}^k$ is the sample space.

Definition 1 A *sharp* hypothesis H states that θ belongs to a sub-manifold Θ_H of smaller dimension than Θ.

The subset Θ_H has null Lebesgue measure whenever H is sharp. A probability density on the parameter space is an ordering system, notwithstanding having every point probability zero. In the FBST construction, all sets of same nature are treated accordingly in the same way. As a consequence, the sets that define sharp hypotheses keep having nil probabilities. As opposed to changing the nature of H by assigning positive probability to it, the tangential set T of points, having posterior density values higher than any θ in Θ_H, is considered. H is rejected if the posterior probability of T is *large*. The formalization of these ideas is presented below.

Let us consider a standard parametric statistical model; i.e., for an integer m, the parameter is $\theta \in \Theta \subset \mathcal{R}^m$, $g(\bullet)$ a probability prior density over Θ, x is the observation (a scalar or a vector), and $L_x(\bullet)$ is the likelihood generated by data x. Posterior to the observation of x, the sole relevant entity for the evaluation of the Bayesian evidence ev is the posterior probability density for θ given x, denoted by

$$g_x(\theta) = g(\theta|x) \propto g(\theta)L_x(\theta).$$

Of course, one is restricted to the case where the posterior probability distribution over Θ is absolutely continuous; i.e., $g_x(\theta)$ is a density over Θ. For simplicity, H is used for Θ_H in the sequel.

Definition 2 (evidence) Consider a sharp hypothesis $H : \theta \in \Theta_H$ and

$$g^* = \sup_H g_x(\theta) \text{ and } T = \{\theta \in \Theta : g_x(\theta) > g^*\}.$$

The *Bayesian evidence value against* H is defined as the posterior probability of the tangential set, i.e.,

$$\underline{ev} = Pr\{\theta \in T|x\} = \int_T g_x(\theta)d\theta.$$

One must note that the evidence value supporting H, $ev = 1 - \underline{ev}$, is not an evidence against A, the alternative hypothesis (which is not sharp anyway). Equivalently, $\underline{ev}$ is not evidence in favor of A, although it is against H.

Definition 3 (test) The *FBST* (Full Bayesian Significance Test) is the procedure that rejects H whenever $ev = 1 - \underline{ev}$ is small.

The following example illustrates the use of the FBST and two standard tests, McNemar and Jeffreys' Bayes Factor. Irony et al. (2000) discuss this inference problem introduced by McNemar (1955).

Example 1 McNemar vs. FBST Two professors, Ed and Joe, from the Department of Dentistry evaluated the skills of 224 students in dental fillings preparation. Each student was evaluated by both professors. The evaluation result could be approval (A) or disapproval (F). The Department wants to check whether the professors are equally exigent. Table 1 presents the data.

This is a four-fold classification with probabilities p_{11}, p_{12}, p_{21}, and p_{22}. Using standard notation, the hypothesis to be tested is $H : p_{1\cdot} = p_{\cdot 1}$ which is equivalent to $H : p_{12} = p_{21}$ (against $A : p_{12} \neq p_{21}$). In order to have the likelihood function readily available, we will consider a uniform prior, i.e., a *Dirichlet* density with parameter $(1, 1, 1, 1)$.

The McNemar exact significance for this data set is $pv = .064$. Recall that this test is based in a partial likelihood function, a binomial with $p = p_{12}(p_{12} + p_{21})^{-1}$ and $n = 66$. With the normal approximation, the pv become .049 with the partial likelihood used by McNemar, the FBST evidence is $ev = .045$. The value of the Bayes Factor under the same uniform prior is $BF = .953$. If one assigns probability 1/2 to the sharp hypothesis H, its posterior probability attains $\pi = .488$. Hence, the posterior probability π barely differs from 1/2, the probability previously assigned to H, while pv and ev seem to be more conclusive against H. While, in the three dimension full model, $ev = 0.265$ may seem to be a not low value and the test cannot be performed without a criterion. In other

Full Bayesian Significant Test (FBST). Table 1 Results of the evaluation of 224 students

	Joe		
Ed	A	F	Total
A	62	41	103
F	25	96	121
Total	87	137	224

words, a decision is not made until *ev* is compared to a "critical value." The derivation of such a criterion – resulting from the identification of the FBST as a genuine Bayes procedure – is the subject of Madruga et al. (2001).

The strong disagreement among the values of *ev*, *pv*, and *BF* seldom occurs in situations where Θ is a subset of the real line. The speculation is that this is related to the elimination of nuisance parameters: By conditioning in McNemar case and by marginalization in the Bayes Factor case. In higher dimension, elimination of nuisance parameters seems to be problematic, as pointed by Basu (1977).

FBST Theory

From a theoretical perspective, on the other hand, it may be propounded that if the computation of *ev* is to have any inferential meaning, then it ought to proceed to a declaration of significance (or not). To this – in a sense – simultaneously NPW and Fisherian viewpoint can be opposed the identification of ev as an estimator of the indicator function $\phi = I(\theta \in \Theta_H)$. In fact, Madruga et al. (2001) show that there are loss functions the minimization of which makes *ev* a Bayes estimator of ϕ (see Hwang et al. 1992).

Madruga et al. (2001) prove that the FBST procedure is the posterior minimization of an expected loss λ defined as follows:

$$\lambda(\text{Rejection of } H, \theta) = a\{1 - I[\theta \in T]\} \text{ and}$$

$$\lambda(\text{Acceptance of } H, \theta) = b + dI[\theta \in T].$$

Here, a, b and d are positive real numbers. The operational FBST procedure is given by the criterion according to which H is to be rejected if, and only if, the evidence *ev* is smaller than $c = (b+d)/(a+d)$. One should notice that the evidence *ev* is the Bayesian formal test statistic and that positive probability for H is never required. A complete discussion of the above approach can be found in Pereira et al. (2008).

Final Remarks

The following list states several desirable properties attended by *ev*:

1. *ev* is a probability value derived from the posterior distribution on the full parameter space.
2. Both *ev* and FBST possesses versions which are invariant for alternative parameterizations.
3. The need of approximations in the computation of *ev* is restricted to numerical maximization and integration.
4. FBST does not violate the Likelihood Principle.
5. FBST neither requires nuisance parameters elimination nor the assignment of positive prior probabilities to sets of zero Lebesgue measure.
6. FBST is a formal Bayes test and therefore has critical values obtained from considered loss functions.
7. *ev* is a possibilistic support for sharp hypotheses, complying with the Onus Probandi juridical principle (In Dubio Pro Reo rule), Stern (2003).
8. Derived from the full posterior distribution, *ev* is a homogeneous computation calculus with the same two steps: constrained optimization and integration with the posterior density.
9. Computing time was not a great burden whenever FBST was used. The sophisticated numerical algorithms used could be considered a more serious obstacle to the popularization of the FBST.

ev was developed to be the Bayesian *pv* alternative, while maintaining the most desirable (known or perceived) properties in practical use. The list presented above seems to respond successfully to the challenge: the FBST is conceptually simple and elegant, theoretically coherent, and easily implemented for any statistical model, as long as the necessary computational procedures for numerical optimization and integration are available.

About the Author

Dr Carlos Pereira is a Professor and Head, Department of Statistics, University of São Paulo, Brazil. He is Past President of the Brazilian Statistical Society (1998–1990). He was the Director of the Institute of Mathematic and Statistics, São Paulo, Brazil (1994–1998). He was also Director of the Bioinformatic Scientific Center, University of São Paulo (2006–2009). He is an Elected member of the International Statistical Institute. He has authored and co-authored more than 150 papers and 4 books, including *Bayesian Analysis* (in Portuguese) in 1982 – the first Bayesian book published in Latin America. Professor Pereira has received the Ralph Bradley award from Florida State University in 1980. He was a research engineer at IEOR in Berkeley at the University of California (1986–1988). He was Associate editor of *Entropy*, *Environmetrics*, and *Brazilian J of Probability and Statistics*. Currently, he is the Statistical editor of the *Brazilian J of Psychiatry*. He was a member of both the Environmetrics Society and Board of Directors of *Entropy*.

Cross References

►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Significance Testing: An Overview

References and Further Reading

Basu D (1977) On the elimination of nuisance parameters. JASA 72:355–366

Hwang JT, Casella G, Robert C, Wells MT, Farrel RG (1992) Estimation of accuracy in testing. Ann Stat 20:490–509

Irony TZ, Pereira CA de B, Tiwari RC (2000) Analysis of opinion swing: comparison of two correlated proportions. Am Stat 54(1):57–62

Madruga MR, Esteves LG, Wechsler S (2001) On the Bayesianity of Pereira-Stern tests. Test 10:291–299

McNemar Q (1955) Psychological statistics. Wiley, New York

Pereira CA de B, Stern JM (1999) Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy 1: 69–80

Pereira CA de B, Stern JM, Wechsler S (2008) Can a significance test be genuinely Bayesian? Bayesian Anal 3(1):79–100

Stern JM (2007) Cognitive constructivism, eigen-solutions, and sharp statistical hypotheses. Cybernetics Human Knowing 14(1):9–36

Functional Data Analysis

Hans-Georg Müller
Professor of Statistics
University of California-Davis, Davis, CA, USA

Functional data analysis (FDA) refers to the statistical analysis of data samples consisting of random functions or surfaces, where each function is viewed as one sample element. Typically, the random functions contained in the sample are considered to be independent and smooth. FDA methodology is essentially nonparametric, utilizes smoothing methods, and allows for flexible modeling. The underlying random processes generating the data are sometimes assumed to be (non-stationary) ▶Gaussian processes.

Functional data are ubiquitous and may involve samples of density functions (Kneip and Utikal 2001) or hazard functions (Chiou and Müller 2009). Application areas include growth curves, econometrics, evolutionary biology, genetics and general kinds of longitudinal data. FDA methodology features functional principal component analysis (Rice and Silverman 1991), warping and curve registration (Gervini and Gasser 2004) and functional regression (Ramsay and Dalzell 1991). Theoretical foundations and asymptotic analysis of FDA are closely tied to perturbation theory of linear operators in Hilbert space (Bosq 2000). Finite sample implementations often require to address ill-posed problems with suitable regularization. A broad overview of applied aspects of FDA can be found in the textbook Ramsay and Silverman (2005).

The basic statistical methodologies of ANOVA, regression, correlation, classification and clustering that are available for scalar and vector data have spurred analogous developments for functional data. An additional aspect is that the time axis itself may be subject to random distortions and adequate functional models sometimes need to reflect such time-warping. Another issue is that often the random trajectories are not directly observed. Instead, for each sample function one has available measurements on a time grid that may range from very dense to extremely sparse. Sparse and randomly distributed measurement times are frequently encountered in longitudinal studies. Additional contamination of the measurements of the trajectory levels by errors is also common. These situations require careful modeling of the relationship between the recorded observations and the assumed underlying functional trajectories (Rice and Wu 2001; James and Sugar 2003; Yao et al. 2005). Initial analysis of functional data includes exploratory plotting of the observed functions in a "spaghetti plot" to obtain an initial idea of functional shapes, check for ▶outliers and identify "landmarks." Preprocessing may include outlier removal and curve alignment (registration) to adjust for time-warping.

Basic objects in FDA are the mean function μ and the covariance function G. For square integrable random functions $X(t)$,

$$\mu(t) = E(Y(t)),\ G(s,t) = \mathrm{cov}\{X(s), X(t)\},\quad s,t \in \mathcal{T}, \tag{1}$$

with auto-covariance operator $(Af)(t) = \int_{\mathcal{T}} f(s)G(s,t)\,ds$. This linear operator of Hilbert-Schmidt type has orthonormal eigenfunctions ϕ_k, $k = 1,2,\ldots$, with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, such that $A\,\phi_k = \lambda_k\,\phi_k$. The foundation for functional principal component analysis is the Karhunen-Loève representation of random functions $X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k\,\phi_k(t)$, where $A_k = \int_{\mathcal{T}}(Y(t) - \mu(t))\phi_k(t)\,dt$ are uncorrelated centered random variables with $var(A_k) = \lambda_k$.

Estimators employing smoothing methods (local least squares or splines) have been developed for various sampling schemes (sparse, dense, with errors) to obtain a data-based version of this representation, where one regularizes by truncating at a finite number K of included components. The idea is to borrow strength from the entire sample of functions rather than estimating each function separately. The functional data are then represented by the subject-specific vectors of score estimates $\hat{A}_k$, $k = 1,\ldots,K$, which can be used to represent individual trajectories and

for subsequent statistical analysis. Useful representations are alternatively obtained with pre-specified fixed basis functions, notably B-splines and wavelets.

Functional regression models may include one or several functions among the predictors, responses, or both. For pairs (X, Y) with centered random predictor functions X and scalar responses Y, the linear model is

$$E(Y|X) = \int_{\mathcal{T}} X(s)\beta(s)\, ds.$$

The regression parameter function β is usually represented in a suitable basis, for example the eigenbasis, with coefficient estimates determined by ►least squares or similar criteria. A variant, which is also applicable for classification purposes, is the generalized functional linear model $E(Y|X) = g\{\mu + \int_{\mathcal{T}} X(s)\beta(s)\, ds\}$ with link function g. The link function (and an additional variance function if applicable) is adapted to the (often discrete) distribution of Y; the components of the model can be estimated by quasi-likelihood.

The class of useful functional regression models is large. A flexible extension of the functional linear model is the functional additive model. Writing centered predictors as $X = \sum_{k=1}^{\infty} A_k \phi_k$, it is given by

$$E(Y|X) = \sum_{k=1}^{\infty} f_k(A_k)\phi_k$$

for smooth functions f_k with $E(f_k(A_k)) = 0$. Of practical relevance are models with varying domains, with more than one predictor function, and functional (autoregressive) time series models. In addition to the functional trajectories themselves, their derivatives are of interest to study the dynamics of the underlying processes.

Acknowledgments

Research partially supported by NSF Grant DMS-0806199.

About the Author

Hans-Georg Müller is Professor at the Department of Statistics, University of California, Davis, USA. For additional information, papers, and software go to http://www.stat.ucdavis.edu/mueller/.

Cross References

►Components of Statistics

References and Further Reading

Bosq D (2000) Linear processes in function spaces: theory and applications. Springer, New York

Chiou J-M, Müller H-G (2009) Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. J Am Stat Assoc 104:572–585

Gervini D, Gasser T (2004) Self-modeling warping functions. J Roy Stat Soc B Met 66:959–971

Hall P, Hosseini-Nasab M (2006) On properties of functional principal components analysis. J Roy Stat Soc B Met 68:109–126

James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. J Am Stat Assoc 98:397–408

Kneip A, Utikal KJ (2001) Inference for density families using functional principal component analysis. J Am Stat Assoc 96: 519–542

Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis. J Roy Stat Soc B Met 53:539–572

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer series in statistics. Springer, New York

Rice JA, Silverman BW (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. J Roy Stat Soc B Met 53:233–243

Rice JA, Wu CO (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57:253–259

Yao F, Müller H-G, Wang J-L (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577–590

Functional Derivatives in Statistics: Asymptotics and Robustness

Luisa Turrin Fernholz
Professor Emerita of Statistics
Temple University, Philadelphia, PA, USA

Introduction

Given a sample $X_1, \ldots, X_n$ of i.i.d. random variables with common distribution function (df) F and empirical df F_n, a statistic $S(X_1, \ldots, X_n)$ is called a *statistical functional* if it can be written in terms of a functional T, independent of n, such that $S(X_1, \ldots, X_n) = T(F_n)$ for all $n \geq 1$. The domain of T contains at least the population df F and the empirical df F_n for all $n \geq 1$. In this setting the statistic $T(F_n)$ estimates the parameter $T(F)$.

The sample mean is a statistical functional since $\overline{X} = 1/n \sum_1^n X_i = \int x dF_n(x) = T_1(F_n)$ which estimates the parameter $T_1(F) = \int x dF(x)$. The statistical functional corresponding to the sample median is $T_2(F_n) = F_n^{-1}(1/2) = \text{med}\{X_1, \ldots, X_n\}$ estimating the population median $T_2(F) = F^{-1}(1/2)$. Most statistics of interest are statistical functionals. They can be defined explicitly, such as T_1 and T_2, or implicitly, such as maximum likelihood type estimators or M-estimators which are solutions of equations in θ of the form $\int \psi(x, \theta)\, dF_n(x) = 0$.

Statistical functionals were introduced by von Mises (1947), who proposed the use of a functional derivative

called the Volterra derivative along with the corresponding Taylor expansion to obtain the asymptotic distribution of a statistic. However, the technical details were obscure with intractable notation and complicated regularity conditions. Consequently, the results appeared difficult to implement, and the von Mises theory was neglected until the late 1960s and 1970s with the surge of ▶robust statistics associated mainly with the work of Huber (1964, 1977) and Hampel (1968, 1974). For these new statistics, the statistical functional setting was found to be optimal for the study of robustness properties and the von Mises approach seemed to provide a natural environment for deriving the asymptotic distribution of the proposed robust estimates. During these robustness years the functional analysis concepts of differentiability and continuity were used to investigate the robustness aspects of the new statistics in addition to the asymptotics. In particular, the introduction of the influence function made a connection between robustness and classical asymptotics.

The Influence Function

Given a statistical functional T and a df F, the *influence function* of T at F is the real valued function $IF_{T,F}$ defined by

$$IF_{T,F}(x) = \lim_{t\to 0^+} \frac{T((1-t)F + t\Delta_x) - T(F)}{t},$$

where Δ_x is the d.f. of the point mass one at x. This function is normalized by setting $IF(x) = IF_{T,F}(x) - E_F(IF_{T,F}(X))$ so that $E_F(IF(X)) = 0$.

The influence function has played an important role in robust statistics. It was introduced by Hampel (1974), who observed that for large n, $IF_{T,F}(x)$ measures the effect on $T(F_n)$ of a single additional observation with value x. A bounded influence function indicates robustness of the statistic. For example, for the sample mean $T_1(F_n)$ as defined above, the influence function is $IF(x) = x - T_1(F)$. For the sample median $T_2(F_n)$, if $f = F'$, we have

$$IF(x) = \left[-1/(2f(T_2(F))\right] I_{\{x<(T_2(F)\}}(x) + \left[1/(2f(T_2(F))\right] I_{\{x\geq T_2(F)\}}(x).$$

Hence, the sample median with bounded influence function is more robust than the sample mean whose influence function is not bounded. A complete treatment of the robustness measures derived from the influence function can be found in Hampel et al. (1986).

In the framework of statistical functionals, the influence function can be viewed as a weak form of a functional derivative. Stronger derivatives were defined to analyze the asymptotic behavior of a statistic, but in all these derivatives the influence function is the crucial ingredient. It also provides a link between robustness and asymptotics as will be shown below.

Functional Derivatives

Consider a statistical functional T with domain an open set which lies in a normed vector space and contains a df F. A continuous linear functional T'_F is the *derivative* of T at F when

$$\lim_{t\to 0} \frac{T(F+tH) - T(F) - T'_F(tH)}{t} = 0, \qquad (1)$$

for H in subsets of the domain of T.

If (1) holds pointwise for each H, then T'_F is the *Gâteaux derivative*.

If (1) holds uniformly for all H in compact subsets of the domain of T, then T'_F is the *Hadamard derivative*.

If (1) holds uniformly for all H in bounded subsets of the domain of T, then T'_F is the *Fréchet derivative*.

Clearly Fréchet differentiability implies Hadamard differentiability, which implies Gâteaux differentiability. In all cases, the influence function is the central ingredient for any derivative since $T'_F(H) = \int IF_{T,F}(x)\, dH(x)$. Now, consider the Taylor expansion of T at F:

$$T(F+tF) - T(F) = T'_F(tH) + Rem$$

with

$$Rem = Rem(T, H, t) = o(t).$$

This remainder tends to zero either pointwise or uniformly according to whether F is Gâteaux, Hadamard, or Fréchet differentiable. For Hadamard derivatives see Reeds (1976) or Fernholz (1983) and for Fréchet derivatives see Huber (1981) or Serfling (1981).

When $t = 1/\sqrt{n}$ and $H = \sqrt{n}\,(F_n - F)$, the linear term of the Taylor expansion of T is

$$\int IF_{T,F}(x)\, d(F_n - F)(x) = \frac{1}{n}\sum_1^n IF(X_i),$$

where IF has been normalized, and the von Mises expansion of T at F is

$$T(F_n) = T(F) + \frac{1}{n}\sum_1^n IF(X_i) + Rem$$

or

$$\sqrt{n}(T(F_n) - T(F)) = \frac{1}{\sqrt{n}}\sum_1^n IF(X_i) + \sqrt{n}\, Rem.$$

When T is Hadamard or Fréchet differentiable, $\sqrt{n}\, Rem \to 0$ in probability, so that under certain regularity conditions for F we have the ▶asymptotic normality,

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2).$$

In this case the influence function gives the asymptotic variance $\sigma^2 = E_F[IF(X)]^2$.

Remarks

The derivative used by von Mises for these calculations was similar to the Gâteaux derivative, so several strong regularity conditions had to be imposed on T to obtain its asymptotic normality. With the Hadamard or Fréchet derivatives these extra conditions are not needed.

It is important to note that the influence function plays a key role in these von Mises calculations. Note also that the influence function provides a link between robustness and asymptotics, and for this reason the von Mises approach via the influence function has become a useful method for obtaining asymptotic normality results.

The use of the Hadamard and Fréchet derivatives translates the problem of asymptotics into a problem of functional differentiability. Since Hadamard and Fréchet derivatives enjoy the chain rule property, we can show that a statistic $T(F_n)$ is asymptotically normal if the functional T is a composition of Hadamard or Fréchet differentiable functional components, where each component has a simple form. For references see Reeds (1976) or Fernholz (1983).

Higher Order Derivatives

The influence function is also called the *first kernel* since higher order derivatives can be defined for a real valued function T. If we set $\varphi_1(x) = IF(x)$ for the first kernel, the *second kernel* is

$$\varphi_2(x, y) = \frac{\partial^2}{\partial s\, \partial t} T\big(F(1 - s - t) + t\Delta_x + s\Delta_y\big)|_{t=0, s=0},$$

and in general, the *kernel of order* k is

$$\varphi_k(x_1, x_2, \ldots, x_k) = \frac{\partial^k}{\partial t_1\, \partial t_2 \ldots \partial t_k} T\left(F\left(1 - \sum_1^k t_i\right) + t_1\Delta_{x_1} + \ldots + t_k\Delta_{x_k}\right)\Big|_{(0,\ldots,0)}.$$

These kernels constitute the main ingredients for general Fréchet, Hadamard or Gâteaux higher order derivatives of T at F and for the corresponding higher order Taylor expansions. Hence, for $k \geq 2$ the k-th order von Mises expansion of $T(F_n)$ at F is:

$$T(F_n) - T(F) = \frac{1}{n}\sum_i \varphi_1(X_i) + \frac{1}{2n^2}\sum_{i,j}\varphi_2(X_i, X_j) + \cdots + \frac{1}{k!n^k}\sum_{i_1,\ldots,i_k}\varphi_k(X_{i_1}, \ldots, X_{i_k}) + Rem_k,$$

where, under certain differentiability conditions for T, the remainder of order k satisfies $Rem_k = o_P(n^{-k/2})$.

Higher order von Mises expansions were used to study the asymptotic distribution of a statistic when it is not normal (see von Mises 1947; Filippova 1972; Reeds 1976). These expansions are also useful to study the bias of a statistic since $E_F(T(F_n)) = T(F) + E_F(Rem_1)$, where

$$Rem_1 = \frac{1}{2n^2}\sum_{i,j}\varphi_2(X_i, X_j) + \cdots + \frac{1}{k!n^k}\sum_{i_1,\ldots,i_k}\varphi_k(X_{i_1}, \ldots, X_{i_k}) + Rem_k.$$

Results in this direction can be found in Sen (1988) and Fernholz (2001).

Multivariate Functionals

The formal von Mises calculations outlined above can be carried out for functionals of p variables after generalizing some basic rules of elementary calculus for the case of functional derivatives. Thus, if $T : \mathbb{R}^p \to \mathbb{R}$ and we have p samples of sizes $n_1, n_2, \ldots, n_p$ from the populations $F_1, \ldots, F_p$ respectively, we can consider the corresponding empirical df's $F_{n_1}, \ldots, F_{n_p}$. Then, the multivariate statistical functional $T(F_{n_1}, \ldots, F_{n_p})$ has p first order partial derivatives given by the corresponding multivariate influence function $\varphi_1 = (\varphi_{11}, \varphi_{12}, \ldots, \varphi_{1p})$, where for $1 \leq i \leq p$ the components are

$$\varphi_{1i}(x) = \frac{\partial T(F_1, \ldots, F_{i-1}, (1-\varepsilon)F_i + \varepsilon\Delta_x, F_{i+1}, \ldots, F_p)}{\partial \varepsilon}\Big|_{\varepsilon=0}.$$

Higher order partial derivatives can be found with the corresponding higher order von Mises expansions. For details, examples, and applications see Filippova (1972), Reeds (1976), and Fernholz (2001).

Statistical Functionals and the Bootstrap

Statistical functionals played a key role in the development of the bootstrap (see ▶Bootstrap Methods) introduced by Efron (1979). The "plug in" principle of Efron is essentially the study of a statistic in the setting of statistical functionals. After the bootstrap was introduced, the functional derivatives provided the answer for one of the basic asymptotic questions regarding the consistency of the bootstrap estimators $T(F_n^*)$. Does the bootstrap work when the von Mises method works? That is, does

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathcal{D}} N(0, \sigma^2) \text{ imply}$$
$$\sqrt{n}\left(T\left(F_n^*\right) - T(F_n)\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2) \ ?$$

The affirmative answer was given by R. Gill (1989) where he used von Mises expansions with Hadamard derivatives to show the asymptotic consistency of the bootstrap.

Smoothed Versions of Statistical Functionals

Using the convolution of F_n with a smooth df kernel sequence K_n we can obtain the smoothed version $\widetilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i)$ of F_n. For a given statistical functional $T(F_n)$ estimating $T(F)$, we can consider the corresponding smoothed functional $T(\widetilde{F}_n)$ which, for continuous populations, may give a better estimate for $T(F)$. Some robustness aspects of $T(\widetilde{F}_n)$ can be analyzed through the influence function of T, and under reasonable regularity conditions for K_n, the ►asymptotic normality of the smoothed version $T(\widetilde{F}_n)$ can be obtained when T is Hadamard differentiable. See Fernholz (1991, 1993).

About the Author

Luisa Turrin Fernholz is Professor Emerita of Statistics at Temple University, Philadelphia, PA (USA). She is currently the director of the Minerva Research Foundation and a member of the Advisory Council Committee for the Department of Mathematics at Princeton University as well as a member of the School of Mathematics Council for the Institute for Advanced Study, at Princeton, NJ (USA). She has previously held faculty positions at Princeton University, University of Pennsylvania, and the University of Buenos Aires. She is an elected member of the International Statistical Institute and a member of the ASA, the IMS, and the Bernoulli Society. She authored the research monograph *Von Mises Calculus for Statistical Functionals* (Springer Verlag, Lecture Notes in Statistics, Vol. 19, 1983). She co-edited several statistics volumes on Data Analysis and Robustness, and has published articles on probability theory as well as asymptotic expansions, robustness, functional derivatives, bias and variance reduction, and bootstrap applications, among other topics.

Cross References

►Bootstrap Methods
►Multivariate Technique: Robustness
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Cabrera J, Fernholz LT (1999) Target estimation for bias and mean square error reduction. Ann Statist 27(3):1080–1104
Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Statist 7:1–26
Fernholz LT (1983) Von Mises calculus for statistical functionals. Lecture notes in statistics, vol 19. Springer, New York
Fernholz LT(1991) Almost sure convergence of smoothed empirical distribution functions. Scand J Statist 18:255–262
Fernholz LT (1993) Smoothed versions of statistical functionals. In: Morgenthaler S, Ronchetti E, Stahel W (eds) New directions in statistical data analysis and robustness, Birkhauser, London, pp 61–72
Fernholz LT (2001) On multivariate higher order von Mises expansions. Metrika 53(2):123–140
Filippova AA (1962) Mises theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. Theory Prob Appl 7:24–57
Gill RD (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises Method, (part I). Scand J Statist 16:97–128
Hampel F (1974) The influence curve and its role in robust estimation. J Am Statist Assoc 69:383–393
Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) Robust statistics. The approach based on the influence function. Wiley, New York
Huber P (1964) Robust estimation of a location parameter. Ann Math Statist 35:73–101
Huber P (1977) Robust Statistical Procedures, vol 27, Regional conference series in applied mathematics. SIAM, Philadelphia
Huber P (1981) Robust statistics. Wiley, New York
Reeds JA (1976) On the definition of von Mises functionals. PhD dissertation, Harvard University, Cambridge
Sen PK (1988) Functional jackknifing: rationality and general asymptotics. Ann Statist 16:450–469
von Mises R (1947) On the asymptotic distribution of differentiable statistical functions. Ann Math Statist 18:309–348

Fuzzy Logic in Statistical Data Analysis

Efendi N. Nasibov
Professor, Head of Department of Computer Science,
Faculty of Science and Arts
Dokuz Eylul University, Imzir, Turkey

Probability and Statistics with Fuzziness

Fuzzy logic and fuzzy sets theory first discussed in 1965 by Zadeh (Zadeh 1965). In classical sets theory, classifications are precise and the subject either belongs to a set or not. On the contrary, in fuzzy sets theory, the subject located in the border both belongs and does not belong to a set simultaneously. As a mathematical representation, in classical

set theory, if the object is the member of a set, it takes the membership value of 1; otherwise it takes the membership value of 0. However, in fuzzy logic, objects could have membership degrees between 0 and 1. In fuzzy logic, for example, a 30-year-old person could be the member of both the "young people" set with a membership degree of 0.6 and the "not young people" set with a membership degree of 0.4.

The relation between "fuzzy sets theory" and "statistics and probability theory" is and important research areas. In probability theory, realization of events is based on the classical 0–1 logic, i.e., an event occurs or does not occur. When the boundaries of classes that reflect the events are precise, such logic is valid. For example, when a dice has been rolled, the event of "coming up 1 or 2" is a precise event and it has a precise probability. But the event of "coming up a little number" is an imprecise event since its boundaries can not be stated; consequently, its probability can not be designated. In such situations, using probability theory together with fuzzy logic and fuzzy sets theory provides more admissible results.

Another important utilization of fuzzy logic and fuzzy sets theory is in statistical data analysis. With improvement of fuzzy sets theory, many studies have been made to combine statistical analysis methods and fuzzy sets theory. An analysis in which fuzzy logic is used is more robust than the classical logic. Furthermore, more reliable results can be obtained by a fuzzy approach (Rubin 1998).

There are many instances where fuzzy logic is used in statistical data analysis, including clustering, classification, regression, ►principal component analysis (PCA), independent component analysis (ICA), ►multidimensional scaling, ►time series, hypothesis tests, and confidence intervals (Coppi et al. 2006; Pop 2004; Taheri 2003; Mares 2007).

Fuzzy Clustering and Classification

Bellman et al. (1966) and Ruspini (1969) are the pioneers who used fuzzy sets theory in cluster analysis. Afterwards, many approaches were proposed on the use of fuzzy logic in cluster analysis. The most widely used approaches are based on fuzzy partitioning.

Fuzzy partitioning: The fuzzy partitioning of the data set $X = x_1, x_2, \ldots, x_n$ into fuzzy clusters $C_1, C_2, \ldots, C_c (1 < c < n)$ is denoted by the matrix $U_f = (u_{ij}) = (\mu_{C_i}(x_j))$, which satisfies the conditions given below:

$$0 \le u_{ij} \le 1, \quad \forall i \in \{1, 2, \ldots, c\}, \quad \forall j \in \{1, 2, \ldots, n\}, \tag{1}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n, \quad \forall i \in \{1, 2, \ldots, c\} \tag{2}$$

where u_{ij} is the membership degree of the element x_j to the cluster C_i. In most cases, the following normalization condition as well as (1) and (2) is required for fuzzy partitioning:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j \in \{1, 2, \ldots, n\}. \tag{3}$$

The first solution algorithms for the clustering approach based on fuzzy partitioning were proposed by Dunn (1973) and improved by Bezdek (1973).

In the most widely used fuzzy c-means (FCM) algorithm, the optimal fuzzy partitioning is obtained by minimizing the following function:

$$J_f(U_f, v_1, \ldots, v_c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d(v_i, x_j)^2 \tag{4}$$

where c is the predetermined number of clusters, $d(v_i, x_j)$ is the distance between the cluster center v_i and the object x_j, and $m (m > 1)$ is the fuzziness index. The solution of (1)–(4) is found through the iterative computation of membership degrees and cluster centers:

$$u_{ij} = \frac{d(v_i, x_j)^{-2/(m-1)}}{\sum_{t=1}^{c} d(v_t, x_j)^{-2/(m-1)}}, \; i = 1, \ldots, c; \; j = 1, \ldots, n \tag{5}$$

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \; i = 1, \ldots, c \tag{6}$$

The FCM algorithm is successful in finding spherical-shaped cluster structures. The Gustafson-Kessel algorithm based on FCM can find ellipsoidal cluster structures by using a covariance matrix. Fuzzy maximum likelihood estimation (FMLE) and the expectation maximization (EM) algorithms are also widely used fuzzy clustering algorithms (Doring et al. 2006).

Another approach for using fuzzy logic in cluster analysis is based on fuzzy neighborhood relations (FDBSCAN, FJP, FN-DBSCAN). In such an approach, the data are handled as fuzzy points and the classes are formed as crisp level sets based on the fuzziness level. Different clustering structures are obtained in different fuzziness levels. This approach could also be conceived as hierarchical clustering. The main point is to find the optimal hierarchy level and the optimal cluster structure convenient to this hierarchy level. In the fuzzy joint points based algorithms such as FJP, NRFJP, MFJP, such a problem has been solved by using an integrated cluster validity mechanism (Nasibov and Ulutagay 2007). The superiority of such algorithms over the FCM-based algorithms is not only the possibility

for finding arbitrarily shaped rather than only spherical-shaped clusters, but also not needing to determine the number of clusters in advance. On the other hand, using a fuzzy neighborhood relation among data can increase the robustness of clustering algorithm (Nasibov and Ulutagay 2009).

Classification is referred to as a supervised classification while clustering is referred to as an unsupervised classification. In a fuzzy supervised classification, the fuzzy partition X of elements is given in advance. One must specify the class of a datum x^* which is handled afterward. To do this, many approaches are used, including fuzzy inference system (FIS), fuzzy linear discriminant analysis, fuzzy k-nearest neighbors, and fuzzy-Bayesian classifier.

Fuzzy Regression

Fuzzy regression analysis is done by applying fuzzy logic techniques to classical regression models. There is no need for the assumptions of classical regression to hold for fuzzy regression. Moreover, fuzzy regression does not require normally distributed data, stability tests, or large samples. Classical regression analysis is able to respond to the needs of numerical science working with precise information. But in the social sciences, in which the personal perceptions are important, it is not easy to estimate the assumed appropriate and consistent estimators, because the concerned data are composed of imprecise, i.e., fuzzy data. In such situations, fuzzy logic can provide approximate ideas to reach adequate conclusions. There are various fuzzy regression models based on either fuzziness of the values of independent/dependent variables or fuzziness of the regression function (Näther 2006).

Usually, the fuzzy regression equation is as follows:

$$\tilde{y}_i = \tilde{b}_0 \oplus \tilde{b}_1 \odot \tilde{x}_{i1} \oplus \ldots \oplus \tilde{b}_p \odot \tilde{x}_{ip},\ i = 1, \ldots, n \qquad (7)$$

where $\oplus$ and $\odot$ are addition and multiplication processes on fuzzy numbers, $(\tilde{b}_0, \tilde{b}_1, \ldots, \tilde{b}_p)$ are fuzzy regression coefficients, $\tilde{y}_i$ is fuzzy response, and $(\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{ip})$ are fuzzy explanatory variables.

The first study of fuzzy logic in regression analysis was made by Tanaka as fuzzy linear regression model (Tanaka et al. 1979, 1980, 1982). In Tanaka's approach, regression line is formed as a fuzzy linear function of data. Linear programming has been used to determine the parameters of this fuzzy function.

Another approach to fuzzy logic in regression analysis minimizes the sum of squares error between the observed and the predicted data which take fuzzy values. In determining the distance between fuzzy data, various fuzzy distances can be used and various models can be constructed (Diamond 1988; D'Urso 2003; Kim and Bishu 1998; Nasibov 2007).

As a third approach, Fuzzy c-Regression Models (FcRM), which arose from the technique of fuzzy clustering application on regression, can be specified. This approach, also called the switching regression, was proposed by Hathaway and Bezdek (Hathaway and Bezdek 1993). In this approach, all data are partitioned into distinct clusters since it is easier to express the structure with partial lines instead of a single regression function. The process works as in the FCM algorithm. The only difference is that, not only the membership degrees, but also the parameters of regression lines of the clusters are updated instead of cluster centers. The optimal value of the parameters has been found using the weighted least squares approach. For synthesis of the results, Fuzzy Inference Systems (FIS) such as Mamdani, Takagi-Sugeno-Kang (TSK), etc., can be used (Jang et al. 1997).

Fuzzy Principal Component Analysis

►Principal component analysis (PCA) is a preferred analysis method to reduce the dimension of the feature space and to extract information. PCA determines the linear combinations that describe maximum variability among the original data measurements. However, it is influenced by ►outliers, loss of data, and poor linear combinations. To resolve this problem, PCA models are created using fuzzy logic and the results are handled more efficiently than classical principal component analysis (Sarbu and Pop 2005). As with fuzzy regression, whole data sets are divided into fuzzy subsets in order to create better PCA models. Thus, the influence of outliers, which have minimum membership degree to clusters, is reduced.

The fuzzy covariance matrix for cluster A_i is constructed as follows:

$$C_{kl}^{(i)} = \frac{\sum_{j=1}^{n} [A_i(x^j)]^2 (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)}{\sum_{j=1}^{n} [A_i(x^j)]^2}, \qquad (8)$$

where $A_i(x^j)$ indicates the membership degree of an object x^j to the cluster A_i and is inversely proportional with the distance between the object and the independent component.

One of the first studies about the fuzzy PCA was performed by Yabuuch and Watada in the construction of the principal component model using fuzzy logic for the elements in the fuzzy groups (Yabuuch and Watada 1997). The fuzzy PCA allows us to analyze the features of vague data samples. Hence, the fuzzy PCA gives more reliable results. Afterwards, the local fuzzy PCA method is used to reduce the dimension of feature vectors effectively. In this method, data space is partitioned into clusters using fuzzy clustering and then PCA is applied by constructing a fuzzy covariance

matrix (Lee 2004). In the study performed by Hsieh and Yang, fuzzy clustering is applied to find the hidden information in a DNA sequence by combining PCA and fuzzy adaptive resonance theory (fuzzy-ART) (Hsieh et al. 2008).

Fuzzy Independent Component Analysis

The recently developed and widely used independent component analysis (ICA) method is used to find linear form of non-Gaussian and statistically independent variables, and to extract information from databases (Hyvarinen et al. 2001). ICA is closely related to the blind source separation (BSS) method. The measurements $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of m unknown source signals ($\mathbf{s} = (s_1, s_2, \ldots, s_m)$) composed by unknown linear mixtures ($\mathbf{A}$) are performed:

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{9}$$

For the computational ease, $m = n$ is assumed. Hence, the matrix $\mathbf{A}$ is estimated by using the advantage of being an independent and non-Gaussian of source data and, using the matrix $\mathbf{W}$ (inverse of $\mathbf{A}$), source signals can be calculated from the equation below:

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{10}$$

The most widely used ICA algorithm is the Fast-ICA algorithm in terms of ease of use and speed. Honda et al. (2000) improved the Fast-ICA algorithm as the fuzzy Fast-ICA algorithm. In the fuzzy Fast-ICA algorithm, the Fuzzy c-Varieties (FCV) clustering method, which separates data into linear clusters using linear variability, is applied and then the local independent components in the fuzzy clusters are estimated by Fast-ICA algorithm.

Honda and Ichihashi (2008) have also proposed the fuzzy local ICA model as the improved version of the local ICA model. In the fuzzy local ICA model, fuzzy clustering, PCA, and multiple regression analysis are used simultaneously.

Gait biometrics have great advantages in comparison with the widely used biometrics such as face, fingerprint, and iris. In order to recognize gait, Lu et al. have developed a simple method based on human silhouette using genetic fuzzy vector machine (GFVM) and independent component analysis (Lu and Zhang 2007).

Fuzzy Time Series

The term "fuzzy time series" was first coined by Song and Chissom (Song and Chissom 1993ab, 1994).

Let $Y(t) \in R^1 (t = \ldots, 0, 1, 2, \ldots)$ be the universe of discourse on which fuzzy sets $f_i(t)(i = 1, 2, \ldots)$ are defined. Let $F(t)$ be a collection on $f_i(t)(i = 1, 2, \ldots)$. Then, $F(t)$ is called a fuzzy time series on $Y(t)(t = \ldots, 0, 1, 2, \ldots)$. In other words, fuzzy time series $F(t)$ is a chronological sequence of imprecise or fuzzy data ordered by time. Fuzzy time series are regarded as realizations of fuzzy random processes. In the fuzzy time series, fuzzy data as well as time-dependent dynamic relation can be considered as fuzzy:

$$F(t) = F(t-1) \circ \tilde{R}(t-1, t) \tag{11}$$

n^{th}-order fuzzy time series forecasting model, can be represented as follows:

$$F(t-1), F(t-2), \ldots, F(t-n) \rightarrow F(t) \tag{12}$$

For modeling of fuzzy time series, fuzzy ARMA, ARIMA processes, or fuzzy artificial neural networks are applied (Tseng et al. 2001; Zhang et al. 1998). Fuzzy time series can be analyzed and forecast by specifying an underlying fuzzy random process with the aid of generally applicable numerical methods.

Statistical Hypothesis Tests and Confidence Intervals

The main purpose of the traditional hypothesis test is to separate $\theta \in \Theta$ parameter space into two regions such as ω and $\Theta \backslash \omega$. The null and alternative hypotheses are as follows:

$$\begin{cases} H_0 : \theta \in \omega & (\textit{null hypothesis}) \\ H_1 : \theta \in \Theta \backslash \omega, & (\textit{alternative hypothesis}) \end{cases} \tag{13}$$

If the boundaries of ω and $\Theta \backslash \omega$ regions are assumed to be fuzzy, the fuzzy hypothesis test can be constructed as follows (Coppi et al. 2006):

$$\begin{cases} H_0 : \mu_\omega(\theta), & (\textit{null hypothesis}) \\ H_1 : \mu_{\Theta \backslash \omega(\theta)}, & (\textit{alternative hypothesis}) \end{cases} \tag{14}$$

Data handled in daily life are usually imprecise, i.e., fuzzy. For instance, water level of the river may not be fully measured due to fluctuations. In such a case, the well-known crisp hypothesis tests will not give reliable results.

Different approaches related to statistical hypothesis tests have been developed using fuzzy sets theory. First, Casals et al. (1986ab) and Casals and Gil (1989) have developed the ►Neyman-Pearson Lemma and Bayes method for statistical hypothesis tests with fuzzy data. There are two approaches to analyze statistical hypothesis tests: (1) observations are ordinary (crisp) and hypotheses are fuzzy (Arnold 1998), (2) both observations and hypotheses are fuzzy (Wu 2005). There may be some problems in applying classical statistical hypothesis to fuzzy observations. For instance, θ might be "approximately one" or θ might be "very large" and so on, where θ is any tested parameter. Bayes method might be useful for such types of hypothesis tests (Taheri and Behboodian 2001). However, if the fuzzy data are observed, the most appropriate method will be to apply fuzzy set theory to establish the statistical model.

In some approaches to using fuzzy logic in hypothesis tests, the estimators as fuzzy numbers are obtained using confidence intervals. If the estimator is a fuzzy number, the test statistic in hypothesis testing will also be a fuzzy number. Thus, the critical value at the hypothesis test is a fuzzy number. The result of this approach might be more realistic than a crisp hypothesis test (Buckley 2005). These results may be evaluated with probability theory (Hryniewicz 2006).

Fuzzy sets theory through the fuzzy random variables is applied to statistical confidence intervals for unknown fuzzy parameters. When the sample size is sufficiently large, an approximate fuzzy confidence interval could be constructed through a central limit theorem (Wu 2009). In case of fuzzy data, an interval estimation problem is formulated and the relation between fuzzy numbers and random intervals is found in (Coral et al. 1988; Gil 1992).

About the Author

Dr Efendi Nasibov is a Professor and Head, Department of Computer Science, Dokuz Eylul University, Turkey. He was a Professor and Head, Theoretical Statistics Division of the Department of Statistics, Dokuz Eylul University, Turkey (2006–2009). He was also the Head of the Department of Decision Making Models and Methods (2003–2009) of the Institute of Cybernetics, National Academy of Sciences of Azerbaijan. He is an elected member of the Academy of Modern Sciences named after Lotfi Zadeh, Baku, Azerbaijan. He has authored and co-authored more than 130 papers and 5 books.

Cross References

►Cluster Analysis: An Introduction
►Confidence Interval
►Data Analysis
►Expert Systems
►Forecasting: An Overview
►Fuzzy Set Theory and Probability Theory: What is the Relationship?; Fuzzy Sets: An Introduction
►Fuzzy Sets: An Introduction
►Hierarchical Clustering
►Multicriteria Clustering
►Neyman-Pearson Lemma
►Principal Component Analysis
►Statistical Methods for Non-Precise Data

References and Further Reading

Arnold BF (1998) Testing fuzzy hypothesis with crisp data. Fuzzy Set Syst 94:323–333
Bellman RE, Kalaba RE, Zadeh LA (1966) Abstraction and pattern classification. J Math Anal Appl 2:581–586
Bezdek JC (1973) Fuzzy mathematics in pattern classification. PhD Thesis, Cornell University, Ithaca, New York
Bezdek JC (1974) Cluster validity with fuzzy sets. J Cybernetics 3:58–73
Buckley JJ (2005) Fuzzy statistics: hypothesis testing. Soft Comput 9:512–518
Casals MR, Gil MA (1989) A note on the operativeness of Neyman–Pearson tests with fuzzy information. Fuzzy Set Syst 30:215–220
Casals MR, Gil MA, Gil P (1986a) On the use of Zadeh's probabilistic definition for testing statistical hypotheses from fuzzy information. Fuzzy Set Syst 20:175–190
Casals MR, Gil MA, Gil P (1986b) The fuzzy decision problem: An approach to the problem of testing statistical hypotheses with fuzzy information. Euro J Oper Res 27:371–382
Coppi R, Gil MA, Kiers HAL (2006) The fuzzy approach to statistical analysis. Comput Stat Data Anal 51:1–14
Corral N, Gil MA (1988) A note on interval estimation with fuzzy data. Fuzzy Set Syst 28:209–215
D'Urso P (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput Stat Data Anal 42:47–72
Diamond P (1988) Fuzzy least squares. Inform Sci 46:141–157
Doring C, Lesot MJ, Kruse R (2006) Data analysis with fuzzy clustering methods. Comput Stat Data Anal 51:192–214
Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernetics 3:32–57
Gil MA (1992) A note on the connection between fuzzy numbers and random intervals. Stat Prob Lett 13:311–319
Hathaway RJ, Bezdek JC (1993) Switching regression models and fuzzy clustering. IEEE Trans Fuzzy Syst 3:195–204
Honda K, Ichihashi H (2008) Fuzzy local ICA for extracting independent components related to external criteria. Appl Math Sci 2(6):275–291
Honda K, Ichihashi H, Ohue M, Kitaguchi K (2000) Extraction of local independent components using fuzzy clustering. In Proceedings of 6th International Conference on Soft Computing, pp 837–842
Hryniewicz O (2006) Possibilistic decisions and fuzzy statistical tests. Fuzzy Set Syst 157:2665–2673
Hsieh KL, Yang IC (2008) Incorporating PCA and fuzzy-ART techniques into achieve organism classification based on codon usage consideration. Comput Biol Med 38:886–893
Hyvarinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Englewood Cliffs
Kim B, Bishu RR (1998) Evaluation of fuzzy linear regression models by comparing membership functions. Fuzzy Set Syst 100:343–352
Lee KY (2004) Local fuzzy PCA based GMM with dimension reduction on speaker identification. Pattern Recogn Lett 25: 1811–1817
Lu J, Zhang E (2007) Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion. Pattern Recogn Lett 28:2401–2411
Mares M (2007) Fuzzy data in statistics. Kybernetica 43(4): 491–502
Nasibov EN (2007) Fuzzy least squares regression model based on weighted distance between fuzzy numbers. Automat Contr Comput Sci 41(1):10–17

Nasibov EN, Ulutagay G (2007) A new unsupervised approach for fuzzy clustering. Fuzzy Set Syst 158:2118–2133
Nasibov EN, Ulutagay G (2009) Robustness of density-based clustering methods with various neighborhood relations. Fuzzy Set Syst 160:3601–3615
Näther W (2006) Regression with fuzzy random data. Comput Stat Data Anal 51:235–252
Pop HF (2004) Data analysis with fuzzy sets: a short survey. Studia University of Babes-Bolyai, Informatica XLIX(2):111–122
Rubin SH (1998) A fuzzy approach towards inferential data mining. Comput Ind Eng 35(1–2):267–270
Ruspini EH (1969) A new approach to clustering. Inform Contr 15:22–32
Sarbu C, Pop HF (2005) Principal component analysis versus fuzzy principal component analysis: a case study: the quality of Danube water (1985–1996). Talanta 65:1215–1220
Song Q, Chissom BS (1993a) Forecasting enrollments with fuzzy time series-part I. Fuzzy Set Syst 54:1–9
Song Q, Chissom BS (1993b) Fuzzy time series and its models. Fuzzy Set Syst 54:269–277
Song Q, Chissom BS (1994) Forecasting enrollments with fuzzy time series-part II. Fuzzy Set Syst 62:1–8
Taheri SM (2003) Trends in fuzzy statistics. Austr J Stat 32(3):239–257
Taheri SM, Behboodian J (2001) A Bayesian approach to fuzzy hypothesis testing. Fuzzy Set Syst 123:39–48
Tanaka H, Okuda T, Asai K (1979) Fuzzy information and decision in statistical model. In Gupta MM et al. (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 303–320
Tanaka H, Uejima S, Asai K (1980) Fuzzy linear regression model. IEEE Trans Syst Man Cybernet 10:2933–2938
Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybernet 12:903–907
Tseng FM, Tzeng GH, Yu HC, Yuan BJC (2001) Fuzzy ARIMA model for forecasting the foreign exchange market. Fuzzy Set Syst 118:9–19
Wu HC (2005) Statistical hypotheses testing for fuzzy data. Inform Sci 175:30–56
Wu HC (2009) Statistical confidence intervals for fuzzy data. Expert Syst Appl 36:2670–2676
Yabuuch Y, Watada J (1997) Fuzzy principal component analysis and its application. Biomedical Fuzzy Hum Sci 3(1):83–92
Zadeh LA (1965) Fuzzy sets. Inform Contr 8(3):338–353
Zhang GP, Eddy PB, Hu YM (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14:35–62

Fuzzy Set Theory and Probability Theory: What is the Relationship?

Lotfi A. Zadeh
Professor Emeritus
University of California-Berkeley, Berkeley, CA, USA

Relationship between probability theory and fuzzy set theory is associated with a long history of discussion and debate. My first paper on fuzzy sets was published in 1965 (Zadeh 1965). In a paper published in 1966, Loginov suggested that the membership function of a fuzzy set may be interpreted as a conditional probability (Loginov 1966). Subsequently, related links to probability theory were suggested and analyzed by many others (Coletti and Scozzafava 2004; Freeling 1981; Hisdal 1986a, b; Nurmi 1977; Ross et al. 2002; Singpurwalla and Booker 2004; Stallings 1977; Thomas 1995; Viertl 1987; Yager 1984). Among such links are links to set-valued random variables (Goodmanm and Nguyen 1985; Orlov 1980; Wang and Sanchez 1982) and to the Dempster–Shafer theory (Dempster 1967; Shafer 1976). A more detailed discussion of these links may be found in my l995 paper "Probability theory and fuzzy logic are complementary rather than competitive," (Zadeh 1995).

In reality, probability theory and fuzzy set theory are distinct theories with different agendas. Scientific theories originate in perceptions. Primitive perceptions such as perceptions of distance, direction, weight, loudness, color, etc. crystallize in early childhood. Among basic perceptions which crystallize at later stages of development are those of likelihood, count, class, similarity and possibility. Fundamentally, probability theory may be viewed as a formalization of perceptions of likelihood and count; fuzzy set theory may be viewed as a formalization of perceptions of class and similarity; and possibility theory may be viewed as a formalization of perception of possibility. It should be noted that perceptions of likelihood and possibility are distinct. Fuzzy set theory and possibility theory are closely related (Zadeh 1978). A key to a better understanding of the nature of the relationship between probability theory and fuzzy set theory is the observation that probability theory is rooted in perceptions of likelihood and count while fuzzy set theory is rooted in perceptions of class and similarity.

In debates over the nature of the relationship between probability theory and fuzzy set theory, there are four schools of thought. The prevailing view within the Bayesian community is that probability theory is sufficient for dealing with uncertainty and imprecision of any kind, implying that there is no need for fuzzy set theory. An eloquent spokesman for this school of thought is an eminent Bayesian, Professor Dennis Lindley. Here is an excerpt of what he had to say on this subject.

> *The only satisfactory description of uncertainty is probability. By this I mean that every uncertainty statement must be in the form of a probability; that several uncertainties must be combined using the rules of probability; and that the calculus of probabilities is adequate to handle all situations involving*

uncertainty… probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate… anything that can be done with fuzzy logic, belief functions, upper and lower probabilities, or any other alternative to probability can better be done with probability (Lindley 1987).

The second school of thought is that probability theory and fuzzy set theory are distinct theories which are complementary rather than competitive. This is the view that is articulated in my l995 Technometrics paper (Zadeh 1995). The third school of thought is that standard probability theory, call it PT, is in need of generalization through addition to PT of concepts and techniques drawn from fuzzy set theory and, more generally, from fuzzy logic, with the understanding that fuzzy set theory is a branch of fuzzy logic. Basically, fuzzy logic, FL, is a precise system of reasoning, computation and deduction in which the objects of discourse are fuzzy sets, that is, classes in which membership is a matter of degree. Thus, in fuzzy logic everything is, or is allowed to be, a matter of degree.

It is important to observe that any bivalent-logic-based theory, T, may be generalized through addition of concepts and techniques drawn from fuzzy logic. Such generalization is referred to as FL-generalization. The view that standard probability theory, PT, can be enriched through FL-generalization is articulated in my 2002 paper "Toward a perception-based theory of probabilistic reasoning" (Zadeh 2002), 2005 paper "Toward a generalized theory of uncertainty (GTU) – an outline" (Zadeh 2005) and 2006 paper "Generalized theory of uncertainty (GTU) – principal concepts and ideas" (Zadeh 2006). The result of FL-generalization, call it PTp, is a generalized theory of probability which has a key capability – the capability to deal with information which is described in a natural language and, more particularly, with perception-based probabilities and relations which are described in a natural language. What is not widely recognized is that many, perhaps most, real-world probabilities are perception-based. Examples: What is the probability that Obama will succeed in solving the financial crisis? What is the probability that there will be a significant increase in the price of oil in the near future? Such probabilities are perception-based and non-numerical. Standard probability theory provides no facilities for computation and reasoning with non-numerical, perception-based probabilities.

The fourth school of thought is that FL-generalization of probability theory should be accompanied by a shift in the foundations of probability theory from bivalent logic to fuzzy logic. This is a radical view which is put forth in my 2004 paper "Probability theory and fuzzy logic – a radical view" (Zadeh 2004).

Is probability theory sufficient for dealing with any kind of uncertainty and imprecision? Professor Lindley's answer is: Yes. In a paper published in 1986 entitled "Is probability theory sufficient for dealing with uncertainty in AI: A negative view," (Zadeh 1986) I argued that the answer is: No. In contradiction to Professor Lindley's assertion, here are some simple examples of problems which do not lend themselves to solution through the use of standard probability theory.

In these examples X is a real-valued variable.
X is larger than approximately a
X is smaller than approximately b

What is the probability that X is approximately c?

Usually X is larger than approximately a
Usually X is smaller than approximately b

What is the probability that X is approximately c?

Usually X is much larger than approximately a
Usually X is much smaller than approximately b

What is the probability that X is approximately c?
What is the expected value of X?

Usually it takes Robert about an hour to get home from work
Robert left work at about 5 pm

What is the probability that Robert is home at 6:l5 pm?

In these examples, question-relevant information is described in natural language. What these examples underscore is that, as was alluded to earlier, standard probability theory does not provide methods of deduction and computation with information described in natural language. Lack of this capability is a serious limitation of standard probability theory, PT. To add this capability to PT it is necessary to FL-generalize PT through addition to PT of concepts and techniques drawn from fuzzy logic.

What would be gained by going beyond FL-generalizat ion of PT, and shifting the foundations of PT from bivalent logic to fuzzy logic? There is a compelling reason for such a shift. At this juncture, most scientific theories, including probability theory, are based on bivalent logic. In bivalent-logic-based theories, the basic concepts are defined as bivalent concepts, with no shades of truth allowed. In reality, most basic concepts are fuzzy, that is, are a matter of degree. For example, in probability theory the concept of independence is defined as a bivalent concept, meaning that two events A and B are either independent or not independent,

with no degrees of independence allowed. But what is quite obvious is that the concept of independence is fuzzy rather than bivalent. The same applies to the concepts of event, stationarity and more generally, to most other basic concepts within probability theory. A shift in the foundations of probability theory would involve a redefinition of bivalent concepts as fuzzy concepts. Such redefinition would enhance the ability of probability theory to serve as a model of reality.

What is widely unrecognized at this juncture is that (a) the capability of probability theory to deal with real-world problems can be enhanced through FL-generalization. Even more widely unrecognized is that (b) the ability of probability theory to serve as a model of reality can be further enhanced through a shift in the foundations of probability theory from bivalent logic to fuzzy logic. But as we move further into the age of machine intelligence and automated decision-making the need for (a) and (b) will become increasingly apparent. I believe that eventually (a) and (b) will gain acceptance.

Acknowledgments

Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant, The Ministry of Communications and Information Technology of Azerbaijan and the BISC Program of UC Berkeley.

About the Author

Lotfi Zadeh was born in 1921 in Baku, Azerbaijan. After his PhD from Columbia University in 1949 in Electrical Engineering, he taught at Columbia for ten years till 1959 where he was a full professor. He joined the Electrical Engineering Department at the University of California, Berkeley in 1959 and served as its chairman from 1963 to 1968. Presently, he is a professor in the Graduate school serving as the Director of BISC (Berkeley Initiative in Soft Computing). He authored a seminal paper in fuzzy sets in 1965. This landmark paper initiated a new direction, which over the past three decades led to a vast literature, and a rapidly growing number of applications ranging from consumer products to subway trains and decision support systems. For this seminal contribution he received the Oldenburger medal from the American Society of Mechanical Engineers in 1993, the IEEE Medal of Honor in 1995, the Okawa prize in 1996, the B. Bolzano Medal from the Academy of Sciences of the Czech Republic, and the Benjamin Franklin Medal in Electrical Engineering. He is a member of the National Academy of Engineering and a Foreign Member of the Polish, Finnish, Korean, Bulgarian, Russian and Azerbaijan Academy of Sciences. He has single-authored over two hundred papers and serves on the editorial boards of over fifty journals. Dr. Zadeh is a recipient of twenty-six honorary doctorates.

Cross References

►Fuzzy Logic in Statistical Data Analysis; Fuzzy Sets: An Introduction
►Fuzzy Sets: An Introduction
►Philosophy of Probability
►Probability Theory: An Outline

References and Further Reading

Coletti G, Scozzafava R (2004) Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Set Syst 144(1):227–249
Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–329
Dubois D, Nguyen HT, Prade H (2000) Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. In: Dubois D, Prade H (eds) Fundamentals of fuzzy sets. The handbooks of fuzzy sets series. Kluwer, Boston, MA, pp 343–438
Freeling ANS (1981) Possibilities versus fuzzy probabilities – Two alternative decision aids. Tech. Rep. 81–6, Decision Science Consortium Inc., Washington, DC
Goodman IR, Nguyen HT (1985) Uncertainty models for knowledge-based systems. North Holland, Amsterdam
Hisdal E (1986) Infinite-valued logic based on two-valued logic and probability. Part 1.1: Difficulties with present-day fuzzy-set theory and their resolution in the TEE model. Int J Man-Mach Stud 25(1):89–111
Hisdal E (1986) Infinite-valued logic based on two-valued logic and probability. Part 1.2: Different sources of fuzziness. Int J Man-Mach Stud 25(2):113–138
Lindley DV (1987) The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. Statistical Science 2:17–24
Loginov VJ (1966) Probability treatment of Zadeh membership functions and their use in pattern recognition, Eng Cybern 68–69
Nurmi H (1977) Probability and fuzziness: some methodological considerations. Unpublished paper presented at the sixth research conference on subjective probability, utility, and decision making, Warszawa
Orlov AI (1980) Problems of optimization and fuzzy variables. Znaniye, Moscow
Ross TJ, Booker JM, Parkinson WJ (eds) (2002) Fuzzy logic and probability applications: bridging the gap. Society for Industrial and Applied Mathematics, Philadelphia, PA
Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton, NJ
Singpurwalla ND, Booker JM (2004) Membership functions and probability measures of fuzzy sets. J Am Stat Assoc 99:467
Stallings W (1977) Fuzzy set theory versus Bayesian statistics. IEEE Trans Syst Man Cybern, SMC-7:216–219
Thomas SF (1995) Fuzziness and probability, ACG Press, Wichita KS
Viertl R (1987) Is it necessary to develop a fuzzy Bayesian inference? In: Viertl R (ed) Probability and Bayesian statistics. Plenum, New York, pp 471–475

Wang PZ, Sanchez E (1982) Treating a fuzzy subset as a projectable random set. In: Gupta MM, Sanchez E (eds) Fuzzy information and decision processes. North Holland, Amsterdam, pp 213–220
Yager RR (1984) Probabilities from fuzzy observations. Inf Sci 32:1–31
Zadeh LA (1965) Fuzzy sets. Inform Contr 8:338–353
Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Set Syst 1:3–28
Zadeh LA (1986) Is probability theory sufficient for dealing with uncertainty in AI: a negative view. In: Kanal LN, Lemmer JF (eds) Uncertainty in artificial intelligence. North Holland, Amsterdam
Zadeh LA (1995) Probability theory and fuzzy logic are complementary rather than competitive. Technometrics 37:271–276
Zadeh LA (2002) Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J Stat Plan Inference 105:233–264
Zadeh LA (2004) Probability theory and fuzzy logic – a radical view. J Am Stat Assoc 99(467):880–881
Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – an outline. Inf Sci 172:1–40
Zadeh LA (2006) Generalized theory of uncertainty (GTU) – principal concepts and ideas. Comput Stat Data Anal 51:15–46

Fuzzy Sets: An Introduction

Madan Lal Puri
Professor Emeritus
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia
Indiana University, Bloomington, IN, USA

Some of the basic properties and implications of the concepts of fuzzy set theory are presented. The notion of a fuzzy set is seen to provide a convenient point of departure for the construction of a conceptual framework which parallels in many respects the framework used in the case of ordinary sets but is more general than the latter. The material presented is from the basic paper of Zadeh (1965) who introduced the notion of fuzzy sets. The reader is also referred to Rosenfeld (1982) for a brief survey of some of the concepts of fuzzy set theory and its application to pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles).

Introduction

In everyday life we often deal with imprecisely defined properties or quantities–e.g., "a few books," "a long story," "a popular teacher," "a tall man," etc. More often than not, the classes of objects which we encounter in the real physical world do not have precisely defined criteria of membership. For example, consider the class of animals. This class clearly includes dogs, horses, birds, etc. as its members, and clearly excludes rocks, fluids, plants, etc. However, such objects as starfish, bacteria, etc. have an ambiguous status with respect to the class of animals. The same kind of ambiguity arises in the case of a number such as 10 in relation to the "class" of all numbers which are much greater than 1.

Clearly, the class of all real numbers which are much greater than 1, or "the class of tall men" do not constitute classes in the usual mathematical sense of these terms. Yet, the fact remains that such imprecisely defined "classes" play an important role in human thinking, particularly, in the domain of pattern recognition, communication of information, decision theory, control theory and medical diagnosis, among others.

The purpose of this note is to provide in a preliminary way some of the basic properties and implications of a concept which is being used more and more in dealing with the type of "classes"mentioned above. The concept in question is that of a "fuzzy set" with a continuum of grades of membership, the concept introduced by Zadeh (1965) in order to allow imprecisely defined notions to be properly formulated and manipulated.

Over the past 20–25 years there has been a tremendous growth of literature on fuzzy sets amounting by now to over 2,000 papers and several textbooks; there is even a journal devoted to this subject.

This note is intended to provide a brief survey of some of the basic concepts of fuzzy sets and related topics.

We begin with some basic definitions.

Definitions

Let $\mathcal{X}$ be a space of points (objects), with a generic element of $\mathcal{X}$ denoted by x. Thus, $\mathcal{X} = \{x\}$.

A fuzzy set (class) A in $\mathcal{X}$ is characterized by a *membership (characteristic) function* $f_A(x)$ which associates with each point x in $\mathcal{X}$ a real number in the interval $[0,1]$, with the value of $f_A(x)$ at x representing the "grade of membership" or "the degree of membership" of x in A. The key idea in fuzzy set theory is that an element has a "degree of membership" in a fuzzy set, and we usually assume that this degree is a real number between 0 and 1. The nearer the value of $f_A(x)$ to unity, the higher the degree of membership of x in A. In the case of the "fuzzy set" of tall men, the elements are men, and their degrees of membership depend on their heights; e.g., a man who is 5 ft tall might have degree 0, a man who is $6\frac{1}{2}$ ft tall might have degree 1, and men of intermediate heights might have intermediate degrees. Analogous remarks apply to such fuzzy sets as the set of young women, the set of rich people, the set of first rate mathematicians, and so on. When A is a set in the ordinary sense of the term, its membership function can take

on only two values 0 and 1, with $f_A(x) = 1$ if $x \in A$ or 0 if $x \notin A$. Thus, in this case, $f_A(x)$ reduces to the familiar characteristic function or indicator function of a so-called crisp set A.

It may be noted that the notion of a fuzzy set is completely nonstatistical in nature, and the rules which are commonly used for manipulating fuzzy set memberships are not the same as the rules for manipulating probabilities.

The Algebra of Fuzzy Subsets

The rules for combining and manipulating fuzzy subsets of $\mathcal{X}$ (Blurrian algebra) should reduce to the rules of ordinary subset algebra when subsets are crisp. This motivates the fuzzy subset algebra introduced by Zadeh (1965), where $\leq$, sup ($\vee$) and inf ($\wedge$) play the roles of $\subseteq$, $\cup$, and $\cap$, respectively. We say that

1. A fuzzy set is empty iff (if and only if) its membership function is $\equiv 0$ on $\mathcal{X}$.
2. Two fuzzy sets A and B are *equal* (and we write $A = B$) iff $f_A(x) = f_B(x)\ \forall\ x \in \mathcal{X}$. (Instead of writing $f_A(x) = f_B(x)\ \forall\ x \in \mathcal{X}$, we shall write $f_A = f_B$).
3. The *complement* of a fuzzy set A is denoted by A' and is defined as $f'_A = 1 - f_A$.
4. $A \subset B$ iff $f_A \leq f_B$.
5. The *union* of two fuzzy sets A and B with respect to respective membership functions $f_A(x)$ and $f_B(x)$ is a fuzzy set C, (and we write $C = A \cup B$) whose membership function is related to those of A and B by

$$f_C(x) = \max[f_A(x), f_B(x)] \quad \forall\ x \in \mathcal{X} \text{ i.e., } f_C = f_A \vee f_B. \tag{I}$$

(Note that the union has the associative property, i.e., $A \cup (B \cup C) = (A \cup B) \cup C$. Also note that a more intuitive way of defining the union is the following: The union of A and B is the smallest fuzzy set containing both A and B. More precisely, if D is any fuzzy set which contains both A and B, then it also contains the union of A and B.)

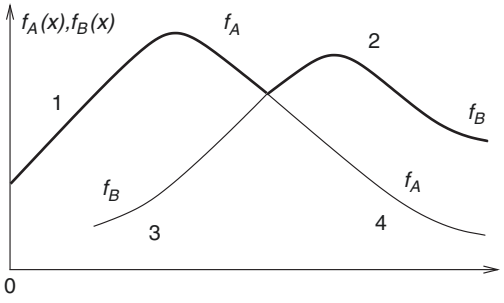
6. The *intersection* of two fuzzy sets A and B with respect to their respective membership functions $f_A(x)$ and $f_B(x)$ is a fuzzy set C (written as $C = A \cap B$) whose membership function f_C is related to those of A and B by $f_C(x) = \min[f_A(x), f_B(x)]\ \forall\ x \in \mathcal{X}$ i.e.,

$$f_C = f_A \wedge f_B.$$

As in the case of union, it is easy to show that the intersection of A and B is the *largest* fuzzy set which is contained in both A and B.

7. A and B are *disjoint* if $A \cap B = C$ is empty, i.e., $f_C(x) \equiv 0\ \forall\ x \in \mathcal{X}$.

Note that $\cap$, like union, has the associative property. Also note that the notion of "belonging" which plays a fundamental role in the case of ordinary sets, does not have the same role in the case of fuzzy sets. Thus, it is not meaningful to speak of a point x "belonging" to a fuzzy set A except in the trivial sense of $f_A(x)$ being positive. Less trivially, one can introduce two levels α and β ($0 < \alpha < 1$, $\beta < \alpha$, $0 < \beta < 1$) and agree to say that (1) $x \in A$ if $f_A(x) \geq \alpha$, (2) $x \notin A$ if $f_A(x) \leq \beta$; and (3) x has an intermediate status relative to A, if $\beta < f_A(x) < \alpha$. This leads to a three valued logic with three truth values: T ($f_A(x) \geq \alpha$), F ($f_A(x) \leq \beta$), and U ($\beta < f_A(x) < \alpha$). Note that the empty and universal (fuzzy) subsets are just the constant functions 0 and 1; they are in fact non-fuzzy!



(Curve segments 1 and 2 comprise the membership function of the union (heavy lines). Curve segments 3 and 4 comprise the membership function of the intersection).

It is clear that these definitions are the extensions of the definitions of $\subseteq$, $\cup$ and $\cap$ for ordinary sets. It is also trivial to verify that they have properties analogous to those of $\subseteq$, $\cup$ and $\cap$, e.g., $A \cup B$ is the $\cap$ of all fuzzy sets $C \supseteq A \subset C$ and $B \subset C$; and $A \cap B$ is the union of all fuzzy sets $C \supseteq C \subset A$ and $C \subset B$. Evidently $\subseteq$ is a partial order relation, and $\cup$ and $\cap$ are commutative, associative and distributive over each other. It is also easy to extend many of the basic identities which hold for ordinary sets to fuzzy sets. As examples, we have the De Morgan's Laws:

$$(A \cup B)' = A' \cap B' \tag{1}$$

$$(A \cap B)' = A' \cup B'. \tag{2}$$

To prove (1), for example, note that the left hand side

$$= 1 - \max[f_A, f_B]$$

$$= \min[1 - f_A, 1 - f_B] = \min[f_{A'}, f_{B'}]$$

$$= \text{Right hand side.}$$

This can easily be verified by testing it for two possible cases: $f_A(x) > f_B(x)$ and $f_A(x) < f_B(x)$. Essentially fuzzy sets in $\mathcal{X}$ constitute a distributive lattice with a 0 and 1 (Birkoff, 1948).

8. The *algebraic product* of A and B is denoted by AB, and is defined in terms of membership functions of A and B by the relation $f_{AB} = f_A \cdot f_B$. Clearly $AB \subset A \cap B$.
9. The *algebraic sum* of A and B is denoted by $A + B$ and is defined as

$$f_{A+B} = f_A + f_B$$

provided $f_A + f_B \leq 1$.

Thus, unlike the algebraic product, the algebraic sum is meaningful only if $f_A(x) + f_B(x) \leq 1 \; \forall \; x \in X$.

10. The *absolute difference* of A and B is denoted by $|A - B|$ and is defined as $f_{|A-B|} = |f_A - f_B|$. Note that in the case of ordinary sets, $|A - B|$ reduces to the relative complement of $A \cap B$ in $A \cup B$. ($|A - B|$ is the symmetric difference $A \triangle B = (A - B) \cup (B - A)$).
11. The dual of algebraic product is the sum $A \oplus B = (A'B')' = A + B - AB$. (Note that for ordinary sets $\cap$ and the algebraic product are equivalent operations, as are $\cup$ and $\oplus$.)

Convex Combination

By a convex combination of two vectors f and g is usually meant a linear combination of f and g of the form $\lambda f + (1 - \lambda)g$ where $0 \leq \lambda \leq 1$. The mode of combining f and g can be generalized to fuzzy sets in the following manner:

Let A, B and C be arbitrary fuzzy sets. *The convex combination* A, B and C is denoted by $(A, B; C)$ and is defined by the relation

$$(A, B; C) = CA + C'B$$

where C' is the complement of C. In terms of membership functions, this means

$$f_{(A,B;C)}(x) = f_C(x)f_A(x) + [1 - f_C(x)]f_B(x), \quad x \in \mathcal{X}.$$

A basic property of the convex combination of A, B and C is expressed as

$$A \cap B \subset (A, B; C) \subset A \cup B \quad \forall \; C.$$

This is an immediate consequence of

$$\min[f_A(x), f_B(x)] \leq \lambda f_A(x) + (1 - \lambda)f_B(x) \leq \max[f_A(x), f_B(x)], \quad x \in \mathcal{X}$$

which holds for all λ in $[0,1]$. It is interesting to observe that given any fuzzy set C satisfying $A \cap B \subset C \subset A \cup B$, one can always find a fuzzy set $D \ni C = (A, B; D)$. The membership function of this set D is given by

$$f_D(x) = \frac{f_C(x) - f_B(x)}{f_A(x) - f_B(x)}, \quad x \in \mathcal{X}.$$

Functions

What about functions? Let f be a function from $\mathcal{X}$ into T, and let A be a fuzzy subset of $\mathcal{X}$ with membership function μ_A. Then, the image of A under f is defined in terms of its membership function μ_A by

$$[f(\mu_A)](y) \equiv \begin{cases} \sup\limits_{f(x)=y} \mu_A(x) \; \forall \; y \in T & \text{i.e., } \sup\limits_{x \in \inf^{-1}(y)} \mu_A(x) \; \forall \; y \in T \\ 0 & \text{if } f^{-1}(y) = \emptyset. \end{cases}$$

Similarly if B is a fuzzy subset of T, then the preimage or inverse image of B under f is defined in terms of its membership function μ_B by

$$[f^{-1}(\mu_B)](x) \equiv \mu_B(f(x)) \quad \forall \; x \in \mathcal{X} \text{ i.e., } f^{-1}(\mu_B) \equiv \mu_B \circ f.$$

Explanation

Let $f : X \to T$. Let B be a fuzzy set in T with membership $\mu_B(y)$. The inverse mapping f^{-1} induces a fuzzy set A in $\mathcal{X}$ whose membership function is defined by

$$\mu_A(x) = \mu_B(y), \quad y \in T$$

for all x in $\mathcal{X}$ which are mapped by f into y.

Consider now the converse problem. Let A be a fuzzy set in $\mathcal{X}$, and as before, let $f : \mathcal{X} \to T$. Question: What is the membership function for the fuzzy set B in T which is induced by this mapping? If f is not $1 : 1$, then an ambiguity arises when two or more distinct points in $\mathcal{X}$, say x_1 and x_2, with different grades of membership in A, are mapped into the same point y in T. In this case, the question is: What grade of membership in B should be assigned to y? To resolve this ambiguity, we agree to assign the larger of the grades of membership to y. More generally, the membership function for B will be defined by

$$\mu_B(y) = \max_{x \in f^{-1}(y)} \mu_A(x), \quad y \in T$$

where $f^{-1}(y) = \{x; x \in X; f(X) = y\}$. Evidently these definitions generalize the standard definitions of the image and the preimage of a subset, and one can verify that these definitions are compatible with fuzzy subset algebra in the usual ways, e.g., one can show that f and f^{-1} have the following properties:

(a) $f^{-1}\left(\bigvee\limits_{i \in I} \mu_{A_i}\right) = \bigvee\limits_{i \in I} f^{-1}(\mu_{A_i})$

(b) $f^{-1}\left(\bigwedge\limits_{i \in I} \mu_{A_i}\right) = \bigwedge\limits_{i \in I} f^{-1}(\mu_{A_i})$

also

(c) $f\left(\underset{i\in I}{\vee}\mu_{A_i}\right) = \underset{i\in I}{\vee} f(\mu_{A_i})$

(d) $f\left(\underset{i\in I}{\wedge}\mu_{A_i}\right) \le \underset{i\in I}{\wedge} f(\mu_{A_i})$

(e) $\overline{f(\mu_A)} \le f(\overline{\mu}_A); \overline{f^{-1}(\mu_B)} = f^{-1}(\overline{\mu}_B).$

$\overline{\quad}$ means complement.

(f) $f(f^{-1}(\mu_B)) \le \mu_B; f^{-1}(f(\mu_A)) \ge \mu_A.$

Proof (a): $f^{-1}\left(\underset{i\in I}{\vee}\mu_A\right) = \left(\underset{i\in I}{\vee}\mu_{A_i}\right)\circ f = \underset{i\in I}{\vee}(\mu_{A_i}\circ f) = \underset{i\in I}{\vee} f^{-1}(\mu_{A_i})$ and so on.

(c)
$$f\left[\underset{i}{\vee}\mu_{A_i}\right](y) = \begin{cases} \sup\limits_{f(x)=y}\left[\underset{i}{\vee}\mu_{A_i}\right](x) \\ 0 \qquad \text{if } x = f^{-1}(y) = \emptyset \end{cases}$$
$$= \begin{cases} \underset{i}{\vee}\sup\limits_{x\in f^{-1}(y)}[\mu_{A_i}(x)] = \underset{i}{\vee} f(\mu_{A_i}) \\ 0 \end{cases}$$

Fuzzy Relations

The concept of a *relation* has a natural extension to fuzzy sets and plays an important role in the theory of such sets and their applications, just as it does in the case of ordinary sets. Ordinarily, a relation is defined as a set of ordered pairs, e.g., the set of all ordered pairs of real numbers x and y such that $x \ge y$. In the context of fuzzy sets, a *fuzzy relation* in $\mathcal{X}$ is a fuzzy set in the product space $\mathcal{X}\times\mathcal{X}$, e.g., the relation denoted by $x \gg y$; $x, y \in \mathbb{R}$ may be regarded as a fuzzy set A in $\mathbb{R}^2$, with the membership function of A, $f_A(x,y)$ having the following (subjective) representative values: $f_A(10,5) = 0; f_A(100,10) = 0.7, f_A(100,1) = 1$, etc.

More generally, one can define an *n–ary fuzzy relation* in $\mathcal{X}$ as a fuzzy set A in the product space $\mathcal{X}\times\mathcal{X}\times\ldots\times\mathcal{X}$. For such relations, the membership function is of the form $f_A(x_1,\ldots,x_n)$ where $x_i \in \mathcal{X}$, $i = 1,\ldots,n$.

In the case of binary fuzzy relations, the composition of two fuzzy relations A and B is denoted by $B\circ A$, and is defined as a fuzzy relation in $\mathcal{X}$ whose membership function is related to those of A and B by

$$f_{B\circ A}(x,y) = \sup_v \min[f_A(x,v), f_B(v,y)]$$
$$= \underset{v}{\vee}[f_A(x,v)\wedge f_B(v,y)]$$

$\forall\, x, y$ and v in $\mathcal{X}$. (Note also that this generalizes the usual definition.) Note that the operation of composition has the associative property: $A\circ(B\circ C) = (A\circ B)\circ C$.

Convexity

A fuzzy set A is *convex* iff the sets Γ_α defined by

$$\Gamma_\alpha = \{x; f_A(x) \ge \alpha\} \tag{3}$$

are convex for all α in the interval $(0,1]$.

An alternative and more direct definition of convexity is the following: A fuzzy set A is *convex* iff

$$f_A[\lambda x_1 + (1-\lambda)x_2] \ge \min[f_A(x_1), f_A(x_2)] \tag{4}$$

for all x_1 and x_2 in $\mathcal{X}$ and all λ in $[0,1]$.

Note that this definition does not imply that the function $f_A(x)$ must be a convex function of x.

It can be seen that the two definitions are equivalent (see Zadeh 1965).

A basic property of convex fuzzy sets is:

Theorem *If A and B are fuzzy convex, then $A\wedge B$ is also fuzzy convex.*

Boundedness

A fuzzy set A is bounded iff the sets $\Gamma_\alpha = \{x; f_A(x) \ge \alpha\}$ are bounded for all $\alpha > 0$; i.e., for all $\alpha > 0$, $\exists$ a finite $R(\alpha) \ni \|x\| \le R(\alpha)$ for all x in Γ_α.

If A is a bounded set, then for all $\varepsilon > 0$, $\exists$ a hyperplane $H \ni f_A(x) \le \varepsilon\ \forall\ x$ on the side of H which does not contain the origin. For example consider the set $\Gamma_\varepsilon = \{x; f_A(x) \ge \varepsilon\}$. By hypothesis this set is contained in a sphere S of radius $R(\varepsilon)$. Let H be any hyperplane supporting S. Then, all points on the side of H which does not contain the origin lie outside or on S, and hence for all such points $f_A(x) \le \varepsilon$.

Preliminary

As a preliminary, let A and B be two bounded fuzzy sets and let H be a hypersurface in $\mathbb{R}^{(n)}$ defined by the equation $h(x) = 0$ with all points for which $h(x) \ge 0$ being on one side of H and all points for which $h(x) \le 0$ being on the other side. Let K_H be a number dependent on $H \ni f_A(x) \le K_H$ on one side of H and $f_B(x) \le K_B$ on the other side. Let $M_H = \inf K_H$. The number $D_H = 1 - M$ is called the *degree of separation* of A and B by H.

In general one is concerned not with a given hypersurface H, but with a family of hypersurfaces $\{H_\lambda\}$, with λ

ranging over $\mathbb{R}^{(m)}$. The problem then is to find a member of this family which realizes the highest degree of separation.

A special case of this problem is one where the H_λ are hyperplanes in $\mathbb{R}^{(n)}$, with λ ranging over $\mathbb{R}^{(n)}$. In this case we define the *degree of separation* of A and B by

$$D = 1 - M \quad \text{where} \quad M = \inf_H M_H.$$

Separation of Convex Fuzzy Sets

The classical separation theorem for ordinary convex sets states, in essence, that if A and B are disjoint convex sets, then there exists a separating hyperplane H such that A is on one side of H and B is on the other side. This theorem can be extended to convex fuzzy sets, without requiring that A and B be disjoint, since the condition of disjointedness is much too restrictive in the case of fuzzy sets. It turns out that the answer is in the affirmative.

Theorem *Let A and B be bounded convex fuzzy sets in $\mathbb{R}^{(n)}$, with maximal grades M_A and M_B respectively i.e., $M_A = \sup_x f_A(x)$, $M_B = \sup_x M_B(x)$. Let M be the maximal grade for the intersection $A \cap B$ (i.e., $M = \sup_x \min[f_A(x), f_B(x)]$). Then $D = 1 - M$. (D is called the degree of separation of A and B by the hyperplane H).*

In other words, the theorem states that the highest degree of separation of two convex fuzzy sets that can be achieved with a hyperplane in $R^{(n)}$ is one minus the maximal grade in the intersection $A \cap B$. Zadeh has applied these types of results in the problems of optimization, pattern discrimination, etc.

Concluding Remarks

The concepts of fuzzy sets and fuzzy functions have been found useful in many applications, notably in pattern recognition, clustering, information retrieval and systems analysis, among other areas (cf. Negoita and Ralescu 1975). Motivated by some of these applications and related problems, Puri and Ralescu (1982, 1983) introduced the integration on fuzzy sets and differentials of fuzzy functions. This led to the study of fuzzy random variables, their expectations, concept of normality for fuzzy random variables and different limit theorems for fuzzy random variables (cf. Puri and Ralescu (1985, 1986, 1991), Klement, Puri and Ralescu (1984, 1986), and Proske and Puri (2002a, b and the references cited in these papers).

About the Author

Professor Puri was ranked the fourth most prolific statistician in the world for his writings in the top statistical journals in a 1997 report by the Natural Sciences and Engineering Research Council of Canada. Among statisticians in universities which do not have separate departments of statistics, Puri was ranked number one in the world by the same report. Puri has received a great many honors for his outstanding contributions to statistics and we mention only a few. Professor Puri twice received the Senior U.S. Scientist Award from Germany's Alexander von Humboldt Foundation, and he was honored by the German government in recognition of past achievements in research and teaching. Madan Puri has been named the recipient of the 2008 Gottfried E. Noether Senior Scholar Award (an annual, international prize honoring the outstanding statisticians across the globe), for "outstanding contributions to the methodology and/or theory and teaching of nonparametric statistics that have had substantial, sustained impact on the subject, its practical applications and its pedagogy." For many years Professor Puri has been highly cited researcher in mathematics according to ISI Web of knowledge ISI HighlyCited. com According to For many years Professor Puri has been highly cited researcher in mathematics according to ISI Web of Knowledge ISI HighlyCited.Com Professor Puri, his greatest honor came in 2003 when under the editorship of Professors Peter Hall, Marc Hallin, and George Roussas, the International Science Publishers published "Selected Collected Works of Madan L. Puri," a series of three volumes, each containing about 800 pages.

Cross References

►Fuzzy Logic in Statistical Data Analysis: Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Statistical Methods for Non-Precise Data

References and Further Reading

Klement EP, Puri ML, Ralescu DA (1984) Law of large numbers and central limit theorem for fuzzy random variables. Cybern Syst Anal 2:525–529
Klement EP, Puri ML, Ralescu DA (1986) Limit theorems for fuzzy random variables. Proc R Soc London 407:171–182
Negoita CV, Ralescu DA (1975) Applications of fuzzy sets to system analysis. Wiley, New York
Proske F, Puri ML (2002a) Central limit theorem for Banach space valued fuzzy random variables. Proc Am Math Soc 130: 1493–1501
Proske F, Puri ML (2002b) Strong law of large numbers for Banach space valued fuzzy random variables. J Theor Probab 15:543–552
Puri ML, Ralescu DA (1982) Integration on fuzzy sets. Adv Appl Math 3:430–434
Puri ML, Ralescu DA (1983) Differentials of fuzzy functions. J Math Anal Appl 91:552–558

Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Probab 13:1373–1379
Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
Puri ML, Ralescu DA (1991) Limit theorems for fuzzy martingales. J Math Anal Appl 160:107–122
Rozenfeld A (1982) How many are a few? Fuzzy sets, fuzzy numbers, and fuzzy mathematics. Math Intell 4:139–143
Singpurwalla ND, Booker JM (2004) Membership functions and probability measures of fuzzy sets. J Am Stat Assoc 99: 867–889
Zadeh LA (1965) Fuzzy sets. Inform Contr 8:338–353