P

Panel Data

Markku Rahiala
Professor (emeritus) of Econometrics
University of Oulu, Oulu, Finland

Panel data (or longitudinal data) are data sets, where information on a number of observational units have been collected at several time points. These observational units are usually called *individuals* or *subjects*. In economic applications they might be households, firms, individual persons, countries, investors or other economic agents. In medical, biological and social applications the subjects might be patients, test animals, individual persons etc.

Panel data are most often used to study unidirectional relationships between some explanatory variables $X_{it} = (x_{1,i,t} \ \ldots \ x_{m,i,t})'$ and a continuous dependent variable y_{it}, where i ($i = 1, \ldots, N$) refers to individual and t ($t = t_{oi}, \ldots, T_i$) refers to time or period. (To simplify notation, we will assume $t_{oi} \equiv 1$ and $T_i \equiv T$.) Panel data have many advantages over cross sectional data or single aggregated time series. One can for instance study the dynamic effects of the covariates on a micro level and one can explore heterogeneities among the individuals. Regression models of the linear type (for suitably transformed variables)

$$y_{it} = \alpha_i + \beta' X_{it} + \varepsilon_{it} \tag{1}$$

are especially popular as model frameworks. The error terms ε_{it} are assumed to be independent of the explanatory variables X_{it}, the processes $\{\varepsilon_{it}\}$ are assumed stationary with zero means and the data from different individuals are assumed independent of each other. The length of the data T is often fairly short, whereas the number of subjects N might be large. The levels of any dependent variable usually show considerable individual variation, which makes it necessary to allow for individually varying intercept terms α_i in model (1). If these intercepts are taken as fixed parameters, the model is called a *fixed effects* model. When N is large, this will however lead to some inferential problems. For instance ML estimators of the variance–covariance structure of the error processes would be inconsistent for fixed T and increasing N. This is why the intercept terms are often treated as mutually independent random variables $\alpha_i \sim \mathrm{IID}(\alpha, \tau^2)$ (α_i also independent of $\{\varepsilon_{it}\}$), whenever the observed subjects can be interpreted as a sample from a larger population of potential subjects. These *random effects* models are special cases of so-called *mixed* models or *variance components* models. If the effects of the covariates X_{it} vary over individuals, the regression coefficients $\beta_{(i)}$ can similarly be interpreted as random variables, $\begin{pmatrix}\alpha_i & \beta_{(i)}'\end{pmatrix}' \sim \mathrm{IID}_{m+1}\left(\begin{pmatrix}\alpha & \beta'\end{pmatrix}', G\right)$. If all the random elements in the model were assumed normally distributed, the whole model for subject i could be written as

$$\begin{aligned} Y_i &= (y_{i1} \ \ldots \ y_{iT})' \\ &= \alpha + X_i\beta + Z_i U_i + \varepsilon_i, \quad \varepsilon_i \text{ independent of } U_i, \\ U_i &\sim \mathrm{NID}_{m+1}(0, G), \quad \varepsilon_i = (\varepsilon_{i1} \ \ldots \ \varepsilon_{iT})' \nu \\ &\sim \mathrm{NID}_T(0, R), \end{aligned} \tag{2}$$

where $U_i = \begin{pmatrix}\alpha_i & \beta_{(i)}'\end{pmatrix}' - \begin{pmatrix}\alpha & \beta'\end{pmatrix}'$, $X_i = (X_{i1} \ \ldots \ X_{iT})'$, $Z_i = (\mathbf{1}_T \quad X_i)$ and $\mathbf{1}_T = (1 \ \ldots \ 1)'$. This is a standard form of a mixed model leading to GLS estimators for α and β once the parameters incorporated in the matrices G and R have been estimated.

To take account of possible unobserved changes in the general (either economic or biological) environment, one can include an additive term $\gamma_{(i)}' F_t$ in model (1), where F_t denotes an r-dimensional vector of common factors and $\gamma_{(i)}$ is a vector containing the factor loadings for individual i. These common factors will induce dependencies between the y_{it}-observations from different individuals. (See e.g., Pesaran 2006.)

In model (1), all the explanatory variables were assumed *exogenous*. However, econometric models quite often include also lagged values of the dependent variable as regressors. Much of economic theory starts from the assumption that the economic agents are optimizing an *intertemporal* utility function, and this assumption often induces an autoregressive, dynamic model

$$\begin{aligned} y_{it} &= \alpha_i + \phi_1 y_{i,t-1} + \ldots + \phi_p y_{i,t-p} + \beta' X_{it} + \varepsilon_{it}, \\ \varepsilon_{it} &\sim \mathrm{IID}(0, \sigma^2) \end{aligned} \tag{3}$$

for the dependent variable y_{it}. In case of random intercepts α_i, the lagged values of the dependent variable and

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,
© Springer-Verlag Berlin Heidelberg 2011

the combined error terms $(\alpha_i - \alpha) + \varepsilon_{it}$ will be correlated. This would lead to inconsistent least squares estimators for the ϕ -parameters. The problem can be circumvented by the GMM estimation method (Generalized Method of Moments). (See e.g., the Appendices in Arellano 2003.) Once the order p of the autoregressive model (3) has been correctly specified, it will be easy to find valid instruments for the GMM estimationt among the lagged differences of the dependent variable. Model (3) can be straightforwardly extended to the vector-valued case. (See e.g., Hsiao 2003, Chap. 4.7.)

If the dependent variable y_{it} is *discrete* (either a count variable or measured on a nominal or ordinal scale), one can combine the basic idea of model (1) and the concept of ►*generalized linear models* by assuming that the covariates X_{it} and the heterogeneity terms α_i affect the so-called *linear predictors* analogously to (1),

$$\eta_{it} = g\left(E\left(y_{it} \mid X_{it}, \alpha_i\right)\right) = \alpha_i + \beta' X_{it} \qquad (4)$$

and by assuming that conditionally on X_{it} and α_i , y_{it} follows a distribution belonging to the exponential family of distributions.1 (See e.g. Diggle et al. 2001, Chap. 11, or Fitzmaurice et al. 2004, Chap. 12.) Function g is called the *link* function, and models (4) are called *generalized linear mixed models* (GLMM). If for instance y_{it} would be dichotomous obtaining the values 0 or 1, *logit* link function would lead to the model

$$P\left(y_{it} = 1 \mid X_{it}, \alpha_i\right) = \exp(\alpha_i + \beta' X_{it}) / \left(1 + \exp(\alpha_i + \beta' X_{it})\right).$$

The resulting likelihood function contains complicated integral expressions, but they can be effectively approximated by numerical techniques.

About the Author

Markku Rahiala received his Ph.D. in statistics in 1985 at the University of Helsinki. He is Professor of econometrics, University of Oulu since 1998. He was Associate editor of the *Scandinavian Journal of Statistics* (1989–1995). He is Member of the Econometric Society since 1985.

Cross References

►Data Analysis
►Event History Analysis
►Linear Mixed Models
►Medical Statistics
►Multilevel Analysis
►Nonsampling Errors in Surveys
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Repeated Measures
►Sample Survey Methods
►Social Network Analysis
►Statistical Analysis of Longitudinal and Correlated Data
►Testing Variance Components in Mixed Linear Models

References and Further Reading

Arellano M (2003) Panel data econometrics. Oxford University Press, Oxford

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL (2001) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford

Fitzmaurice GM, Laird NM, Ware JH (2004) Applied longitudinal analysis. Wiley, Hoboken

Hsiao C (2003) Analysis of panel data, 2nd edn. Cambridge University Press, Cambridge

Pesaran MH (2006) Estimation and inference in large heterogeneous panels with multifactor error structure. Econometrica 74: 967–1012

Parametric and Nonparametric Reliability Analysis

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

Introduction

Reliability is "the probability that a piece of equipment (component, subsystem or system) successfully performs its intended function for a given period of time under specified (design) conditions" (Martz and Waller 1982). Failure means that an item does not perform its required functions. To evaluate the performance of an item, to predict its failure time and to find its failure pattern is the subject of Reliability.

Mathematically we can define reliability, $R(t)$, as follows:

$$R(t) = \Pr(T > t) = 1 - \int_0^t f(\tau)\, d\tau, \quad (t \geq 0)$$

where T denotes the failure time of the system or component, and $f(t)$ the failure probability distribution.

The main entity in performing accurate reliability analysis depends on having properly identified a classical discrete or continuous probability distribution that will characterize the behavior of the failure data. In practice, scientists and engineers either assume one, such as the exponential, Weibull, Poisson, etc., or a perform goodness of fit test to properly identify the failure distribution and then proceed with the reliability analysis. It is possible that the assumed failure distribution is not the correct one and

furthermore, the goodness of fit test methodology failed to identify a classical probability distribution. Thus, proceeding with the reliability analysis will result in misleading and incorrect results.

In this brief document we discuss a nonparametric reliability procedure when one cannot identify a classical failure distribution, $f(t)$, to characterize the failure data of the system. The method is based on estimating the failure density through the concept of distribution-free kernel density method. Utilizing such methods on the subject area offers significant computation difficulties. Therefore, in order to use this method, one must be able to obtain the optimal bandwidth for the kernel density estimate. Here, we recommend a six-step procedure which one can apply to compute the optimal nonparametric probability distribution that characterizes the failure times. Some useful references on the subject matter are Bean and Tsokos (1980, 1982), Liu and Tsokos (2001, 2002a, b), Qiao and Tsokos (1994, 1995), Rust and Tsokos (1981), Silverman (1986), and Tsokos and Rust (1980). First we briefly discuss the parametric approach to reliability using the popular three-parameter Weibull probability distribution as the failure model. Some additional useful failure models can be found in Tsokos (1998, 1995).

Parametric Approach to Reliability

The two-parameter Weibull probability distribution is universally used to characterize the failure times of a system or component to study its reliability behavior instead of the three-parameter Weibull model. Recently, methods have been developed along with effective algorithms for which one can obtain estimates of the three-parameter Weibull probability distribution. Here we will use the three-parameter Weibull failure model.

The three-parameter Weibull failure model is given by

$$\hat{f}(t) = \frac{\hat{c}\,(t-\hat{a})^{\hat{c}-1}}{\hat{b}^{\hat{c}}} \exp\left\{-\left(\frac{t-\hat{a}}{\hat{b}}\right)^{\hat{c}}\right\},$$

where $\hat{a}$, $\hat{b}$ and $\hat{c}$ are the maximum likelihood estimates to the location, scale and shape parameters, respectively. For calculation of the estimates of a, b and c, see Qiao and Tsokos (1994, 1995).

Thus, the parametric reliability estimation $\hat{R}_p(t)$ of the three-parameter Weibull model is given by

$$\hat{R}_p(t) = \exp\left\{-\left(\frac{t-\hat{a}}{\hat{b}}\right)^{\hat{c}}\right\}.$$

The goodness of fit criteria that one can use in identifying the appropriate classical failure probability distribution to characterize the failure times is the popular ►Kolmogorov–Smirnov test.

Briefly, it tests the null hypothesis that the data $\{t_j\}_{j=1}^n$ is from some specified classical probability distribution against the alternative hypothesis that it is from another probability distribution. That is,

$$\begin{cases} H_0 : \{t_j\}_{j=1}^n \sim F(t), \\ H_1 : \{t_j\}_{j=1}^n \not\sim F(t). \end{cases}$$

Let $F_n^*(t)$ be the empirical distribution function for the failure data. The Kolmogorov–Smirnov statistic is defined by

$$D_n = \sum_t \left|F_n^*(t) - F(t)\right|.$$

The statistic D_n can be easily calculated from the following formula:

$$D_n = \max\left\{\max_{1\le i\le n}\left[\frac{i}{n} - F\left(t_{(i)}\right)\right], \max_{1\le i\le n}\left[\frac{i-1}{n} - F\left(t_{(i)}\right)\right]\right\},$$

where $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ is the order statistic of $\{t_j\}_{j=1}^n$.

Let $D_{n,\alpha}$ be the upper α-percent point of the distribution of D_n, that is,

$$P\{D_n > D_{n,\alpha}\} \le \alpha.$$

Tables for the exact critical values $D_{n,\alpha}$ are available. See Miller (1956) and Owen (1962), among others, to make the appropriate decision.

Nonparametric Approach to Reliability

Let $\{t_j\}_{j=1}^n$ be the failure data characterized by the probability density function $f(t)$. Then the nonparametric probability density estimation $\hat{f}$ can be written as

$$\hat{f}_{\hat{h}}(t) = \frac{1}{n\hat{h}} \sum_{j=1}^{n} K\left(\frac{t-t_j}{\hat{h}}\right)$$

where $K(t)$ is the kernel and assumed to be Gaussian given by

$$K(t) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}$$

and $\hat{h}$ is the estimate of the optimal bandwidth. There are several other choices for the kernel, namely, Epanechnikov, Corine, biweight, triweight, triangle and uniform. For further information regarding the selection process, see Silverman (1986). The most important element in $\hat{f}_{\hat{h}}(t)$ being effective to characterize the failure data is the bandwidth, $\hat{h}$. Given below is a procedure that works fairly well in obtaining optimal $\hat{h}$ and then $\hat{f}_{\hat{h}}(t)$ for the failure data,

(Bean and Tsokos 1982; Silverman 1986). This procedure is summarized below.

- Calculate S^2 and T_4 using the failure data $t_1, t_2, \ldots, t_n$:

$$S^2 = (n-1)^{-1} \sum_{j=1}^{n} (t_j - \bar{t})^2 \text{ and } T_4 = \frac{1}{n} \sum_{j=1}^{n} (t_j - \bar{t})^4.$$

- Determine a value for U_2 and U_4 as defined below:

$$U_2 \approx S^2 \text{ and } U_4 \approx \frac{n^3}{(n-1)(n^2-2n+3)} \left(T_4 - \frac{3(n-1)(2n-3)}{n^3} S^2 \right).$$

- Find estimates of the parameters μ and σ:

$$\hat{\mu} = \sqrt[4]{\frac{3U_2^2 - U_4}{2}} \text{ and } \hat{\sigma} = \sqrt{U_2 - \hat{\mu}^2}.$$

- Calculate $\int_{-\infty}^{\infty} f''^2(t)\, dt$ from the following:

$$\int_{-\infty}^{\infty} f''^2(t)\, dt = \frac{3}{16\sqrt{\pi}\hat{\sigma}^5} + \frac{1}{4\sqrt{\pi}\hat{\sigma}^5} e^{-\frac{\hat{\mu}^2}{\hat{\sigma}^2}} \left(\frac{3}{4} - \frac{3\hat{\mu}^2}{\hat{\sigma}^2} + \frac{\hat{\mu}^4}{\hat{\sigma}^4} \right).$$

- Find h_{opt} from the following:

$$h_{opt} = 2^{-\frac{1}{5}} \pi^{-\frac{1}{10}} n^{-\frac{1}{5}} \left\{ \int_{-\infty}^{\infty} f''^2(t) dt \right\}^{-\frac{1}{5}}.$$

- Obtain the estimate of the nonparametric failure distribution of the data:

$$\hat{f}(t) = \frac{1}{n h_{opt}} \sum_{j=1}^{n} K\left(\frac{t - t_j}{h_{opt}} \right).$$

There are other methods for dealing with the optimal bandwidth selection, see Bean and Tsokos (1982) and Silverman (1986).

The nonparametric estimate of the failure probability distribution $\hat{R}_{np}(t)$ can be obtained,

$$\begin{aligned} \hat{R}_{np}(t) &= \int_t^{\infty} \hat{f}(\tau)\, d\tau \\ &= \int_t^{\infty} \frac{1}{n\hat{h}} \sum_{j=1}^{n} K\left(\frac{\tau - t_j}{\hat{h}} \right) d\tau \\ &= \frac{1}{n\hat{h}} \sum_{j=1}^{n} \int_t^{\infty} K\left(\frac{\tau - t_j}{\hat{h}} \right) d\tau \\ &= \frac{1}{n} \sum_{j=1}^{n} \int_{-\infty}^{\frac{t-t_j}{\hat{h}}} K(\tau)\, d\tau. \end{aligned}$$

To evaluate the integral in the above equation, let

$$\Phi(t) = \int_{-\infty}^{t} K(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx.$$

Then we have

$$\begin{aligned} \hat{R}_{np}(t) &= \frac{1}{n} \sum_{j=1}^{n} \left[1 - \Phi\left(\frac{t - t_j}{\hat{h}} \right) \right] \\ &= 1 - \frac{1}{n} \sum_{j=1}^{n} \Phi\left(\frac{t - t_j}{\hat{h}} \right). \end{aligned}$$

To calculate $\Phi(t)$, let

$$\int_0^u e^{-x^2} dx = e^{-u^2} \sum_{k=0}^{\infty} \frac{2^k \cdot u^{2k+1}}{(2k+1)!!}.$$

It follows that we can write

$$\begin{aligned} \Phi(t) &= \int_0^t K(x)\, dx + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}} dx + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2} \int_0^{\frac{t}{\sqrt{2}}} e^{-\tau^2} d\tau + 0.5 \\ &= \frac{1}{\sqrt{\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \sum_{k=0}^{\infty} \frac{2k \left(\frac{t}{\sqrt{2}} \right)^{2k+1}}{(2k+1)!!} + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!!} + 0.5. \end{aligned}$$

Note that $\Phi(t) \approx 0$ when $t < -4$, and $\Phi(t) \approx 1$ when $t > 4$. Then we need to carry out the summation in the interval $|t| \le 4$.

Since the sum converges quite fast, when $|t| < 4$, we have overcome the numerical difficulty in the calculation of the nonparametric reliability. An efficient numerical procedure is given below to evaluate the above summation.

Step 1 Construct a subroutine for calculating $\Phi(t)$ as follows:

1. Notations: At input, t stores the point; at output, p stores $\Phi(t)$. Other values for the computation: cc, each term of the sum; tt stores the value $t * t$, to save computer time.
2. Let $p = t$, $cc = t$, $tt = t * t$.
3. For $k = 1, 2, 3, \ldots$, perform (4) through (5) that follows.
4. If $|cc| <$ tolerance (we use 10^{-4} for tolerance), then

$$p = \frac{1}{\sqrt{2\pi}} e^{-\frac{tt}{2}} \cdot p + 0.5,$$

output p and exit. Otherwise continue with (5).

5. $cc = cc \cdot tt/(2k+1)$, $p = p + cc$.

Step 2 Find the optimal bandwidth $\hat{h}$ from the six-step procedure introduced in section "▶Parametric Approach to Reliability".

Step 3 The reliability function at any given point t is given by

$$\hat{R}_{np}(t) = 1 - \frac{1}{n}\sum_{j=1}^{n} \Phi\left(\frac{t - t_j}{\hat{h}}\right).$$

Several applications of real data, along with Monte Carlo simulations and the nonparametric kernel probability estimate of reliability, give very good results in comparison with the parametric version of the reliability function.

About the Author

For biography *see* the entry ►Mathematical and Statistical Modeling of Global Warming.

Cross References

►Bayesian Reliability Modeling
►Degradation Models in Reliability and Survival Analysis
►Imprecise Reliability
►Kolmogorov-Smirnov Test
►Nonparametric Density Estimation
►Ordered Statistical Data: Recent Developments
►Tests of Fit Based on The Empirical Distribution Function
►Weibull Distribution

References and Further Reading

Bean S, Tsokos CP (1980) Developments in nonparametric density estimation. Int Stat Rev 48(3):267–287
Bean S, Tsokos CP (1982) Bandwidth selection procedures for kernel density estimates. Comm Stat A Theor 11(9):1045–1069
Liu K, Tsokos CP (2001) Kernel estimates of symmetric density function. Int J Appl Math 6(1):23–34
Liu K, Tsokos CP (2002a) Nonparametric reliability modeling for parallel systems. Stochastic Anal Appl 20(1):185–197
Liu K, Tsokos CP (2002b) Optimal bandwidth selection for a nonparametric estimate of the cumulative distribution function. Int J Appl Math 10(1):33–49
Martz HF, Waller RA (1982) Bayesian reliability analysis. Wiley Series in probability and mathematical statistics: applied probability and statistics. Wiley, Chichester
Miller LH (1956) Table of percentage points of Kolmogorov statistics. J Am Stat Assoc 51:111–121
Owen DB (1962) Handbook of statistical tables. Addison-Wesley, Reading, MA
Qiao H, Tsokos CP (1994) Parameter estimation of the Weibull probability distribution. Math Comput Simulat 37(1):47–55
Qiao H, Tsokos CP (1995) Estimation of the three parameter Weibull probability distribution. Math Comput Simulat 39(1–2):173–185
Rust A, Tsokos CP (1981) On the convergence of kernel estimators of probability density functions. Ann Inst Stat Math 33(2):233–246
Silverman BW (1986) Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman and Hall, London
Tsokos CP (1995) Reliability growth: nonhomogeneous Poisson process. In: Balakrishnan N, and Cohen AC (eds) Recent advances in life-testing and reliability. CRC, Boca Raton, pp 319–334
Tsokos CP (1998) Ordinary and Bayesian approach to life testing using the extreme value distribution. In: Basu AP, Basu SK, Mukhopadhyay S (eds) Frontiers in reliability analysis volume 4 of Series on Quality, Reliability and Engineering Statistics, World Scientific, Singapore, pp 379–395
Tsokos CP, Rust A (1980) Recent developments in nonparametric estimation of probability density. In: Applied stochastic processes (Proc. Conf., Univ. Georgia, Athens, Ga., 1978), Academic, New York, pp 269–281

Parametric Versus Nonparametric Tests

David J. Sheskin
Professor of Psychology
Western Connecticut State University, Danbury, CT, USA

A common distinction made with reference to statistical tests/procedures is the classification of a procedure as *parametric* versus *nonparametric*. This distinction is generally predicated on the number and severity of assumptions regarding the population that underlies a specific test. Although some sources use the term *assumption free* (as well as *distribution free*) in reference to nonparametric tests, the latter label is misleading, in that nonparametric tests are not typically assumption free. Whereas *parametric statistical tests* make certain assumptions with respect to the characteristics and/or parameters of the underlying population distribution upon which a test is based, *nonparametric tests* make fewer or less rigorous assumptions. Thus, as Marascuilo and McSweeney (1977) suggest, nonparametric tests should be viewed as *assumption freer* tests. Perhaps the most common assumption associated with parametric tests that does not apply to nonparametric tests is that data are derived from a normally distributed population.

Many sources categorize a procedure as parametric versus nonparametric based on the level of measurement a set of data being evaluated represents. Whereas parametric tests are typically employed when interval or ratio data are evaluated, nonparametric tests are used with rank-order (ordinal) and categorical data. Common example of parametric procedures employed with interval or ratio data are ►*Student's t tests*, *analysis of variance* procedures (see ►Analysis of Variance), and the *Pearson product moment coefficient of correlation* (see ►Correlation Coefficient). Examples of commonly

described nonparametric tests employed with rank-order data are the *Mann–Whitney U test*, *Wilcoxon's signed–ranks and matched–pairs signed ranks tests*, the *Kruskal–Wallis one-way analysis of variance by ranks*, the *Friedman two–way analysis of variance by ranks*, and *Spearman's rank order correlation coefficient*. Examples of commonly described nonparametric tests employed with categorical data are ►*chi-square tests* such as the *goodness-of-fit test*, *test of independence*, and *test of homogeneity* and the *McNemar test*.

Researchers are in agreement that since ratio and interval data contain a greater amount of information than rank-order and categorical data, if ratio or interval data are available it is preferable to employ a parametric test for an analysis. One reason for preferring a parametric test is that the latter type of test generally has greater *power* than its nonparametric analog (i.e., a parametric test is more likely to reject a false null hypothesis). If, however, one has reason to believe that one or more of the assumptions underlying a parametric test have been saliently violated (e.g., the assumption of underlying normal population distributions associated with the *t test for independent samples*), many sources recommend that a nonparametric test (e.g., a rank-order test that does not assume population normality such as the *Mann–Whitney U test*, which can also be used to evaluate data involving two independent samples) will provide a more reliable analysis of the data. Yet other sources argue it is still preferable to employ a parametric test under the latter conditions, on the grounds that parametric tests are, for the most part, *robust*. A *robust test* is one that still allows a researcher to obtain reasonably reliable conclusions even if one or more of the assumptions underlying the test are violated. As a general rule, when a parametric test is employed in circumstances when one or more of its assumptions are thought to be violated, an adjusted probability distribution is employed in evaluating the data. Sheskin (2007, p. 108) notes that in most instances, the debate concerning whether a researcher should employ a parametric or nonparametric test for a specific experimental design turns out to be of little consequence, since in most cases data evaluated with both a parametric test and its nonparametric analog will result in a researcher reaching the same conclusions.

Cross References

►Analysis of Variance
►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Asymptotic Relative Efficiency in Testing
►Chi-Square Test: Analysis of Contingency Tables
►Explaining Paradoxes in Nonparametric Statistics
►Fisher Exact Test
►Frequentist Hypothesis Testing: A Defense
►Kolmogorov-Smirnov Test
►Measures of Dependence
►Mood Test
►Multivariate Rank Procedures: Perspectives and Prospectives
►Nonparametric Models for ANOVA and ANCOVA Designs
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Permutation Tests
►Randomization Tests
►Rank Transformations
►Robust Inference
►Scales of Measurement and Choice of Statistical Methods
►Sign Test
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Inference
►Statistical Inference: An Overview
►Student's t-Tests
►Validity of Scales
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Marascuilo LA, McSweeney M (1977) Nonparatmetric and distribution-free methods for the social sciences. Brooks/Cole, Monterey, CA

Sheskin DJ (2007) Handbook of parametric and nonparametric statistical procedures, 4th edn. Chapman and Hall/CRC, Boca Raton

Pareto Sampling

Lennart Bondesson
Professor Emeritus
Umeå University, Umeå, Sweden

Introduction

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be a population of units. To get information about the population total Y of some interesting y-variable, unequal probability sampling is a widely applied method. Often a random sample of a fixed number, n, of distinct units is to be selected with prescribed inclusion probabilities π_i, $i = 1, 2, \ldots, N$, for the units in $\mathcal{U}$. These π_i should sum to n and are usually chosen to be proportional to some auxiliary variable. In this form unequal

probability sampling is called fixed size πps sampling (ps = proportional to size). By the help of a πps sample and recorded y-values for the units in the sample, the total Y can be estimated unbiasedly by the ►Horvitz–Thompson estimator $\hat{Y}_{HT} = \sum_{i\in s} y_i/\pi_i$, where s denotes the sample. There are many possible fixed size πps sampling designs, see, e.g., Brewer and Hanif (1983) and Tillé (2006).

This article treats Rosén's (1997a,b) Pareto order πps sampling design. Contrary to many other fixed size πps designs, it is very easy to implement. However, it is only approximate. Independently of Rosén, Saavedra (1995) suggested the design. Both were inspired by unpublished work of Ohlsson, cf. Ohlsson (1998).

The Pareto Design

Let U_i, $i = 1,2,\ldots,N$, be independent $U(0,1)$-distributed random numbers. Further, let

$$Q_i = \frac{U_i/(1-U_i)}{p_i/(1-p_i)}, \quad i = 1,2,\ldots,N,$$

be so-called ranking variables. Put $p_i = \pi_i$, $i = 1,2,\ldots,N$, and select as sample those n units that have the smallest Q-values. The factual inclusion probabilities π_i^* do not equal the desired π_i but approximately. For $d = \sum_{i=1}^{N} \pi_i(1-\pi_i)$ not too small ($d > 1$), the agreement is surprisingly good and better the larger d is.

The main advantages of the method are its simplicity, its high ►entropy, and that the U_is can be used as permanent random numbers, i.e., can be reused when at a later occasion the population, more or less altered, is re-sampled. In this way changes can be better estimated.

The name of the method derives from the fact that $u/(1-u) = F^{-1}(u)$, where F^{-1} is the inverse of the special Pareto distribution function $F(x) = x/(1+x)$ on $(0,\infty)$. A general order sampling procedure uses instead $Q_i = F^{-1}(U_i)/F^{-1}(\pi_i)$ for any specified F. As $d \to \infty$, correct inclusion probabilities are obtained but Rosén showed that the Pareto choice gives smallest possible asymptotic bias for them. Ohlsson (1998) uses $Q_i = U_i/\pi_i$, i.e., the distribution function $F(x)$ of a uniform distribution over (0,1).

Probabilistically the Pareto design is very close to Sampford's (1967) design for which the factual inclusion probabilities agree with the desired ones. Let $\mathbf{x}$ be a binary N-vector such that $x_i = 1$ if unit i is sampled and otherwise 0. The probability function $p(\mathbf{x})$ of the Sampford design is given by

$$p(\mathbf{x}) = C\prod_{i=1}^{N} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \times \sum_{k=1}^{N}(1-\pi_k)x_k, \quad \sum_{i=1}^{N} x_i = n,$$

where C is a constant. For the Pareto design the probability function has the same form but the factor $1-\pi_k$ is replaced by c_k, where c_k is given by an integral that is closely proportional to $1-\pi_k$ if d is large (Bondesson et al. 2006).

For the Pareto design the factual inclusion probabilities π_i^* can be calculated in different ways (Aires 1999; Bondesson et al. 2006). By an iterative procedure based on recalculated factual inclusion probabilities, it is possible to adjust the parameters $p_i = \pi_i$ for the Pareto procedure, so that the desired inclusion probabilities π_i are exactly obtained (Aires 2000). The iterative procedure is time consuming. It is also possible to get good improvement by a simple adjustment. The ranking variables Q_i are replaced by the adjusted ranking variables

$$Q_i^{Adj} = Q_i \exp\left(\pi_i(1-\pi_i)\left(\pi_i - \frac{1}{2}\right)/d^2\right).$$

For $N = 6$ and $n = 3$ the following table illustrates the improvement:

	π_i	0.1	0.3	0.4	0.5	0.75	0.95
Pareto	π_i^*	0.0963	0.2916	0.3952	0.5040	0.7610	0.9519
AdjPar	π_i^*	0.0987	0.2987	0.3993	0.5018	0.7510	0.9505

Restricted Pareto Sampling

Pareto sampling can be extended to cases where there are further restrictions on the sample than just fixed sample size (Bondesson 2010). Such restrictions appear if the population is stratified in different ways. The restrictions are usually of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $M \times N$ matrix. Then

$$\sum_{i=1}^{N} x_i \log Q_i = \sum_{i=1}^{N} x_i\left(\log\frac{U_i}{1-U_i} - \log\frac{\pi_i}{1-\pi_i}\right)$$

is minimized with respect to $\mathbf{x}$ given the linear restrictions. This minimization can be done by using a program for combinatorial optimization but it usually also suffices to use linear programming and the simplex algorithm with the additional restrictions $0 \le x_i \le 1$ for all i. Under some conditions asymptotically correct inclusion probabilities are obtained. However, in practice some adjustment is often needed. A simple adjustment is to replace Q_i by

$$Q_i^{Adj} = Q_i \exp\left(\pi_i(1-\pi_i)\left(\pi_i - \frac{1}{2}\right)\left(\mathbf{a}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{a}_i\right)^2\right),$$

where $\boldsymbol{\Sigma} = \mathbf{ADA}^T$ with $\mathbf{D} = \mathrm{diag}(\pi_1(1-\pi_1),\ldots,\pi_N(1-\pi_N))$, and $\mathbf{a}_i$ is the ith column vector in $\mathbf{A}$. This method

is suggested by Bondesson (2010), who also presents another method for improvement, conditional Pareto sampling. For the latter method, the random numbers are conditioned to satisfy $\mathbf{AU} = \frac{1}{2}\mathbf{A1}$.

About the Author

Lennart Bondesson (academically a grand-son of Harald Cramér) received his Ph.D. in 1974. In 1983 he became professor of mathematical statistics in forestry at the Swedish University of Agricultural Sciences in Umeå, and in 1999 professor of mathematical statistics at Department of Mathematics and Mathematical Statistics, Umeå University. He has published more than 80 papers and one research book *Generalized Gamma Convolutions and Related Classes of Distributions and Densities* (Lecture Notes in Statistics 76, Springer, 1992). Professor Bondesson was Editor of *Scandinavian Journal of Statistics* (2001–2003).

Cross References

►Horvitz–Thompson Estimator
►Sampling Algorithms
►Uniform Random Number Generators

References and Further Reading

Aires N (1999) Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. Meth Comput Appl Probab 4:457–469
Aires N (2000) Comparisons between conditional Poisson sampling and Pareto πps sampling designs. J Stat Plann Infer 88:133–147
Bondesson L (2010) Conditional and restricted Pareto sampling; two new methods for unequal probability sampling. Scand J Stat 37, doi: 10.1111/j.1467-9469.2010.000700.x
Bondesson L, Traat I, Lundqvist A (2006) Pareto sampling versus Sampford and conditional Poisson sampling. Scand J Statist 33: 699–720
Brewer KRW, Hanif M (1983) Sampling with unequal probabilities. Lecture Notes in Statistics, No. 15. Springer, New York
Ohlsson E (1998) Sequential Poisson sampling. J Official Stat 14: 149–162
Rosén's B (1997a) Asymptotic theory for order sampling. J Stat Plann Infer 62:135–158
Rosén's B (1997b) On sampling with probability proportional to size. J Stat Plann Infer 62:159–191
Saavedra P (1995) Fixed sample size PPS approximations with a permanent random number. Joint Statistical Meetings American Statistical Association, Orlando, Florida
Sampford MR (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54:499–513
Tillé Y (2006) Sampling algorithms. Springer series in statistics. Springer science + Business Media, New York

Partial Least Squares Regression Versus Other Methods

Smail Mahdi
Professor of Mathematical Statistics
Mathematics and Physics, University of The West Indies, Cave Hill Campus, Barbados

Introduction

The concept of regression originated from genetics and the word *regression* was introduced into statistical literature in the published paper by Sir Francis Galton (1886) on the relationship between the stature of children and their parents. The relationship was found to be approximately linear with an approximate gradient of 2/3, and this suggested that very tall parents tend to have children shorter than themselves and vice versa. This phenomena was referred to as regression or return to mediocrity or to an average value. The general framework of the linear regression model (see ►Linear Regression Models)

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is an $n \times q$ matrix of observations on q dependent variables, $\mathbf{X}$ is an $n \times p$ explicative matrix on p variables, $\boldsymbol{\beta}$ is a $p \times q$ matrix of unknown parameters, and $\boldsymbol{\epsilon}$ is $n \times q$ matrix of errors. The rows of $\boldsymbol{\epsilon}$ are independent and identically distributed, often assumed to be Gaussian. The justification for the use of a linear relationship comes from the fact that the conditional mean of a Gaussian random vector given the value of another Gaussian random vector is linear when the joint distribution is Gaussian. Without loss of generality, we will assume throughout that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mean centered and scaled. The aim is to estimate the unknown matrix $\boldsymbol{\beta}$. The standard approach is to solve in L2 the optimization problem:

$$L(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{\beta}\boldsymbol{X}\|_2.$$

If $\boldsymbol{X}$ has a full rank, then a unique solution exists and is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

When the errors are assumed Gaussian, $\hat{\boldsymbol{\beta}}_{OLS}$ is also the maximum likelihood estimator and therefore, the best unbiased linear estimator (BLUE) of $\boldsymbol{\beta}$. Unfortunately, when some of the dependent variables are collinear, the matrix $\boldsymbol{X}$ does not have full rank and the ordinary least squares estimator may not exist or

becomes inappropriate. To overcome this multicollinearity problem (see ▶Multicollinearity), many other estimators have been proposed in literature. This includes ridge regression estimator (RR), principal component regression estimator (PCR), and more recently the partial least squares regression estimator (PLSR).

Ridge Regression

The slightest multicollinearity of the independent vectors may make the matrix $\boldsymbol{X}^T\boldsymbol{X}$ ill conditioned and increase the variance of the components of $\hat{\boldsymbol{\beta}}_{OLS}$, which can lead to an unstable estimation of $\boldsymbol{\beta}$. Hoerl (1962) proposed the ridge regression method (see ▶Ridge and Surrogate Ridge Regressions) that consists in adding a small positive constant λ to the diagonal of the standardized matrix $\boldsymbol{X}^T\boldsymbol{X}$ to obtain the RR estimator as

$$\hat{\boldsymbol{\beta}}_{RR} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The matrix $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ is always invertible since it is positive definite. Several techniques are available in literature for finding the optimum value of λ. Another solution for the collinearity problem is given by the principal component regression (PCR) technique that is outlined below.

Principal Component Regression

Principal component regression is a ▶principal component analysis (PCA) followed by a linear regression. In this case, the response variables are regressed on the leading principal components of the matrix $\boldsymbol{X}$, which can be obtained from its singular value decomposition (SVD). PCA is a compressing data technique that consists of finding a k-rank, $k = 1, \ldots, p$, projection matrix $\boldsymbol{P}_k$ such that the variance of the projected data $\boldsymbol{X} \times \boldsymbol{P}_k$ is maximized. The columns of the matrix $\boldsymbol{P}_k$ consist of the k unit-norm leading eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. Using the matrix $\boldsymbol{P}_k$, we regress $\boldsymbol{Y}$ on the orthogonal score matrix $\boldsymbol{T}_k$ satisfying $\boldsymbol{X} = \boldsymbol{T}_k\boldsymbol{P}_k$; this yields the k latent-component regression coefficients matrix

$$\hat{\boldsymbol{\beta}}_{PCR} = \boldsymbol{P}_k \left(\boldsymbol{T}_k^T\boldsymbol{T}_k\right)^{-1} \boldsymbol{T}_k^T\boldsymbol{Y}.$$

However, two major problems may occur: firstly the choice of k and secondly, how, if the ignored principal components that correspond to the small eigenvalues are in fact relevant for explaining the covariance structure of the $\boldsymbol{Y}$ variables. For this reason, PLS comes into play to better deal with the multicollinearity problem by creating $\boldsymbol{X}$ latent components for explaining $\boldsymbol{Y}$, through the maximization of the covariance structure between the $\boldsymbol{X}$ and $\boldsymbol{Y}$ spaces.

Partial Least Squares Regression

Partial least squares (PLS), also known as projection method to latent structures, is applied to a broad area of data analysis including ▶generalized linear models, regression, classification, and discrimination. It is a multivariate technique that generalizes and combines ideas from principal component analysis (PCA) and ordinary least squares (OLS) regression methods. It is designed to not only confront situations of correlated predictors but also relatively small samples and even the situation where the number of dependent variables exceeds the number of cases. The original idea came in the work of the Swedish statistician Herman Wold (1966) in the 1960s and became popular in computational chemistry and sensory evaluation by the work of his son Svante who developed the popular software soft independent modeling of class analogies (SIMCA) (Wold 1976). PLS find first two sets of weights $\boldsymbol{\omega} = (w_1, \ldots, w_m)$ and $\boldsymbol{U} = (u_1, \ldots, u_m)$ for $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that $Cov\left(\boldsymbol{t}_l^X, \boldsymbol{t}_l^Y\right)$, where $\boldsymbol{t}_l^X = \boldsymbol{X} \times \boldsymbol{\omega}_l$, $\boldsymbol{t}_l^Y = \boldsymbol{Y} \times \boldsymbol{U}_l$ and $l = 1, \ldots, m$, is maximal. Then, the $\boldsymbol{Y}$ variables are regressed on the latent matrix $\boldsymbol{T}$ whose columns are the latent vectors $\boldsymbol{t}_i$, see, e.g., (Vinzi et al. 2005) for more details. Classically, the ordinary least squares regression (OLS) is used but other methods have been considered along with the corresponding algorithms. We describe the PLS technique through the algorithm, outlined, for instance in Mevik and Wehrens (2007), which is based on the singular value decomposition (SVD) of the cross product $\boldsymbol{X}^T\boldsymbol{Y}$ types. First, we set $\boldsymbol{E} = \boldsymbol{X}$ and $\boldsymbol{F} = \boldsymbol{Y}$. Then, we perform the singular decomposition of $\boldsymbol{E}^T\boldsymbol{F}$ and take the first left-singular vector $\boldsymbol{\omega}$ and the right-singular vector $\boldsymbol{q}$ of $\boldsymbol{E}^T\boldsymbol{F}$ to obtain the scores $\boldsymbol{t}$ and $\boldsymbol{u}$ as follows:

$$\boldsymbol{t} = \boldsymbol{E}\boldsymbol{\omega}$$

and

$$\boldsymbol{u} = \boldsymbol{F}\boldsymbol{q}.$$

The first score then is obtained as $\boldsymbol{t} = \boldsymbol{t}_1 = \dfrac{\boldsymbol{t}}{(\boldsymbol{t}^T\boldsymbol{t})^{1/2}}$. The effect of this score $\boldsymbol{t}_1$ on $\boldsymbol{E}$ and $\boldsymbol{F}$ is obtained by regressing $\boldsymbol{E}$ and $\boldsymbol{F}$ on $\boldsymbol{t}_1$. This gives $\boldsymbol{p} = \boldsymbol{E}^T\boldsymbol{t}$ and $\boldsymbol{q} = \boldsymbol{F}^T\boldsymbol{t}$. This effect is then subtracted from $\boldsymbol{E}$ and $\boldsymbol{F}$ to obtain the deflated matrices, with one rank less, $\boldsymbol{E}'$ and $\boldsymbol{F}'$, see, for example, Wedderburn (1934). The matrices $\boldsymbol{E}'$ and $\boldsymbol{F}'$ are then substituted for $\boldsymbol{E}$ and $\boldsymbol{F}$ and the process is reiterated again from the singular value decomposition of $\boldsymbol{E}^T\boldsymbol{F}$ to obtain the second normalized latent component $\boldsymbol{t}_2$. The process is continued until the required number m of latent components is obtained. The optimal value of m is often obtained by cross-validation or bootstrap, see, for example, Tenenhaus (1998). The obtained latent components $\boldsymbol{t}_l$, $l = 1, \ldots, m$ are

saved after each iteration as column of the latent matrix $\boldsymbol{T}$. Finally, the original Y variables are regressed on the latent components $\boldsymbol{T}$ to obtain the regression estimator

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y}$$

and the estimator

$$\hat{\boldsymbol{Y}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y}$$

of $\boldsymbol{Y}$. Note that several of the earlier PLS algorithms, available in literature, lack robustness. Recently, Simonetti et al. (2006) substituted systematically the least median of squares regression, Rousseeuw (1984), for the least squares regression in Garthwaite (1994) PLS setup to obtain a robust estimation for the considered data set.

Cross References

- ►Chemometrics
- ►Least Squares
- ►Linear Regression Models
- ►Multicollinearity
- ►Multivariate Statistical Analysis
- ►Principal Component Analysis
- ►Ridge and Surrogate Ridge Regressions

References and Further Reading

Garthwaite PH (1994) An interpretation of partial least squares. J Am Stat Assoc 89(425):122–127

Galton F (1886) Regression toward mediocrity in hereditary stature. Nature 15:246–263

Hoerl AE (1962) Application of ridge analysis to regression problems. Chem Eng Prog 58:54–59

Mevik B, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. J Stat Softw 18(2):1–24

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–888

Simonetti B, Mahdi S, Camminatiello I (2006) Robust *PLS* regression based on simple least median squares regression. MTISD'06 Conference, Procida, Italy

Tenenhaus M (1998) La Régréssion *PLS*: Théorie et Pratique. Technip, Paris

Vinzi VE, Lauro CN, Amato S (2005) *PLS* typological regression: algorithms, classification and validation issues. New developments in classification and data analysis: Vichi M, Monari P, Mignani S, Montanari A (eds) Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, Springer, Berlin, Heidelberg

Wedderburn JHM (1934) Lectures on matrices. AMS, vol 17. Colloquium, New York

Wold H (1966) Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR (ed) Multivariate analysis. Academic, New York, pp 391–420

Wold S (1976) Pattern recognition by means of disjoint principal components models. Pattern Recogn 8:127–139

Pattern Recognition, Aspects of

Dražen Domijan[1], Bojana Dalbelo Bašić[2]
[1]Associate Professor
University of Rijeka, Rijeka, Croatia
[2]Professor
University of Zagreb, Zagreb, Croatia

Introduction

Recognition is regarded as a basic attribute of human beings and other living organisms. A pattern is a description of an object (Tou and Gonzalez 1974). Pattern recognition is a process of assigning category labels to a set of patterns. For instance, visual patterns "A," "a," and "***A***" are members of the same category, which is labeled as "letter A" and can easily be distinguished from the patterns, "B," "b," and "***B***," which belong to another category labeled as "letter B." Humans perform pattern recognition very well and the central problem is how to design a system to match human performance. Such systems find practical applications in many domains such as medical diagnosis, image analysis, face recognition, speech recognition, handwritten character recognition, and more (Duda et al. 2001; Fukunaga 1990).

The problem of pattern recognition can be tackled using handcrafted rules or heuristics for distinguishing the category of objects, though in practice such an approach leads to proliferation of the rules and exceptions to the rules, and invariably gives poor results (Bishop 2006). Some classification problems can be tackled by syntactic (linguistic) pattern recognition methods, but most real-word problems are tackled using the machine learning approach. Pattern recognition is an interdisciplinary field involving statistics, probability theory, computer science, machine learning, linguistics, cognitive science, psychology, etc. Pattern recognition systems involve the following phases: sensing, feature generation, feature selection, classifier design, and system performance evaluation (Tou and Gonzalez 1974).

In the previous example, the pattern was a set of black and white pixels. However, patterns might be a set of continuous variables. In statistical pattern recognition, features are treated as random variables. Therefore, patterns are random vectors that are assigned to a class or category with certain probability. In this case, patterns could be conceived as points in a high-dimensional feature space. Pattern recognition is the task of estimating a function that divides the feature space into regions, where each region corresponds to one of the categories (classes).

Such a function is called the decision or discriminant function and the surface that is realized by the function is known as the decision surface.

Feature Generation and Feature Selection

Before the measurement data obtained from the sensor could be utilized for the design of the pattern classifier, sometimes it is necessary to perform several preprocessing steps such as outlier removal, data normalization, and treatment of the missing data. After preprocessing, features are generated from measurements using data reduction techniques, which exploits and removes redundancies in the original data set. A popular way of generating features is to use linear transformations such as the Karhunen–Loeve transformation (▶principal component analysis), independent component analysis, discrete Fourier transform, discrete cosine and sine transforms, and Hadamard and Haar transforms. Important consideration in using these transformations is that they should preserve as much of the information that is crucial for classification task as possible, while removing as much redundant information as possible (Theodoridis and Koutroumbas 2009).

After the features are generated, they could be independently tested for their discriminatory capability. We might select features based on the statistical hypothesis testing. For instance, we may employ t-test or Kruskall–Wallis test to investigate the difference in mean feature values for two classes. Another approach is to construct the characteristic receiver operating curve and to explore how much overlap exists between distributions of feature values for two classes. Furthermore, we might compute class separability measures that take into account correlations between features such as Fisher's discriminant ratio and divergence. Besides testing individual features, we might ask what is the best feature vector or combination of features that gives the best classification performance. There are several searching techniques such as sequential backward or forward selection that can be employed in order to find an answer (Theodoridis and Koutroumbas 2009).

Design of a Pattern Classifier

The principled way to design a pattern classifier would involve characterization of the class probability density functions in the feature space and finding an appropriate discriminant function to separate the classes in this space. Every classifier can make an error by assigning the wrong class label to the pattern. The goal is to find the classifier with the minimal probability of classification error. The best classifier is based on the Bayes decision rule (Fukunaga 1990). The basic idea is to assign a pattern to the class having the highest a posteriori probability for a given pattern. A posteriori probabilities for a given pattern are computed from a priori class probabilities and conditional density functions. However, in practice, it is often difficult to compute a posteriori probabilities as a priori class probabilities are not known in advance.

Therefore, although the Bayesian classifiers are optimal they are rarely used in practice. Instead, classifiers are designed directly from the data. A simple and computationally efficient approach to the design of the classifier is to assume that the discriminant function is linear. In that case, we can construct a decision hyperplane through the feature space defined by

$$g(x) = w^T x + w_0 = 0$$

where $w = [w_1, w_2, \ldots]^T$ are unknown coefficients or weights, $x = [x_1, x_2, \ldots]$ is a feature vector, and w_0 is a threshold or bias. Finding unknown weights is called learning or training. In order to find an appropriate value for the weights, we can use iterative procedures such as the perceptron learning algorithm. The basic idea is to compute error or cost function, which measures the difference between actual classifier output and desired output. Error function is used to adjust current values of the weights. This process is repeated until perceptron converges, that is, until all patterns are correctly classified. This is possible if patterns are linearly separable. After the perceptron converges, new patterns could be classified according to the simple rule:

$$\text{If } w^T x + w_0 > 0 \text{ assign } x \text{ to class } \omega_1$$
$$\text{If } w^T x + w_0 < 0 \text{ assign } x \text{ to class } \omega_2$$

where $\omega 1$ and $\omega 2$ are class labels.

The problem with linear classifiers is that they lead to suboptimal performance when classes are not linearly separable. An example of pattern recognition problem that is not linearly separable is the logical predicate XOR where the patterns (0,1) and (1,0) belong to class $\omega 1$, and the patterns (0,0) and (1,1) belong to $\omega 2$. It is not possible to draw a straight line (linear decision boundary) in two-dimensional feature space that will discriminate between these two classes. One approach to deal with such problems is to build a linear classifier that will minimize the mean square error between the desired and the actual output of the classifier. This can be achieved using least mean square (LMS) or Widrow–Hoff algorithm for weight

adjustment. Another approach is to design a nonlinear classifier (Bishop 1995). Examples of nonlinear classifiers are multilayer perceptron trained with error backpropagation, radial basis functions network, k-nearest neighbor classifier, and decision trees. In practice, it is possible to combine outputs from several different classifiers in order to achieve better performance. Classification is an example of so-called supervised learning in which each feature vector has a preassigned target class.

Performance Evaluation of the Pattern Classifier

An important task in the design of a pattern classifier is how well it will perform when faced with new patterns. This is an issue of generalization. During learning, the classifier builds a model of the environment to which it is exposed. The model might vary in complexity. A complex model might offer a better fit to the data, but might also capture more noise or irrelevant characteristics in the data, and thus be poor in the classification of new patterns. Such a situation is called over-fitting. On the other hand, if the model is of low complexity, it might not fit the data well. The problem is how to select an appropriate level of complexity that will enable the classifier to fit the observed data well, while preserving enough flexibility to classify unobserved patterns. This is known as a bias-variance dilemma (Bishop 1995).

The performance of the designed classifier is evaluated by counting the number of errors committed during a testing with a set of feature vectors. Error counting provides an estimation of classification error probability. The important question is how to choose a set of feature vectors that will be used for building the classifier and a set of feature vectors that will be used for testing. One approach is to exclude one feature vector from the sample, train the classifier on all other vectors, and then test the classifier with the excluded vector. If misclassification occurs, error is counted. This procedure is repeated N times with different excluded feature vectors every time. The problem with this procedure is that it is computationally demanding. Another approach is to split the data set into two subsets: (1) a training sample used to adjust (estimate) classifier parameters and (2) a testing sample that is not used during training but is applied to the classifier following completion of the training. The problem with this approach is that it reduces the size of the training and testing samples, which reduces the reliability of the estimation of classification error probability (Theodoridis and Koutroumbas 2009).

Acknowledgment

We would like to thank Professor Sergios Theodoridis for helpful comments and suggestions that significantly improved our presentation.

Cross References

►Data Analysis
►Fuzzy Sets: An Introduction
►ROC Curves
►Significance Testing: An Overview
►Statistical Pattern Recognition Principles

References and Further Reading

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford, UK

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic, San Diego, CA

Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Elsevier

Tou JT, Gonzalez RC (1974) Pattern recognition Principles. Addison-Wesley, Reading, MA

Permanents in Probability Theory

Ravindra B. Bapat
Professor, Head New Delhi Centre
Indian Statistical Institute, New Delhi, India

Preliminaries

If A is an $n \times n$ matrix, then the permanent of A, denoted by per A, is defined as

$$\text{per } A = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)},$$

where S_n is the set of permutations of $1, 2, \ldots, n$. Thus the definition of the permanent is similar to that of the determinant except that all the terms in the expansion have a positive sign.

Example: Consider the matrix

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 1 & 2 & 3 \\ -2 & 4 & 1 \end{bmatrix}.$$

Then

$$\text{per } A = 4 + 6 + 12 - 12 + 24 - 1 = 33.$$

Permanents find numerous applications in probability theory, notably in the theory of discrete distributions and ▶order statistics. There are two main advantages of employing permanents in these areas. Firstly, permanents serve as a convenient notational device, which makes it feasible to write complex expressions in a compact form. The second advantage, which is more important, is that some theoretical results for the permanent lead to statements of interest in probability theory. This is true mainly of the properties of permanents of entrywise nonnegative matrices.

Although the definition of the permanent is similar to that of the determinant, many of the nice properties of the determinant do not hold for the permanent. For example, the permanent is not well behaved under elementary transformations, except under the transformation of multiplying a row or column by a constant. Similarly, the permanent of the product of two matrices does not equal the product of the permanents in general. The Laplace expansion for the determinant holds for the permanent as well and is a convenient tool for dealing with the permanent. Thus, if $A(i,j)$ denotes the submatrix obtained by deleting row i and column j of the $n \times n$ matrix A, then

$$\text{per } A = \sum_{k=1}^{n} a_{ik} \text{per } A(i,k) = \sum_{k=1}^{n} a_{ki} \text{per } A(k,i), \quad i = 1, 2, \ldots, n.$$

We refer to van Lint and Wilson (2001) for an introduction to permanents.

Combinatorial Probability

Matrices all of whose entries are either 0 or 1, the so called $(0,1)$-matrices, play an important role in combinatorics. Several combinatorial problems can be posed as problems involving counting certain permutations of a finite set of elements and hence can be represented in terms of $(0,1)$-matrices. We give two well-know examples in combinatorial probability.

Consider n letters and n envelopes carrying the corresponding addresses. If the letters are put in the envelopes at random, what is the probability that none of the letters goes into the right envelope? The probability is easily seen to be

$$\frac{1}{n!}\text{per}\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix},$$

which equals $1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}$. The permanent in the above expression counts the number of *derangements* of n symbols.

Another problem, which may also be posed as a probability question, is the *problème des ménages* (Kaplansky and Riordan 1946). In how many ways can n couples be placed at a round table so that men and women sit in alternate places and no one is sitting next to his or her spouse? This number equals $2n!$ times the permanent of the matrix $J_n - I_n - P_n$, where J_n is the $n \times n$ matrix of all ones, I_n is the $n \times n$ identity matrix, and P_n is the full cycle permutation matrix, having ones at positions $(1,2), (2,3), \ldots, (n-1,n), (n,1)$ and zeroes elsewhere. The permanent can be expressed as

$$\text{per } (J_n - I_n - P_n) = \sum_{i=0}^{n} (-1)^i \frac{2n}{2n-i} \binom{2n-i}{i} (n-i)!.$$

Discrete Distributions

The densities of some discrete distributions can be conveniently expressed in terms of permanents. We illustrate by an example of multiparameter ▶multinomial distribution.

We first consider the multiparameter binomial. Suppose n coins, not necessarily identical, are tossed once, and let X be the number of heads obtained. Let p_i be the probability of heads on a single toss of the i-th coin, $i = 1, 2, \ldots, n$. Let $\mathbf{p}$ be the column vector with components $p_1, \ldots, p_n$, and let $\mathbf{1}$ be the column vector of all ones. Then it can be verified that

$$Prob(X = x) = \frac{1}{x!(n-x)!} \text{per } [\underbrace{\mathbf{p}, \ldots, \mathbf{p}}_{r} \underbrace{(\mathbf{1} - \mathbf{p}), \ldots, (\mathbf{1} - \mathbf{p})}_{n-r}]. \tag{1}$$

Now consider an experiment which can result in any of r possible outcomes, and suppose n trials of the experiment are performed. Let p_{ij} be the probability that the experiment results in the j-th outcome at the i-th trial, $i = 1, 2, \ldots, n; j = 1, 2, \ldots, r$. Let P denote the $n \times r$ matrix (p_{ij}) which is row stochastic. Let $P_1, \ldots, P_r$ be the columns of P. Let X_j denote the number of times the j-th outcome is obtained in the n trials, $j = 1, 2, \ldots, r$. If $k_1, \ldots, k_r$ are nonnegative integers summing to n, then as a generalization of (1) we have (Bapat 1990)

$$Prob(X_1 = k_1, \ldots, X_r = k_r) = \frac{1}{k_1! \cdots k_r!} \text{per } [\underbrace{P_1, \ldots, P_1}_{k_1}, \ldots, \underbrace{P_r, \ldots, P_r}_{k_r}].$$

Similar representations exist for the multiparameter negative binomial.

Order Statistics

Permanents provide an effective tool in dealing with order statistics corresponding to independent random variables, which are not necessarily identically distributed. Let $X_1, \ldots, X_n$ be independent random variables with distribution functions $F_1, \ldots, F_n$ and densities $f_1, \ldots, f_n$ respectively. Let $Y_1 \le \cdots \le Y_n$ denote the corresponding order statistics. We introduce the following notation. For a fixed y, let

$$\hat{f}(y) = \begin{bmatrix} f_1(y) \\ \vdots \\ f_n(y) \end{bmatrix} \text{ and } \hat{F}(y) = \begin{bmatrix} F_1(y) \\ \vdots \\ F_n(y) \end{bmatrix}.$$

For $1 \le r \le n$, the density function of Y_r is given by Vaughan and Venables (1972)

$$g_r(y) = \frac{1}{(r-1)!(n-r)!}\text{per}\,[\underbrace{\hat{f}(y)}_{1}\ \underbrace{\hat{F}(y)}_{r-1}\ \underbrace{\mathbf{1}-\hat{F}(y)}_{n-r}], \infty < y < \infty$$

For $1 \le r \le n$, the distribution function of Y_r is given by Bapat and Beg (1989)

$$Prob(Y_r \le y) = \sum_{i=r}^{n} \frac{1}{i!(n-i)!}\text{per}\,[\underbrace{\hat{F}(y)}_{i}\ \underbrace{\mathbf{1}-\hat{F}(y)}_{n-i}], \quad \infty < y < \infty$$

The permanental representation can be used to extend several recurrence relations for order statistics from the i.i.d. case to the case of nonidentical, independent random variables. Using the Alexandroff inequality for the permanent of a nonnegative matrix, it can be shown (Bapat 1990) that for any y, the sequence $Prob(Y_i \le y | Y_{i-1} \le y)$, $i = 2, \ldots, n$ is nonincreasing.

For additional material on applications of permanents in order statistics we refer to Balakrishnan (2007) and the references contained therein.

Acknowledgments

The support of the JC Bose Fellowship, Department of Science and Technology, Government of India, is gratefully acknowledged.

About the Author

Past President of the Indian Mathematical Society (2007–2008), Professor Bapat joined the Indian Statistical Institute, New Delhi, in 1983, where he holds the position of Head, Delhi Centre at the moment. He held visiting positions at various Universities in the U.S., including University of Connecticut and Oakland University, and visited several Institutes abroad in countries including France, Holland, Canada, China and Taiwan for collaborative research and seminars. The main areas of research interest of Professor Bapat are nonnegative matrices, matrix inequalities, matrices in graph theory and generalized inverses. He has published more than 100 research papers in these areas in reputed national and international journals. He has written books on Linear Algebra, published by Hindustan Book Agency, Springer and Cambridge University Press. He has recently written a book on Mathematics for the general readers, in Marathi, which won the state government award for best literature in Science for 2004. He is Elected Fellow of the Indian Academy of Sciences, Bangalore and Indian National Science Academy, Delhi.

Cross References

►Multinomial Distribution
►Order Statistics
►Probability Theory: An Outline
►Random Permutations and Partition Models
►Univariate Discrete Distributions: An Overview

References and Further Reading

Balakrishnan N (2007) Permanents, order statistics, outliers, and robustness. Rev Mat Complut 20(1):7–107
Bapat RB (1990) Permanents in probability and statistics. Linear Algebra Appl 127:3–25
Bapat RB, Beg MI (1989) Order statistics for nonidentically distributed variables and permanents. Sankhya A 51(1):79–93
Kaplansky I, Riordan J (1946) The probléme des ménages. Scripta Math 12:113–124
van Lint JH, Wilson RM (2001) A course in combinatorics, 2nd edn. Cambridge University Press, Cambridge
Vaughan RJ, Venables WN (1972) Permanent expressions for order statistic densities J R Stat Soc B 34:308–310

Permutation Tests

Markus Neuhäuser
Professor
Koblenz University of Applied Sciences, Remagen, Germany

A permutation test is illustrated here for a two-sample comparison. The notation is as follows: Two independent groups with sample sizes n and m have independently and

identically distributed values $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively, $n + m = N$. The means are denoted by $\overline{X}$ and $\overline{Y}$, and the distribution functions by F and G. These distribution functions of the two groups are identical with the exception of a possible location shift: $F(t) = G(t - \theta)$ for all $t, -\infty < \theta < \infty$. The null hypothesis states $H_0 : \theta = 0$, whereas $\theta \neq 0$ under the alternative H_1.

In this case Student's t test (see ▶Student's t-Tests) can be applied. However, if F and G were not normal distributions, it may be better to avoid using the t distribution. An alternative method is to use the permutation null distribution of the t statistic.

In order to generate the permutation distribution all possible permutations under the null hypothesis have to be generated. In the two-sample case, each permutation is a possible (re-)allocation of the N observed values to two groups of sizes n and m. Hence, there are $\binom{N}{n}$ possible permutations. The test statistic is calculated for each permutation. The null hypothesis can then be accepted or rejected using the permutation distribution of the test statistic, the p-value being the probability of the permutations giving a value of the test statistic as or more supportive of the alternative than the observed value. Thus, inference is based upon how extreme the observed test statistic is relative to other values that could have been obtained under the null hypothesis.

Under H_0 all permutations have the same probability. Hence, the p-value can simply be computed as the proportion of the permutations with a test statistic's value as or more supportive of the alternative than the observed value.

The order of the permutations is important, rather than the exact values of the test statistic. Therefore, modified test statistics can be used (Manly 2007, pp. 16–17). For example, the difference $\overline{X} - \overline{Y}$ can be used instead of the t statistic

$$t = \frac{\overline{X} - \overline{Y}}{S \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where S is the estimated standard deviation.

This permutation test is called Fisher–Pitman permutation test or randomization test (see ▶Randomization Tests), it is a nonparametric test (Siegel 1956; Manly 2007). However, at least an interval measurement is required for the Fisher-Pitman test because the test uses the numerical values $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ (Siegel 1956).

The permutation distribution depends on the observed values, therefore a permutation test is a conditional test, and a huge amount of computing is required. As a result, the Fisher–Pitman permutation test was hardly ever applied before the advent of fast PCs, although it was proposed in the 1930s and its high efficiency was known since decades. Nowadays, the test is often recommended and implemented in standard software such as SAS (for references see Lehmann 1975, or Neuhäuser and Manly 2004).

Please note that the randomization model of inference does not require randomly sampled populations. For a permutation test it is only required that the groups or treatments have been assigned to the experimental units at random (Lehmann 1975).

A permutation test can be applied with other test statistics, too. Rank-based statistics such as Wilcoxon's rank sum can also be used as test statistic (see the entry about the Wilcoxon–Mann–Whitney test). Rank tests can also be applied for ordinal data. Moreover, rank tests had advantages in the past because the permutation null distribution can be tabulated. Nowadays, with modern PCs and fast algorithms, permutation tests can be carried out with any suitable test statistic. However, rank tests are relatively powerful in the commonly-occurring situation where the underlying distributions are non-normal (Higgins 2000). Hence, permutation tests on ranks are still useful despite the fact that more complicated permutation tests can be carried out (Neuhäuser 2005).

When the sample sizes are very large, the number of possible permutations can be extremely large. Then, a ▶simple random sample of M permutations can be drawn in order to estimate the permutation distribution. A commonly used value is $M = 10,000$. Please note that the original observed values must be one of the M selected permutations (Edgington and Onghena 2007, p. 41).

It should be noted that permutation procedures do have some disadvantages. First, they are computer-intensive, although the computational effort seems to be no longer a severe objection against permutation tests. Second, conservatism is often the price for exactness. For this reason, the virtues of permutation tests continue to be debated in the literature (Berger 2000).

When a permutation test is performed for other situations than the comparison of two samples the principle is analogue. The Fisher-Pitman test can also be carried out with the ANOVA F statistic. Permutation tests can also be applied in case of more complex designs. For example, the residuals can be permuted (ter Braak 1992; Anderson 2001).

Permutation tests are also possible for other situations than the location-shift model. For example, Janssen (1997) presents a permutation test for the ▶Behrens-Fisher problem where the population variances may differ.

About the Author

For biography *see* the entry ►Wilcoxon-Mann-Whitney Test.

Cross References

►Behrens–Fisher Problem
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Randomization
►Randomization Tests
►Statistical Fallacies: Misconceptions, and Myths
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Anderson MJ (2001) Permutation tests for univariate or multivariate analysis of variance and regression. Can J Fish Aquat Sci 58: 626–639

Berger VW (2000) Pros and cons of permutation tests in clinical trials. Stat Med 19:1319–1328

Edgington ES, Onghena P (2007) Randomization tests, 4th edn. Chapman and Hall/CRC, Boca Raton

Higgins JJ (2000) Letter to the editor. Am Stat 54, 86

Janssen A (1997) Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. Stat Probab Lett 36:9–21

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

Manly BFJ (2007) Randomization, bootstrap and Monte Carlo methods in biology, 3rd edn. Chapman and Hall/CRC, London

Neuhäuser M (2005) Efficiency comparisons of rank and permutation tests. Stat Med 24:1777–1778

Neuhäuser M, Manly BFJ (2004) The Fisher-Pitman permutation test when testing for differences in mean and variance. Psychol Rep 94:189–194

Siegel S (1956) Nonparametric statistics for the behavioural sciences. McGraw-Hill, New York

ter Braak CJF (1992) Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel KH, Rothe G, Sendler W (eds) Bootstraping and related techniques. Springer, Heidelberg, pp 79–85

Pharmaceutical Statistics: Bioequivalence

Francis Hsuan
Professor Emeritus
Temple University, Philadelphia, PA, USA

In the terminology of pharmacokinetics, the *bioavailability* (BA) of a drug product refers to the rate and extent of its absorbed active ingredient or active moiety that becomes available at the site of action. A new/test drug product (T) is considered *bioequivalent* to an existing/reference product (R) if there is no significant difference in the bioavailabilities between the two products when they are administered at the same dose under similar conditions in an appropriately designed study. Studies to demonstrate either *in-vivo* or *in-vitro* bioequivalence are required for government regulatory approvals of generic drug products, or new formulations of existing products with known chemical entity.

Statistically an *in-vivo* bioequivalence (BE) study employs a crossover design with T and R drug products administered to healthy subjects on separate occasions according to a pre-specified randomization schedule, with ample washout between the occasions. The concentration of the active ingredient of interest in blood is measured over time per subject in each occasion, resulting in multiple concentration-time profiles curves for each subject. From which a number of bioavailability measures, such as area under the curve (AUC) and peak concentration (C_{max}) in either raw or logarithmic scale, are then computed, statistically modeled, and analyzed for bioequivalence. Let Y_{ijkt} be a log-transformed bioavailability measure of subject j in treatment sequence i at period t, having the treatment k. In a simplest 2×2 crossover design with two treatment sequences T/R ($i=1$) and R/T ($i=2$), is assumed to follow a mixed-effects ANOVA model

$$Y_{ijkt} = \mu_0 + \omega_i + \phi_k + \pi_t + S_{ij} + \varepsilon_{ijkt}$$

where μ_0 is an overall constant, ω_i, ϕ_k and π_t are, respectively, (fixed) effects of sequence i, formulation k and period t, S_{ij} is the (random) effect of subject j in sequence i, and ε_{ijkt} the measurement error. In this model, the between-subject variation is captured by the random component $S_{ij} \sim N\left(0, \sigma_B^2\right)$ and the within-subject variation $\varepsilon_{ijkt} \sim N\left(0, \sigma_{Wk}^2\right)$ is allowed to depend on formulation k. Bioequivalence of T and R is declared when the null hypothesis $H_0 : \phi_T \le \phi_R - \varepsilon$ or $\phi_T \ge \phi_R + \varepsilon$ can be rejected against the alternative $H_1 : -\varepsilon \le \phi_T - \phi_R \le \varepsilon$ at some significance level α, where $\varepsilon = \log(1.25) = 0.2231$ and $\alpha = 0.05$ are set by regulatory agencies. This last sentence is referred to as the criterion for *average BE*. A common method to establish average BE is to calculate the shortest $(1-2\alpha)\times 100\%$ confidence interval of $\delta = \phi_T - \phi_R$ and show that it is contained in the equivalence interval $(-\varepsilon, \varepsilon)$.

Establishing average BE using any nonreplicated $k \times k$ crossover design can be conducted along the same line as described above. For certain types of drug products, such as those with Narrow Therapeutic Index (NTI), questions were raised regarding the adequacy of the average BE method (e.g., Blakesley et al. 2004). To address this issue,

other criteria and/or statistical methods for bioequivalence have been proposed and studied in the literature. In partic-h ular, *population BE* (PBE) assesses the difference between T and R in both means and variances of bioavailability measures, and *individual BE* (IBE) assesses, in addition, the variation in the average T and R difference among individuals. Some of these new concepts, notably individual bioequivalence, would require high-order crossover designs such as TRT/RTR. Statistical designs and analyses for population BE and individual BE are described in detail in a statistical guidance for industry (USA FDA, 2001) and several monographs (Chow and Shao 2002; Wellek 2003). Hsuan and Reeve (2003) proposed a unified procedure to establish IBE using any high-order crossover design and a multivariate ANOVA model. Recently the USA FDA (2007) proposes a process of making available the public guidance(s) on how to design bioequivalence studies for specific drug products.

About the Author

Dr. Francis Hsuan has been a faculty member in the Department of Statistics at the Fox School, Temple University, for more than 25 years. He received his Ph.D. in Statistics from Cornell University and his B.S. in Physics from the National Taiwan University. He was also a visiting scholar at the Harvard School of Public Health from 1998 to 1999, working on projects related to the analysis of longitudinal categorical data with informative missingness.

Cross References

►Biopharmaceutical Research, Statistics in
►Equivalence Testing
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences

References and Further Reading

Blakesley V, Awni W, Locke C, Ludden T, Granneman GR, Braverman LE (2004) Are bioequivalence studies of levothyroxine sodium formulations in euthyroid volunteers reliable? Thyroid 14:191–200

Chow SC, Shao J (2002) Statistics in drug research: methodologies and recent developments. Marcell Dekker, New York

Hsuan F, Reeve R (2003) Assessing individual bioequivalence with high-order crossover designs: a unified procedure. Stat Med 22:2847–2860

FDA (2001) Guidance for industry on statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, USA, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070244.pdf

FDA (2007) Guidance for industry on bioequivalence recommendations for specific products. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, USA, http://www.fda.gov/downloads/Drugs/GuidancehComplianceRegulatoryInformation/Guidances/ucm072872.pdf

Wellek S (2003) Testing statistical hypotheses of equivalence. Chapman and Hall/CRC, Florida, USA

Philosophical Foundations of Statistics

Inge S. Helland
Professor
University of Oslo, Oslo, Norway

The philosophical foundations of statistics involve issues in theoretical statistics, such as goals and methods to meet these goals, and interpretation of the meaning of inference using statistics. They are related to the philosophy of science and to the ►philosophy of probability.

As with any other science, the philosophical foundations of statistics are closely connected to its history, which again is connected to the men with whom different philosophical directions can be associated. Some of the most important names in this connection are Thomas Bayes (1702–1761), Ronald A. Fisher (1890–1962) and Jerzy Neyman (1894–1981).

The standard statistical paradigm is tied to the concept of a statistical model, an indexed family of probability measures $P^{\theta}(\cdot)$ on the observations, indexed by the parameters θ. Inference is done on the parameter space. This paradigm was challenged by Breiman (2001), who argued for an algorithmical, more intuitive model concept. Breiman's tree models are still much in use, together with other algorithmical bases for inference, for instance within chemometry. For an attempt to explain some of these within the framework of the ordinary statistical paradigm, see Helland (2010).

On the other hand, not all indexed families of distributions lead to sensible models. McCullagh (2002) showed that several absurd models can be produced.

The standard statistical model concept can be extended by implementing some kind of model reduction (Wittgenstein 1961: "The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience") or by, e.g., adjoining a symmetry group to the model (Helland 2004, 2010).

To arrive at methods of inference, the model concept must be supplemented by certain principles. In this connection, an experiment is ordinarily seen as given by a statistical model together with some focus parameter. Most

statisticians agree at least to some variants of the following three principles: The conditionality principle (When you choose an experiment randomly, the information in this large experiment, including the ►randomization, is not more than the information in the selected experiment.), the sufficiency principle (Experiments with equal values of a sufficient statistics have equal information.) and the likelihood principle (All the information about the parameter is contained in the likelihood for the parameter, given the observations.). Birnbaum's famous theorems says that likelihood principle follows from the conditionality principle together with the sufficiency principle (for some precisely defined version of these principles). This, and the principles themselves are discussed in detail in Berger and Wolpert (1984).

Berger and Wolpert (1984) also argue that the likelihood principle "nearly" leads to Bayesian inference, as the only mode of inference which really satisfies the likelihood principle. This whole chain of reasoning has been countered by Kardaun et al. (2003) and by leCam (1984), who states that he prefers to be a little "unprincipled."

To arrive at statistical inference, whether it is point estimates, confidence intervals (credibility intervals for Bayesians) or hypothesis testing, we need some decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview). Such decision theory may be formulated differently by different authors. The foundation of statistical inference from a Bayesian point of view is discussed by Good (1988), Lindley (2000) and Savage (1972). From the frequentist point of view it is argued by Efron (1986) that one should be a little more informal; but note that a decision theory may be very useful also in this setting.

The philosophy of foundations of statistics involves many further questions which have direct impact on the theory and practice of statistics: Conditioning, randomization, shrinkage, subjective or objective priors, reference priors, the role of information, the interpretation of probabilities, the choice of models, optimality criteria, nonparametric versus parametric inference, the principle of an axiomatic foundation of statistics etc. Some papers discussing these issues are Cox (1997) with discussion, Kardaun et al. (2003) and Efron (1978, 1979). The last paper takes up the important issue of the relationship between statistical theory and the ongoing revolution of computers.

The struggle between ►Bayesian statistics and frequentist statistics is not so hard today as it used to be some years ago, partly since it has been realized that the two schools often lead to similar results. As hypothesis testing and confidence intervals are concerned, the frequentist school and the Bayesian school must be adjoined by the Fisherian or fiducian school, although largely out of fashion today. The question of whether these three schools in some sense can agree on testing, is addressed by Berger (2003).

The field of design of experiments also has its own philosophical foundation, touching upon practical issues like randomization, blocking and replication, and linked to the philosophy of statistical inference. A good reference here is Cox and Reid (2000).

About the Author

Inge Svein Helland is Professor, University of Oslo (1996–present). He was Head of Department of Mathematical Sciences, Agricultural University of Norway (1987–1989) and Head of Statistics Division, Department of Mathematics, University of Oslo (1999–2001). He was Associate Editor, *Scandinavian Journal of Statistics* (1994–2001). Professor Helland has (co-)authored about 65 papers, reports, manuscripts, including the book *Steps Towards a Unified Basis for Scientific Models and Methods* (World Scientific, Singapore, 2010).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Foundations of Probability
►Likelihood
►Model Selection
►Philosophy of Probability
►Randomization
►Statistical Inference: An Overview

References and Further Reading

Berger JO (2003) Could fisher, jeffreys and neyman have agreed on testing? Stat Sci 18:1–32
Berger JO, Wolpert RL (1984) The likelihood principle. Lecture Notes, Monograph Series, vol 6, Institute of Mathematical Statistics, Hayward
Breiman L (2001) Statistical modelling: the two cultures. Stat Sci 16:199–231
Cox DR (1997) The current position of statistics: a personal view. Int Stat Rev 65:261–290
Cox DR, Reid N (2000) The theory of design of experiments. Chapman and Hall/CRC, Boca Raton, FL
Efron B (1978) Controverses in the foundations of statistics. Am Matem Month 85:231–246
Efron B (1979) Computers and the theory of statistics: thinking the unthinkable. Siam Rev 21:460–480
Efron B (1986) Why isn't everyone Bayesian? Am Stat 40:1–5
Good IJ (1988) The interface between statistics and philosophy of science. Stat Sci 3:386–412
Helland IS (2004) Statistical inference under symmetry. Int Stat Rev 72:409–422

Helland IS (2010) Steps towards a unified basis for scientific models and methods. World Scientific, Singapore
Kardaun OJWF, Salomé D, Schaafsma W, Steerneman AGM, Willems JC, Cox DR (2003) Reflections on fourteen cryptic issues concerning the nature of statistical inference. Int Stat Rev 71: 277–318
leCam L (1984) Discussion in Berger and Wolpert. 182–185
Lindley DV (2000) The philosophy of statistics. Statistician 49: 293–337
McCullagh P (2002) What is a statistical model? Ann Stat 30: 1225–1310
Savage LJ (1972) The foundations of statistics. Dover, New York
Wittgenstein L (1961) Tractatus Logico-Philosophicus (tr. Pears and McGuinnes). Routledge Kegan Paul, London

Philosophy of Probability

Martin Peterson
Associate Professor
Eindhoven University of Technology, Eindhoven, Netherlands

The probability calculus was created in 1654 by Pierre de Fermat and Blaise Pascal. The philosophy of probability is the philosophical inquiry into the semantic and epistemic properties of this mathematical calculus. The question at the center of the philosophical debate is *what it means to say* that the probability of an event or proposition equals a certain numerical value, or in other words, what the *truth-conditions* for a probabilistic statement are. There is significant disagreement about this, and there are two major camps in the debate, viz., *objectivists* and *subjectivists*.

Objectivists maintain that statements about probability refer to some features of the external world, such as the relative frequency of some event. Probabilistic statements are thus objectively true or false, depending on whether they correctly describe the relevant features of the external world. For example, some objectivists maintain that it is true that the probability that a coin will land heads is $\frac{1}{2}$ if and only if the relative frequency of this type of event is $\frac{1}{2}$ (Other objective interpretations will be discussed below).

Subjectivists disagree with this picture. They deny that statements about probability should be understood as claims about the external world. On their view, they should rather be understood as claims about the speaker's degree of belief that an event will occur. Consider, for example, Mary's probability that her suitor will propose. What is Mary's probability that this event will occur? It seems rather pointless to count the number of marriage proposals that other people get, because this does not tell us anything about the probability that *Mary* will be faced with a marriage proposal. If it is true that Mary's probability is, say, 1/2 then it is true because of her mental state, i.e., her degree of belief that her suitor will propose. Unfortunately, nothing follows from this about whether her suitor actually will propose or not. Mary's probability that her suitor will propose may be high even if the suitor feels that marriage is totally out of the question.

Both the objective and subjective interpretations are compatible with Kolmogorov's axioms. These axioms come out as true, irrespective of whether we interpret them along the lines suggested by objectivists and subjectivists. However, more substantial questions about what the probability of an event is may depend on which interpretation is chosen. For example, Mary's subjective belief that her suitor will propose might be low, although the objective probability is quite high.

The notion of subjective probability is closely related to ►Bayesian statistics, presented elsewhere in this book. (In recent years some authors have, however, also developed objectivist accounts of Bayesian statistics.)

In what follows we shall give a more detailed overview of some of the most well-known objective and subjective interpretations, viz. the relative-frequency view, the propensity interpretation, the logical interpretation, and the subjective interpretation. We shall start, however, with the classical interpretation. It is strictly speaking neither an objective nor a subjective interpretation, since it is salient about many of the key questions discussed by objectivists and subjectivists.

The *classical interpretation*, advocated by Laplace, Pascal, Bernoulli and Leibniz, holds the probability of an event to be a fraction of the total number of possible ways in which the event can occur. Hence, the probability that you will get a six if you roll a six-sided die is 1/6. However, it takes little reflection to realize that this interpretation presupposes that all possible outcomes are equally likely. This is not always a plausible assumption, as Laplace and others were keen to stress. For an extreme example, consider the weather in the Sahara desert. It seems to be much more probable that it will be sunny in the Sahara desert tomorrow than not, but according to the classical interpretation the probability is $\frac{1}{2}$ (since there are two possible outcomes, sun or no sun). Another problem with the classical interpretation is that it is not applicable when the number of possible outcomes is infinite. Then the probability of every possibility would be zero, since the ratio between any finite number and infinity is zero.

The *frequency interpretation*, briefly mentioned above, holds that the probability of an event is the ratio between the numbers of times the event has occurred divided by the total number of observed cases. Hence, if you toss a coin 1,000 times and it lands heads up 508 times then the relative frequency, and thus the probability,

would be 508/1000 = 0.508. A major challenge for anyone seeking to defend the frequency interpretation is to specify which reference class is the relevant one and why. For example, suppose I toss the coin another 1,000 times, and that it lands heads up on 478 occasions. Does this imply that the probability has changed from 0.508 to 0.486? Or is the "new" probability 508 + 486/2.000? The physical constitution of the coin is clearly the same.

The problem of identifying the relevant reference class becomes particularly pressing as the frequency interpretation is applied to unique events, i.e., events that only occur once, such as the US presidential election in 2000 in which George W Bush won over Al Gore. The week before the election the probability was – according to many political commentators – about 50% that Bush would win. Now, to which reference class does this event belong? If this was a *unique* event the reference class has, by definition, just one element, viz., the event itself. So according to the frequency interpretation the probability that Bush was going to win was 1/1, since Bush actually won the election. This cannot be the right conclusion.

Venn famously argued that the frequency interpretation makes sense only if the reference class is taken to be infinitely large. More precisely put, he pointed out that one should distinguish sharply between the underlying *limiting* frequency of an event and the frequency *observed* so far. The limiting frequency is the proportion of successful outcomes *would* get if one were to repeat *one and the same* experiment infinitely many times. So even though the US presidential election in 2000 did *actually* take place just once, we can nevertheless *imagine* what would have happened had it been repeated infinitely many times.

Of course, we cannot actually toss a coin infinitely many times, but we could imagine doing so. Therefore, the limiting frequency is often thought of as an abstraction, rather than as a series of events that take place in the real world. This point has some important philosophical implications. First, it seems that one can never be sure that a limiting relative frequency exists. When tossing a coin, the relative frequency of heads will perhaps never converge towards a specific number. In principle, it could oscillate forever. Moreover, the limiting relative frequency seems to be inaccessible from an epistemic point of view, even in principle. If you observe that the relative frequency of a coin landing heads up is close to 1/2 in a series of ten million tosses, this does not exclude that the true long-run frequency is much lower or higher than 1/2. No finite sequence of observations can prove that the limiting frequency is even close to the observed frequency.

According to the *propensity interpretation*, probabilities should be identified with another feature of the external world, namely the propensity (or disposition or tendency) of an object to give rise to a certain effect. For instance, symmetrical coins typically have a propensity to land heads up about every second time they are tossed, which means that their probability of doing so is about one in two.

The propensity interpretation was developed by Popper in the 1950s. His motivation for developing this view was that it avoids the problem of assigning probabilities to unique events faced by the frequency view. Even an event that cannot take place more than once can nevertheless have a certain propensity (or tendency) to occur. However, Popper's version of the theory also sought to connect propensities with long-run frequencies whenever the latter existed. Thus, his theory is perhaps best thought of as a hybrid between the two views. Contemporary philosophers have proposed "pure" versions of the propensity interpretation, which make no reference what so ever to long-run frequencies.

A well-known objection to the propensity interpretation is Humphreys' paradox. To state this paradox, recall that conditional probabilities can be "inverted" by using ►Bayes' theorem. Thus, if we know the probability of A given B we can calculate the probability of B given A, given that we know the priors. The point made by Humphreys is that propensities cannot be inverted in this sense. Suppose, for example, that we know the probability that a train will arrive on time at its destination given that it departs on time. Then it makes sense to say that if the train departs on time, it has a propensity to arrive on time at its destination. However, even though it makes sense to speak of the inverted probability, i.e., the probability that the train departed on time given that it arrived on time, it makes no sense to speak of the corresponding inverted propensity. No one would admit that the on-time arrival of the train has a propensity to make it depart on time a few hours earlier.

The thrust of Humphreys' paradox is thus the following: Even though we may not know exactly what a propensity (or disposition or tendency) is, we do know that propensities have a temporal direction. If A has a propensity to give rise to B, then A cannot occur after B. In this respect, propensities function very much like causality; if A causes B, then A cannot occur after B. However, probabilities lack this temporal direction. What happens now can tell us something about the probability of past events, and reveal information about past causes and propensities, although the probability in itself is a non-temporal concept. Hence, it seems that it would be a mistake to identify probabilities with propensities.

The *logical interpretation* of probability was developed by Keynes and Carnap. Its basic idea is that probability is a logical relation between a hypothesis and the evidence

supporting it. More precisely put, the probability relation is best thought of as a generalization of the principles of deductive logic, from the deterministic case to the indeterministic one. For example, if an unhappy housewife claims that the probability that her marriage will end in a divorce is 0.9, this means that the evidence she has at hand (no romantic dinners, etc.) entails the hypothesis that the marriage will end in a divorce to a certain degree, which can be represented by the number 0.9. Coin tossing can be analyzed along the same lines. The evidence one has about the shape of the coin and past outcomes entails the hypothesis that it will land heads up to a certain degree, and this degree is identical with the probability of the hypothesis being true.

Carnap's analysis of the logical interpretation is quite sophisticated, and cannot be easily summarized here. However, a general difficulty with logical interpretations is that they run a risk of being too dependent on evidence. Sometimes we wish to use probabilities for expressing mere guesses that have no correlation whatsoever to any evidence. For instance, I think the probability that it will be sunny in Rio de Janeiro tomorrow is 0.4. This guess is not based on any meteorological evidence. I am just guessing – the set of premises leading up to the hypothesis that it will be sunny is empty; hence, there is no genuine "entailment" going on here. So how can the hypothesis that it will be sunny in Rio de Janeiro tomorrow be entailed to degree 0.4, or any other degree?

It could be replied that pure guesses are irrational, and that it is therefore not a serious problem if the logical interpretation cannot handle this example. However, it is not evident that this is a convincing reply. People do use probabilities for expressing pure guesses, and the probability calculus can easily be applied for checking whether a set of such guesses are coherent or not. If one thinks that the probability for sun is 0.4 it would for instance be correct to conclude that the probability that it will not be sunny is 0.6. This is no doubt a legitimate way of applying the probability calculus. But if we accept the logical interpretation we cannot explain why this is so, since this interpretation defines probability as a *relation* between a (non-empty) set of evidential propositions and a hypothesis.

Let us now take closer look at the *subjective* interpretation. The main idea is that probability is a kind of mental phenomenon. Probabilities are not part of the external world; they are entities that human beings somehow create in their minds. If you claim that the probability for sun tomorrow is, say, 0.8 this merely means that your subjective degree of belief that it will be sunny tomorrow is strong and that the strength of this belief can be represented by the number 0.8. Of course, whether it *actually* will rain tomorrow depends on objective events in the external world, rather than on your beliefs. So it is *probable* that it will rain tomorrow just in case you believe that it will rain to a certain degree, irrespective of what the weather is actually like tomorrow. However, this should not be taken to mean that any subjective degree of belief is a probability. Advocates of the subjective approach stress that for a partial belief to qualify as a probability, one's degrees of belief must be compatible with the axioms of the probability calculus.

Subjective probabilities can vary across people. Mary's degree of belief that it will rain tomorrow might be strong, at the same time as your degree of belief is much lower. This just means that your mental dispositions are different. When two decision makers hold different subjective probabilities, they just happen to believe something to different degrees. It does not follow that at least one person has to be wrong. Furthermore, if there were no humans around at all, i.e., if all believing entities were to be extinct, it would simply be false that some events happen with a certain probability, including quantum-mechanical events. According to the pioneering subjectivist Bruno de Finetti, "Probability does not exist."

Subjective views have been around for almost a century. de Finetti's pioneering work was published in 1931. Ramsey presented a similar subjective theory in a paper written in 1926 and published posthumously in 1931. However, most modern accounts of subjective probability start off from Savage's theory, presented in 1954, which is more precise from a technical point of view. The key idea in all three accounts is to introduce an ingenious way in which subjective probabilities can be measured. The measurement process is based on the insight that the degree to which a decision maker believes something is closely linked to his or her behavior. Imagine, for instance, that we wish to measure Mary's subjective probability that the coin she is holding in her hand will land heads up the next time it is tossed. First, we ask her which of the following very generous options she would prefer.

(a) "If the coin lands heads up you win a trip to Bahamas; otherwise you win nothing"
(b) "If the coin *does not* land heads up you win a trip to Bahamas; otherwise you win nothing"

Suppose Mary prefers A to B. We can then safely conclude that she thinks it is *more probable* that the coin will land heads up rather than not. This follows from the assumption that Mary prefers to win a trip to Bahamas rather than nothing, and that her preference between uncertain prospects is entirely determined by her beliefs and desires with respect to her prospects of winning the trip to Bahamas. If she on the other hand prefers B to A, she thinks it is *more probable* that the coin will not land heads

up, for the same reason. Furthermore, if Mary is indifferent between A and B, her subjective probability that the coin will land heads up is exactly 1/2. This is because no other probability would make both options come out as equally attractive, irrespective of how strongly she desires a trip to Bahamas, and irrespective of which decision rule she uses for aggregating her desires and beliefs into preferences.

Next, we need to generalize the measurement procedure outlined above such that it allows us to always represent Mary's degree of belief with precise numerical probabilities. To do this, we need to ask Mary to state preferences over a *much larger* set of options and then *reason backwards*. Here is a rough sketch of the main idea: Suppose that Mary wishes to measure her subjective probability that her etching by Picasso worth $20,000 will be stolen within one year. If she considers $1,000 to be a fair price for insuring her Picasso, that is, if that amount is the highest price she is prepared to pay for a gamble in which she gets $20,000 if the event *S*: "The Picasso is stolen within a year" takes place, and nothing otherwise, then Mary's subjective probability for *S* is $\frac{1{,}000}{20{,}000} = 0.05$, given that she forms her preferences in accordance with the principle of maximizing expected monetary value. If Mary is prepared to pay up to $2,000 for insuring her Picasso, her subjective probability is $\frac{2{,}000}{20{,}000} = 0.1$, given that she forms her preferences in accordance with the principle of maximizing expected monetary value.

Now, it seems that we have a general method for measuring Mary's subjective probability: We just ask her how much she is prepared to pay for "buying a contract" that will give her a fixed income if the event we wish to assign a subjective probability to takes place. The highest price she is prepared to pay is, by assumption, so high that she is indifferent between paying the price and not buying the contract. (This assumption is required for representing probabilities with precise numbers; if buying and selling prices are allowed to differ we can sometimes use intervals for representing probabilities. See e.g., Borel and Baudain 1962 and Walley 1991.)

The problem with this method is that very few people form their preferences in accordance with the principle of maximizing expected monetary value. Most people have a decreasing marginal utility for money. However, since we do not know anything about Mary's utility function for money we cannot replace the monetary outcomes in the examples with the corresponding utility numbers. Furthermore, it also makes little sense to *presuppose* that Mary uses a specific decision rule, such as the expected utility principle, for forming preferences over uncertain prospects. Typically, we do not know anything about how people form their preferences.

Fortunately, there is a clever solution to all these problems. The main idea is to impose a number of structural conditions on preferences over uncertain options. The structural conditions, or axioms, merely restrict what *combinations* of preferences it is legitimate to have. For example, if Mary strictly prefers option A to option B in the Bahamas example, then she must not strictly prefer B to A. Then, the subjective probability function is established by reasoning backwards while taking the structural axioms into account: Since the decision maker preferred some uncertain options to others, and her preferences over uncertain options satisfy a number of structural axioms, the decision maker behaves *as if* she were forming her preferences over uncertain options by first assigning subjective probabilities and utilities to each option, and thereafter maximizing expected utility. A peculiar feature of this approach is, thus, that probabilities (and utilities) are derived from "within" the theory. The decision maker does not prefer an uncertain option to another *because* she judges the subjective probabilities (and utilities) of the outcomes to be more favorable than those of another. Instead, the well-organized structure of the decision maker's preferences over uncertain options logically imply that they can be described *as if* her choices were governed by a subjective probability function and a utility function, constructed such that a preferred option always has a higher expected utility than a non-preferred option. These probability and utility functions need not coincide with the ones outlined above in the Bahamas example; all we can be certain of is that there exist *some* functions that have the desired technical properties.

Cross References

►Axioms of Probability
►Bayes' Theorem
►Bayesian Statistics
►Foundations of Probability
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Philosophical Foundations of Statistics
►Probability Theory: An Outline
►Probability, History of

References and Further Reading

Borel E, Baudain M (1962) Probabilities and life. Dover Publishers, New York

de Finetti B (1931/89) Probabilism: a critical essay on the theory of probability and on the value of science. Erkenntnis 31:169–223 (trans: de Finetti B (1931) Probabilismo. Logos 14:163–219)

Humphreys P (1985) Why propensities cannot be probabilities. Philos Rev 94:557–570

Jeffrey R (1983) The logic of deciosion, 2nd edn. (significant improvements from 1st edn). University of Chicago Press, Chicago
Kreps DM (1988) Notes on the theory of choice, Westview Press, Boulder
Laplace PS (1814) A philosophical essay on probabilities (English transl 1951). Dover Publications Inc, New York
Mellor DH (1971) The Matter of Chance. Cambridge University Press, Cambridge
Popper K (1957) The propensity interpretation of the calculus of probability and the quantum theory. In: Körner S (ed) The colston papers, vol 9, pp 65–70
Ramsey FP (1926) Truth and probability in Ramsey, 1931. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays, Ch. VII, Kegan, Paul, Trench, Trubner & Co, London, Harcourt, Brace and Company, New York, pp 156–198
Savage LJ (1954) The foundations of statistics. Wiley, New York. 2nd edn. 1972, Dover
Walley P (1991) Statistical reasoning with imprecise probabilities, monographs on statistics and applied probability. Chapman & Hall, London

Point Processes

David Vere-Jones
Professor Emeritus
Victoria University of Wellington, Wellington, New Zealand

"Point Processes" are locally finite (i.e., no finite accumulation points) families of events, typically occurring in time, but often with additional dimensions (marks) to describe their locations, sizes and other characteristics.

The subject originated in attempts to develop life-tables. The first such studies included one by Newton and another by his younger contemporary Halley.

Newton's study was provoked by his life-long religious concerns. In his book, "Chronology of Ancient Kingdoms Amended" he set out to estimate the dates of various biblical events by counting the number of kings or other rulers between two such biblical events and allowing each ruler a characteristic length of reign. The value he used seems to have been arrived at by putting together all the observations from history that he could find (he included rulers from British, classical, European and biblical histories and even included Methuselah's quoted age among his data points) and taking some average of their lengths of rules, but he acknowledged himself that his methods were informal and personal.

Halley, by contrast, was involved in a scientific exercise, namely the development of actuarial tables for use in calculating pensions and annuities. Indeed, he was requested by the newly established Royal Society to prepare such a table from records in Breslau, a city selected because it had been less severely affected by the plague than most of the larger cities in Europe, so that its mortality data were felt more likely to be typical of those of cities and periods in normal times.

Another important early stimulus for point process studies was the development of telephone engineering. This application prompted Erlang's early studies of the Poisson process (see ▶Poisson Processes), which laid down many of the concepts and procedures, such as ▶renewal processes, and forward and backward recurrence times, subsequently entering into point process theory. It was also the context of Palm's (1943) deep studies of telephone-traffic issues. Palm was the first to use the term "point process" (Punkt-Prozesse) itself.

A point process can be treated in many different ways: for example, as a sequence of delta functions, as a sequence of time intervals between event occurrences; as an integer-valued random measure; or as the sum of a regular increasing component and a jump-type martingale.

Treating each point as a delta-function in time yields a time series which has generalized functions as realizations, but in other respects has many similarities with a continuous time series. In particular, stationarity, ergodicity, and a "point process spectrum" can be developed through this approach.

If attention is focused on the process of intervals between successive points, the paradigm example is the renewal process, where the intervals between events are independent, and, save possibly for the first interval, identically distributed.

Counting the number of events, say $N(A)$ falling into a pre-specified set A (an interval, or more general Borel set), leads to treating the point process as an integer-valued random measure.

An important underlying theorem, first enunciated and analyzed by Slivnyak (1962), asserts the equivalence of the counting and interval based approaches.

Random measures form a generalization of point processes which are of considerable importance in their own right. Their first and second order moment measures

$$M_1(A) = E[N(A)]$$
$$M_2(A \times B) = E[N(A)N(B)]$$

and the associated signed measure, the covariance measure

$$C_2(A \times B) = M_2(A \times B) - M_1(A)M_1(B)$$

form non-random measures which have been extensively studied.

A third point of view originated more recently from martingale ideas applied to point processes, and has proved a rich source of both new models and new methods of analysis. The key here is to define the point process in terms of its conditional intensity (or conditional hazard) function, representing the conditional rate of occurrence of events, given the history (record of previously occurring events and any other relevant information) up to the present.

These ideas are closely linked to the martingale representation of point processes. This takes the form of an increasing step function, with unit steps at the time points when the events occur, less a continuous increasing part given by the integral of the conditional intensity.

The Hawkes' processes [introduced by Hawkes (1971a, b)] form an important class of point processes introduced and defined by their conditional intensities, which take the general form

$$\lambda(t) = \mu + \sum_{i:t_i - t} g(t - t_i)$$

where μ is a non-negative constant ("the arrival rate") and g is a non-negative integrable function ("the infectivity function").

The archetypal point process is the simple Poisson process, where the intervals between successive events in time are independent, and, (with the possible exception of the initial interval) are identically and exponentially distributed with a common mean, say m. For this process, the conditional intensity is constant, equal to the rate of occurrence (intensity) of the Poisson process, here $1/m$. Processes with continuous conditional intensities are sometimes referred to as "processes of Poisson type" since they behave locally like Poisson processes over intervals small enough for the conditional intensity to be considered approximately constant.

For the Poisson process itself, and also Poisson cluster processes, where cluster centers follow a simple Poisson process, and the clusters are independent subprocesses, identically distributed relative to their cluster centers, it is possible to write down simple expressions for the characteristic functional

$$\Phi[h] = E\left[e^{i \int h(t)\, dN(t)}\right]$$

where the carrying function h is integrable against M_1, or the essentially equivalent probability generating functional

$$G[\xi] = E[\Pi \xi(t_i)]$$

where ξ plays the role of h in the characteristic functional.

For example, the Poisson process with continuous intensity m has probability generating functional

$$G[h] = \exp\left\{ - m \int [1 - h(u)]\, du \right\}.$$

Such functionals provide a comprehensive summary of the process and its attributes, especially the moment structure.

However, the usefulness of characteristic or generating functionals in practice is restricted by the relatively few examples for which they can be obtained in convenient closed form. Nevertheless, where available, their form is essentially independent of the space (phase space) in which the points are located. By contrast, finding extensions of the conditional intensity to spaces of more than one dimension has proved extremely difficult, on account of the absence of a clear linear ordering, a problem which equally affects other attempts to extend martingale ideas to higher dimensions.

About the Author

Dr. David Vere-Jones is Emeritus Professor, Victoria University of Wellington (since 2000). He is Past President, New Zealand Statistical Association (1981–1983), Past President of Interim Executive of International Association for Statistical Education (1991–1992), Founding President of the New Zealand Mathematical Society (1974). He received the Rutherford Medal, New Zealand's top science award, in 1999 for "outstanding and fundamental contributions to research and education in probability, statistics and the mathematical sciences, and for services to the statistical and mathematical communities, both within New Zealand and internationally." Professor Vere-Jones was also awarded the NZ Science and Technology Gold Medal (2000), and the NZSA Campbell Award for 2009, in recognition for his contributions to the statistical sciences. He has (co-)authored over 100 refereed publications, and three books. In 2001 (April 19–21), a Symposium in Honor of David Vere-Jones on the Occasion of His 65th Birthday was held in Wellington.

"David Vere-Jones is New Zealand's leading resident mathematical statistician. He has made major contributions to the theory of statistics, its applications, and to the teaching of statistics in New Zealand. He is highly regarded internationally and is involved in numerous international activities. One of his major research areas has been concerned with Point Processes... . A substantial body of the existing theory owes its origins to him, either directly or via his students. Of particular importance and relevance to New Zealand is his pioneering work on the applications of point process theory to seismology" (NZMS Newsletter 24, August 1982).

Cross References

►Khmaladze Transformation
►Martingales
►Non-Uniform Random Variate Generations
►Poisson Processes
►Renewal Processes
►Spatial Point Pattern
►Spatial Statistics
►Statistics of Extremes
►Stochastic Processes

References and Further Reading

Cox DR, Isham V (1980) Point Processes. Chapman and Hall, London

Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Methuen, London

Daley DJ, Vere-Jones D (1988) An introduction to the theory of point processes, 1st edn. Springer, New York; 2nd. edn. (2002) vols 1 and 2, Springer, New York

Hawkes AG (1971a) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58:83–90

Hawkes AG (1971b) Point spectra of some mutually exciting point processes. J R Stat Soc 33:438–443

Palm C (1943) Intensitasschwankungen im fernsprechverkehr. Ericsson Technics 44:1–189

Slivnyak IM (1962) Some properties of stationary flows of homogeneous random events. Teor. Veroyantnostei I Primen 7:347–352, Translation in Theory of Probability and Applications, 7:36–341

Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields. Wiley, Chichester

Stoyan D, Kendall WS, Mecke J (1995) Stochastic geometry. 2nd edn. Wiley, Chichester; 1st edn. Academie, Berlin, 1987

Poisson Distribution and Its Application in Statistics

Lelys Bravo de Guenni
Professor
Universidad Simón Bolívar, Caracas, Venezuela

The Poisson distribution was first introduced by the French Mathematician Siméon-Denis Poisson (1781–1840) to describe the probability of a number of events occurring in a given time or space interval, with the probability of occurrence of these events being very small. However, since the number of trials is very large, these events do actually occur.

It was first published in 1837 in his work *Recherches sur la probabilité des jugements en matières criminelles et matière civile* (Research on the Probability of Judgments in Criminal and Civil Matters). In this work, the behavior of certain random variables X that count the number of occurrences (or *arrivals*) of such events in a given interval in time or space was described. Some examples of these events are infant mortality in a city, the number of misprints in a book, the number of bacteria on a plate, the number of activations of a geiger counter, and so on.

Assuming that λ is the expected value of such arrivals in a time interval of fixed length, the probability of observing exactly k events is given by the probability mass function

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k = 0, 1, 2, \ldots$. The parameter distribution λ is a positive real number, which represents the average number of events occurring during a fixed time interval. For example, if the event occurs on average three times per second, in 10 s the event will occur on average 30 times and $\lambda = 30$. When the number of trials n is large and the probability of occurrence of the event λ/n approaches to zero, the ►binomial distribution with parameters n and $p = \lambda/n$ can be approximated to a Poisson distribution with parameter λ. The binomial distribution gives the probability of x successes in n trials.

If X is a random variable with a Poisson distribution, the expected value of X and the variance of X are both equal to λ. To estimate λ by maximum likelihood, given a sample $k_1, k_2, \ldots, k_n$, the log-likelihood function

$$L(\lambda) = \log \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!}$$

is maximized with respect to λ and the resulting estimate is $\hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$. This is an unbiased estimator since the expected value of each k_i is equal to λ; and it is also an efficient estimator since its estimator variance achieves the Cramer–Rao lower bound. From the Bayesian inference perspective, a conjugate prior distribution for the parameter λ is the Gamma distribution. Suppose that λ follows a Gamma prior distribution with parameters α and β, such that

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad \lambda > 0$$

If a sample $k_1, k_2, \ldots, k_n$ of size n is observed, the posterior probability distribution for λ is given by

$$p(\lambda|k_1, k_2, \ldots, k_n) \sim Gamma\left(\alpha + \sum_{i=1}^{n} k_i, \beta + n\right).$$

When $\alpha \to$ and $\beta \to 0$, we have a diffuse prior distribution, and the posterior expected value of λ ($E[\lambda|k_1, k_2, \ldots, k_n]$) approximates to the maximum likelihood estimator $\hat{\lambda}$.

A use for this distribution was not found until 1898, when an individual named Bortkiewicz (O'Connor and

Robertson 1950) was asked by the Prussian Army to investigate accidental deaths of soldiers attributed to being kicked by horses. In 1898, he published *The Law of Small Numbers*. In this work he was the first to note that events with low frequency in a large population followed a Poisson distribution even when the probabilities of the events varied. Bortkiewicz studied the distribution of 122 men kicked to death by horses among 14 Prussian army corps over 20 years. This famous data set has been used in many statistical textbooks as a classical example on the use of the Poisson distribution (see Yule and Kendall [1950] or Fisher [1954]). He found that in about half of every army corps–year combination, there were no deaths for horse kicking. For other combinations of corps–years, the number of deaths were from 1 to 4. Although the probability of horse kick deaths might vary from corps and years, the overall observed frequencies were very close to the expected frequencies estimated by using a Poisson distribution.

In epidemiology, the Poisson distribution has been used as a model for deaths. In one of the oldest textbooks of statistics published by Bowley in 1901 (cited by Hill [2002]), he fitted a Poisson distribution to deaths from splenic fever in the years 1875–1894 and showed a reasonable agreement with the theory. At that time, splenic fever was a synonym for present-day anthrax.

A more extensive use of the Poisson distribution can be found within the Poisson generalized linear models, usually called the *Poisson regression models* (see ►Poisson Regression). These models are used to model count data and contingency tables. For contingency tables, the Poisson regression model is best known as the *log-linear model*. In this case, the response variable Y has a Poisson distribution and usually, the logarithm of its expected value ($E[Y]$) is expressed as a linear predictor $X\beta$ where X is a $n \times p$ matrix of explanatory variables and β is a parameter vector of size p. In this case, the *link* function $g(.)$, which relates the expected value of the response variable Y with the linear predictor is the logarithm function, in such a way that the mean of the response variable $\mu = g^{-1}(X\beta)$.

Poisson regression can also be used to model what is called the *relative risk*. This is the ratio between the counts and an *exposure* factor E. For example, to model the relative risk of disease in a region, we make $\eta = Y/E$, where Y is the observed number of cases and E is the expected number of cases, which depends on the number of persons at risk. The usual model for Y is

$$Y|\eta \sim \text{Poisson}(E\eta)$$

where η is the true relative risk of disease (Banerjee et al. (2004)). When applying the Poisson model to data, the main assumption is that the variance is equal to the mean. In many cases, this may not be assumed since the variance of counts are usually greater than the mean. In this case, we have *overdispersion*. One way to deal with this problem is to use the negative binomial distribution, which is a two-parameter family that allows the mean and variance to be fitted separately. In this case, the mean of the Poisson distribution λ is assumed a random variable as drawn from a Gamma distribution.

Another common problem with Poisson regression is excess zeros: if there are two processes at work, one determining whether there are zero events or any events, and a Poisson process (see ►Poisson Processes) determining how many events there are, there will be more zeros than a Poisson regression would predict. An example would be the distribution of cigarettes smoked in an hour by members of a group where some individuals are nonsmokers. These data sets can be modeled as *zero inflated Poisson models*, where p is the probability of observing zero counts, and $1 - p$ is the probability of observing a count variable modeled as a Poisson(λ).

About the Author

For biography *see* the entry ►Normal Scores.

Cross References

►Contagious Distributions
►Dispersion Models
►Expected Value
►Exponential Family Models
►Fisher Exact Test
►Geometric and Negative Binomial Distributions
►Hypergeometric Distribution and Its Application in Statistics
►Modeling Count Data
►Multivariate Statistical Distributions
►Poisson Processes
►Poisson Regression
►Relationships Among Univariate Statistical Distributions
►Spatial Point Pattern
►Statistics, History of
►Univariate Discrete Distributions: An Overview

References and Further Reading

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis of spatial data. Chapman and Hall/CRC, Boca Raton
Fisher RA (1954) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Hill G (2002) Horse kicks, antrax and the poisson model for deaths. Chronic Dis Can 23(2):77
O'Connor JJ, Robertson EF (1950) http://www-groups.dcs.st-and.ac.uk/history/Mathematicians/Bortkiewicz.html
Yule GU, Kendall MG (1950) An introduction to the theory of statistics. Charles Griffin, London

Poisson Processes

MR Leadbetter
Professor
University of North Carolina, Chapel Hill, NC, USA

Introduction

Poisson Processes are surely ubiquitous in the modeling of point events in widely varied settings, and anything resembling a brief exhaustive account is impossible. Rather we aim to survey several "Poisson Habitats" and properties, with glimpses of underlying mathematical framework for these processes and close relatives. We refer to three (of many) authoritative works (Cox and Lewis 1966; Daley and Vere-Jones 1988; Kallenberg 1986) for convenient detailed accounts of Poisson Process and general related theory, tailored to varied mathematical tastes.

In this entry we first consider Poisson Processes in their classical setting as series of random events (►point processes) on the real line (e.g., in time), their importance in one dimension for stochastic modeling being rivaled only by the Wiener Process - both being basic (for different purposes) in their own right, and as building blocks for more complex models. Classical applications of the Poisson process abound - exemplified by births, radioactive disintegrations, customer arrival times in queues, instants of new cases of disease in an epidemic, and (a favorite of ours), the crossings of a high level by a stationary process. The classical format for Poisson Processes will be described in section "►Poisson Processes on the Real Line", and important variants in section "►Important Variants".

Unlike many situations in which generalizations to more than one dimension seem forced, the Poisson process has important and natural extensions to two and higher dimensions, as indicated in section "►Poisson Processes in Higher Dimensions". Further, an even more attractive feature (at least to those with theoretical interests) is the fact that Poisson Processes can be defined on spaces with very little structure, as indicated in section "►Poisson Processes on Abstract Spaces".

Poisson Processes on the Real Line

A *Point Process* on the real line is simply a sequence of events occurring in time (or some other 1-dimensional (e.g., distance) parameter) according to a probabilistic mechanism. One way to describe the probability structure is to define a sequence $0 \le \tau_1 < \tau_2 < \tau_3 \ldots < \infty$ where the τ_i are the "times" of occurrences of interest ("events" of the point process). They are assumed to be random variables (written here as distinct, i.e., strictly increasing, when the point process is termed "simple" but successive τ_i can be taken to be equal if "multiple events" are to be considered.) It is assumed that the points τ_i tend to infinity as $i \to \infty$ so that there are no "accumulation points". One may define a point process as a *random set* $\{\tau_i : i = 1, 2, \ldots\}$ of such points (cf Ryll-Nardzewski 1961). Alternatively the occurrence times τ_i are a family of random variables for $i = 1, 2, \ldots$ and may be discussed within the framework of *random sequences* (discrete parameter stochastic processes). However often the important quantities for a point process on the real line are the random variables $N(B)$ which are the (random) numbers of events (τ_i) in sets B of interest (usually Borel sets). When $B = (0, t]$ we write N_t for $N((0, t])$, i.e., the number of events occurring from time zero up to and including time t. $\{N_t\}$ is thus a family of random variables for positive t – or a non-negative integer-valued continuous parameter stochastic process on the positive real line. Likewise $\{N(B)\}$ defines a non-negative integer-valued stochastic process indexed by the (Borel) sets B (finite for bounded B).

Here (as is customary) we focus on the "counting" r.v.s $\{N_t\}$ or $\{N(B)\}$ rather than the consideration of more geometric properties of the sets $\{\tau_i\}$. Note that these two families are essentially equivalent since knowledge of $N(B)$ for each Borel set determines that of the sub-family $\{N_t\}(B = (0, t])$ and the converse is also true since $N(B)$ is a measure defined on the Borel sets and is determined by its values on the intervals $(0, t]$. Further their distributions are determined by those of the occurrence times τ_k and conversely, in view of equality of the events $(\tau_k > t)$, $(N_t < k)$.

We (finally!) come to the subject of this article – the Poisson Process. In its simplest context this is defined as a family $N(B)$, (or N_t) as above on the positive real line by the requirement that each N_t be Poisson, $P\{N_t = r\} = e^{-\lambda t}(\lambda t)^r/r!$, $r = 0, 1, 2, \ldots$ and that "increments" $(N_{t_2} - N_{t_1})$, $(N_{t_4} - N_{t_3})$, are independent for $0 < t_1 < t_2 \le t_3 < t_4$. Equivalently $N(B)$ is Poisson with mean $\lambda m(B)$ where m denotes Lebesgue measure, and $N(B_1)$, $N(B_2)$ are independent for disjoint B_1, B_2. $\lambda = \mathcal{E}N_1$ is the expected number of events per unit time, or the *intensity* of the Poisson process.

That the Poisson – rather than some other – distribution plays a central role is due in part to the long history of its use to describe rare events – such as the classical number of deaths by horse kicks in the Prussian army. But more significantly the very simplest modeling of a point process would surely require independence of increments which holds as noted above for the Poisson Process. Further this process has the stationarity property that the distribution of the "increment" $(N_{t+h} - N_t)$ depends only on the length

h of the interval, not on its starting point t. Moreover the probability of an event in a small interval of length h is approximately λh and the probability of more than one in that interval is of smaller order, i.e.,

$$P\{(N_{t+h} - N_t) = 1\} = \lambda h + o(h),$$
$$P\{(N_{t+h} - N_t) \geq 2\} = o(h) \text{ as } h \to o.$$

It turns out that the Poisson Process as defined is the *only* point process exhibiting stationarity and these two latter properties [see e.g., Durrett (2005) for proof] which demonstrates what Kingman aptly describes as the "inevitability" of the Poisson distribution, in his volume (Kingman 1993).

The Poisson Process has extensive properties which are well described in many works [e.g., Cox and Lewis (1966) and Daley and Vere-Jones (1988)] For example the equivalence of the events $(\tau_k > t)$, $(N_t < k)$ noted above readily shows that the inter-arrival times $(\tau_k - \tau_{k-1})$ are i.i.d. exponential random variables with mean $\lambda^{-1}(\tau_0 = 0)$, and τ_k itself is the sum of the first k of these, thus distributed as $(2\lambda)^{-1}\chi^2_{2k}$. On the other hand we have the famous apparent paradox that the random interval which contains a fixed point t_0 is distributed as the sum of two independent such exponential variables – the time from the preceding event plus that to the following event. This and a host of other useful properties may be conveniently found in Cox and Lewis (1966) and Daley and Vere-Jones (1988). Finally, the above discussion has focused on Poisson Processes on the positive real line. It is a simple matter to add an independent Poisson Process on the negative real line to obtain one on the entire real line $(-\infty, \infty)$.

Important Variants

The stationarity requirement of a constant intensity λ may be generalized to include a time varying intensity function $\lambda(t)$ for which the number of events $N(B)$ in a (Borel) set B is still Poisson but with mean $\Lambda(B) = \int_B \lambda(t)dt$, keeping independence of $N(B_1)$, $N(B_2)$ for disjoint B_1, B_2. Then N_t is Poisson with mean $\int_0^t \lambda(t)dt$. More generally one may take Λ to be a "measure" on the Borel sets but not necessarily of this integral (absolutely continuous) form which unlike the simple Poisson Process above, does not necessarily prohibit the occurrence of more than one event at the same instant, (multiple events) and may allow positive probability of an event occurring at a given fixed time point. Further one may consider random versions of the intensity e.g., with $\lambda(t)$ being itself a stochastic process ("*stochastic intensity*") as for example the blood pressure of an individual (varying randomly in time) leading to (conditionally) Poisson chest pain incidents. The resulting point processes are termed *doubly stochastic Poisson* or *Cox Processes*, and are widely used in medical trials e.g., of new treatments. For other widely used variants of Poisson Processes (e.g., "Mixed" and "Compound" Poisson processes) as well as extensive theory of point process properties, we recommend the very readable volume (Daley and Vere-Jones (1988)).

Poisson Processes in Higher Dimensions

Point processes (especially Poisson) have also been traditionally very useful in modeling point events in space and space-time dimensions. The locations of ore deposits in two or three spatial dimensions and the occurrences of earthquakes in two dimensions and perhaps time ("spatio-temporal") are important examples. The mathematical framework extends naturally from one dimension, $N(B)$ being the number of point events in the two- or 3-dimensional (Borel) set B, and notational extensions such as $N_t(B)$ for the number of events in a spatial set B up to time t.

Not infrequently a time parameter is considered simply as equivalent to the addition of just one more spatial dimension, but the obvious differences in the questions to be asked for space and time suggest that the notation reflect the different character of the parameters. Further natural dependence structure (correlation assumptions, mixing conditions, long range dependence) may differ for spatial and time parameters. Further "*coordinatewise mixing*" (introduced in Leadbetter and Rootzen (1998)) seems promising in current investigation to facilitate point process theory in higher dimensions, where the parameters have different roles. A reader interested in the theory and applications in higher dimensions should consult (Daley and Vere-Jones 1988) and the wealth of references therein.

Poisson Processes on Abstract Spaces

There is substantial development of point process (and more general "random measure") theory in more abstract spaces, usually with an intricate topological structure (see Kallenberg (1986)). However for discussion of existence and useful basic modeling properties, the topological assumptions are typically solely used for definition of a natural simple measure-theoretic structure without any necessary underlying topology - though useful for deeper considerations such as weak convergence, beyond pure modeling. Further a charming property of Poisson processes in particular is that they may be defined simply on spaces with very little structure as we now indicate.

Specifically let S be a space, and $\mathcal{S}$ a σ-ring (here a σ-field for simplicity) of subsets of S. For a givenprobability

space $(\Omega, \mathcal{F}, P)$, a *random measure* is defined to be any family of non-negative- (possibly infinite) valued random variables $N_\omega(B)$ for each $B \in \mathcal{S}$ which is a measure (countably additive) on $\mathcal{S}$ for each $\omega \in \Omega$. A point process is a random measure for which each $N_\omega(B)$ is integer-valued (or $+\infty$). In this very general context one may construct a Poisson Process (see ►Poisson Processes) by simple steps (cf Kallenberg (1986) and Kingman (1993)) which we indicate. Define i.i.d. random elements $\tau_1, \tau_2, \ldots, \tau_n$ on S for each positive integer n with common distribution $\nu = P\tau_j^{-1}$ yielding a point process on S consisting of a finite number (n) of points. By regarding n as random having a Poisson distribution with mean $a > 0$ (or mixing the (joint) distributions of this point process with Poisson weights) one obtains a finite valued Poisson process $\{N(B)\}$ with the finite intensity measure $\mathcal{E}N(B) = a\nu(B)$. Finally if λ is a σ-finite measure on $\mathcal{S}$ we may write $\mathcal{S} = \bigcup_1^\infty S_i$ where $\lambda(S_i) < \infty$. Let $\{N_i(B), i = 1, 2, \ldots\}$ be point processes with the finite intensity measures $\mathcal{E}N_i(B) = \lambda(B \cap S_i)$. The superposition of these point processes gives a Poisson Process with intensity λ.

Relatives of the Poisson Process such as those above (Mixed, Compound, Doubly Stochastic...) may be constructed in a similar way to the one-dimensional framework. These sometimes require small or modest additional assumptions about the space S such as measurability of singleton sets, and the separation of two of its points by measurable sets. One may also obtain many general results analogous results to those of one dimension by assuming the existence of a *countable semiring* which covers S, and plays the role of bounded sets on which the point process is finite-valued, in this general non-topological context. Finally, as noted, the reference Kingman (1993) gives an account of Poisson processes primarily in this general framework, along with the basic early paper (Kendall 1974). In a topological setting the monograph (Kallenberg 1986) gives a comprehensive development of random measures, motivating our own non-topological approach.

About the Author

M.R. Leadbetter obtained his B.Sc. (1953), M.Sc. (1954), University of New Zealand, and Ph.D. (1963), UNC-Chapel Hill. He is a Fellow of American Statistical Association, Institute of Mathematical Statistics and member of the International Statistical Institute. He has published a book with Harald Cramér, *Stationary and Related Stochastic Processes* (Wiley, 1967). Professor Leadbetter was awarded an Honorary Doctorate, from Lund University. His name is associated with the Theorem of Cramér and Leadbetter.

Cross References

►Erlang's Formulas
►Lévy Processes
►Markov Processes
►Non-Uniform Random Variate Generations
►Point Processes
►Poisson Distribution and Its Application in Statistics
►Probability on Compact Lie Groups
►Queueing Theory
►Radon–Nikodým Theorem
►Renewal Processes
►Spatial Point Pattern
►Statistical Modelling in Market Research
►Stochastic Models of Transport Processes
►Stochastic Processes
►Stochastic Processes: Classification
►Univariate Discrete Distributions: An Overview

References and Further Reading

Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Methuen Monograph, London
Daley D, Vere-Jones D (1988) An introduction to the theory of point processes. Springer, New York
Durrett R (2005) Probability: theory and examples, 3rd edn. Duxbury, Belmont, CA
Kallenberg O (1986) Random measures, 4th edn. Academic, New York
Kendall DG (1974) Foundations of a theory of random sets. In: Kendall DG, Harding EF (eds) Stochastic geometry. Wiley, London
Kingman JFC (1993) Poisson processes. Clarendon, Oxford
Leadbetter MR, Rootzen H (1998) On extreme values in stationary random fields. In: Karatzas et al. (eds) Stochastic processes and related topics, volume in memory of Stamatis Cambanis, Birkhäuser
Ryll-Nardzewski C (1961) Remarks on processes of calls. Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability 2:455–465

Poisson Regression

Gerhard Tutz
Professor
Ludwig-Maximilians-Universität München, Germany

Introduction

The Poisson regression model is a standard model for count data where the response variable is given in the form of event counts such as the number of insurance claims within a given period of timeor the number of cases with a

specific disease in epidemiology. Let $(Y_i, \boldsymbol{x}_i)$ denote n independent observations, where $\boldsymbol{x}_i$ is a vector of explanatory variables and Y_i is the response variable. It is assumed that the response given $\boldsymbol{x}_i$ follows a Poisson distribution which has probability function

$$P(Y_i = r) = \begin{cases} \frac{\lambda_i^r}{r!} e^{-\lambda_i} & \text{for } r \in \{0, 1, 2, \dots\} \\ 0 & \text{otherwise.} \end{cases}$$

Mean and variance of the Poisson distribution are given by $E(Y_i) = var(Y_i) = \lambda_i$. Equality of the mean and variances is often referred to as the *equidispersion property* of the Poisson distribution. Thus, in contrast to the normal distribution, for which mean and variance are unlinked, the Poisson distribution implicitly models stronger variability for larger means, a property which is often found in real life data. The support of the Poisson distribution is $0, 1, 2, \dots$, which makes it an appropriate distribution model for count data.

A Poisson regression model assumes that the conditional mean $\mu_i = E(Y_i|\boldsymbol{x}_i)$ is determined by

$$\mu_i = h\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \quad \text{or equivalently} \quad g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where g is a known link function and $h = g^{-1}$ denotes the response function. Since the Poisson distribution is from the simple exponential family the model is a *generalized linear model* (GLM, see ▶Generalized Linear Models). The most widely used model uses the canonical link function by specifying

$$\mu_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \quad \text{or} \quad \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Since the logarithm of the conditional mean is linear in the parameters the model is called a *log-linear* model. The log-linear version of the model is particularly attractive because interpretation of parameters is very easy. The model implies that the conditional mean given $\boldsymbol{x}^T = (x_1, \dots, x_p)$ has a multiplicative form given by

$$\mu(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{\beta}) = e^{x_1 \beta_1} \dots e^{x_p \beta_p}.$$

Thus e^{β_j} represents the multiplicative effect on $\mu(\boldsymbol{x})$ if variable x_j changes by one unit to $x_j + 1$.

Inference

Since the model is a generalized linear model inference may be based on the methods that are available for that class of models (see for example McCullagh and Nelder 1989). One obtains for the derivative of the log-likelihood, which is the so-called score function

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{h'\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{h(\boldsymbol{x}_i \boldsymbol{\beta})} \left(y_i - h\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right),$$

and the Fisher matrix $F(\boldsymbol{\beta}) = E(-\partial h / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{h'(\boldsymbol{x}_i^T \boldsymbol{\beta})^2}{h(\boldsymbol{x}_i^T \boldsymbol{\beta})}$. Under regularity conditions, $\hat{\boldsymbol{\beta}}$ defined by $s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ is consistent and asymptotically normal distributed,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, F(\boldsymbol{\beta})^{-1}),$$

where $F(\boldsymbol{\beta})$ may be replaced by $F(\hat{\boldsymbol{\beta}})$ to obtain standard errors.

Goodness-of fit and tests on the significance of parameters based on deviance are provided within the framework of GLMs.

Extensions

In many applications count data are overdispersed, with conditional variance exceeding conditional mean. Several extensions of the basic model that account for overdispersion are available, in particular *quasi-likelihood methods* and more general distribution models like the Gamma–Poisson or *negative binomial model*. Quasi-likelihood uses the same estimation equations as maximum likelihood estimates, which are computed by solving

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial \mu_i}{\partial \eta} \frac{y_i - \mu_i}{v(\mu_i)} = \mathbf{0},$$

where $\mu_i = h(\eta_i)$ and $v(\mu_i)$ is the variance function. But instead of assuming the variance function of the Poisson model $v(\mu_i) = \mu_i$ one uses a more general form. For example, Poisson with overdispersion uses $v(\mu_i) = \phi \mu_i$ for some unknown constant ϕ. The case $\phi > 1$ represents *overdispersion* of the Poisson model, the case $\phi < 1$, which is rarely found in applications, is called *underdispersion*. Alternative variance functions usually continue to model the variance as a function of the mean. The variance function $v(\mu_i) = \mu_i + \gamma \mu_i^2$ with additional parameter γ corresponds to the variance of the negative binomial distribution.

It may be shown that the asymptotic properties of quasi-likelihood estimates are similar to that for GLMs. In particular one obtains asymptotically a normal distribution with the covariance given in the form of a pseudo-Fisher matrix, see McCullagh (1983), and McCullagh and Nelder (1989).

A source book for the modeling of count data which includes many applications is Cameron and Trivedi (1998). An econometric view on count data is outlined in Winkelmann (1997) and Kleiber and Zeileis (2008).

About the Author

Prof. Dr. Gerhard Tutz works at the Department of Statistics, Ludwig-Maximilians University Munich. He served as Head of the department for several years. He coauthored

the book *Multivariate Statistical Modeling Based on Generalized Linear Models* (with Ludwig Fahrmeir, Springer, 2001).

Cross References

►Dispersion Models
►Generalized Linear Models
►Geometric and Negative Binomial Distributions
►Modeling Count Data
►Poisson Distribution and Its Application in Statistics
►Robust Regression Estimation in Generalized Linear Models
►Statistical Methods in Epidemiology

References and Further Reading

Cameron AC, Trivedi PK (1998) Regression analysis of count data. econometric society monographs no. 30. Cambridge University Press, Cambridge
Kleiber C, Zeileis A (2008) Applied Econometrics with R. Springer, New York
McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York
Winkelmann R (1997) Count data models: econometric theory and application to labor mobility, 2nd edn. Springer, Berlin

Population Projections

Janez Malačič
Professor, Faculty of Economics
University of Ljubljana, Ljubljana, Slovenia

Population projections are a basic tool that demographers use to forecast a future population. They can be produced in the form of a *prognosis* or as *prospects*. The first is the most likely future development according to the expectations of the projections' author(s) and is produced in a single variant. The second type is based more on an "if-then" approach and is calculated in more variants. Usually, there are three or four variants, namely, low, medium, high, and constant variants. In practice, the medium variant is the most widely used and is taken as the most likely or accurate variant, that is, as a prognosis.

Population projections can be produced by *mathematical* or *analytical methods*. The *mathematical methods* use extrapolation(s) of various mathematical functions. For example, a census population can be extrapolated for a certain period into the future based on a linear, geometric, exponential, or some other functional form. The functional form is chosen on the basis of (1) past population development(s), (2) the developments of a neighboring and other similar populations, as well as (3) on the basis of general and particular demographic knowledge. In the great majority of cases, the mathematical methods are used for short- and midterm periods in the future. For population projections of small settlements and regions, only mathematical methods can be used in any reasonable way. Exceptionally, for very long-term periods (several centuries) a logistic curve can be used for a projection of the total population.

Analytical methods for population projections are much more complex. The population development in the projection period is decomposed at the level of basic components. These components are mortality, fertility, and migration. For each component, a special hypothesis of future development is produced. Very rich and complex statistical data are needed for analytical population projections. They are considered suitable for a period of 10–25 years into the future. They cannot be used to project populations in small areas because such life tables (see ►Life Table) and some other data will not be available. Analytical population projections offer very detailed information on particular population structures at present and in the future as well as data on the development of basic population components (e.g., mortality, fertility, and migration). They are also the basis for several other derived projections like those of households, of the active, rural, urban, and pensioned population.

Suppose that we have the census population divided by gender and age (in five-year age groups, say) as our starting point: ${}_{x+5}V^{t}_{m,x}$ and ${}_{x+5}V^{t}_{f,x}$, where V stands for population size, x for age, t for time, and m and f for male and female. To make a projection, we need three hypotheses. *The mortality hypothesis* is constructed on the basis of life table indicators. We take survival ratios ${}_{x+5}P_{m,x}$ and ${}_{x+5}P_{f,x}$ for all five-year age groups $(0-4, 5-9, \ldots, 80-84, 85+)$. Our hypothesis can be that mortality is constant, declining, or increasing, or we can have a combination of all three for each gender and each age group. Then we apply a population projection model aging procedure in the following form (spelled out for males):

$$ {}_{5}V^{t}_{m,0} * {}_{5}P_{m,0} = {}_{5}V^{t+5}_{m,5} \rightarrow {}_{5}V^{t+5}_{m,5} * {}_{5}P_{m,0} = {}_{5}V^{t+10}_{m,10}, \text{ etc.} $$

Evidently, in this case constant mortality hypothesis was used. The aging procedure would be applied for both genders and for all five-year age groups.

In the next step, we would construct a *fertility hypothesis*. This one can also be that fertility is constant, declining, increasing, or a combination of all three. It provides the newborn population for each year in the projection period. A set of different fertility indicators are available. The most convenient indicators are age-specific fertility rates, ${}_{x+5}f_{x}$,

where x is equal to 15, 20, 25, 30, 35, 40, and 45. In principle, we calculate the future number of births (N) for seven five-year age-groups with the formula:

$$ {}_5N_x^{t-(t+5)} = 5\left(\left({}_5V_{f,x}^{t} + {}_5V_{f,x}^{t+5}\right)/2 * {}_5f_x^{t}\right). $$

The number of births, $N^{t-(t+5)}$, should be subgrouped by gender. We can apply the demographic "constant" alternative and suppose that 485 girls are born per 1000 births. Then we calculate

$$ {}_5V_{m,0}^{t+5} = {}_5P_{m,r} * N_m^{t-(t+5)} \text{ and } {}_5V_{f,0}^{t+5} = {}_5P_{f,r} * N_f^{t-(t+5)} $$

${}_5P_m$, and ${}_5P_{f,r}$ are survival ratios for newborn boys and girls. In the case of a closed population or a population with zero migration, our projection is finished.

However, real populations have in- and out-migration. To cover this case, we should construct a *migration hypothesis*. The procedure is similar to the mortality and fertility hypotheses. Suppose we use net migration rates for five-year age groups, separately by gender. In principle, we calculate age-specific net migration factors (for males), nm is net migration:

$$ {}_5F_{m,x} = 1 + ({}_5nm_{m,x}/1,000). $$

The population aging procedure changes slightly:

$$ {}_5V_{m,x+5}^{t+5} = {}_5V_{m,x}^{t} * {}_5P_{m,x} * {}_5F_{m,x}. $$

The procedure for the female population is parallel to the procedure for the male population. The most serious problem in practice is that age-specific migration data may be unavailable or of poor quality.

Such simple analytical population projection procedures have been improved considerably in the literature during recent decades. Probably the most important improvement is the construction of *probabilistic population projections*, for which considerable progress has been made in recent decades (Lutz et al. 1998). Many analytical population projections for countries and regions are now supplemented by several national probabilistic population forecasts.

About the Author

Dr. Janez Malačić is a Professor of demography and labor economics, Faculty of Economics, Ljubljana University, Slovenia. He is a Former President of the Slovenian Statistical Society (1985–1987), a Former President of the Society of Yugoslav Statistical Societies (1986–1988), and a member of the IUSSP (from 1986). He has authored two books and more than 150 papers. His papers have been published in eight languages.

Cross References

►Actuarial Methods
►African Population Censuses
►Census
►Demographic Analysis: A Stochastic Approach
►Demography
►Life Table
►Survival Data

References and Further Reading

Eurostat, European Commission (2007) Work session on demographic projections. Bucharest, 10–12 October 2007. Methodologies and working papers. 370 p

Lutz W, Goldstein JR (guest eds) (2004) How to deal with uncertainty in population forecasting? Int Stat Rev 72(1–2):1–106, 157–208

Lutz W, Vaupel JW, Ahlburg DA (eds) (1999) Frontiers of population Forecasting. A Supplement to vol 24, 1998, population and Development Review. The Population Council, New York

Portfolio Theory

Harry M. Markowitz
Professor, Winner of the Nobel Memorial Prize in Economic Sciences in 1990
University of California, San Diego, CA, USA

Portfolio Theory considers the trade-off between some measure of risk and some measure of return on the portfolio-as-a-whole. The measures used most frequently in practice are expected (or mean) return and variance or, equivalently, standard deviation. This article discusses the justification for the use of mean and variance, sources of data needed in a mean-variance analysis, how mean-variance tradeoff curves are computed, and semi-variance as an alternative to variance.

Mean-Variance Analysis and its Justification

While the idea of trade-off curves goes back at least to Pareto, the notion of a trade-off curve between risk and return (later dubbed the efficient frontier) was introduced in Markowitz (1952). Markowitz proposed expected return and variance as both a hypothesis about how investors act and as a rule for guiding action in fact. By Markowitz (1959) he had given up the notion of mean and variance as a hypothesis but continued to propose them as criteria for action.

Tobin (1958) said that the use of mean and variance as criteria assumed either a quadratic utility function or a

Gaussian probability distribution. This view is sometimes ascribed to Markowitz, but he never justified the use of mean and variance in this way. His views evolved considerably from Markowitz (1952) to Markowitz (1959). Concerning these matters Markowitz (1952) should be ignored. Markowitz (1959) accepts the views of Von Neumann and Morgenstern (1944) when probability distributions are known, and L.J. Savage (1954) when probabilities are not known. The former asserts that one should maximize expected utility; the latter asserts that when probabilities are not known one should maximize expected utility using probability beliefs when objective probabilities are not known.

Markowitz (1959) conjectures that a suitably chosen point from the efficient frontier will approximately maximize expected utility for the kinds of utility functions that are commonly proposed for cautious investors, and for the kinds of probability distributions that are found in practice. Levy and Markowitz (1979) expand on this notion considerably. Specifically, Levy and Markowitz show that for such probability distributions and utility functions there is typically a correlation between the actual expected utility and the mean-variance approximation between 0.95 and of 0.99. They also show that the Pratt (1964) and Arrow (1971) objection to quadratic utility does not apply to the kind of approximations used by Levy and Markowitz, or in Markowitz (1959).

Models of Covariance

If covariances are computed from historical returns with more securities than there are observations, e.g., 5,000 securities and 60 months of observations, then the covariance matrix will be singular. A preferable alternative is to use a model of covariance where the return on the *i*th security is assumed to obey the following relationship

$$r_i = \alpha_i + \sum \beta_{ik} f_k + u_i$$

where the u_i are independent of each other and the f_k. The f_k may be either factors or scenarios or some of each. These ideas are carried out in, for example, Sharpe (1963), Rosenberg (1974) and Markowitz and Perold (1981a, 1981b).

Estimation of Parameters

Covariance matrices are sometimes estimated from historical returns and sometimes from factor or scenario models such as the one-factor model of Sharpe, the many-factor model of Rosenberg, or the scenario models of Markowitz and Perold cited above.

Expected returns are estimated in a great variety of ways. I do not believe that anyone suggests that, in practice, historical average returns should be used as the expected returns of individual stocks. The Ibbotson and Sinquefield (2007) series are frequently used to estimate expected returns for asset classes. Black and Litterman (1991, 1992) propose a very interesting Bayesian approach to the estimation of expected returns. Richard Michaud (1989) proposes to use estimates for asset classes based on what he refers to as a "resampled frontier". Additional methods for estimating expected return are based on statistical methods for "disentangling" various anomalies, see Jacobs and Levy (1988), or estimates based on factors that Graham and Dodd (1940) might use: see Lakonishok et al. (1994), Ohlson (1979), and Bloch et al. (1993). The last mentioned paper is based on results obtained by back-testing many alternate hypotheses concerning how to achieve excess returns. When many estimation methods are tested, the expected future return for the best of the lot should not be estimated as if this were the only procedure tested. Estimates should be corrected for "data mining." See Markowitz and Xu (1994).

Computation of M-V Efficient Sets

The set of mean-variance efficient portfolios is piecewise linear. The critical line algorithm (CLA) traces out this set, one linear piece at a time, without having to search for optima. CLA is described in Appendix A of Markowitz (1959) and, less compactly, in Markowitz and Todd (2000).

Downside Risk

"Semi-variance" or downside risk is like variance, but only considers deviations below the mean or below some target return. It is proposed by Markowitz (1959) Chap. 9 and championed by Sortino and Satchell (2001). It is used less frequently in practice than variance.

About the Author

Professor Markowitz has applied computer and mathematical techniques to various practical decision making areas. In finance: in an article in 1952 and a book in 1959 he presented what is now referred to as MPT, "modern portfolio theory." This has become a standard topic in college courses and texts on investments, and is widely used by institutional investors for asset allocation, risk control and attribution analysis. In other areas: Dr. Markowitz developed "sparse matrix" techniques for solving very large mathematical optimization problems. These techniques are now standard in production software for optimization programs. Dr. Markowitz also designed and supervised the development of the SIMSCRIPT programming language. SIMSCRIPT has been widely used for programming computer simulations of systems like factories, transportation systems and communication networks.

In 1989 Dr. Markowitz received The John von Neumann Award from the Operations Research Society of America for his work in portfolio theory, sparse matrix techniques and SIMSCRIPT. In 1990 he shared The Nobel Prize in Economics for his work on portfolio theory.

Cross References

►Actuarial Methods
►Business Statistics
►Copulas in Finance
►Heteroscedastic Time Series
►Optimal Statistical Inference in Financial Engineering
►Semi-Variance in Finance
►Standard Deviation
►Statistical Modeling of Financial Markets
►Variance

References and Further Reading

Arrow K (1971) Aspects of the theory of risk bearing. Markham Publishing Company, Chicago

Black F, Litterman R (1991) Asset allocation: combining investor views with market equilibrium. J Fixed Income 1(2):7–18

Black F, Litterman R (1992) Global portfolio optimization. Financ Anal J 48(5):28–43

Bloch M, Guerard J, Markowitz H, Todd P, Xu G (1993) A comparison of some aspects of the US and Japanese equity markets. Jpn World Econ 5:3–26

Graham B, Dodd DL (1940) Security analysis, 2nd edn. McGraw-Hill, New York

Ibbotson RG, Sinquefield RA (2007) Stocks, bonds, bills and inflation yearbook. Morningstar, New York

Jacobs BI, Levy KN (1988) Disentangling equity return regularities: new insights and investment opportunities. Financ Anal J 44(3):18–44

Lakonishok J, Shleifer A, Vishny RW (1994) Contrarian investment, extrapolation and risk. J Financ 49(5):1541–1578

Levy H, Markowitz HM (1979) Approximating expected utility by a function of mean and variance. Am Econ Rev 69(3):308–317

Markowitz HM (1952) Portfolio selection. J Financ 7(1):77–91

Markowitz HM (1959) Portfolio selection: efficient diversification of investments. Wiley, New York (Yale University Press, 1970, 2nd edn, Basil Blackwell, 1991)

Markowitz HM, Perold AF (1981a) Portfolio analysis with factors and scenarios. J Financ 36(14):871–877

Markowitz HM, Perold AF (1981b) Sparsity and piecewise linearity in large portfolio optimization problems. In: Duff IS (ed) Sparse matrices and their uses. Academic Press, New York, pp 89–108

Markowitz HM, Todd P (2000) Mean-variance analysis in portfolio choice and capital markets. Frank J. Fabozzi Associates, New Hope [revised reissue of Markowitz (1987) (with chapter by Peter Todd)]

Markowitz HM, Xu GL (1994) Data mining corrections. J Portfolio Manage 21:60–69

Michaud RO (1989) The Markowitz optimization enigma: is optimized optimal? Financ Anal J 45(1):31–42

Ohlson JA (1979) Risk return, security-valuation and the stochastic behavior of accounting numbers. J Financ Quant Anal 14(2):317–336

Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136

Rosenberg B (1974) Extra-market components of covariance in security returns. J Financ Quant Anal 9(2):263–273

Savage LJ (1954) The foundations of statistics. Wiley, New York (Second revised edn. Dover, New York)

Sharpe WF (1963) A simplified model for portfolio analysis. Manage Sci 9(2):277–293

Sortino F, Satchell S (2001) Managing downside risk in financial markets: theory, practice and implementation. Butterworth-Heinemann, Burlington

Tobin J (1958) Liquidity preference as behavior towards risk. Rev Econ Stud 25(1):65–86

Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. 3rd edn. (1953), Princeton University Press, Princeton

Posterior Consistency in Bayesian Nonparametrics

Jayanta K. Ghosh[1], R. V. Ramamoorthi[2]
[1]Professor of Statistics
Purdue University, West Lafayette, IN, USA
[2]Professor
Michigan State University, East Lansing, MI, USA

Bayesian Nonparametrics (see ►Bayesian Nonparametric Statistics) took off with two papers of Ferguson (Ferguson 1974, 1983) and followed by Antoniak (Antoniak 1974). However consistency or asymptotics were not major issue in those papers, which were more concerned with taking the first steps towards a usable, easy to interpret prior with easy to choose hyperparameters and a rich support. Unfortunately, the fact that the Dirichlet sits on discrete distributions diminished the early enthusiasm.

The idea of consistency came from Laplace and informally may be defined as : Let $\mathcal{P}$ be a set of probability measures on a sample space $\mathcal{X}$, Π be a prior on $\mathcal{P}$. The posterior is said to be consistent at a true value P_0 if the following holds: For sample sequences with P_0 probability 1, the posterior probability of any neighborhood U of P_0 converges to 1.

The choice of neighborhoods U determines the strength of consistency. One choice, when the sample space is separable metric, is weak neighborhoods U of P_0. When elements of $\mathcal{P}$ have densities, L_1 neighborhoods of P_0 is often the relevant choice. If the family $\mathcal{P}$ is parametrized by θ, then these notions easily translate to θ, via continuity requirements of the map $\theta \mapsto P_\theta$.

The early papers on consistency were by Freedman (Freedman 1963, 1965) on multinomials with (countably) infinite classes. They were very interesting but provided a

somewhat negative picture, namely, that in a topological sense, for most priors (i.e., outside a class of first category) consistency fails to hold. This lead Freedman to consider tail-free priors including the Dirichlet prior for the countably many probabilities of the countably infinite multinomial. Around the same time, in her posthumous paper (Schwartz 1965) of 1965, arising from her thesis at Berkeley, written under Le Cam, Schwartz showed among other things that, if the prior assigned positive probability to all Kullback-Leibler neighborhoods of the true density, then consistency holds in the sense of weak convergence. This is an important result which showed that this is the right notion of support in these problems, not the more usual one adopted by Freedman.

These early papers were followed by Ferguson's Dirichlet process on the set of all probability measures along with Antoniak's study of mixtures. Even though the set of discrete measures had full measure under the Dirichlet process, it still enjoyed the property of consistency at all distributions. The paper by Diaconis and Freedman (Diaconis and Freedman 1986), along with the discussions revived interest in consistency issues. Diaconis and Freedman showed that with Dirichlet process prior consistency can go awry in the presence of a location parameter. Barron, in his discussion of the paper provided insight as to why it would be unreasonable to expect consistency in this example. Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) discuss several different explanation for lack of consistency if one uses the Dirichlet Process in a semiparametric problem with a location parameter.

Other major contributions have been made by Barron (Barron 1988), Barron, Schervish and Wasserman (Barron et al. 1999), Walker (Walker 2004) and Coram and Lalley (Coram and Lalley 2006). A thorough review up to 2008 is available in Choi and Ramamoorthi (Choi and Ramamoorthi 2008). Contributions to rates of convergence have been made by Ghosal, Ghosh and van der Vaart, (Ghosal et al. 2000), Ghosal and van der Vaart (Ghosal and van der Vaart 2001), Shen and Waseerman (2001), Kleijn and van der Vaart (Kleijn and van der Vaart 2006) and (van der Vaart and van Zanten 2009), (van der Vaart and van Zanten 2008). see also the book by Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) for many basic early results on consistency and other aspects of, like choice of priors and consistency for density estimation, semiparametric problems and survival analysis.

Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) deal only with nonparametric estimation problems. Work on nonparametric testing and consistency problems there have begun only recently. A survey is available in Tokdar, Chakravarti and Ghosh (Tokdar et al. 2010).

About the Authors

For biography of Professor Ghosh *see* the entry ►Bayesian P-Values.

Professor R.V. Ramamoorthi obtained his doctoral degree from the Indian Statistical Institute. His early work was on sufficiency and decision theory, a notable contribution there being the joint result with Blackwell showing that Bayes sufficiency is not equivalent to Fisherian sufficiency. Over the last few years his research has been on Bayesian nonparamterics where J.K. Ghosh and he authored one of the first books on the topic, *Bayesian Nonparametrics* (Springer 2003).

"This is the first book to present an exhaustive and comprehensive treatment of Bayesian nonparametrics. Ghosh and Ramamoorthi present the theoretical underpinnings of nonparametric priors in a rigorous yet extremely lucid style…It is indispensable to any serious Bayesian. It is bound to become a classic in Bayesian nonparametrics." (Jayaram Sethuraman, Review Of Bayesian Nonparametrics, *Sankhya*, 2004, 66, 208–209).

Cross References

►Bayesian Nonparametric Statistics
►Bayesian Statistics
►Nonparametric Statistical Inference

References and Further Reading

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174
Barron A (1988) The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Department Statistics, University of Illinois, Champaign
Barron A, Schervish MJ, Wasserman L (1999) The consistency of posterior distributions in nonparametric problems. Ann Stat 27:536–561
Choi T, Ramamoorthi RV (2008) Remarks on consistency of posterior distributions. In: Pushing the limits of contemporary statistics: contributions in honor of Ghosh JK, vol 3 of Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood, pp 170–186
Coram M, Lalley SP (2006) Consistency of Bayes estimators of a binary regression function. Ann Stat 34:1233–1269
Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. Ann Stat 14:1–67. With a discussion and a rejoinder by the authors
Ferguson TS (1974) Prior distributions on spaces of probability measures. Ann Stat 2:615–629
Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In: Recent advances in statistics. Academic, New York, pp 287–302
Freedman DA (1963) On the asymptotic behavior of Bayes' estimates in the discrete case. Ann Math Stat 34:1386–1403
Freedman DA (1965) On the asymptotic behavior of Bayes estimates in the discrete case II. Ann Math Stat 36:454–456
Ghosal S, Ghosh JK, van der Vaart AW (2000) Convergence rates of posterior distributions. Ann Stat 28:500–531

Ghosal S, van der Vaart AW (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Stat 29:1233–1263
Ghosh JK, Ramamoorthi RV (2003) Bayesian nonparametrics. Springer Series in Statistics. Springer, New York
Kleijn BJK, van der Vaart AW (2006) Misspecification in infinite-dimensional Bayesian statistics. Ann Stat 34:837–877
Schwartz L (1965) On Bayes procedures. Z Wahrscheinlichkeitstheorie und Verw Gebiete 4:10–26
Shen X, Wasserman L (2001) Rates of convergence of posterior distributions. Ann Stat 29(3):687–714
Tokdar S, Chakrabarti A, Ghosh J (2010) Bayesian nonparametric goodness of fit tests. In: M-H Chen, DK Dey, P Mueller, D Sun, K Ye (eds) Frontiers of statistical decision making and Bayesian analysis. Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood
van der Vaart AW, van Zanten JH (2008) Reproducing kernel Hilbert spaces of Gaussian priors. In: Pushing the limits of contemporary statistics: contributions in honor of Ghosh JK, vol 3 of Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood, pp 200–222
van der Vaart AW, van Zanten JH (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. Ann Stat 37:2655–2675
Walker S (2004) New approaches to Bayesian consistency. Ann Stat 32:2028–2043

Power Analysis

Kevin R. Murphy
Professor
Pennsylvania State University, University Park, PA, USA

One of the most common applications of statistics in the social and behavioral science is in testing null hypotheses. For example, a researcher wanting to compare the outcomes of two treatments will usually do so by testing the hypothesis that in the population there is no difference in the outcomes of the two treatments. The power of a statistical test is defined as the likelihood that a researcher will be able to reject a specific null hypothesis when it is in fact false.

Cohen (1988), Lipsey (1990), and Kraemer and Thiemann (1987) provided excellent overviews of the methods, assumptions, and applications of power analysis. Murphy and Myors (2003) extended traditional methods of power analysis to tests of hypotheses about the size of treatment effects, not merely tests of whether or not such treatment effects exist.

The power of a null hypothesis test is a function of sample size (n), effect size (ES), and the standard used to define statistical significance (α), and the equations that define this relation can be easily rearranged to solve for any of four quantities (i.e., power, n, ES, and α), given the other three. The two most common applications of statistical power analysis are in: (1) determining the power of a study, given n, ES, and α, and (2) determining how many observations will be needed (i.e., n required), given a desired level of power, an ES estimate, and the α value. Both of these methods are widely used in designing studies; one widely-accepted convention is that studies should be designed so that they achieve power levels of 0.80 or greater (i.e., so that they have at least an 80% chance of rejecting a false null hypothesis; Cohen 1988; Murphy and Myors 2003).

There are two other applications of power analysis that are less common, but no less informative. First, power analysis may be used to evaluate the sensitivity of studies. That is, power analysis can indicate what sorts of effect sizes might be reliably detected in a study. If one expects the effect of a treatment to be small, it is important to know whether the study will detect that effect, or whether the study as planned only has sufficient sensitivity to detect larger effects. Second, one may use power analysis to make rational decisions about the criteria used to define "statistical significance."

Power analyses are included as part of several statistical analysis packages (e.g., SPSS provides Sample Power, a flexible and powerful program) and it is possible to use numerous websites to perform simple power analyses. Two notable software packages designed for power analysis are:

- *G*Power* (Faul et al. 2007; http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/ is distributed as a freeware program that is available for both Macintosh and Windows environments. It is simple, fast, and flexible.
- *Power and Precision*, distributed by Biostat, was developed by leading researchers in the field (e.g., J. Cohen). This program is very flexible, covers a large array of statistical tests, and provides power analyses and confidence intervals for most tests.

About the Author

Kevin Murphy is a Professor of Psychology and Information Sciences and Technology at the Pennsylvania State University. He has served as President of the Society for Industrial and Organizational Psychology and Editor of *Journal of Applied Psychology*. He is the author if eleven books, including *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests* (with Brett Myors, Erlbaum 2009), and over 150 articles and chapters.

Cross References

►Effect Size
►Presentation of Statistical Testimony
►Psychology, Statistics in
►Sample Size Determination
►Significance Testing: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Significance

References and Further Reading

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale

Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Meth 39:175–191

Kraemer HC, Thiemann S (1987) How many subjects? Sage, Newbury Park

Lipsey MW (1990) Design sensitivity. Sage, Newbury Park

Murphy K, Myors B (2009) Statistical power analysis: a simple and general model for traditional and modern hypothesis tests, 3rd edn. Erlbaum, Mahwah

Preprocessing in Data Mining

Edgar Acuña
Professor
University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico

Introduction

►Data mining is the process of extracting hidden patterns in a large dataset. Azzopardi (2002) breaks the data mining process into five stages:

(a) *Selecting the domain* – data mining should be assessed to determine whether there is a viable solution to the problem at hand and a set of objectives should be defined to characterize these problems.
(b) *Selecting the target data* – this entails the selection of data that is to be used in the specified domain; for example, selection of subsets of features or data samples from larger databases.
(c) *Preprocessing the data* – this phase is primarily aimed at preparing the data in a suitable and useable format, so that a knowledge extraction process can be applied.
(d) *Extracting the knowledge/information* – during this stage, the types of data mining operations (association rules, regression, supervised classification, clustering, etc.), the data mining techniques, and data mining algorithms are chosen and the data is then mined.
(e) *Interpretation and evaluation* – this stage of the data mining process is the interpretation and evaluation of the discoveries made. It includes filtering information that is to be presented, visualizing graphically, or locating the useful patterns and translating the patterns discovered into an understandable form.

In the process of data mining, many patterns are found in the data. Patterns that are interesting for the miner are those that are easily understood, valid, potentially useful, and novel (Fayyad et al. 1996). These patterns should validate the hypothesis that the user seeks to confirm. The quality of patterns obtained depends on the quality of the data analyzed. It is common practice to prepare data before applying traditional data mining techniques such as regression, association rules, clustering, and supervised classification.

Section "►Reasons for Applying Data Preprocessing" of this article provides a more precise justification for the use of data preprocessing techniques. This is followed by a description in section "►Techniques for Data Preprocessing" of some of the data preprocessing techniques currently in use.

Reasons for Applying Data Preprocessing

Pyle (1999) suggests that about 60% of the total time required to complete a data mining project should be spent on data preparation since it is one of the most important contributors to the success of the project. Transforming the data at hand into a format appropriate for knowledge extraction has a significant influence on the final models generated, as well as on the amount and quality of the knowledge discovered during the process. At the same time, the effect caused by changes made to a dataset during data preprocessing can either facilitate or complicate even further the knowledge discovery process; thus changes made must be selected with care.

Today's real-world datasets are highly susceptible to noise, missing and inconsistent data due to human errors, mechanical failures, and to their typically large size. Data affected in this manner is known as "dirty." During the past decades, a number of techniques have been developed to preprocess data gathered from real-world applications before the data is further processed for other purposes.

Cases where data mining techniques are applied directly to raw data without any kind of data preprocessing are still frequent; yet, data preprocessing has been recommended as an obligatory step. Data preprocessing techniques should never be applied blindly to a dataset, however. Prior to any data preprocessing effort, the dataset should be explored and characterized. Two methods for

exploring the data prior to preprocessing are *data characterization* and *data visualization*.

Data Characterization

Data characterization describes data in ways that are useful to the miner and begins the process of understanding what is in the data. Engels and Theusinger (1998) describe the following characteristics as standard for a given dataset: the number of classes, the number of observations, the percentage of missing values in each attribute, the number of attributes, the number of features with numeric data type, and the number of features with symbolic data type. These characteristics can provide a first indication of the complexity of the problem being studied.

In addition to the above-mentioned characteristics, parameters of location and dispersion can be calculated as single-dimensional measurements that describe the dataset. Location parameters are measurements such as minimum, maximum, arithmetic mean, median, and empirical quartiles. On the other hand, dispersion parameters such as range, standard deviation, and quartile deviation provide measurements that indicate the dispersion of values of the feature.

Location and dispersion parameters can be divided in two classes: those that can deal with extreme values and those that are sensitive to them. A parameter that can deal well with extreme values is called robust. Some statistical software packages provide the computation of robust parameters in addition to the traditional non-robust parameters. Comparing robust and non-robust parameter values can provide insight to the existence of ►outliers during the data characterization phase.

Data Visualization

Visualization techniques can also be of assistance during this exploration and characterization phase. Visualizing the data before preprocessing it can improve the understanding of the data, thereby increasing the likelihood that new and useful information will be gained from the data. Visualization techniques can be used to identify the existence of missing values, and outliers, as well as to identify relationships among attributes. These techniques can, in effect, assist in ranking the "impurity" of the data and in selecting the most appropriate data preprocessing technique to apply.

Techniques for Data Preprocessing

Applying the correct data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Lu et al. (1996), Pyle (1999), and Azzopardi (2002) present descriptions of common techniques for preparing data for analysis. The techniques described by both authors can be summarized as follows:

(a) *Data cleaning* – filling in missing values, smoothing noisy data, removing outliers, and resolving inconsistencies.
(b) *Data reduction* – reducing the volume of data (but preserving the patterns) by removing repeated observations and applying *instance selection* as well as *feature selection* techniques. Discretization of continuous attributes is also a way of data reduction.
(c) *Data transformation* – converting text and graphical data to a format that can be processed, normalizing or scaling the data, aggregation, and generalization.
(d) *Data integration* – correcting differences in coding schemes due to the combining of several sources of data.

Data Cleaning

Data cleaning provides methods to deal with dirty data. Since dirty datasets can cause problems for data exploration and analysis, data cleaning techniques have been developed to clean data by filling in missing values (value imputation), smoothing noisy data, identifying and/or removing outliers, and resolving inconsistencies. Noise is a random error or variability in a measured feature, and several methods can be applied to remove it. Data can also be smoothed by using regression to find a mathematical equation to fit the data. Smoothing methods that involve *discretization* are also methods of data reduction since they reduce the number of distinct values per attribute. Clustering methods can also be used to remove noise by detecting outliers.

Data Integration

Some studies require the integration of multiple databases, or files. This process is known as *data integration*. Since attributes representing a given concept may have different names in different databases, care must be taken to avoid causing inconsistencies and redundancies in the data. Inconsistencies are observations that have the same values for each of the attributes but that are assigned to different classes. Redundant observations are observations that contain the same information.

Attributes that have been derived or inferred from others may create redundancy problems. Again, having a large amount of redundant and inconsistent data may slow down the knowledge discovery process for a given dataset.

Data Transformation

Many data mining algorithms provide better results if the data has been normalized or scaled to a specific range before these algorithms are applied. The use of normalization techniques is crucial when distance-based algorithms are applied, because the distance measurements taken on by attributes that assume many values will generally outweigh distance measurements taken by attributes that assume fewer values. Other methods of *data transformation* include data aggregation and generalization techniques. These methods create new attributes from existing information by applying summary operations to data or by replacing raw data by higher-level concepts. For example, monthly sales data may be aggregated to compute annual sales.

Data Reduction

The increased size of current real-world datasets has led to the development of techniques that can reduce the size of the dataset without jeopardizing the data mining results. The process known as *data reduction* obtains a reduced representation of the dataset that is much smaller in volume, yet maintains the integrity of the original data. This means that data mining on the reduced dataset should be more efficient yet produce similar analytical results. Han and Kamber (2006) mention the following strategies for data reduction:

(a) *Dimension reduction*, where algorithms are applied to remove irrelevant, weakly relevant, or redundant attributes.
(b) *Data compression*, where encoding mechanisms are used to obtain a reduced or compressed representation of the original data. Two common types of data compression are wavelet transforms and ►principal component analysis.
(c) *Numerosity reduction*, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which store only the model parameters instead of the actual data), or nonparametric methods such as clustering and the use of histograms.
(d) *Discretization and concept hierarchy generation*, where raw data values for attributes are replaced by ranges or higher conceptual levels. For example, concept hierarchies can be used to replace a low-level concept such as age, with a higher-level concept such as young, middle-aged, or senior. Some detail may be lost by such data generalizations.
(e) *Instance selection*, where a subset of best instances of the whole dataset is selected. Some of the instances are more relevant than others to perform a data mining, and working only with an optimal subset of instances, it will be more cost-and time-efficient. Variants of the classical sampling techniques can be used.

Final Remarks

Acuna (2009) has developed Drep, an R package for data preprocessing and visualization. Drep performs most of the data preprocessing techniques mentioned in this article. Currently, research is being done in order to apply preprocessing methods to data streams, see Aggarwal (2007) for more details.

About the Author

Dr. Edgar Acuña, is a Professor, Department of Mathematical Sciences, University of Puerto Rico at Mayaguez. He is also the leader of the group in Computational and statistical Learning from databases at the University of Puerto Rico. He has authored and co-authored more than 20 papers mainly on data preprocessing. He is the author of book (in Spanish) *Analisis Estadistico De Datos usando* Minitab (John Wiley & Sons, 2002). In 2003, he was honored with the Power Hitter Award in Business and Technology. In 2008, he was selected as a Fullbright visiting Scholar. Currently, he is an Associate editor for the *Revista Colombiana de Estadistica*.

Cross References

►Box–Cox Transformation
►Data Mining
►Multi-Party Inference and Uncongeniality
►Outliers

References and Further Reading

Acuna E (2009) Dprep: data preprocessing and visualization functions for classification. URL http://cran.r-hproject.org/package=dprep. R package version 2.1
Aggarwal CC (ed) (2007) Data streams: models and algorithms. Springer, New York
Azzopardi L (2002) "Am I Right?" asked the classifier: preprocessing data in the classification process. Comput Inform Syst 9:37–44
Engels R, Theusinger C (1998) Using a data metric for preprocessing advice for data mining applications. In: Proceedings of 13th European conference on artificial intelligence, pp 430–434
Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining, Chapter 1, AAAI Press/MIT Press, pp 1–34
Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. Morgan Kaufman Publishers
Lu H, Sun S, Lu Y (1996) On preprocessing data for effective classification. ACM SIGMOD'96 workshop on research issues on data mining and knowledge discovery, Montreal, QC
Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, San Francisco

Presentation of Statistical Testimony

Joseph L. Gastwirth
Professor of Statistics and Economics
George Washington University, Washington, DC, USA

Introduction

Unlike scientific research, where we have the luxury of carrying out new studies to replicate and determine the domain of validity of prior investigations, the law also considers other social goals. For example, in many nations one spouse cannot be forced to testify against the other. A main purpose of the law is to resolve a dispute, so a decision needs to be made within a reasonable amount of time after the charge is filed. Science is primarily concerned with determining the true mechanism underlying a phenomenon. Typically no limits are placed on the nature of the experiment or approach an investigator may take to a problem. In most uses of statistics in legal proceedings, the relevant events happened several years ago; rarely will you be able to collect additional data. (One exception occurs in cases concerned with violations of laws protecting intellectual property, such as the Lanham Act in the USA, which prohibits a firm from making a product that infringes on an established one. Surveys of potential consumers are conducted to estimate the percentage who might be confused as to the source of the product in question.) Often, the data base will be one developed for administrative purposes, e.g., payroll or attendance records, which you will need to rely on.

Civil cases differ from criminal cases in that the penalty for violating a civil statue is not time in prison but rather compensation for the harm done. Consequently, the burden of proof a plaintiff has in discrimination or tort case is to show that the defendant caused the alleged harm by "the preponderance of the evidence" rather than the stricter standard of "beyond a reasonable doubt" used in criminal cases. Thus, many statistical studies are more useful in civil cases.

A major difference between presenting testimony in court or a regulatory hearing and giving a talk at a major scientific conference is that expert witnesses, like all others, are only allowed to answer questions put to them by the lawyers. Thus, particular findings or analyses that you believe are very important may not be submitted as evidence if the lawyer who hired you does not ask you about them when you are testifying. Unless the judge, or in some jurisdictions a juror, asks you a question related to that topic you are not allowed to discuss it.

Courts have also adopted criteria to assess the reliability of scientific evidence as well as some traditional ways of presenting and analyzing some types of data. (The leading case is *Daubert v. Merrell-Dow Pharmaceuticals Inc.*, 509 U.S. 579 (1993). The impact of this case and two subsequent ones on scientific evidence is described by Berger (2000) and Rosenblum (2000). Some of the criteria courts consider are: can the methodology was subject to peer review, can it be replicated and tested and whether the potential error rates are known and considered in the expert's report and testimony.) This may limit the range of analyses you can use in the case at hand; although subsequently it can stimulate interesting statistical research. A potentially more serious threat to the credibility of your research and subsequent testimony case is due to the fact that the lawyers provide you with the data and background information. Thus, you may not even be informed that other information or data sets exist.

A related complication can arise when the lawyer hires both a consulting expert who has complete access to all the data as he or she is protected by the "work product" rule and then hires a "testifying expert". This second expert may only be asked to analyze the data favorable to the defendant and not told that any other data exists. Sometimes, the analytic approach, e.g., regression analysis, may be suggested to this expert because the lawyer already knows the outcome. If one believes an alternative statistical technique would be preferable or at least deserves exploration, the expert may be constrained as to the choice of methodology.

This entry describes examples of actual uses of statistical evidence, along with suggestions to aid courts in understanding the implications of the data. Section "►Presenting the Data or Summary Tables That will be Analyzed" discusses the presentation of data and the results of statistical tests. One dataset illustrates the difficulty of getting lawyers and judges to appreciate the statistical concept of "power", the ability of a test to detect a real or important difference. As a consequence an analysis that used a test with *no* power was accepted by a court. In section "►A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))" will show how in a subsequent case I was able to present more detailed data, which helped make the data clearer to the court. The last section offers some suggestions for improving the quality of statistical analyses and their presentation.

Presenting the Data or Summary Tables That will be Analyzed

In the classic *Castenada v. Partida* (430 U.S. 482, 97 S. Ct. 1272 (1997)) case concerning whether Mexican-Americans

were discriminated against in the jury selection process, the Court summarized the data by years, i.e., the data for *all* juries during the year were aggregated and the minority fraction compared to their fraction of the total population as well as the subgroup eligible for jury service. (The data is reported in footnote 7 of the opinion as well as in the texts: Finkelstein and Levin (2000) and Gastwirth (1988).) The data showed a highly significant difference between the Mexican-American fraction of jurors (39%) and both their fraction of the total population (79.1%) and of adults with some schooling (65%). From a statistical view the case is important as it established that formal statistical hypothesis testing would be used rather than intuitive judgments about whether the difference between the percentage of jurors who were from the minority group differed sufficiently from the percentage of minorities eligible for jury service. When the lower courts followed the methodology laid out in *Castenada*, they also adopted the tradition of presenting yearly summaries of the data in discrimination cases.

Unlike jury discrimination cases, which are typically brought by a defendant in a criminal case rather than the minority juror who was dismissed, in equal employment cases the plaintiff is the individual who suffered the alleged discriminatory act. In the United States the plaintiff has 180 days from the time of the alleged act, e.g., not being hired or promoted or of being laid off, to file a formal complaint with the Equal Employment Opportunity Commission (EEOC). Quite often after receiving notice of the complaint, the employer will modify their system to mitigate the effect of the employment practice under scrutiny on the minority group in question. The impact of this "change" in policy on statistical analysis has often been overlooked by courts. In particular, if a change occurs during the year the charge occurs and employer may change their policy and include the post-charge minority hires or promotions in their analysis.

Let me use data from a case, *Capaci v. Katz & Besthoff* (525 F. Supp. 317 (E.D. La. 1981), *aff'd in part, rev'd in part*, 711 F.2d 647 (5th Cir. 1983)), in which I was an expert for the plaintiffs to illustrate this. On January, 11, 1973 the plaintiff filed a charge of discrimination against women in promotions in the pharmacy department. One way such discriminatory practices may be carried out is to require female employees to work longer at the firm before promotion than males. Therefore, a study comparing the length of time males and female Pharmacists served before they were promoted to Chief Pharmacist was carried out. The time frame considered started in July 1, 1965 the effective date of the Civil Rights Act until the date of the charge. The times each Pharmacist who was promoted had served until they received their promotion are reported in Table 1. Only their initials of the employees are given.

Presentation of Statistical Testimony. Table 1 Months of service for male and female pharmacists employed at K&B during the period July 1, 1965 thru January 11, 1973 before receiving a promotion to chief pharmacist

Females: 229; 453.
Males: 5; 7; 12; 14; 14; 14; 18; 21; 22, 23; 24; 25; 25; 34; 34; 37; 47; 49; 64; 67; 69; 125; 192; 483.

Applying the Wilcoxon test (see ▶Wilcoxon–Mann–Whitney Test), incorporating ties yielded a statistically significant difference (p-value = 0.02). The average number of months the two females worked until they were promoted was 341, while the average male worked for 59 months before their promotion. The corresponding medians were 341 and 25 months, respectively. The defendant's expert presented the data, given in Table 1, broken out into two time periods ending at the end of a year. The first was from 1965 until the end of 1973 and the second was from 1974 until 1978. The defendant's data includes three more females and eight more males in the defendant's data because their expert included essentially the first year's data *subsequent* to the charge. Furthermore, the three females fell into the seniority categories of 20–29, 30–39 and 70–79 months, i.e., they had much less seniority than the two females who were promoted *prior* to the complaint

The defendant's expert did not utilize the Wilcoxon test; rather he analyzed all the data sets with the *median* test and found no significant difference differences. In contrast with the Wilcoxon analysis of the data in Table 1, the median test did not find the difference in time to promotion data in the pre-charge period to be statistically significant.

Because only two females were promoted from July 1, 1965 until the charge was filed in January 1973, the median test has *zero* power of detecting a difference between the two samples. Thus, I suggested that the plaintiffs' lawyer ask the following series of questions to the defendant's expert on cross exam:

1. What is the difference between the average time to promotion of the male and female pharmacists in Table 1.

   Expected Answer: about 20 years. The actual difference was 23 years as the mean female took 341 months while the mean male took 59 months to be promoted.
2. Suppose the difference in the two means or averages was *50* years, would the median test have found a

statistically significant difference between the times that females had to work before being promoted than males?

Expected Answer: No.

3. Suppose the difference in the two means or averages was *100* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

4. Suppose the difference in the two means or averages was *1,000* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

5. Suppose the difference in the two means or averages was *one million* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

   My thought was that the above sequence of questions would have shown the judge the practical implication of finding a non-significant result with a test that did not have any power, in the statistical sense. Unfortunately, after the lawyer asked the first question, she jumped to the last one. By doing so, the issue was not made clear to the judge. When I asked why, I was told that she felt that the other expert realized the point. Of course, the questions were designed to explain the practical meaning of statistical "power" to the trial judge, not the expert. A while later while describing the trial to another, more experienced lawyer, he told me that after receiving the No answers to the five questions he would have turned to the expert and asked him:

6. As your statistical test could not detect a difference of a million years between the times to promotion of male and female employees, just how long would my client and other females have to work without receiving a promotion before your test would find a statistically significant difference?

This experience motivated me to look further into the power properties of nonparametric tests, especially in the unbalanced sample size setting (Gastwirth and Wang 1987; Freidlin and Gastwirth 2000a). The data from the *Capaci* case is discussed by Finkelstein and Levin (2000, p. 344) and Gastwirth (1988, p. 312) and the need to be cautious when a test with low power accepts the null hypothesis is emphasized by Zeisel and Kaye (1997, p. 88).

To further illustrate the change in practices the employer made one could examine the data for 1974–1978. It turns out the mean (median) time to promotion for males was 65.725 (35) and for females was 11.66 (15). Thus, *after* the charge males had to work at least a year *more* than females before they were promoted to Chief Pharmacist. This is an example of a phenomenon I refer to as "A Funny Thing Happens on the Way to the Courtroom." From both a "common sense" standpoint as well as legal one, the employment actions in the period leading up to the complaint have the most relevance for determining what happened when the plaintiff was being considered for promotion. (Similar issues of timing occur in contract law, where the meaning and conditions of a contract at the time it was signed are used to determine whether it has been properly carried out by both parties. In product liability law, a manufacturer is not held liable for risks that were not known when the product was sold to the plaintiff but were discovered subsequently.) Indeed, quite often a plaintiff applies for promotion on several occasions and only after being denied it on all of them, files a formal charge. (For example, in *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977 (1988) the plaintiff had applied for promotion four times. The opinion indicates that the Justices felt that she was unfairly denied promotion on her fourth attempt.)

Comment 1: Baldus and Cole (1987, p. 190) refer to the dispute concerning the Wilcoxon and Median tests in the *Capaci* case in a section concerning the difference between practical and statistical significance. Let me quote them:

> ► Exclusive concern for the ►statistical significance of a disparity encourages highly technical skirmishes between plaintiff's and defendants' experts who may find competing methods of computing statistical significance advantageous in arguing their respective positions (citing the *Capaci* case). The importance of such skirmishes maybe minimized by limiting the role of a test of significance to that of aiding in the interpretation of a measure of impact whose practical significance may be evaluated on nontechnical grounds.

Thus, even the authors of perhaps a commonly cited text on statistical proof of discrimination at the time did not appreciate the importance of the theory of hypothesis testing and the role of statistical power in choosing between tests. More importantly, a difference in the median time to promotion of 341 − 25 = 316 or about 15 years (or the difference in the average times of 23.5 years) would appear to me to be of practical significance. Thus, well respected authors as well as the judiciary allowed the defendant's expert to use

the median test, which had no power to detect a difference in time to promotion, to obfuscate a practically meaningful difference.

Comment 2: The expert for the defendant was a social scientist rather than a statistician. Other statisticians who have faced experts of this type have mentioned to me that often non-statisticians haven't had sufficient training in our subject to know how one should properly choose a procedure. They may select the first procedure that comes to mind or choose a method that helps their client even though it is quite inappropriate and their ignorance of statistical theory makes it difficult for the lawyer you are working for to get them to admit that their method is not as powerful (or accurate or reliable) as yours. An example of this occurred in a case concerning sex discrimination in pay when an "expert" compared the wages of small groups of roughly similar males and females with the t-test. It is well known that typically income and wage data are quite skewed and that the distribution of the two-sample t-test in small samples depends on the assumption of normality. I provided this information to the lawyer who then asked the other "expert" whether he had ever done any research using income or wage data (Ans. No) and whether he had ever carried out any research or read literature on the *t*-test and its properties (Ans. No). Thus, it was difficult for the lawyer to get this "expert" to admit that using the t-test in such a situation is questionable and the significance levels might not be reliable. On a more positive note, the *Daubert* (509 U.S. 579 (1993)) opinion listed several criteria for courts to evaluate expert scientific testimony, one of which is that the method used has a known error rate. Today one might be able to fit a skewed distribution to the data and then show by simulation that a nominal 0.05 level test has an actual level (α) of 0.10 or more. Similarly, if the one must use the t-test in such a situation one could conduct a simulation study to obtain "correct" critical values that will ensure that a nominal 0.05 level test has true level between 0.04 and 0.06. (Although I use the 0.05 level for illustration, I agree with Judge Posner (2001) that it should not be used as a talisman. Indeed, Judge P. Higginbotham's statement in *Vuyanich v. Republic National Bank*, 505 F. Supp. 224 (N.D. Texas 1980) that the p-value is a sliding-scale makes more sense than a simple yes-no dichotomy of significance or not in the legal context as the statistical evidence is only part of the story. The two-sided p-value 0.06 on the post-charge time until promotion data from the Capaci case illustrates the wisdom of the statements of these judges. Not only do the unbalanced sample sizes diminish the power of two-sample tests, the change in the promotion practices of the defendant subsequent to the charge are quite clear from the change in difference in average waiting times until promotion of males and females as well as the change from a significant difference in favor of males before the charge to a nearly significant change in favor of females after the charge.)

Another way of demonstrating that an expert does not possess relevant knowledge is for the lawyer to show them a book or article that states the point you wish to make and ask the expert to read it and then say they agree or disagree with the point. (Dr. Charles Mann told the author that he has been able to successfully use this technique.) If that expert disagrees, a follow-up question can inquire why or on what grounds does he or she disagree with it.

The tradition of reporting data by year also makes it more difficult to demonstrate that a change occurred after a charge was filed. In *Valentino v. USPS* (674 F.2d 56 (DC Circ. 1982)) the plaintiff had applied for a promotion in 1976 and filed the charge in May, 1976. The data is reported in Table 2 and has reanalyzed by Freidlin and Gastwirth (2000b) and Kadane and Woodworth (2004), suggested that females received fewer promotions than expected in the two previous years but after 1976 they received close to their expected number. Let me recall the data and analysis the yearly summaries enabled us to present.

The data for each year were analyzed by the Mantel-Haenszel test applied to the 2×2 tables for each grade grouping in Tables 2 and 3. This was done because the Civil Service Commission reported its data in this fashion. Notice that during two time periods, the number of grade advancements awarded to females for each year is significantly less than expected. After 1976, when the charge was filed the female employees start to receive their expected number of promotions.

My best recollection is that the promotion the plaintiff applied for was the 34th competitive one filled in 1976. Unfortunately, data on all of the applicants was not available even though EEOC guidelines do require employers to preserve records for at least 6 months. Of the 17 or so positions for which data on the actual applicants was available every one was given to a male candidate. Since females did receive their expected number of promotions over the entire year it would appear that the defendant changed its practices after the charge and consequently prevailed in the case. (The data discussed in the cited references considered employees in job categories classified by their level in the system. The district court accepted the defendant's criticism that since each job level contains positions in a variety of occupations, the plaintiffs should have stratified the data by occupation; see 511 F. Supp 917, 939 (D.C. DC, 1981). Later in the opinion, at 511 F. Supp. 951, the opinion accepted a regression analysis submitted by the defendant

Presentation of Statistical Testimony. Table 2 Number of employees and promotions they received: from the *Valentino v. U.S.P.S Case*

	Grade 17–19		Grade 20–22		Grade 23–25		Grade 26–28		Grade 29–31	
Time period	M	F	M	F	M	F	M	F	M	F
06/74–03/75	229	73	360	48	703	33	236	7	82	1
	67	5	74	9	132	2	28	1	8	0
03/75–01/76	205	89	373	43	716	36	277	9	85	1
	40	6	39	5	41	1	19	0	7	0
01/76–01/77	233	101	396	52	727	36	271	9	85	2
	31	10	32	4	54	5	28	2	5	0
01/77–01/78	200	86	377	52	680	35	262	8	89	3
	43	18	80	9	57	6	18	1	14	0
01/78–01/79	196	90	325	50	685	37	252	9	78	3
	29	8	45	7	42	3	14	1	6	1

Key to symbols: F=females; M=males; for any time period and grade group the number of promotions is below the number of employees. For example, in grades 17–19 during 06/74–03/75 period 67 out of 229 eligible males were promoted compared to 5 out of 73 eligible females

Presentation of Statistical Testimony. Table 3 The results obtained from Mantel-Haenszel test for equality of promotion rates applied to the stratified data for each period in Table 2

Year	Observed	Expected	p-value (two-sided)
1974–1975	17	34.1	0.0006
1975–1976	12	21.16	0.020
1976–1977	21	20.32	0.885
1977–1978	34	33.23	0.869
1978–1979	20	21.66	0.674

that used grade level as a predictor of salary noting that 'level" is a good proxy for occupation. While upholding the ultimate finding that U.S.P.S did not discriminate against women, the appellate opinion, fn. 15 at 71, did not accept the district court's criticism that a regression submitted by plaintiffs should have included the level of a position. The reason is that in a promotion case, it is advancement in one's job level that is the issue. Thus, courts do accept regressions that include the major other job-related characteristics such as experience, education, special training and objective measures of productivity.) There was one unusual aspect of the case; the Postal Service had changed the system of reviewing candidates for promotions as of January 1, 1976. As no other charges of discrimination in promotion had been filed in either 1974 or 1975, the analysis of data for those years is considered background information unless the plaintiff can demonstrate that the same process carried over into the time period when the promotion in question was made. (In *Evans v. United Airlines*, the Supreme Court stated that since charges of employment discrimination need to be filed within 180 days of the charge, earlier data is useful background information but is not sufficient by itself to establish a *prima facie* case of discrimination. If there is a continuing violation, however, then the earlier data can be used in conjunction with more recent data by the plaintiffs. The complex legal issues involved in determining when past practices have continued into the time period relevant to a particular case are beyond the scope of the present paper.)

When data is reported in yearly periods, invariably some post-charge data will be included in the data for the year in which the charge was file and statisticians should examine it to see if there is evidence of a change in employment practice subsequent to the charge. In the *Capaci* case, the defendant included eleven months of post-charge data in their first time period, 1965–1973. The inclusion of three additional females promoted in that period lessens the impact of the fact that only two female Pharmacists were promoted during the previous seven and a half years. Similarly, reporting the data by year in *Valentino* enabled the

defendant to monitor the new promotion system begun at the start of 1976 subsequent to the complaint, so females did receive their expected number of promotions for the 1976, the year of the charge.

A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))

The plaintiff in the case alleged that he had been denied a promotion to a GS-14 level position in the computer division of a government agency and filed a formal complaint in 1977. As the files on the actual applicants were unavailable, we considered all individuals employed in GS-13 level jobs whose records indicated that they met the Civil Service criteria for a promotion to the next job-level to be the pool of "qualified applicants." (These qualifications were that they were employed in an appropriate computer-related position and had at lest one year of experience at the previous level (GS-13) or its equivalent.) There were about ten opportunities for promotion to a GS-14 post during the several years prior to the complaint. About three of the successful GS-14 job applicants were outside candidates who had a Ph.D. in computer science; all of whom were white. Since they had a higher level of education than the internal candidates, they are excluded from Table 4, which gives the number of employees who were eligible and the number promoted, by race, for the ten job announcements.

Although the data is longitudinal in nature and technically one might want to apply a survival test such as the log-rank procedure, it was easier to analyze the data by the Mantel-Haenszel (MH) test that combines the observed minus expected numbers from the individual the individual 2 × 2 tables (this is, of course, the way the log-rank test is also computed and the resulting statistics are the same). Although there were only 18 promotions awarded to internal candidates during the period under consideration *none* of them went to a black. Moreover, the exact p-value of the MH test was 0.007, a clearly significant result. The analysis can be interpreted as a test for the effect of race controlling for eligibility by Feinberg (1989, p. 100) and has been discussed by Agresti (1996) in the STATXACT manual (2003, p. 597) where it is shown that the *lower* end of a 95% confidence interval of the odds a white employee receives a promotion relative to a minority employee is about two. Thus, we can conclude that the odds of a black employee had of being promoted were half those of a white, which is clearly of practical as well as statistical significance.

In order to demonstrate that the most plausible potential explanation of the promotion data, the white employees had greater seniority, data was submitted that showed that by 1977 the average black employee had worked over a year more at the GS-13 level than the average white one. Thus, if seniority were a major factor that could explain why whites received the earlier promotions, it should have worked in favor of the black employees in the later part of the period. The defendant did not suggest an alternative analysis of this data but concentrated on post-charge data but Judge A. Robinson observed at the trial that their analysis should have considered a time frame around the time of the complaint.

Presentation of Statistical Testimony. Table 4
Promotion data for GS-14 positions obtained by internal job candidates, by Race, from *Hogan v. Pierce* (From Plaintiff's exhibit on file with D.C. District Court. Reproduced in Gastwirth (1988, p. 266) and in the STATXACT manual (Version 6, p. 597))

	Whites		Blacks	
Date of Promotion	Eligible	Promoted	Eligible	Promoted
July 1974	20	4	7	0
August 1974	17	4	7	0
Sept. 1974	15	2	8	0
April 1975	18	1	8	0
May 1975	18	1	8	0
Oct. 1975	30	1	10	0
Nov. 1975	31	2	10	0
Feb. 1976	31	1	10	0
March 1976	31	1	10	0
Nov. 1977	34	1	13	0

A Few Suggestions to Statisticians Desiring to Increase the Validity of and Weight Given to Statistical and Scientific Evidence

Mann (2000), cited and endorsed by Mallows (2003), noted that if your analysis does not produce results the lawyer who hired you desired; you will likely be replaced. Indeed, he notes that many attorneys act as though they will be able to find a statistician who will give them the results they want and that "regrettably they may often be correct." This state of affairs unfortunately perpetuates the statement that there are "lies, damn lies and statistics" attributed to both B. Disraeli and M. Twain. Statisticians

have suggested questions that judges may ask experts (Feinberg et al. 1995) and discussed related ethical issues arising in giving testimony (Meier 1986; Kadane 2005). Here we mention a few more suggestions for improving statistical testimony and close with a brief discussion questioning the wisdom of a recent editorial that appeared in *Science*.

1. Mann (2000) is correct in advising statisticians who are asked to testify inquire as to the existence of other data or the analysis of any other previous expert the lawyer consulted. I would add that these are especially important considerations if you are asked to examine data very shortly before a trial as you will have a limited time to understand the data and how it was generated so you will need to totally rely on the data and background information the lawyer provides. In civil cases, it is far preferable for an expert to be involved during the period when discovery is being carried out. Both sides are asking the other for relevant data and information and it is easier for you to tell the lawyer the type of data that you feel would help answer the relevant questions. If the lawyers ignore your requests or don't give you a sensible justification (let me give an example of a business justification. You are asked to work for a defendant in an equal employment case concerning the fairness of a layoff carried out in a plant making brakes for cars and SUVs. Data on absenteeism would appear to be quite useful. The employer knows that the rate of absenteeism in the plant was "higher" than normal for the industry and might be concerned that if this information was put into the public record, they might become involved in many suits arising from accidents involving cars with brakes made there. Since the corporation is a profit-making organization not a scientific one, they will carry out a "cost-benefit" analysis to decide whether you will be given the data), you should become concerned.

   The author's experience in *Valentino* illustrates this point. One reason the original regression equation we developed for the plaintiffs was considered incomplete was because it did not contain relevant information concerning any special degrees the employees had. This field was omitted from the original file we were given and we only learned about it a week or so before trial. After finding that employees with business, engineering or law degrees received about the same salary increase, we then included a single indicator for having a degree in one of the three areas. To assist the court, it would have been preferable to use separate indicators for each degree. (Both opinions, see 674 F.2d 56, 70 fn. 21, downplayed this factor because the defendant's expert testified that it was unreliable as the information depended on whether applicants listed any specialty on their form. More importantly, use of the single indicator variable for three different degrees may not have made it clear that they all had a similar effect on salary and that the indicator was restricted to employees in these three specialties.) Given the very short time available to incorporate the new, but important information that did reduce the estimated coefficient on sex by a meaningful amount, although it remained significant, one did not have time to explore how best to present the analysis.
2. When you are approached by counsel, you should ask for permission to write an article about any interesting statistical problems or data arising in the case. (Of course, data and exhibits submitted formally into evidence are generally considered to be in the public domain so one can use them in scholarly writings. Sometimes, however, the data in the exhibits are summaries and either the individual data or summaries of smaller sub-groups are required to illustrate the methodology.) While the lawyers and client might request that some details of the case not be reported so that the particular case or parties involved are not identifiable, you should be given permission to show the profession how you used or adapted existing methodology. If you perceive that the client or lawyer want to be very restrictive about the use and dissemination of the data and your analysis, you should think seriously about becoming involved. I realize that it is much easier for an academic statistician to say "no" in this situation, than statisticians who do consulting for a living; especially if they have a long-term business relationship with a law firm.
3. Statisticians are now being asked by judges to advise them in "*Daubert*" hearings, which arise when one party in the case challenges the scientific validity or reliability of the expert testimony the other side desires to use. This is a useful service that all scientific professions can provide the legal system. Before assessing a proposed expert's testimony and credentials, it is wise for you to look over the criteria the court suggested in the *Daubert* opinion as well as some examples of decisions in such matters. Because courts take a skeptical view of "new techniques" that have not been subject to peer review but appear to have been developed specifically for the case at hand, one should not fault an expert who does not use the most recently developed method but uses a previously existing method that has been generally regarded in the field as appropriate for

the particular problem. The issue is not whether you would have performed the same analysis the expert conducted but whether the analysis is appropriate and statistically sound. Indeed, Kadane (1990) conducted a Bayesian analysis of data arising in an equal employment and confirmed it with a frequentist method used in ►biostatistics that had been suggested for the problem by Gastwirth and Greenhouse (1987). He noted that it is reassuring when two approaches lead to the same inference. In my experience, this often occurs when data sets of moderate to large sample size are examined and one uses some common sense in interpreting any difference. (By common sense I mean that if one statistician analyzes a data set with test A and obtains a p-value of 0.04, i.e., a statistically significant one, but the other statistician uses another appropriate test B and obtains a p-value of 0.06, a so-called non-significant one, the results are really not meaningfully different.)

4. A "*Daubert*" hearing can be used to provide the judge with questions to ask an expert about the statistical methodology utilized. (This also gives the court's expert the opportunity to find the relevant portions of books and articles that can be shown the expert, thereby implementing the approach described in Comment 2 of section "►A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))".) This may be effective in demonstrating to a court that a non-statistician really does not have a reasonable understanding of some basic concepts such as power or the potential effect of violations of the assumptions underlying the methods used.
5. One task experts are asked to perform is to criticize or raise questions about the findings and methodology of the testimony given by the other party's expert. This is one place where is easy for one to become overly partisan and lose their scientific objectivity. This important problem is discussed by (Meier 1986). Before raising a criticism, one should ask whether it is important. Could it make a difference in either the conclusion or the weight given to it? In addition to checking that the assumptions underlying the analysis they are presenting are basically satisfied, one step an expert can carry out to protect their own statistical analyses from being unfairly criticized is to carry out a ►sensitivity analysis. Many methods for assessing whether an omitted variable could change an inference have been developed (Rosenbaum 2002) and van Belle (2002) has discussed the relative importance of violations of the assumptions underlying commonly used techniques.
6. While this entry focused on statistical testimony in civil cases, similar issues arise in the use of statistical evidence in criminal cases. The reader is referred to Aitken and Taroni (2004), Balding (2005) for the commonly used statistical methods used in this setting. Articles by a number of experts in both civil and criminal cases appear in Gastwirth (2000) provide additional perspectives on issues arising in the use of statistical evidence in the legal setting.

Acknowledgments

This entry was adapted from Gastwirth (2005). The author wishes to thank Prof. M. Lovic for his helpful suggestions.

About the Author

Professor Joseph L. Gastwirth is a Fellow of ASA (1970), IMS (1972), and AAS (1971), and an Elected member of ISI (1980). He was President of the Washington Statistical Society (1982–1983). He has served three times on the American Statistical Association's Law and Justice Statistics Committee and currently is its Chair. He has written over 150 articles concerning both the theory and application of statistics; especially in the legal setting. In 1985, he received a Guggenheim Fellowship, which led to his two-volume book *Statistical Reasoning in Law and Public Policy* (1988), and the Shiskin Award (1998) for Research in Economic Statistics for his development of statistical methods for measuring economic inequality and discrimination. He edited the book *Statistical Science in the Courtroom* (2000) and in 2002, his article with Dr. B. Freidlin using change-point methods to analyze data arising in equal employment cases shared the "Applications Paper of the Year" award of the American Statistical Association. He has served on the editorial boards of several major statistical journals and recently, he was appointed Editor of the journal, *Law, Probability and Risk*. On August 1, 2009, a workshop celebrating 45 years of statistical activity of Professor Gastwirth was organized at the George Washington University to "recognize his outstanding contributions to the development of nonparametric and robust methods and their use in genetic epidemiology, his pioneering research in statistical methods and measurement for the analysis of data arising in the areas of economic inequality, health disparities and equal employment, and in other legal applications".

Cross References

►Frequentist Hypothesis Testing: A Defense
►Null-Hypothesis Significance Testing: Misconceptions
►Power Analysis
►P-Values

►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Evidence
►Statistics and the Law
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York
Aitken C, Taroni F (2004) Statistics and the evaluation of evidence for forensic scientists, 2nd edn. Wiley, Chichester, UK
Balding DJ (2005) Weight of the evidence for forensic DNA profiles. Wiley, Chichester, UK
Baldus DC, Cole JWL (1987) Statistical proof of discrimination: cumulative supplement. Shepard's/McGraw-Hill, Colorado Springs
Berger MA (2000) The supreme court's trilogy on the admissibility of expert testimony. In: Reference manual on scientific evidence, 2nd edn. Federal Judicial Center, Washington, DC, pp 9–38
Feinberg SE (1989) The evolving role of statistical assessments as evidence in courts. Springer, New York
Feinberg SE, Krislov SH, Straf M (1995) Understanding and evaluating statistical evidence in litigation. Jurimetrics J 36:1–32
Finkelstein MO, Levin B (2000) Statistics for lawyers, 2nd edn. Springer, New York
Freidlin B, Gastwirth JL (2000a) Should the median test be retired from general use? Am Stat 54:161–164
Freidlin B, Gastwirth JL (2000b) Change point tests designed for the analysis of hiring data arising in equal employment cases. J Bus Econ Stat 18:315–322
Gastwirth JL, Greenhouse SW (1987) Estimating a common relative risk: application in equal employment. J Am Stat Assoc 82:38–45
Gastwirth JL, Wang JL (1987) Nonparametric tests in small unbalanced samples: application in employment discrimination cases. Can J Stat 15:339–348
Gastwirth JL (1988) Statistical reasoning in law and public policy vol. 1 statistical concepts and issues of fairness. Academic, Orlando, FL
Gastwirth JL (ed) (2000) Statistical science in the courtroom. Springer, NY
Gastwirth JL (2005) Some issues arising in the presentation of statistical testimony. Law, Probability and Risk 4:5–20
Kadane JB (1990) A statistical analysis of adverse impact of employer decisions. J Am Stat 85:925–933
Kadane JB (2005) Ethical issues in being an expert witness. Law, Probability and Risk 4:21–23
Kadane JB, Woodworth GG (2004) Hierarchical models for employment decisions. J Bus Econ Stat 22:182–xx
Mallows C (2003) Parity: implementing the telecommunications act of 1996. Stat Sci 17:256–270
Mann CR (2000) Statistical consulting in the legal environment. In: Gastwirth JL (ed) Statistical science in the courtroom. Springer, New York, pp 245–262
Meier P (1986) Damned liars and expert witnesses. J Am Stat Assoc 81:269–276
Rosenbaum PR (2002) Observational studies, 2nd edn. Springer, New York
Rosenblum M (2000) On the evolution of analytic proof, statistics and the use of experts in EEO litigation. In: Gastwirth JL (ed) Statistical science in the courtroom. Springer, New York, pp 161–194
van Belle G (2002) Statistical rules of thumb. Wiley, New York
Zeisel H, Kaye DH (1997) Prove it with figures: empirical methods in law and litigation. Springer, New York

Principal Component Analysis

Ian Jolliffe
Professor
University of Exeter, Exeter, UK

Introduction

Large or massive data sets are increasingly common and often include measurements on many variables. It is frequently possible to reduce the number of variables considerably while still retaining much of the information in the original data set. Principal component analysis (PCA) is probably the best known and most widely used dimension-reducing technique for doing this. Suppose we have n measurements on a vector $\mathbf{x}$ of p random variables, and we wish to reduce the dimension from p to q, where q is typically much smaller than p. PCA does this by finding linear combinations, $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \ldots, \mathbf{a}_q'\mathbf{x}$, called *principal components*, that successively have maximum variance for the data, subject to being uncorrelated with previous $\mathbf{a}_k'\mathbf{x}$s. Solving this maximization problem, we find that the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$ are the eigenvectors of the covariance matrix, $\mathbf{S}$, of the data, corresponding to the q largest eigenvalues (see ►Eigenvalue, Eigenvector and Eigenspace). The eigenvalues give the variances of their respective principal components, and the ratio of the sum of the first q eigenvalues to the sum of the variances of all p original variables represents the proportion of the total variance in the original data set accounted for by the first q principal components. The familiar algebraic form of PCA was first presented by Hotelling (1933), though Pearson (1901) had earlier given a geometric derivation. The apparently simple idea actually has a number of subtleties, and a surprisingly large number of uses, and has a vast literature, including at least two comprehensive textbooks (Jackson 1991; Jolliffe 2002).

An Example

As an illustration we use an example that has been widely reported in the literature, and which is originally due to

Principal Component Analysis. Table 1 Principal Component Analysis Vectors of coefficients for the first two principal components for data from Yule et al. (1969)

Variable	a_1	a_2
x_1	0.34	0.39
x_2	0.34	0.37
x_3	0.35	0.10
x_4	0.30	0.24
x_5	0.34	0.32
x_6	0.27	−0.24
x_7	0.32	−0.27
x_8	0.30	−0.51
x_9	0.23	−0.22
x_{10}	0.36	−0.33

Yule et al. (1969). The data consist of scores, between 0 and 20, for 150 children aged $4^1/_2 - 6$ years from the Isle of Wight, on ten subtests of the Wechsler Pre-School and Primary Scale of Intelligence. Five of the tests were "verbal" tests and five were 'performance' tests. Table 1 gives the vectors $\mathbf{a}_1, \mathbf{a}_2$ that define the first two principal components for these data.

The first component is a linear combination of the ten scores with roughly equal weight (maximum 0.36, minimum 0.23) given to each score. It can be interpreted as a measure of the overall ability of a child to do well on the full battery of ten tests, and represents the major (linear) source of variability in the data. On its own it accounts for 48% of the original variability. The second component contrasts the first five scores (verbal tests) with the five scores on the performance tests. It accounts for a further 11% of the total variability. The form of this second component tells us that once we have accounted for overall ability, the next most important (linear) source of variability in the tests scores is between those children who do well on the verbal tests *relative to* the performance tests and those children whose test score profile has the opposite pattern.

Covariance or Correlation

Principal components successively maximize variance, and can be found from the eigenvalues/eigenvectors of a covariance matrix. Often a modification is adopted, in order to avoid two problems. If the p variables are measured in a mixture of units, then it is difficult to interpret the principal components. What is meant by a linear combination of weight, height and temperature, for example? Furthermore, if we measure temperature and weight in °F and pounds respectively, we may get completely *different principal components* from those obtained from the *same data* but using °C and kilograms. To avoid this arbitrariness, we standardize each variable to have zero mean and unit variance. Finding linear combinations of these standardized variables that successively maximize variance, subject to being uncorrelated with previous linear combinations, leads to principal components defined by the eigenvalues and eigenvectors of the correlation matrix, rather than the covariance matrix, of the original variables. When all variables are measured in the same units, covariance-based PCA may be appropriate, but even here they can be uninformative when a few variables have much larger variances than the remainder. In such cases the first few components are dominated by the high-variance variables and tell us little that could not have been deduced by inspection of the original variances. Circumstances exist where covariance-based PCA is of interest but most PCAs encountered in practice are correlation-based. Our example is a case where either approach could be used. The results given above are based on the correlation matrix but, because the variances of all 10 tests are similar, results from a covariance-based analysis would be little different.

How Many Components?

We have talked about q principal components accounting for most of the variation in the p variables? What is meant by "most" and, more generally, how do we decide how many components to keep? There is a large literature on this topic – see, for example, Jolliffe (2002), Chap. 6. Perhaps the simplest procedure is to set a threshold, say 80%, and stop when the first q components account for a percentage of total variation greater than this threshold. In our example the first two components accounted for only 59% of the variation. The threshold is often set higher than this – 70% to 90% are the usual sort of values, but it depends on the context of the data set and can be higher or lower. Other techniques are based on the values of the eigenvalues or on the differences between consecutive eigenvalues. Some of these simple ideas, as well as more sophisticated ones (Jolliffe 2002, Chap. 6) have been borrowed from factor analysis (see ►Factor Analysis and Latent Variable Modelling). This is unfortunate because the different objectives of PCA and factor analysis (see below for more on this) mean that typically fewer dimensions should be retained in

factor analysis than in PCA, so the factor analysis rules are often inappropriate. It should also be noted that although it is usual to discard low-variance principal components, they can sometimes be useful in their own right, for example in finding ►outliers (Jolliffe 2002, Chap. 10) and in quality control (Jackson 1991).

Confusion with Factor Analysis

There is much confusion between principal component analysis and factor analysis, partly because some widely used software packages treat PCA as a special case of factor analysis, which it is not. There are several technical differences between PCA and factor analysis, but the most fundamental difference is that factor analysis explicitly specifies a model relating the observed variables to a smaller set of underlying unobservable factors. Although some authors express PCA in the framework of a model, its main application is as a descriptive, exploratory technique, with no thought of an underlying model. This descriptive nature means that distributional assumptions are unnecessary to apply PCA in its usual form. It can be used, although caution may be needed in interpretation, on discrete and even binary data, as well as continuous variables. One notable feature of factor analysis is that it is generally a two-stage procedure; having found an initial solution, it is rotated towards simple structure. This idea can be borrowed and used in PCA; having decided to keep q principal components, we may rotate within the q-dimensional subspace defined by the components in a way that makes the axes as easy as possible to interpret. This is one of number of techniques that attempt to simplify the results of PCA by post-processing them in some way, or by replacing PCA with a modified technique (Jolliffe 2002, Chap. 11).

Uses of Principal Component Analysis

There are many variations on the basic use of PCA as a dimension reducing technique whose results are used in a descriptive/exploratory manner – see Jackson (1991), Jolliffe (2002). PCA is often used a first step, reducing dimensionality before undertaking another technique, such as multiple regression, cluster analysis (see ►Cluster Analysis: An Introduction), discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) or independent component analysis.

Extensions to Principal Component Analysis

PCA has been extended in many ways. For example, one restriction of the technique is that it is linear. A number of non-linear versions have therefore been suggested. These include the Gifi approach to multivariate analysis. Another area in which many variations have been proposed is when the data are time series, so that there is dependence between observations as well as between variables (Jolliffe 2002, Chap. 12). There are many other extensions and modifications, and the list continues to grow.

Acknowledgments

This article is a revised and shortened version of an entry that appeared in The Encyclopedia of Statistics in Behavioral Science, published by Wiley.

About the Author

Professor Ian Jolliffe is Honorary Visiting Professor at the University of Exeter. Before his early retirement he was Professor of Statistics at the University of Aberdeen. He received the International Meetings in Statistical Climatology Achievement Award in 2004 and was elected a Fellow of the American Statistical Association in 2009. He has authored, co-authored or co-edited four books, including *Principal Component Analysis* (Springer, 2nd edition, 2002) and *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (jointly edited with D B Stephenson, Wiley, 2003, 2nd edition due 2011). He has also published over 80 papers in peer-reviewed journal and is currently Associate Editor for *Weather and Forecasting*.

Cross References

►Analysis of Multivariate Agricultural Data
►Data Analysis
►Eigenvalue, Eigenvector and Eigenspace
►Factor Analysis and Latent Variable Modelling
►Fuzzy Logic in Statistical Data Analysis
►Multivariate Data Analysis: An Overview
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Multivariate Technique: Robustness
►Partial Least Squares Regression Versus Other Methods

References and Further Reading

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441, 498–520
Jackson JE (1991) A user's guide to principal components. Wiley, New York
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
Yule W, Berger M, Butler S, Newham V, Tizard J (1969) The WPPSI: an empirical evaluation with a British sample. Brit J Educ Psychol 39:1–13

Principles Underlying Econometric Estimators for Identifying Causal Effects

James J. Heckman
Winner of the Nobel Memorial Prize in Economic Sciences in 2000, Henry Schultz Distinguished Service Professor of Economics
The University of Chicago, Chicago, IL, USA
University College Dublin, Dublin, Ireland

This paper reviews the basic principles underlying the identification of conventional econometric evaluation estimators for causal effects and their recent extensions. Heckman (2008) discusses the econometric approach to causality and compares it to conventional statistical approaches. This paper considers alternative methods for identifying causal models.

The paper is in four parts. The first part presents a prototypical economic choice model that underlies econometric models of causal inference. It is a framework that is useful for analyzing and motivating the economic assumptions underlying alternative estimators. The second part discusses general identification assumptions for leading econometric estimators at an intuitive level. The third part elaborates the discussion of matching in the second part. Matching is widely used in applied work and makes strong informational assumptions about what analysts know relative to what the people they analyze know. The fourth part concludes.

A Prototypical Policy Evaluation Problem

Consider the following prototypical policy problem. Suppose a policy is proposed for adoption in a country. It has been tried in other countries and we know outcomes there. We also know outcomes in countries where it was not adopted. From the historical record, what can we conclude about the likely effectiveness of the policy in countries that have not implemented it?

To answer questions of this sort, economists build models of counterfactuals. Consider the following model. Let Y_0 be the outcome of a country (e.g., GDP) under a no-policy regime. Y_1 is the outcome if the policy is implemented. $Y_1 - Y_0$ is the "treatment effect" or causal effect of the policy. It may vary among countries. We observe characteristics X of various countries (e.g., level of democracy, level of population literacy, etc.). It is convenient to decompose Y_1 into its mean given X, $\mu_1(X)$ and deviation from mean U_1. One can make a similar decomposition for Y_0:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0. \end{aligned} \tag{1}$$

Additive separability is not needed, but it is convenient to assume it, and I initially adopt it to simplify the exposition and establish a parallel regression notation that serves to link the statistical literature on treatment effects with the economic literature. (Formally, it involves no loss of generality if we define $U_1 = Y_1 - E(Y_1 \mid X)$ and $U_0 = U_0 - E(Y_0 \mid X)$.)

It may happen that controlling for the X, $Y_1 - Y_0$ is the same for all countries. This is the case of homogeneous treatment effects given X. More likely, countries vary in their responses to the policy even after controlling for X.

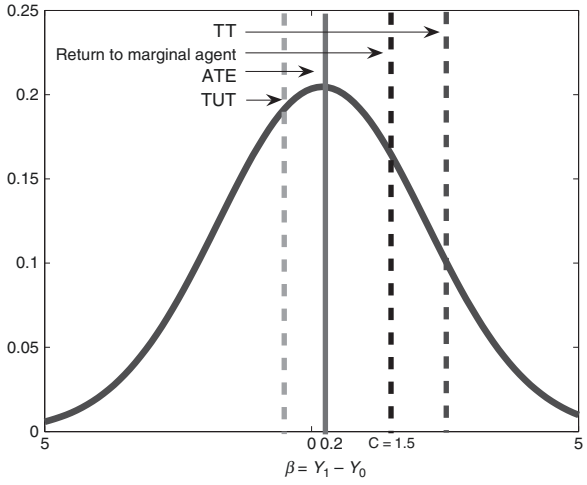
Figure 1 plots the distribution of $Y_1 - Y_0$ for a benchmark X. It also displays the various conventional treatment parameters. I use a special form of a "generalized Roy" model with constant cost C of adopting the policy (see Heckman and Vytlacil 2007a, for a discussion of this model). This is called the "extended Roy model." I use this model because it is simple and intuitive. (The precise parameterization of the extended Roy model used to generate the figure and the treatment effects is given at the base of Fig. 1.) The special case of homogeneity in $Y_1 - Y_0$ arises when the distribution collapses to its mean. It would be ideal if one could estimate the distribution of $Y_1 - Y_0$ given X and there is research that does this.

More often, economists focus on some mean of the distribution in the literature and use a regression framework to interpret the data. To turn (1) into a regression model, it is conventional to use the switching regression framework. (Statisticians sometimes attribute this representation to Rubin (1974, 1978), but it is due to Quandt (1958, 1972). It is implicit in the Roy (1951) model. See the discussion of this basic model of counterfactuals in Heckman and Vytlacil (2007a)). Define $D = 1$ if a country adopts a policy; $D = 0$ if it does not. Substituting (1) into this expression, and keeping all X implicit, one obtains

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned} \tag{2}$$

This is the Roy-Quandt "switching regression" model. Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon \tag{3}$$



TT = 2.666, TUT = –0.632
Return to Marginal Agent = $C = 1.5$
ATE = $\mu_1 - \mu_0 = \bar{\beta} = 0.2$

The Model

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$

General Case

$(U_1 \neq U_0) \not\perp\!\!\!\perp D$
ATE ≠ TT ≠ TUT

The Researcher Observes (Y, D, C)

$Y = \alpha + \beta D + U_0$ where $\beta = Y_1 - Y_0$

Parameterization

$\alpha = 0.67 \quad (U_1, U_0) \sim N(\mathbf{0}, \Sigma) \quad D^* = Y_1 - Y_0 - C$

$\bar{\beta} = 0.2 \quad \Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \quad C = 1.5$

Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 1 Distribution of gains, the Roy economy (Heckman et al. 2006)

where $\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$ and $\varepsilon = U_0$. I will also use the notation that $\upsilon = U_1 - U_0$, letting $\bar{\beta} = \mu_1 - \mu_0$ and $\beta = \bar{\beta} + \upsilon$. Throughout this paper I use treatment effect and regression notation interchangeably. The coefficient on D is the treatment effect. The case where β is the same for every country is the case conventionally assumed. More elaborate versions assume that β depends on X ($\beta(X)$) and estimates interactions of D with X. The case where β varies even after accounting for X is called the "random coefficient" or "heterogenous treatment effect" case. The case where $\upsilon = U_1 - U_0$ depends on D is the case of essential heterogeneity analyzed by Heckman et al. (2006). This case arises when treatment choices depend at least in part on the idiosyncratic return to treatment. A great deal of attention has been focused on this case in recent decades and I develop the implications of this model in this paper.

An Index Model of Choice and Treatment Effects: Definitions and Unifying Principles

I now present the model of treatment effects developed in Heckman and Vytlacil (1999, 2001, 2005, 2007a,b) and Heckman et al. (2006), which relaxes the normality, separability and exogeneity assumptions invoked in the traditional economic selection models. It is rich enough to generate all of the treatment effects in the program evaluation literature as well as many other policy parameters. It does not require separability. It is a nonparametric generalized Roy model with testable restrictions that can be used to unify the treatment effect literature, identify different treatment effects, link the literature on treatment effects to the literature in structural econometrics and interpret the implicit economic assumptions underlying ▶instrumental variables, regression discontinuity design methods, control functions and matching methods.

Y is the measured outcome variable. It is produced from the switching regression model (2). Outcomes are general nonlinear, nonseparable functions of observables and unobservables:

$$Y_1 = \mu_1(X, U_1) \tag{4}$$
$$Y_0 = \mu_0(X, U_0). \tag{5}$$

Examples of models that can be written in this form include conventional latent variable models for discrete choice that are generated by a latent variable crossing a threshold: $Y_i = \mathbf{1}\left(Y_i^* \geq 0\right)$, where $Y_i^* = \mu_i(X) + U_i$, $i = 0, 1$. Notice that in the general case, $\mu_i(X, U_i) - E(Y_i \mid X) \neq U_i$, $i = 0, 1$.

The individual treatment effect associated with moving an otherwise identical person from "0" to "1" is $Y_1 - Y_0 = \Delta$ and is defined as the causal effect on Y of a *ceteris paribus* move from "0" to "1". To link this framework to the literature on economic choice models, I characterize the

decision rule for program participation by an index model:

$$D^* = \mu_D(Z) - V; \qquad D = 1 \quad \text{if} \quad D^* \geq 0; \qquad D = 0 \quad \text{otherwise}, \tag{6}$$

where, from the point of view of the econometrician, (Z, X) is observed and (U_1, U_0, V) is unobserved. The random variable V may be a function of (U_1, U_0). For example, in the original Roy Model, μ_1 and μ_0 are additively separable in U_1 and U_0 respectively, and $V = -[U_1 - U_0]$. In the original formulations of the generalized Roy model, outcome equations are separable and $V = -[U_1 - U_0 - U_C]$, where U_C arises from the cost function. Without loss of generality, I define Z so that it includes all of the elements of X as well as any additional variables unique to the choice equation.

I invoke the following assumptions that are weaker than those used in the conventional literature on structural econometrics or the recent literature on semiparametric selection models and at the same time can be used both to define and to identify different treatment parameters. (A much weaker set of conditions is required to define the parameters than is required to identify them. See Heckman and Vytlacil (2007b, Appendix B).) The assumptions are:

(A-1) (U_0, U_1, V) *are independent of* Z *conditional on* X *(Independence);*

(A-2) $\mu_D(Z)$ *is a nondegenerate random variable conditional on* X *(Rank Condition);*

(A-3) *The distribution of* V *is continuous; (Absolutely continuous with respect to Lebesgue measure.)*

(A-4) *The values of* $E|Y_1|$ *and* $E|Y_0|$ *are finite (Finite Means);*

(A-5) $0 < \Pr(D = 1 \mid X) < 1$.

(A-1) assumes that V is independent of Z given X and is used below to generate counterfactuals. For the definition of treatment effects one does not need either (A-1) or (A-2). The definitions of treatment effects and their unification do not require any elements of Z that are not elements of X or independence assumptions. However, an analysis of instrumental variables requires that Z contain at least one element not in X. Assumptions (A-1) or (A-2) justify application of instrumental variables methods and nonparametric selection or control function methods. Some parameters in the recent IV literature are defined by an instrument so I make assumptions about instruments up front, noting where they are not needed. Assumption (A-4) is needed to satisfy standard integration conditions. It guarantees that the mean treatment parameters are well defined. Assumption (A-5) is the assumption in the population of both a treatment and a control group for each X. Observe that there are no exogeneity requirements for X. This is in contrast with the assumptions commonly made in the conventional structural literature and the semiparametric selection literature (see, e.g., Powell 1994).

A counterfactual "no feedback" condition facilitates interpretability so that conditioning on X does not mask the effects of D. Letting X_d denote a value of X if D is set to d, a sufficient condition that rules out feedback from D to X is:

(A-6) *Let* X_0 *denote the counterfactual value of* X *that would be observed if* D *is set to* 0. X_1 *is defined analogously. Assume* $X_d = X$ *for* $d = 0, 1$. *(The* X_D *are invariant to counterfactual manipulations.)*

Condition (A-6) is not strictly required to formulate an evaluation model, but it enables an analyst who conditions on X to capture the "total" or "full effect" of D on Y (see Pearl 2000). This assumption imposes the requirement that X is an external variable determined outside the model and is not affected by counterfactual manipulations of D. However, the assumption allows for X to be freely correlated with U_1, U_0 and V so it can be endogenous.

In this notation, $P(Z)$ is the probability of receiving treatment given Z, or the "propensity score" $P(Z) \equiv \Pr(D = 1 \mid Z) = F_{V|X}(\mu_D(Z))$, where $F_{V|X}(\cdot)$ denotes the distribution of V conditional on X. (Throughout this paper, I will refer to the cumulative distribution function of a random vector A by $F_A(\cdot)$ and to the cumulative distribution function of a random vector A conditional on random vector B by $F_{A|B}(\cdot)$. I will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot \mid b)$.) I denote $P(Z)$ by P, suppressing the Z argument. I also work with U_D, a uniform random variable ($U_D \sim \text{Unif}[0,1]$) defined by $U_D = F_{V|X}(V)$. (This representation is valid whether or not (A-1) is true. However, (A-1) imposes restrictions on counterfactual choices. For example, if a change in government policy changes the distribution of Z by an external manipulation, under (A-1) the model can be used to generate the choice probability from $P(z)$ evaluated at the new arguments, i.e., the model is invariant with respect to the distribution Z.) The separability between V and $\mu_D(Z)$ or $D(Z)$ and U_D is conventional. It plays a crucial role in justifying instrumental variable estimators in the general models analyzed in this paper.

Vytlacil (2002) establishes that assumptions (A-1)–(A-5) for the model of Eqs. (2)–(6) are equivalent to the assumptions used to generate the LATE model of Imbens and Angrist (1994). Thus the nonparametric selection model for treatment effects developed by Heckman

and Vytlacil is implied by the assumptions of the Imbens and Angrist instrumental variable model for treatment effects. The Heckman and Vytlacil approach is more general and links the IV literature to the literature on economic choice models. The latent variable model is a version of the standard sample selection bias model. This weaves together two strands of the literature often thought to be distinct (see e.g., Angrist and Krueger 1999). Heckman et al. (2006) develop this parallelism in detail. (The model of Eqs. (4)–(6) and assumptions (A-1)–(A-5) impose two testable restrictions on the distribution of (Y, D, Z, X). First, it imposes an index sufficiency restriction: for any set $\mathcal{A}$ and for $j = 0, 1$,

$$\Pr(Y_j \in \mathcal{A} \mid X, Z, D = j) = \Pr(Y_j \in \mathcal{A} \mid X, P(Z), D = j).$$

Z (given X) enters the model only through the propensity score $P(Z)$ (the sets of $\mathcal{A}$ are assumed to be measurable). This restriction has empirical content when Z contains two or more variables not in X. Second, the model also imposes monotonicity in p for $E(YD \mid X = x, P = p)$ and $E(Y(1-D) \mid X = x, P = p)$. Heckman and Vytlacil (2005, Appendix A) develop this condition further, and show that it is testable.

Even though this model of treatment effects is not the most general possible model, it has testable implications and hence empirical content. It unites various literatures and produces a nonparametric version of the selection model, and links the treatment literature to economic choice theory.)

Definitions of Treatment Effects in the Two Outcome Model

The difficulty of observing the same individual in both treated and untreated states leads to the use of various population level treatment effects widely used in the biostatistics literature and often applied in economics. (Heckman et al. (1999) discuss panel data cases where it is possible to observe both Y_0 and Y_1 for the same person.) The most commonly invoked treatment effect is the Average Treatment Effect (ATE): $\Delta^{\text{ATE}}(x) \equiv E(\Delta \mid X = x)$ where $\Delta = Y_1 - Y_0$. This is the effect of assigning treatment randomly to everyone of type X assuming full compliance, and ignoring general equilibrium effects. (See, e.g., Imbens (2004).) The average impact of treatment on persons who actually take the treatment is Treatment on the Treated (TT): $\Delta^{\text{TT}}(x) \equiv E(\Delta \mid X = x, D = 1)$. This parameter can also be defined conditional on $P(Z)$: $\Delta^{\text{TT}}(x, p) \equiv E(\Delta \mid X = x, P(Z) = p, D = 1)$. (These two definitions of treatment on the treated are related by integrating out the conditioning p variable: $\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{TT}}(x, p) dF_{P(Z)|X,D}(p|x, 1)$ where $F_{P(Z)|X,D}(\cdot|x, 1)$ is the distribution of $P(Z)$ given $X = x$ and $D = 1$.)

The mean effect of treatment on those for whom $X = x$ and $U_D = u_D$, the Marginal Treatment Effect (MTE), plays a fundamental role in the analysis of the next subsection:

$$\Delta^{\text{MTE}}(x, u_D) \equiv E(\Delta \mid X = x, U_D = u_D). \tag{7}$$

This parameter is defined independently of any instrument. I separate the definition of parameters from their identification. The MTE is the expected effect of treatment conditional on observed characteristics X and conditional on U_D, the unobservables from the first stage decision rule. For u_D evaluation points close to zero, $\Delta^{\text{MTE}}(x, u_D)$ is the expected effect of treatment on individuals with the value of unobservables that make them most likely to participate in treatment and who would participate even if the mean scale utility $\mu_D(Z)$ is small. If U_D is large, $\mu_D(Z)$ would have to be large to induce people to participate.

One can also interpret $E(\Delta \mid X = x, U_D = u_D)$ as the mean gain in terms of $Y_1 - Y_0$ for persons with observed characteristics X who would be indifferent between treatment or not if they were randomly assigned a value of Z, say z, such that $\mu_D(z) = u_D$. When Y_1 and Y_0 are value outcomes, MTE is a mean willingness-to-pay measure. MTE is a choice-theoretic building block that unites the treatment effect, selection, matching and control function literatures.

A third interpretation is that MTE conditions on X and the residual defined by subtracting the expectation of D^* from D^*: $\tilde{U}_D = D^* - E(D^* \mid Z, X)$. This is a "replacement function" interpretation in the sense of Heckman and Robb (1985a) and Matzkin (2007), or "control function" interpretation in the sense of Blundell and Powell (2003). (These three interpretations are equivalent under separability in D^*, i.e., when (6) characterizes the choice equation, but lead to three different definitions of MTE when a more general nonseparable model is developed. See Heckman and Vytlacil (2007b).) The additive separability of Eq. 6 in terms of observables and unobservables plays a crucial role in the justification of instrumental variable methods.

The LATE parameter of Imbens and Angrist (1994) is a version of MTE. I define LATE independently of any instrument after first presenting the Imbens–Angrist definition. Define $D(z)$ as a counterfactual choice variable, with $D(z) = 1$ if D would have been chosen if Z had been set to z, and $D(z) = 0$ otherwise. Let $\mathcal{Z}(x)$ denote the support of the distribution of Z conditional on $X = x$. For any $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ such that $P(z) > P(z')$, LATE is $E(\Delta \mid X = x, D(z) = 1, D(z') = 0) = E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0)$, the mean gain to persons who would be induced to switch from $D = 0$ to $D = 1$ if Z were

manipulated externally from z' to z. In an example of the returns to education, z' could be the base level of tuition and z a reduced tuition level. Using the latent index model, Heckman and Vytlacil (1999, 2005) show that LATE can be written as

$$\begin{aligned} E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0) \\ = E\left(Y_1 - Y_0 \mid X = x, u'_D < U_D < u_D\right) \\ = \Delta^{\text{LATE}}\left(x, u_D, u'_D\right) \end{aligned}$$

for $u_D = \Pr(D(z) = 1) = P(z)$, $u'_D = \Pr(D(z') = 1) = P(z')$, where assumption (A-1) implies that $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$ and $\Pr(D(z') = 1) = \Pr(D = 1 \mid Z = z')$.

Imbens and Angrist define the LATE parameter as the probability limit of an estimator. Their analysis conflates issues of definition of parameters with issues of identification. The representation of LATE given here allows analysts to separate these two conceptually distinct matters and to define the LATE parameter more generally. One can in principle evaluate the right hand side of the preceding equation at any u_D, u'_D points in the unit interval and not only at points in the support of the distribution of the propensity score $P(Z)$ conditional on $X = x$ where it is identified. From assumptions (A-1), (A-3), and (A-4), $\Delta^{\text{LATE}}\left(x, u_D, u'_D\right)$ is continuous in u_D and u'_D and $\lim_{u'_D \uparrow u_D} \Delta^{\text{LATE}}\left(x, u_D, u'_D\right) = \Delta^{\text{MTE}}(x, u_D)$. (This follows from Lebesgue's theorem for the derivative of an integral and holds almost everywhere with respect to Lebesgue measure. The ideas of the marginal treatment effect and the limit form of LATE were first introduced in the context of a parametric normal generalized Roy model by Björklund and Moffitt (1987), and were analyzed more generally in Heckman (1997). Angrist et al. (2000) also define and develop a limit form of LATE.)

Heckman and Vytlacil (1999) use assumptions (A-1)–(A-5) and the latent index structure to develop the relationship between MTE and the various treatment effect parameters shown in the first three lines of Table 1a. They present the formal derivation of the parameters and associated weights and graphically illustrates the relationship between ATE and TT. All treatment parameters may be expressed as weighted averages of the MTE:

$$\begin{aligned} &\text{Treatment Parameter } (j) \\ &= \int \Delta^{\text{MTE}}(x, u_D)\ \omega_j(x, u_D)\ du_D \end{aligned}$$

where $\omega_j(x, u_D)$ is the weighting function for the MTE and the integral is defined over the full support of u_D. Except for the OLS weights, the weights in the table all integrate to one, although in some cases the weights for IV may be negative (Heckman et al. 2006).

In Table 1a, $\Delta^{\text{TT}}(x)$ is shown as a weighted average of Δ^{MTE}:

$$\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D,$$

where

$$\omega_{TT}(x, u_D) = \frac{1 - F_{P|X}(u_D \mid x)}{\int_0^1 \left(1 - F_{P|X}(t \mid x)\right) dt} = \frac{S_{P|X}(u_D \mid x)}{E(P(Z) \mid X = x)}, \tag{8}$$

and $S_{P|X}(u_D \mid x)$ is $\Pr(P(Z) > u_D \mid X = x)$ and $\omega_{\text{TT}}(x, u_D)$ is a weighted distribution. The parameter $\Delta^{\text{TT}}(x)$ oversamples $\Delta^{\text{MTE}}(x, u_D)$ for those individuals with low values of u_D that make them more likely to participate in the program being evaluated. Treatment on the untreated (TUT) is defined symmetrically with TT and oversamples those least likely to participate. The various weights are displayed in Table 1b. A central theme of the analysis of Heckman and Vytlacil is that under their assumptions all estimators and estimands can be written as weighted averages of MTE. This allows them to unify the treatment effect literature using a common functional MTE (u_D).

Observe that if $E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(Y_1 - Y_0 \mid X = x)$, so $\Delta = Y_1 - Y_0$ is mean independent of U_D given $X = x$, then $\Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{LATE}}$. Therefore in cases where there is no heterogeneity in terms of unobservables in MTE (Δ constant conditional on $X = x$) or agents do not act on it so that U_D drops out of the conditioning set, marginal treatment effects are average treatment effects, so that all of the evaluation parameters are the same. Otherwise, they are different. Only in the case where the marginal treatment effect is the average treatment effect will the "effect" of treatment be uniquely defined.

Figure 2a plots weights for a parametric normal generalized Roy model generated from the parameters shown at the base of Fig. 2b. The model allows for costs to vary in the population and is more general than the extended Roy model used to construct Fig. 1. The weights for IV depicted in Fig. 2b are discussed in Heckman et al. (2006) and the weights for OLS are discussed in the next section. A high u_D is associated with higher cost, relative to return, and less likelihood of choosing $D = 1$. The decline of MTE in terms of higher values of u_D means that people with higher u_D have lower gross returns. TT overweights low values of u_D (i.e., it oversamples U_D that make it likely to have $D = 1$). ATE samples U_D uniformly. Treatment on the Untreated

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 1

(a) Treatment effects and estimands as weighted averages of the marginal treatment effect

$\text{ATE}(x) = E(Y_1 - Y_0 \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$

$\text{TT}(x) = E(Y_1 - Y_0 \mid X = x, D = 1) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$

$\text{TUT}(x) = E(Y_1 - Y_0 \mid X = x, D = 0) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TUT}}(x, u_D)\, du_D$

Policy relevant treatment effect $(x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{PRTE}}(x, u_D)\, du_D$ for two policies a and a' that affect the Z but not the X

$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}}^J(x, u_D)\, du_D$, given instrument J

$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{OLS}}(x, u_D)\, du_D$

(b) Weights (Heckman and Vytlacil 2005)

$\omega_{\text{ATE}}(x, u_D) = 1$

$\omega_{\text{TT}}(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x) dp\right] \frac{1}{E(P \mid X = x)}$

$\omega_{\text{TUT}}(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$

$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'},X}(u_D) - F_{P_a,X}(u_D)}{\Delta \bar{P}}\right]$

$\omega_{\text{IV}}^J(x, u_D) = \left[\int_{u_D}^1 (J(Z) - E(J(Z) \mid X = x)) \int f_{J,P|X}(j, t \mid X = x)\, dt\, dj\right] \frac{1}{\text{Cov}(J(Z), D \mid X = x)}$

$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1 \mid X = x, U_D = u_D)\, \omega_1(x, u_D) - E(U_0 \mid X = x, U_D = u_D)\, \omega_0(x, u_D)}{\Delta^{\text{MTE}}(x, u_D)}$

$\omega_1(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x)\, dp\right] \left[\frac{1}{E(P \mid X = x)}\right]$

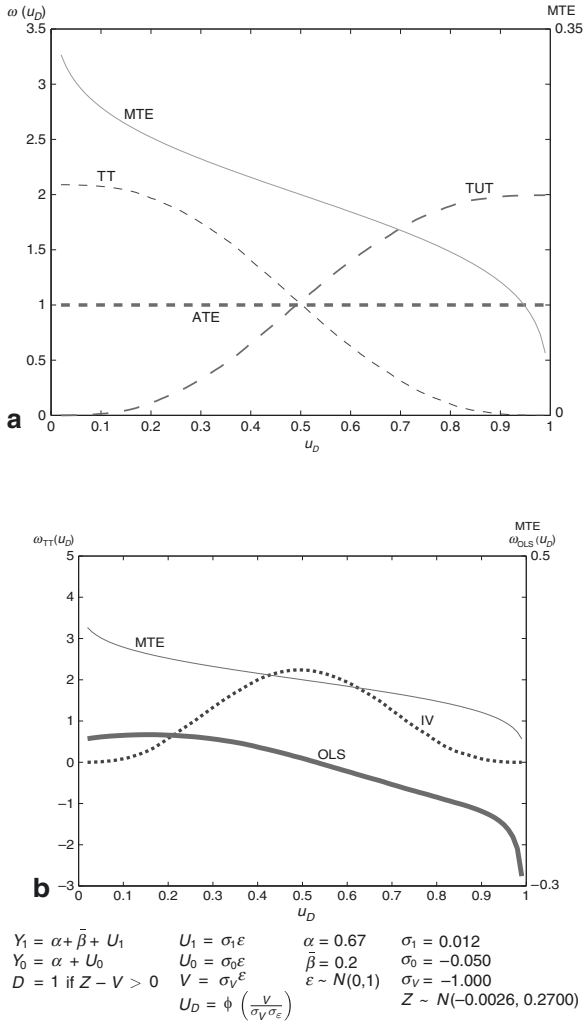
$\omega_0(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$

$(E(Y_1 - Y_0 \mid X = x, D = 0))$, or TUT, oversamples the values of U_D which make it unlikely to have $D = 1$.

Table 2 shows the treatment parameters produced from the different weighting schemes for the model used to generate the weights in Fig. 2a and 2b. Given the decline of the MTE in u_D, it is not surprising that TT>ATE>TUT. This is the generalized Roy version of the principle of diminishing returns. Those most likely to self select into the program benefit the most from it. The difference between TT and ATE is a sorting gain: $E(Y_1 - Y_0 \mid X, D = 1) - E(Y_1 - Y_0 \mid X)$, the average gain experienced by people who sort into treatment compared to what the average person would experience. Purposive selection on the basis of gains should lead to positive sorting gains of the kind found in the table. If there is negative sorting on the gains, then TUT ≥ ATE ≥ TT.

The Weights for a Generalized Roy Model

Heckman et al. (2006) show that all of the weights for treatment effects and IV estimators can be estimated over the available support. Since the MTE can be estimated by the method of Local Instrumental variables, we can form each treatment effect and each IV estimand as an integral to two estimable functions (subject to support). For the case of continuous Z, I plot the weights associated with the MTE for IV. This analysis draws on Heckman et al. (2006), who derive the weights. Figure 3 plots $E(Y \mid P(Z))$ and MTE for the extended Roy models generated by the parameters displayed at the base of the figure. In cases where $\beta \perp\!\!\!\perp D$, $\Delta^{\text{MTE}}(u_D)$ is constant in u_D. This is trivial when β is a constant. When β is random but selection into D does not depend on β, MTE is still flat. The more interesting case termed "essential heterogeneity" by



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 2(a) Weights for the marginal treatment effect for different parameters (Heckman and Vytlacil 2005) (**b**) Marginal treatment effect vs linear instrumental variables and ordinary least squares weights (Heckman and Vytlacil 2005)

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 2 Treatment parameters and estimands in the generalized Roy example

Treatment on the treated	0.2353
Treatment on the untreated	0.1574
Average treatment effect	0.2000
Sorting gain[a]	0.0353
Policy relevant treatment effect (PRTE)	0.1549
Selection bias[b]	−0.0628
Linear instrumental variables[c]	0.2013
Ordinary least squares	0.1725

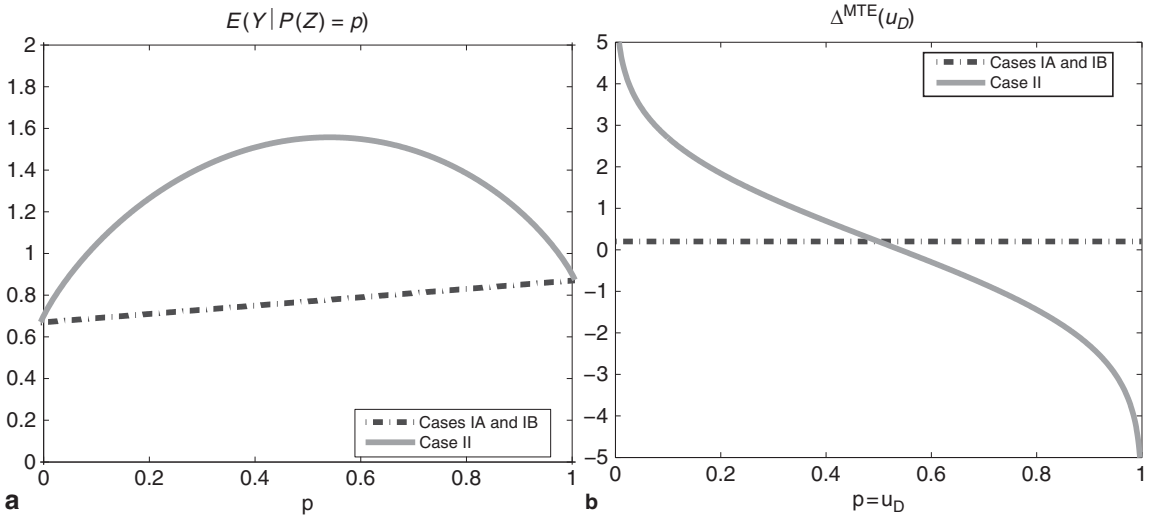
[a] $TT - ATE = E(Y_1 - Y_0 \mid D = 1) - E(Y_1 - Y_0)$
[b] $OLS - TT = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$
[c] Using Propensity Score $P(Z)$ as the instrument.
Note: The model used to create Table 2 is the same as those used to create Fig. 2a and b. The PRTE is computed using a policy t characterized as follows:
If $Z > 0$ then $D = 1$ if $Z(1 + t) - V > 0$.
If $Z \leq t$ then $D = 1$ if $Z - V > 0$.
For this example t is set equal to 0.2.

Heckman and Vytlacil has $\beta \not\perp\!\!\!\perp D$. The left hand side (Fig. 3a) depicts $E(Y \mid P(Z))$ in the two cases. The first case makes $E(Y \mid P(Z))$ linear in $P(Z)$. The second case is nonlinear in $P(Z)$. This arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y \mid P(Z))$ is presented in the right panel (Fig. 3b). It is a constant for the first case (flat MTE) but declining in $U_D = P(Z)$ for the case with selection on the gain. A simple test for linearity in $P(Z)$ in the outcome equation reveals whether or not the analyst is in cases I and II ($\beta \perp\!\!\!\perp D$) or case III ($\beta \not\perp\!\!\!\perp D$). (Recall that we keep the conditioning on X implicit.) These cases are the extended Roy counterparts to $E(Y \mid P(Z) = p)$ and MTE shown for the generalized Roy model in Figs. 4a and 4b.

MTE gives the mean marginal return for persons who have utility $P(Z) = u_D$. Thus, $P(Z) = u_D$ is the margin of indifference. Those with low u_D values have high returns. Those with high u_D values have low returns. Figure 3 highlights that in the general case MTE (and LATE) identify average returns for persons at the margin of indifference at different levels of the mean utility function ($P(Z)$).

Figure 5 plots MTE and LATE for different intervals of u_D using the model generating Fig. 3. LATE is the chord of $E(Y \mid P(Z))$ evaluated at different points. The relationship between LATE and MTE is depicted in the right panel (b) of Fig. 5. LATE is the integral under the MTE curve divided by the difference between the upper and lower limits.

Treatment parameters associated with the second case are plotted in Fig. 6. The MTE is the same as that presented in Fig. 3. ATE has the same value for all p. The effect of treatment on the treated for $P(Z) = p$, $\Delta^{TT}(p) = E(Y_1 - Y_0 \mid D = 1, P(Z) = p)$ declines in p (equivalently it declines in u_D). Treatment on the untreated given p, $\text{TUT}(p) = \Delta^{TUT}(p) = E(Y_1 - Y_0 \mid D = 0, P(Z) = p)$ also declines in p.



Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice model

$$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$$

Case IA	Case IB	Case II
$U_1 = U_0$	$U_1 - U_0 \perp\!\!\!\perp D$	$U_1 - U_0 \not\perp\!\!\!\perp D$
$\bar{\beta}$ =ATE=TT=TUT=IV	$\bar{\beta}$ =ATE=TT=TUT=IV	$\bar{\beta}$ =ATE≠TT≠TUT≠IV

Parameterization

Cases IA, IB, and II	Cases IB and II	Case II
$\alpha = 0.67$	$(U_1, U_0) \sim N(\mathbf{0}, \Sigma)$	$D^* = Y_1 - Y_0 - \gamma Z$
$\bar{\beta} = 0.2$	with $\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$	$Z \sim N(\mu_Z, \Sigma_Z)$
		$\mu_Z = (2, -2)$ and $\Sigma_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$
		$\gamma = (0.5, 0.5)$

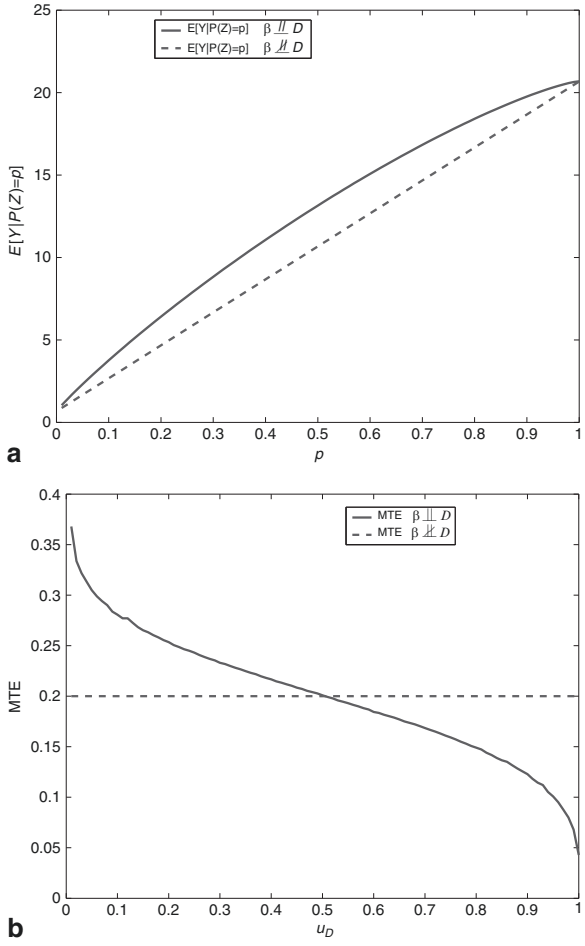
Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 3 Conditional expectation of Y on $P(Z)$ and the marginal treatment effect (MTE) the extended Roy economy (Heckman et al. 2006)

$$LATE(p, p') = \frac{\Delta^{TT}(p')p' - \Delta^{TT}(p)p}{p' - p}, \qquad p' \neq p$$

$$MTE = \frac{\partial[\Delta^{TT}(p)p]}{\partial p}.$$

One can generate all of the treatment parameters from $\Delta^{TT}(p)$.

Matching on $P = p$ (which is equivalent to nonparametric regression given $P = p$) produces a biased estimator of $TT(p)$. Matching assumes a flat MTE (average return



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 4 (**a**) Plot of the $E(Y|P(Z) = p)$, (**b**) Plot of the identified marginal treatment effect from Fig. 2a (the derivative). Note: Parameters for the general heterogeneous case are the same as those used in Fig. 2a and 2b. For the homogeneous case we impose $U_1 = U_0$ ($\sigma_1 = \sigma_0 = 0.012$). (Heckman and Vytlacil 2005)

equals marginal return) as we develop below. Therefore it is systematically biased for $\Delta^{TT}(p)$ in a model with essential heterogeneity, where $\beta \not\perp\!\!\!\perp D$. Making observables alike makes the unobservables dissimilar. Holding p constant across treatment and control groups understates $TT(p)$ for low values of p and overstates it for high values of p. I develop this point further in section "►Matching", where I discuss the method of matching. First I present a unified approach that integrates all evaluation estimators in a common framework.

The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators

This section reviews the main principles underlying the evaluation estimators commonly used in the econometric literature. I assume two potential outcomes (Y_0, Y_1). $D = 1$ if Y_1 is observed, and $D = 0$ corresponds to Y_0 being observed. The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{9}$$

The *evaluation problem* arises because for each person we observe either Y_0 or Y_1 but not both. Thus in general it is not possible to identify the individual level treatment effect $Y_1 - Y_0$ for any person. The typical solution to this problem is to reformulate the problem at the population level rather than at the individual level and to identify certain mean outcomes or quantile outcomes or various distributions of outcomes as described in Heckman and Vytlacil (2007a). For example, a commonly used approach focuses attention on average treatment effects, such as ATE $= E(Y_1 - Y_0)$.

If treatment is assigned or chosen on the basis of potential outcomes, so

$$(Y_0, Y_1) \not\perp\!\!\!\perp D,$$

where $\not\perp\!\!\!\perp$ denotes "is not independent" and "$\perp\!\!\!\perp$" denotes independent, we encounter the problem of selection bias. Suppose that we observe people in each treatment state $D = 0$ and $D = 1$. If $Y_j \not\perp\!\!\!\perp D$, then the observed Y_j will be selectively different from randomly assigned Y_j, $j = 0, 1$. Thus $E(Y_0 \mid D = 0) \neq E(Y_0)$ and $E(Y_1 \mid D = 1) \neq E(Y_1)$. Using unadjusted data to construct $E(Y_1 - Y_0)$ will produce one source of evaluation bias:

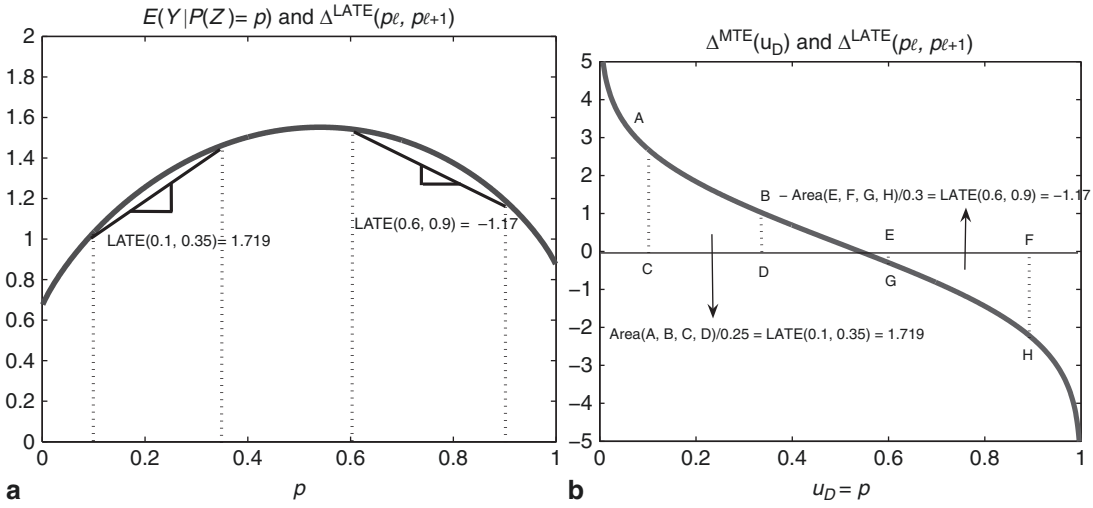
$$E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) \neq E(Y_1 - Y_0).$$

The selection problem underlies the evaluation problem. Many methods have been proposed to solve both problems.

The method with the greatest intuitive appeal, which is sometimes called the "gold standard" in evaluation analysis, is the method of random assignment. Nonexperimental methods can be organized by how they attempt to approximate what can be obtained by an ideal random assignment. If treatment is chosen at random with respect to (Y_0, Y_1), or if treatments are randomly assigned and there is full compliance with the treatment assignment,

(R-1) $(Y_0, Y_1) \perp\!\!\!\perp D$.

It is useful to distinguish several cases where (R-1) will be satisfied. The first is that agents (decision makers whose choices are being analyzed) pick outcomes that are random with respect to (Y_0, Y_1). Thus agents may not know



$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E(Y|P(Z) = p_{\ell+1}) - E(Y|P(Z) = p_\ell)}{p_{\ell+1} - p_\ell} = \frac{\int_{p_\ell}^{p_{\ell+1}} \Delta^{\text{MTE}}(u_D)du_D}{p_{\ell+1} - p_\ell}$$

$$\Delta^{\text{LATE}}(0.6, 0.9) = -1.17$$
$$\Delta^{\text{LATE}}(0.1, 0.35) = 1.719$$

Outcomes	Choice model
$Y_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
$Y_0 = \alpha + U_0$	with $D^* = Y_1 - Y_0 - \gamma Z$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \Sigma) \text{ and } Z \sim N(\mu_Z, \Sigma_Z)$$
$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \ \mu_Z = (2, -2) \text{ and } \Sigma_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$$
$$\alpha = 0.67, \bar{\beta} = 0.2, \gamma = (0.5, 0.5)$$

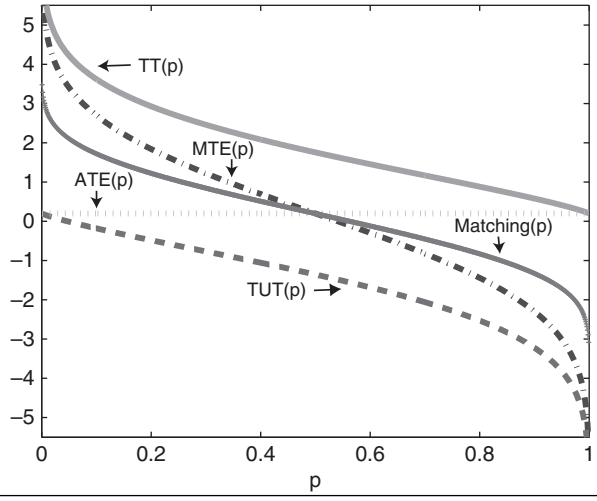
Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 5 The local average treatment effect the extended Roy economy (Heckman et al. 2006)

(Y_0, Y_1) at the time they make their choices to participate in treatment or at least do not act on (Y_0, Y_1), so that $\Pr(D = 1 \mid X, Y_0, Y_1) = \Pr(D = 1 \mid X)$ for all X. Matching assumes a version of (R-1) conditional on matching variables X: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.

A second case arises when individuals are randomly assigned to treatment status even if they would choose to self select into no-treatment status, and they comply with the randomization protocols. Let ξ be randomized assignment status. With full compliance, $\xi = 1$ implies that Y_1 is observed and $\xi = 0$ implies that Y_0 is observed. Then, under randomized assignment,

(R-2) $(Y_0, Y_1) \perp\!\!\!\perp \xi$,

even if in a regime of self-selection, $(Y_0, Y_1) \not\perp\!\!\!\perp D$. If ►randomization is performed conditional on X, we obtain $(Y_0, Y_1) \perp\!\!\!\perp \xi \mid X$.



Parameter	Definition	Under assumptions[a]
Marginal treatment effect	$E\left[Y_1 - Y_0 \mid D^* = 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\Phi^{-1}(1-p)$
Average treatment effect	$E\left[Y_1 - Y_0 \mid P(Z) = p\right]$	$\bar{\beta}$
Treatment on the treated	$E\left[Y_1 - Y_0 \mid D^* > 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{p}$
Treatment on the untreated	$E\left[Y_1 - Y_0 \mid D^* \leq 0, P(Z) = p\right]$	$\bar{\beta} - \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{1-p}$
OLS/matching on $P(Z)$	$E\left[Y_1 \mid D^* > 0, P(Z) = p\right]$	$\bar{\beta} + \left(\frac{\sigma^2_{U_1} - \sigma_{U_1 U_0}}{\sqrt{\sigma_{U_1 - U_0}}}\right)\left(\frac{1-2p}{p(1-p)}\right)\phi\left(\Phi^{-1}(1-p)\right)$
	$-E\left[Y_0 \mid D^* \leq 0, P(Z) = p\right]$	

Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 6 Treatment parameters and OLS/matching as a function of $P(Z) = p$. Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$. [a] The model in this case is the same as the one presented below Fig. 5. (Heckman et al. 2006)

Let A denote actual treatment status. If the randomization has full compliance among participants, $\xi = 1 \Rightarrow A = 1$; $\xi = 0 \Rightarrow A = 0$. This is entirely consistent with a regime in which a person would choose $D = 1$ in the absence of randomization, but would have no treatment ($A = 0$) if suitably randomized, even though the agent might desire treatment.

If treatment status is chosen by self-selection, $D = 1 \Rightarrow A = 1$ and $D = 0 \Rightarrow A = 0$. If there is imperfect compliance with randomization, $\xi = 1 \not\Rightarrow A = 1$ because of agent choices. In general, $A = \xi D$ so that $A = 1$ only if $\xi = 1$ and $D = 1$. If treatment status is randomly assigned, either through randomization or randomized self-selection,

(R-3) $(Y_0, Y_1) \perp\!\!\!\perp A$.

This version of randomization can also be defined conditional on X. Under (R-1), (R-2), or (R-3), the average treatment effect (ATE) is the same as the marginal treatment effect of Björklund and Moffitt (1987) and Heckman and Vytlacil (1999, 2005, 2007a), and the parameters treatment on the treated (TT) ($E(Y_1 - Y_0 \mid D = 1)$) and treatment on the untreated (TUT) ($E(Y_1 - Y_0 \mid D = 0)$). (The marginal treatment effect is formally defined in the next section.) These parameters can be identified from population means:

$$\text{TT} = \text{MTE} = \text{TUT} = \text{ATE} = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

Forming averages over populations of persons who are treated ($A = 1$) or untreated ($A = 0$) suffices to identify this parameter. If there are conditioning variables X, we can define the mean treatment parameters for all X where (R-1) or (R-2) or (R-3) hold.

Observe that even with random assignment of treatment status and full compliance, one cannot, in general, identify the distribution of the treatment effects

$(Y_1 - Y_0)$, although one can identify the marginal distributions $F_1(Y_1 \mid A = 1, X = x) = F_1(Y_1 \mid X = x)$ and $F_0(Y_0 \mid A = 0, X = x) = F_0(Y_0 \mid X = x)$. One special assumption, common in the conventional econometrics literature, is that $Y_1 - Y_0 = \Delta(x)$, a constant given x. Since $\Delta(x)$ can be identified from $E(Y_1 \mid A = 1, X = x) - E(Y_0 \mid A = 0, X = x)$ because A is randomly allocated, in this special case the analyst can identify the joint distribution of (Y_0, Y_1). (Heckman (1992); Heckman et al. (1997).) This approach assumes that (Y_0, Y_1) have the same distribution up to a parameter Δ (Y_0 and Y_1 are perfectly dependent). One can make other assumptions about the dependence across ranks from perfect positive or negative ranking to independence. (Heckman et al. (1997).) The joint distribution of (Y_0, Y_1) or of $(Y_1 - Y_0)$ is not identified unless the analyst can pin down the dependence across (Y_0, Y_1). Thus, even with data from a randomized trial one cannot, without further assumptions, identify the proportion of people who benefit from treatment in the sense of gross gain ($\Pr(Y_1 \geq Y_0)$). This problem plagues all evaluation methods. Abbring and Heckman (2007) discuss methods for identifying joint distributions of outcomes. (See also Aakvik et al. (2005); Carneiro et al. (2001, 2003); and Cunha et al. (2005).)

Assumption (R-1) is very strong. In many cases, it is thought that there is *selection bias* with respect to Y_0, Y_1, so persons who select into status 1 or 0 are selectively different from randomly sampled persons in the population. The assumption most commonly made to circumvent problems with (R-1) is that even though D is not random with respect to potential outcomes, the analyst has access to control variables X that effectively produce a randomization of D with respect to (Y_0, Y_1) given X. This is the method of matching, which is based on the following conditional independence assumption:

(M-1) $\quad (Y_0, Y_1) \perp\!\!\!\perp D \mid X.$

Conditioning on X randomizes D with respect to (Y_0, Y_1). (M-1) assumes that any selective sampling of (Y_0, Y_1) can be adjusted by conditioning on observed variables. (R-1) and (M-1) are different assumptions and neither implies the other. In a linear equations model, assumption (M-1) that D is independent from (Y_0, Y_1) given X justifies application of ▶least squares on D to eliminate selection bias in mean outcome parameters. For means, matching is just nonparametric regression. (Barnow et al. (1980) present one application of matching in a regression setting.) In order to be able to compare X-comparable people in the treatment regime one must assume

(M-2) $\quad 0 < \Pr(D = 1 \mid X = x) < 1.$

Assumptions (M-1) and (M-2) justify matching. Assumption (M-2) is required for *any* evaluation estimator that compares treated and untreated persons. It is produced by random assignment if the randomization is conducted for all $X = x$ and there is full compliance.

Observe that from (M-1) and (M-2), it is possible to identify $F_1(Y_1 \mid X = x)$ from the observed data $F_1(Y_1 \mid D = 1, X = x)$ since we observe the left hand side of

$$F_1(Y_1 \mid D = 1, X = x) = F_1(Y_1 \mid X = x) = F_1(Y_1 \mid D = 0, X = x).$$

The first equality is a consequence of conditional independence assumption (M-1). The second equality comes from (M-1) and (M-2). By a similar argument, we observe the left hand side of

$$F_0(Y_0 \mid D = 0, X = x) = F_0(Y_0 \mid X = x) = F_0(Y_0 \mid D = 1, X = x),$$

and the equalities are a consequence of (M-1) and (M-2). Since the pair of outcomes (Y_0, Y_1) is not identified for anyone, as in the case of data from randomized trials, the joint distributions of (Y_0, Y_1) given X or of $Y_1 - Y_0$ given X are not identified without further information. This is a problem that plagues all selection estimators.

From the data on Y_1 given X and $D = 1$ and the data on Y_0 given X and $D = 0$, since $E(Y_1 \mid D = 1, X = x) = E(Y_1 \mid X = x) = E(Y_1 \mid D = 0, X = x)$ and $E(Y_0 \mid D = 0, X = x) = E(Y_0 \mid X = x) = E(Y_0 \mid D = 1, X = x)$ we obtain

$$E(Y_1 - Y_0 \mid X = x) = E(Y_1 - Y_0 \mid D = 1, X = x) = E(Y_1 - Y_0 \mid D = 0, X = x).$$

Effectively, we have a randomization for the subset of the support of X satisfying (M-2).

At values of X that fail to satisfy (M-2), there is no variation in D given X. One can define the residual variation in D not accounted for by X as

$$\mathcal{E}(x) = D - E(D \mid X = x) = D - \Pr(D = 1 \mid X = x).$$

If the variance of $\mathcal{E}(x)$ is zero, it is not possible to construct contrasts in outcomes by treatment status for those X values and (M-2) is violated. To see the consequences of this violation in a regression setting, use $Y = Y_0 + D(Y_1 - Y_0)$ and take conditional expectations, under (M-1), to obtain

$$E(Y \mid X, D) = E(Y_0 \mid X) + D[E(Y_1 - Y_0 \mid X)].$$

This follows because $E(Y \mid X, D) = E(Y_0 \mid X, D) + DE(Y_1 - Y_0 \mid X, D)$ but from (M-1), $E(Y_0 \mid X, D) = E(Y_0 \mid X)$ and $E(Y_1 - Y_0 \mid X, D) = E(Y_1 - Y_0 \mid X)$. If $Var(\mathcal{E}(x)) > 0$

for all x in the support of X, one can use nonparametric least squares to identify $E(Y_1 - Y_0 \mid X = x) = \text{ATE}(x)$ by regressing Y on D and X. The function identified from the coefficient on D is the average treatment effect. (Under the conditional independence assumption (M-1), it is also the effect of treatment on the treated $E(Y_1 - Y_0 \mid X, D = 1)$ and the marginal treatment effect formally defined in the next section.) If $Var(\mathcal{E}(x)) = 0$, $\text{ATE}(x)$ is not identified at that x value because there is no variation in D that is not fully explained by X. A special case of matching is linear least squares where one can write

$$Y_0 = X\alpha + U \qquad Y_1 = X\alpha + \beta + U,$$

$U_0 = U_1 = U$ and hence under (M-1),

$$E(Y \mid X, D) = X\alpha + \beta D.$$

If D is perfectly predictable by X, one cannot identify β because of a multicollinearity problem (see ▶Multicollinearity). (M-2) rules out perfect collinearity. (Clearly (M-1) and (M-2) are sufficient but not necessary conditions. For the special case of OLS, as a consequence of the assumed linearity in the functional form of the estimating equation, we achieve identification of β if $Cov(X, U) = 0$, $Cov(D, U) = 0$ and (D, X) are not perfectly collinear. These conditions are much weaker than (M-1) and (M-2) and can be satisfied if (M-1) and (M-2) are only identified in a subset of the support of X.) Matching is a nonparametric version of least squares that does not impose functional form assumptions on outcome equations, and that imposes support condition (M-2).

Conventional econometric choice models make a distinction between variables that appear in outcome equations (X) and variables that appear in choice equations (Z). The same variables may be in (X) and (Z) but more typically, there are some variables not in common. For example, the instrumental variable estimator is based on variables that are not in X but that are in Z. Matching makes no distinction between the X and the Z. (Heckman et al. (1998) distinguish X and Z in matching. They consider a case where conditioning on X may lead to failure of (M-1) and (M-2) but conditioning on (X, Z) satisfies a suitably modified version of this condition.) It does not rely on exclusion restrictions. The conditioning variables used to achieve conditional independence can in principle be a set of variables Q distinct from the X variables (covariates for outcomes) or the Z variables (covariates for choices). I use X solely to simplify the notation. The key identifying assumption is the assumed existence of a random variable X with the properties satisfying (M-1) and (M-2).

Conditioning on a larger vector (X augmented with additional variables) or a smaller vector (X with some components removed) may or may not produce suitably modified versions of (M-1) and (M-2). Without invoking further assumptions there is no objective principle for determining what conditioning variables produce (M-1).

Assumption (M-1) is strong. Many economists do not have enough faith in their data to invoke it. Assumption (M-2) is testable and requires no act of faith. To justify (M-1), it is necessary to appeal to the quality of the data.

Using economic theory can help guide the choice of an evaluation estimator. A crucial distinction is the one between the information available to the analyst and the information available to the agent whose outcomes are being studied. Assumptions made about these information sets drive the properties of econometric estimators. Analysts using matching make strong informational assumptions in terms of the data available to them. In fact, all econometric estimators make assumptions about the presence or absence of informational asymmetries, and I exposit them in this paper.

To analyze the informational assumptions invoked in matching, and other econometric evaluation strategies, it is helpful to introduce five distinct information sets and establish some relationships among them. (See also the discussion in Barros (1987), Heckman and Navarro (2004), and Gerfin and Lechner (2002).) (1) An information set $\sigma(I_R)$ with an associated random variable that satisfies conditional independence (M-1) is defined as a *relevant* information set; (2) The minimal information set $\sigma(I_R)$ with associated random variable needed to satisfy conditional independence (M-1), the *minimal relevant* information set; (3) The information set $\sigma(I_A)$ available to the agent at the time decisions to participate are made; (4) The information available to the economist, $\sigma(I_{E^*})$; and (5) The information $\sigma(I_E)$ used by the economist in conducting an empirical analysis. I will denote the random variables generated by these sets as $I_{R^*}, I_R, I_A, I_{E^*}, I_E$, respectively. (I start with a primitive probability space (Ω, σ, P) with associated random variables I. I assume minimal σ-algebras and assume that the random variables I are measurable with respect to these σ-algebras. Obviously, strictly monotonic or affine transformations of the I preserve the information and can substitute for the I.)

Definition 1 *Define $\sigma(I_{R^*})$ as a relevant information set if the information set is generated by the random variable I_{R^*}, possibly vector valued, and satisfies condition (M-1), so*

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid I_{R^*}.$$

Definition 2 *Define $\sigma(I_R)$ as a minimal relevant information set if it is the intersection of all sets $\sigma(I_{R^*})$ and satisfies $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_R$. The associated random variable I_R is a minimum amount of information that guarantees that condition (M-1) is satisfied. There may be no such set.* (Observe that the intersection of all sets $\sigma(I_{R^*})$ may be empty and hence may not be characterized by a (possibly vector valued) random variable I_R that guarantees $(Y_1, Y_2) \perp\!\!\!\perp D \mid I_R$. If the information sets that produce conditional independence are nested, then the intersection of all sets $\sigma(I_{R^*})$ producing conditional independence is well defined and has an associated random variable I_R with the required property, although it may not be unique (e.g., strictly monotonic transformations and affine transformations of I_R also preserve the property). In the more general case of non-nested information sets with the required property, it is possible that no uniquely defined minimal relevant set exists. Among collections of nested sets that possess the required property, there is a minimal set defined by intersection but there may be multiple minimal sets corresponding to each collection.)

If one defines the relevant information set as one that produces conditional independence, it may not be unique. If the set $\sigma(I_{R^*})$ satisfies the conditional independence condition, then the set $\sigma(I_{R^*}, Q)$ such that $Q \perp\!\!\!\perp (Y_0, Y_1) \mid I_{R^*}$ would also guarantee conditional independence. For this reason, I define the relevant information set to be minimal, that is, to be the intersection of all relevant sets that still produce conditional independence between (Y_0, Y_1) and D. However, no minimal set may exist.

Definition 3 *The agent's information set, $\sigma(I_A)$, is defined by the information I_A used by the agent when choosing among treatments. Accordingly, call I_A the agent's information.*

By the agent I mean the person making the treatment decision not necessarily the person whose outcomes are being studied (e.g., the agent may be the parent; the person being studied may be a child).

Definition 4 *The econometrician's full information set, $\sigma(I_{E^*})$, is defined as all of the information available to the econometrician, I_{E^*}.*

Definition 5 *The econometrician's information set, $\sigma(I_E)$, is defined by the information used by the econometrician when analyzing the agent's choice of treatment, I_E, in conducting an analysis.*

For the case where a unique minimal relevant information set exists, only three restrictions are implied by the structure of these sets: $\sigma(I_R) \subseteq \sigma(I_{R^*})$, $\sigma(I_R) \subseteq \sigma(I_A)$, and $\sigma(I_E) \subseteq \sigma(I_{E^*})$. (This formulation assumes that the agent makes the treatment decision. The extension to the case where the decision maker and the agent are distinct is straightforward. The requirement $\sigma(I_R) \subseteq \sigma(I_{R^*})$ is satisfied by nested sets.) I have already discussed the first restriction. The second restriction requires that the minimal relevant information set must be part of the information the agent uses when deciding which treatment to take or assign. It is the information in $\sigma(I_A)$ that gives rise to the selection problem which in turn gives rise to the evaluation problem.

The third restriction requires that the information used by the econometrician must be part of the information that the agent observes. Aside from these orderings, the econometrician's information set may be different from the agent's or the relevant information set. The econometrician may know something the agent doesn't know, for typically he is observing events after the decision is made. At the same time, there may be private information known to the agent but not the econometrician. Matching assumption (M-1) implies that $\sigma(I_R) \subseteq \sigma(I_E)$, so that the econometrician uses at least the minimal relevant information set, but of course he or she may use more. However, using more information is not guaranteed to produce a model with conditional independence property (M-1) satisfied for the augmented model. Thus an analyst can "overdo" it. Heckman and Navarro (2004) and Abbring and Heckman (2007) present examples of the consequences of the asymmetry in agent and analyst information sets.

The possibility of asymmetry in information between the agent making participation decisions and the observing economist creates the potential for a major identification problem that is ruled out by assumption (M-1). The methods of control functions and instrumental variables estimators (and closely related regression discontinuity design methods) address this problem but in different ways. Accounting for this possibility is a more conservative approach to the selection problem than the one taken by advocates of ▶least squares, or its nonparametric counterpart, matching. Those advocates assume that they know the X that produces a relevant information set. Heckman and Navarro (2004) show the biases that can result in matching when standard econometric model selection criteria are applied to pick the X that are used to satisfy (M-1). Conditional independence condition (M-1) cannot be tested without maintaining other assumptions. (These assumptions may or may not be testable. The required "exogeneity" conditions are discussed in Heckman and Navarro (2004). Thus randomization of assignment of treatment status might be used to test (M-1) but this requires that there be full compliance and that the

randomization be valid (no anticipation effects or general equilibrium effects).) Choice of the appropriate conditioning variables is a problem that plagues *all* econometric estimators.

The methods of control functions, replacement functions, proxy variables, and instrumental variables all recognize the possibility of asymmetry in information between the agent being studied and the econometrician and recognize that even after conditioning on X (variables in the outcome equation) and Z (variables affecting treatment choices, which may include the X), analysts may fail to satisfy conditional independence condition (M-1). (The term and concept of control function is due to Heckman and Robb (1985a,b, 1986a,b). See Blundell and Powell (2003) (who call the Heckman and Robb replacement functions control functions). A more recent nomenclature is "control variate." Matzkin (2007) provides a comprehensive discussion of identification principles for econometric estimators.) These methods postulate the existence of some unobservables θ, which may be vector valued, with the property that

(U-1) $(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta,$

but allow for the possibility that

(U-2) $(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z.$

In the event (U-2) holds, these approaches model the relationships of the unobservable θ with Y_1, Y_0 and D in various ways. The content in the control function principle is to specify the exact nature of the dependence on the relationship between observables and unobservables in a nontrivial fashion that is consistent with economic theory. Heckman and Navarro present examples of models that satisfy (U-1) but not (U-2).

The early literature focused on mean outcomes conditional on covariates (Heckman and Robb 1985a, b, 1986a, b) and assumes a weaker version of (U-1) based on conditional mean independence rather than full conditional independence. More recent work analyzes distributions of outcomes (e.g., Aakvik et al. 2005; Carneiro et al. 2001, 2003). This work is reviewed in Abbring and Heckman (2007).

The normal Roy selection model makes distributional assumptions and identifies the joint distribution of outcomes. A large literature surveyed by Matzkin (2007) makes alternative assumptions to satisfy (U-1) in nonparametric settings. Replacement functions (Heckman and Robb 1985a) are methods that proxy θ. They substitute out for θ using observables. (This is the "control variate" of Blundell and Powell (2003). Heckman and Robb (1985a) and Olley and Pakes (1996) use a similar idea. Matzkin (2007) discusses replacement functions.) Aakvik et al. (1999, 2005), Carneiro et al. (2001, 2003), Cunha et al. (2005), and Cunha et al. (2006, 2010) develop methods that integrate out θ from the model assuming $\theta \perp\!\!\!\perp (X, Z)$, or invoking weaker mean independence assumptions, and assuming access to proxy measurements for θ. They also consider methods for estimating the distributions of treatment effects. These are discussed in Abbring and Heckman (2007).

The normal selection model produces partial identification of a generalized Roy model and full identification of a Roy model under separability and normality. It models the conditional expectation of U_0 and U_1 given X, Z, D. In terms of (U-1), it models the conditional mean dependence of Y_0, Y_1 on D and θ given X and Z. Powell (1994) and Matzkin (2007) survey methods for producing semiparametric versions of these models. Heckman and Vytlacil (2007a, Appendix B) or the appendix of Heckman and Navarro (2007) present a prototypical identification proof for a general selection model that implements (U-1) by estimating the distribution of θ, assuming $\theta \perp\!\!\!\perp (X, Z)$, and invoking support conditions on (X, Z).

Central to both the selection approach and the instrumental variable approach for a model with heterogenous responses is the probability of selection. This is an integral part of the Roy model previously discussed. Let Z denote variables in the choice equation. Fixing Z at different values (denoted z), I define $D(z)$ as an indicator function that is "1" when treatment is selected at the fixed value of z and that is "0" otherwise. In terms of a separable index model $U_D = \mu_D(Z) - V$, for a fixed value of z,

$$D(z) = \mathbf{1}[\mu_D(z) \geq V]$$

where $Z \perp\!\!\!\perp V \mid X$. Thus fixing $Z = z$, values of z do not affect the realizations of V for any value of X. An alternative way of representing the independence between Z and V given X due to Imbens and Angrist (1994), writes that $D(z) \perp\!\!\!\perp Z$ for all $z \in \mathcal{Z}$, where $\mathcal{Z}$ is the support of Z. The Imbens and Angrist independence condition for IV is

$$\{D(z)\}_{z \in \mathcal{Z}} \perp\!\!\!\perp Z \mid X.$$

Thus the probabilities that $D(z) = 1, z \in \mathcal{Z}$ are not affected by the occurrence of Z. Vytlacil (2002) establishes the equivalence of these two formulations under general conditions. (See Heckman and Vytlacil (2007b) for a discussion of these conditions.)

The method of instrumental variables (IV) postulates that

(IV-1) $\left(Y_0, Y_1, \{D(z)\}_{z \in \mathcal{Z}}\right) \perp\!\!\!\perp Z \mid X.$ (*Independence*)

One consequence of this assumption is that $E(D \mid Z) = P(Z)$, the propensity score, is random with respect to potential outcomes. Thus $(Y_0, Y_1) \perp\!\!\!\perp P(Z) \mid X$. So are all other functions of Z given X. The method of instrumental variables also assumes that

(IV-2) $E(D \mid X, Z) = P(X, Z)$ *is a nondegenerate function of* Z *given* X. (*Rank Condition*)

Alternatively, one can write that $Var\left(E\left(D \mid X, Z\right)\right) \neq Var\left(E\left(D \mid X\right)\right)$.

Comparing (IV-1) to (M-1) in the method of instrumental variables, Z is independent of (Y_0, Y_1) given X whereas in matching D is independent of (Y_0, Y_1) given X. So in (IV-1), Z plays the role of D in matching condition (M-1). Comparing (IV-2) with (M-2), in the method of IV the choice probability $\Pr(D = 1 \mid X, Z)$ is assumed to vary with Z conditional on X, whereas in matching, D varies conditional on X. Unlike the method of control functions, no explicit model of the relationship between D and (Y_0, Y_1) is required in applying IV.

(IV-2) is a rank condition and can be empirically verified. (IV-1) is not testable as it involves assumptions about counterfactuals. In a conventional common coefficient regression model

$$Y = \alpha + \beta D + U,$$

where β is a constant and where I allow for $Cov(D, U) \neq 0$, (IV-1) and (IV-2) identify β. $\left(\beta = \frac{Cov(Z,Y)}{Cov(Z,D)}.\right)$ When β varies in the population and is correlated with D, additional assumptions must be invoked for IV to identify interpretable parameters. Heckman et al. (2006) and Heckman and Vytlacil (2007b) discuss these conditions.

Assumptions (IV-1) and (IV-2), with additional assumptions in the case where β varies in the population which I discuss in this paper, can be used to identify mean treatment parameters. Replacing Y_1 with $\mathbf{1}\left(Y_1 \leq t\right)$ and Y_0 with $\mathbf{1}\left(Y_0 \leq t\right)$, where t is a constant, the IV approach allows us to identify marginal distributions $F_1(y_1 \mid X)$ or $F_0(y_0 \mid X)$.

In matching, the variation in D that arises after conditioning on X provides the source of randomness that switches people across treatment status. Nature is assumed to provide an experimental manipulation conditional on X that replaces the randomization assumed in (R-1)–(R-3). When D is perfectly predictable by X, there is no variation in it conditional on X, and the randomization assumed to be given by nature in the matching model breaks down. Heuristically, matching assumes a residual $\mathcal{E}\left(X\right) = D - E(D \mid X)$ that is nondegenerate and is one manifestation of the randomness that causes persons to switch status. (It is heuristically illuminating, but technically incorrect to replace $\mathcal{E}\left(X\right)$ with D in (R-1) or R in (R-2) or T in (R-3). In general $\mathcal{E}\left(X\right)$ is not independent of X even if it is mean independent.)

In the IV method, it is the choice probability $E(D \mid X, Z) = P\left(X, Z\right)$ that is random with respect to (Y_0, Y_1), not components of D not predictable by (X, Z). Variation in Z for a fixed X provides the required variation in D that switches treatment status and still produces the required conditional independence:

$$(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) \mid X.$$

Variation in $P(X, Z)$ produces variations in D that switch treatment status. Components of variation in D not predictable by (X, Z) do not produce the required independence. Instead, the predicted component provides the required independence. It is just the opposite in matching. Versions of the method of control functions use measurements to proxy θ in (U-1) and (U-2) and remove spurious dependence that gives rise to selection problems. These are called replacement functions (see Heckman and Robb 1985a) or control variates (see Blundell and Powell 2003).

The methods of replacement functions and proxy variables all start from characterizations (U-1) and (U-2). θ is not observed and (Y_0, Y_1) are not observed directly but Y is observed:

$$Y = DY_1 + (1 - D)\, Y_0.$$

Missing variables θ produce selection bias which creates a problem with using observational data to evaluate social programs. From (U-1), if one conditions on θ, condition (M-1) for matching would be satisfied, and hence one could identify the parameters and distributions that can be identified if the conditions required for matching are satisfied.

The most direct approach to controlling for θ is to assume access to a function $\tau(X, Z, Q)$ that perfectly proxies θ:

$$\theta = \tau(X, Z, Q). \tag{10}$$

This approach based on a perfect proxy is called the method of replacement functions by Heckman and Robb (1985a). In (U-1), one can substitute for θ in terms of observables (X, Z, Q). Then

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, Q.$$

It is possible to condition nonparametrically on (X, Z, Q) and without having to know the exact functional form of τ. θ can be a vector and τ can be a vector of functions. This method has been used in the economics of education for decades (see the references in Heckman and Robb 1985a). If θ is ability and τ is a test score, it is sometimes assumed

that the test score is a perfect proxy (or replacement function) for θ and that one can enter it into the regressions of earnings on schooling to escape the problem of ability bias (typically assuming a linear relationship between τ and θ). (Thus if $\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, one can write

$$\theta = \tau - \alpha_0 - \alpha_1 X - \alpha_2 Q - \alpha_3 Z,$$

and use this as the proxy function. Controlling for T, X, Q, Z controls for θ. Notice that one does not need to know the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to implement the method, one can condition on X, Q, Z.) Heckman and Robb (1985a) discuss the literature that uses replacement functions in this way. Olley and Pakes (1996) apply this method and consider nonparametric identification of the τ function. Matzkin (2007) provides a rigorous proof of identification for this approach in a general nonparametric setting.

The method of replacement functions assumes that (10) is a perfect proxy. In many applications, this assumption is far too strong. More often, θ is measured with error. This produces a factor model or measurement error model (Aigner et al. 1984). Matzkin (2007) surveys this method. One can represent the factor model in a general way by a system of equations:

$$Y_j = g_j(X, Z, Q, \theta, \varepsilon_j), \quad j = 1, \ldots, J. \tag{11}$$

A linear factor model separable in the unobservables writes

$$Y_j = g_j(X, Z, Q) + \alpha_j \theta + \varepsilon_j, \quad j = 1, \ldots, J, \tag{12}$$

where

$$(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j), \varepsilon_j \perp\!\!\!\perp \theta, \quad j = 1, \ldots, J, \tag{13}$$

and the ε_j are mutually independent. Observe that under (11) and (12), Y_j controlling for X, Z, Q only imperfectly proxies θ because of the presence of ε_j. θ is called a factor, α_j factor loadings and the ε_j "uniquenesses" (see e.g., Aigner 1985).

A large literature, reviewed in Abbring and Heckman (2007) and Matzkin (2007) shows how to establish identification of econometric models under factor structure assumptions. Cunha et al. (2010), Schennach (2004) and Hu and Schennach (2008) establish identification in nonlinear models of the form (11). (Cunha et al. (2006, 2010) apply and extend this approach to a dynamic factor setting where the θ_t are time dependent.) The key to identification is multiple, but imperfect (because of ε_j), measurements on θ from the Y_j, $j = 1, \ldots, J$ and X, Z, Q, and possibly other measurement systems that depend on θ. Carneiro et al. (2003), Cunha et al. (2005, 2006) and Cunha and Heckman (2007, 2008) apply and develop these methods. Under assumption (13), they show how to nonparametrically identify the econometric model and the distributions of the unobservables $F_\Theta(\theta)$ and $F_{\xi_j}(\varepsilon_j)$. In the context of classical simultaneous equations models, identification is secured by using covariance restrictions across equations exploiting the low dimensionality of vector θ compared to the high dimensional vector of (imperfect) measurements on it. The recent literature (Cunha et al. 2003; Cunha et al. 2010; Hu and Schennach 2008) extends the linear model to a nonlinear setting.

The recent econometric literature applies in special cases the idea of the control function principle introduced in Heckman and Robb (1985a). This principle, versions of which can be traced back to Telser (1964), partitions θ in (U-1) into two or more components, $\theta = (\theta_1, \theta_2)$, where only one component of θ is the source of bias. Thus it is assumed that (U-1) is true, and (U-1)′ is also true:

$$\text{(U-1)}' \quad (Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta_1,$$

and (U-2) holds. For example, in a normal selection model with additive separability, one can break U_1, the error term associated with Y_1, into two components:

$$U_1 = E(U_1 \mid V) + \varepsilon,$$

where V plays the role of θ_1 and is associated with the choice equation. Under normality, ε is independent of $E(U_1 \mid V)$. Further,

$$E(U_1 \mid V) = \frac{Cov(U_1, V)}{Var(V)} V, \tag{14}$$

assuming $E(U_1) = 0$ and $E(V) = 0$. Heckman and Robb (1985a) show how to construct a control function in the context of the choice model

$$D = \mathbf{1}[\mu_D(Z) > V].$$

Controlling for V controls for the component of θ_1 in (U-1)′ that gives rise to the spurious dependence. The Blundell and Powell (2003, 2004) application of the control function principle assumes functional form (14) but assumes that V can be perfectly proxied by a first stage equation. Thus they use a replacement function in their first stage. Their method does not work when one can only condition on D rather than on $D^* = \mu_D(Z) - V$ instead of directly measuring it. (Imbens and Newey (2002) extend their approach. See the discussion in Matzkin (2007).) In the sample selection model, it is not necessary to identify V. As developed in Heckman and Robb (1985a) and

Heckman and Vytlacil (2007a, b), under additive separability for the outcome equation for Y_1, one can write

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + \underbrace{E(U_1 \mid \mu_D(Z) > V)}_{\text{control function}},$$

so the analyst "expects out" rather than solve out the effect of the component of V on U_1 and thus control for selection bias under the maintained assumptions. In terms of the propensity score, under the conditions specified in Heckman and Vytlacil (2007a), one may write the preceding expression in terms of $P(Z)$:

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + K_1(P(Z)),$$

where $K_1(P(Z)) = E(U_1 \mid X, Z, D = 1)$. It is not literally necessary to know V or be able to estimate it. The Blundell and Powell (2003, 2004) application of the control function principle assumes that the analyst can condition on and estimate V.

The Blundell and Powell method and the method of Imbens and Newey (2002) build heavily on (14) and implicitly make strong distributional and functional form assumptions that are not intrinsic to the method of control functions. As just noted, their method uses a replacement function to obtain $E(U_1 \mid V)$ in the first step of their procedures. The general control function method does not require a replacement function approach. The literature has begun to distinguish between the more general control function approach and the *control variate* approach that uses a first stage replacement function.

Matzkin (2003) develops the method of unobservable instruments which is a version of the replacement function approach applied to ▶nonlinear models. Her unobservable instruments play the role of covariance restrictions used to identify classical simultaneous equations models (see Fisher, 1966). Her approach is distinct from and therefore complementary with linear factor models. Instead of assuming $(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j)$, she assumes in a two equation system that $(\theta, \varepsilon_1) \perp\!\!\!\perp Y_2 \mid Y_1, X, Z$. See Matzkin (2007).

I do not discuss panel data methods in this paper. The most commonly used panel data method is difference-in-differences as discussed in Heckman and Robb (1985a), Blundell et al. (1998), Heckman et al. (1999), and Bertrand et al. (2004), to cite only a few of the key papers. Most of the estimators I have discussed can be adapted to a panel data setting. Heckman et al. (1998) develop difference-in-differences matching estimators. Abadie (2002) extends this work. (There is related work by Athey and Imbens (2006), which exposits the Heckman et al. (1998) difference-in-differences matching estimator.) Separability between errors and observables is a key feature of the panel data approach in its standard application. Altonji and Matzkin (2005) and Matzkin (2003) present analyses of nonseparable panel data methods. Regression discontinuity estimators, which are versions of IV estimators, are discussed by Heckman and Vytlacil (2007b).

Table 3 summarizes some of the main lessons of this section. I stress that the stated conditions are necessary conditions. There are many versions of the IV and control functions principle and extensions of these ideas which refine these basic postulates. See Heckman and Vytlacil (2007b). Matzkin (2007) is an additional reference on sources of identification in econometric models.

I next introduce the generalized Roy model and the concept of the marginal treatment effect which helps to link the econometric literature to the statistical literature. The Roy model also provides a framework for thinking about the difference in information between the agents and the statistician.

Matching

The method of matching is widely-used in statistics. It is based on strong assumptions which often make its application to economic data questionable. Because of its popularity, I single it out for attention. The method of matching assumes selection of treatment based on potential outcomes

$$(Y_0, Y_1) \not\perp\!\!\!\perp D,$$

so $\Pr(D = 1 \mid Y_0, Y_1)$ depends on Y_0, Y_1. It assumes access to variables Q such that conditioning on Q removes the dependence:

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid Q. \qquad \text{(Q-1)}$$

Thus,

$$\Pr(D = 1 \mid Q, Y_0, Y_1) = \Pr(D = 1 \mid Q).$$

Comparisons between treated and untreated can be made at all points in the support of Q such that

$$0 < \Pr(D = 1 \mid Q) < 1. \qquad \text{(Q-2)}$$

The method does not explicitly model choices of treatment or the subjective evaluations of participants, nor is there any distinction between the variables in the outcome equations (X) and the variables in the choice equations (Z) that is central to the IV method and the method of control functions. In principle, condition (Q-1) can be satisfied using a set of variables Q distinct from all or some of the components of X and Z. The conditioning variables do not have to be exogenous.

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 3 Identifying assumptions under commonly used methods

	Identifying assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi, \xi = 1 \Longrightarrow A = 1, \xi = 0 \Longrightarrow A = 0$ (full compliance) Alternatively, if self-selection is random with with respect to outcomes, $(Y_0, Y_1) \perp\!\!\!\perp D$. Assignment can be conditional on X.	Yes	No
Matching	$(Y_0, Y_1) \not\perp\!\!\!\perp D$, but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, $0 < \Pr(D = 1 \mid X) < 1$ for all X So D conditional on X is a nondegenerate random variable	Yes	No
Control functions and extensions	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function) Version (i) Replacement functions (substitute out θ by observables) (Blundell and Powell, 2003; Heckman and Robb, 1985b; Olley and Pakes, 1996). Factor models Carneiro et al., (2003) allow for measurement error in the proxies. Version (ii) Integrate out θ assuming $\theta \perp\!\!\!\perp (X, Z)$ (Aakvik et al., 2005; Carneiro et al., 2003) Version (iii) For separable models for mean response expect θ conditional on X, Z, D as in standard selection models (control functions in the same sense of Heckman and Robb).	Yes	Yes (for semiparametric models)
IV	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z	Yes	Yes

(Y_0, Y_1) are potential outcomes that depend on X

$$D = \begin{cases} 1 \text{ if assigned (or choose) status 1} \\ 0 \text{ otherwise} \end{cases}$$

Z are determinants of D, θ is a vector of unobservables

For random assignments, A is a vector of actual treatment status. $A = 1$ if treated; $A = 0$ if not.

$\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise (Heckman and Vytlacil 2007b)

From condition (Q-1) one recovers the distributions of Y_0 and Y_1 given Q – $\Pr(Y_0 \le y_0 \mid Q = q) = F_0(y_0 \mid Q = q)$ and $\Pr(Y_1 \le y_1 \mid Q = q) = F_1(y_1 \mid Q = q)$ – but not the joint distribution $F_{0,1}(y_0, y_1 \mid Q = q)$, because the analyst does not observe the same persons in the treated and untreated states. This is a standard evaluation problem common to all econometric estimators. Methods for determining which variables belong in Q rely on untested exogeneity assumptions which we discuss in this section.

OLS is a special case of matching that focuses on the identification of conditional means. In OLS linear functional forms are maintained as exact representations or valid approximations. Considering a common coefficient model, OLS writes

$$Y = \pi Q + D\alpha + U, \tag{Q-3}$$

where α is the treatment effect and

$$E(U \mid Q, D) = 0. \tag{Q-4}$$

The assumption is made that the variance-covariance matrix of (Q, D) is of full rank:

$$Var(Q, D) \text{ full rank.} \tag{Q-5}$$

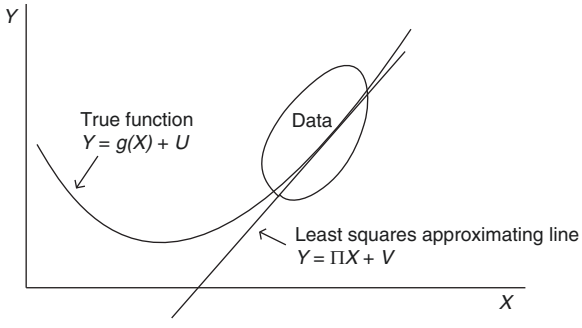
Under these conditions, one can identify α even though D and U are dependent: $D \not\perp\!\!\!\perp U$. Controlling for the observable Q eliminates any spurious mean dependence between D and U: $E(U \mid D) \neq 0$ but $E(U \mid D, Q) = 0$. (Q-3) is the linear regression counterpart to (Q-1). (Q-5) is the linear regression counterpart to (Q-2). Failure of (Q-5) would mean that using a nonparametric estimator one might perfectly predict D given Q, and that $\Pr(D = 1 \mid Q = q) = 1$ or 0. (This condition might be met only at certain values of $Q = q$. For certain parameterizations (e.g., the linear probability model), one may obtain predicted probabilities outside the unit interval.)

Matching can be implemented as a nonparametric method. When this is done, the procedure does not require specification of the functional form of the outcome equations. It enforces the requirement that (Q-2) be satisfied by estimating functions pointwise in the support of Q. Assume that $Q = (X, Z)$ and that X and Z are the same except where otherwise noted. Thus I invoke assumptions (M-1) and (M-2) presented in section "►The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators", even though in principle one can use a more general conditioning set.

Assumptions (M-1) and (M-2) or (Q-1) and (Q-2) rule out the possibility that after conditioning on X (or Q), agents possess more information about their choices than econometricians, and that the unobserved information helps to predict the potential outcomes. Put another way, the method allows for potential outcomes to affect choices but only through the observed variables, Q, that predict outcomes. This is the reason why Heckman and Robb (1985a, 1986b) call the method selection on observables.

Heckman and Vytlacil (2007b) establish the following points. (1) Matching assumptions (M-1) and (M-2) generically imply a flat MTE in u_D, i.e., they assume that $E(Y_1 - Y_0 \mid X = x, U_D = u_D)$ does not depend on u_D. Thus the unobservables central to the Roy model and its extensions and the unobservables central to the modern IV literature are assumed to be absent once the analyst conditions on X. (M-1) implies that all mean treatment parameters are the same. (2) Even if one weakens (M-1) and (M-2) to mean independence instead of full independence, generically the MTE is flat in u_D under the assumptions of the nonparametric generalized Roy model developed in section "►An Index Model of Choice and Treatment Effects: Definitions and Unifying Principles", so again all mean treatment parameters are the same. (3) IV and matching make distinct identifying assumptions even though they both invoke conditional independence assumptions. (4) Comparing matching with IV and control function (sample selection) methods, matching assumes that conditioning on observables eliminates the dependence between (Y_0, Y_1) and D. The control function principle models the dependence. (5) Heckman and Navarro (2004) and Heckman and Vytlacil (2007b) demonstrate that if the assumptions of the method of matching are violated, the method can produce substantially biased estimators of the parameters of interest. (6) Standard methods for selecting the conditioning variables used in matching assume exogeneity. Violations of the exogeneity assumption can produce biased estimators.

Nonparametric versions of matching embodying (M-2) avoid the problem of making inferences outside the support of the data. This problem is implicit in any application of least squares. Figure 7 shows the support problem that can arise in linear least squares when the linearity of the regression is used to extrapolate estimates determined in one empirical support to new supports. Careful attention to support problems is a virtue of any nonparametric method, including, but not unique to, nonparametric matching. Heckman, Ichimura, Smith, and Todd (1998)



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 7 The least squares extrapolation problem avoided by using nonparametric regression or matching (Heckman and Vytlacil 2007b)

show that the bias from neglecting the problem of limited support can be substantial. See also the discussion in Heckman, LaLonde, and Smith (1999).

Summary

This paper exposits the basic economic model of causality and compares it to models in statistics. It exposits the key identifying assumptions of commonly used econometric estimators for causal inference. The emphasis is on the economic content of these assumptions. I discuss how matching makes strong assumption about the information available to economist/statistician.

Acknowledgments

University of Chicago, Department of Economics, 1126 E. 59th Street, Chicago IL 60637, USA. This research was supported by NSF: 97-09-873, 00-99195, and SES-0241858 and NICHD: R01-HD32058-03. I thank Mohan Singh, Sergio Urzua, and Edward Vytlacil for useful comments.

About the Author

Professor Heckman shared the Nobel Memorial Prize in Economics in 2000 with Professor Daniel McFadden for his development of theory and methods for analyzing selective samples. Professor Heckman has also received numerous awards for his work, including the John Bates Clark Award of the American Economic Association in 1983, the 2005 Jacob Mincer Award for Lifetime Achievement in Labor Economics, the 2005 University College Dublin Ulysses Medal, the 2005 Aigner award from the Journal of Econometrics, and Gold Medal of the President of the Italian Republic, Awarded by the International Scientific Committee, in 2008. He holds six honorary doctorates. He is considered to be among the five most influential economists in the world, in 2010. (http://ideas.repec.org/top/top.person.all.html).

Cross References

►Causal Diagrams
►Causation and Causal Inference
►Econometrics
►Factor Analysis and Latent Variable Modelling
►Instrumental Variables
►Measurement Error Models
►Panel Data
►Random Coefficient Models
►Randomization

References and Further Reading

Aakvik A, Heckman JJ, Vytlacil EJ (1999) Training effects on employment when the training effects are heterogeneous: an application to Norwegian vocational rehabilitation programs. University of Bergen Working Paper 0599, University of Chicago

Aakvik A, Heckman JJ, Vytlacil EJ (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. J Econometrics 125:15–51

Abadie A (2002) Bootstrap tests of distributional treatment effects in instrumental variable models. J Am Stat Assoc 97:284–292

Abbring JH, Heckman JJ (2007) Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 5145–5303

Aigner DJ (1985) The residential electricity time-of-use pricing experiments: what have we learned? In: Hausman JA, Wise DA (eds) Social experimentation. University of Chicago Press, Chicago, pp 11–41

Aigner DJ, Hsiao C, Kapteyn A, Wansbeek T (1984) Latent variable models in econometrics. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics, vol 2, chap 23. Elsevier, Amsterdam, pp 1321–1393

Altonji JG, Matzkin RL (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73:1053–1102

Angrist J, Graddy K, Imbens G (2000) The interpretation of instrumental variables estimators in simultaneous equationsmodelswith an application to the demand for fish. Rev Econ Stud 67:499–527

Angrist JD, Krueger AB (1999) Empirical strategies in labor economics. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3A. North-Holland, New York, pp 1277–1366

Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-indifferences models. Econometrica 74:431–497

Barnow BS, Cain GG, Goldberger AS (1980) Issues in the analysis of selectivity bias. In: Stromsdorfer E, Farkas G (eds) Evaluation studies, vol 5. Sage Publications, Beverly Hills, pp 42–59

Barros RP (1987) Two essays on the nonparametric estimation of economic models with selectivity using choice-based samples. PhD thesis, University of Chicago

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-indifferences estimates? Q J Econ 119:249–275

Björklund A, Moffitt R (1987) The estimation of wage gains and welfare gains in self-selection. Rev Econ Stat 69:42–49

Blundell R, Duncan A, Meghir C (1998) Estimating labor supply responses using tax reforms. Econometrica 66:827–861

Blundell R, Powell J (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont LPHM, Turnovsky SJ (eds) Advances in economics and econometrics: theory and applications, eighth world congress, vol 2. Cambridge University Press, Cambridge

Blundell R, Powell J (2004) Endogeneity in semiparametric binary responsemodels. Rev Econ Stud 71:655–679

Carneiro P, Hansen K, Heckman JJ (2001) Removing the veil of ignorance in assessing the distributional impacts of social policies. Swedish Econ Policy Rev 8:273–301

Carneiro P, Hansen K, Heckman JJ (2003) Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. Int Econ Rev 44:361–422

Cunha F, Heckman JJ (2007) Identifying and estimating the distributions of Ex Post and Ex Ante returns to schooling: a survey of recent developments. Labour Econ 14:870–893

Cunha F, Heckman JJ (2008) A new framework for the analysis of inequality. Macroecon Dyn 12:315–354
Cunha F, Heckman JJ, Matzkin R (2003) Nonseparable factor analysis. Unpublished manuscript, University of Chicago, Department of Economics
Cunha F, Heckman JJ, Navarro S (2005) Separating uncertainty fromheterogeneity in life cycle earnings. The 2004 Hicks Lecture. Oxford Economic Papers 57, 191–261
Cunha F, Heckman JJ, Navarro S (2006) Counterfactual analysis of inequality and social mobility. In: Morgan SL, Grusky DB, Fields GS (eds) Mobility and inequality: frontiers of research in sociology and economics, chap 4. Standford University Press, Stanford, pp 290–348
Cunha F, Heckman JJ, Schennach SM (2006) Nonlinear factor analysis. Unpublished manuscript, University of Chicago, Department of Economics, revised 2008
Cunha F, Heckman JJ, Schennach SM (2010) Estimating the technology of cognitive and noncognitive skill formation, Forthcoming. Econometrica
Fisher RA (1966) The design of experiments. Hafner Publishing, New York
Gerfin M, Lechner M (2002) Amicroeconomic evaluation of the active labormarket policy in Switzerland. Econ J 112:854–893
Heckman JJ (1992) Randomization and social policy evaluation. In: Manski C, Garfinkel I (eds) Evaluating welfare and training programs. Harvard University Press, Cambridge, pp 201–230
Heckman JJ (1997) Instrumental variables: a study of implicit behavioral assumptions used inmaking programevaluations. J Hum Resour 32:441–462; addendum published vol. 33 no. 1 (Winter 1998)
Heckman JJ (2008) Econometric causality. Int Stat Rev 76:1–27
Heckman JJ, Ichimura H, Smith J, Todd PE (1998) Characterizing selection bias using experimental data. Econometrica 66: 1017–1098
Heckman JJ, LaLonde RJ, Smith JA (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3A, chap 31. North-Holland, New York, pp 1865–2097
Heckman JJ, Navarro S (2004) Usingmatching, instrumental variables, and control functions to estimate economic choice models. Rev Econ Stat 86:30–57
Heckman JJ, Navarro S (2007) Dynamic discrete choice and dynamic treatment effects. J Econometrics 136:341–396
Heckman JJ, Robb R (1985a) Alternative methods for evaluating the impact of interventions. In: Heckman J, Singer B (eds) Longitudinal analysis of labor market data, vol 10. Cambridge University Press, New York, pp 156–245
Heckman JJ, Robb R (1985b) Alternative methods for evaluating the impact of interventions: an overview. J Econometrics 30: 239–267
Heckman JJ, Robb R (1986a) Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 63–107, reprinted in 2000, Erlbaum, Mahwah
Heckman JJ, Robb R (1986b) Postscript: a rejoinder to Tukey. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 111–114, reprinted in 2000, Erlbaum, Mahwah
Heckman JJ, Smith JA, Clements N (1997) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev Econ Stud 64: 487–536
Heckman JJ, Urzua S, Vytlacil EJ (2006) Understanding instrumental variables in models with essential heterogeneity. Rev Econ Stat 88:389–432
Heckman JJ, Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci 96:4730–4734
Heckman JJ, Vytlacil EJ (2001) Local instrumental variables. In: Hsiao C, Morimune K, Powell JL (eds) Nonlinear statistical modeling: proceedings of the thirteenth international symposium in economic theory and econometrics: essays in honor of Takeshi Amemiya. Cambridge University Press, New York, pp 1–46
Heckman JJ, Vytlacil EJ (2005) Structural equations, treatment effects and econometric policy evaluation. Econometrica 73:669–738
Heckman JJ, Vytlacil EJ (2007a) Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 4779–4874
Heckman JJ, Vytlacil EJ (2007b) Econometric evaluation of social programs, part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 4875–5144
Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. Econometrica 76:195–216
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29
Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–475
Imbens GW, Newey WK (2002) Identification and estimation of triangular simultaneous equations models without additivity. Technical working paper 285, National Bureau of Economic Research
Matzkin RL (2003) Nonparametric estimation of nonadditive randomfunctions. Econometrica 71:1339–1375
Matzkin RL (2007) Nonparametric identification. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam
Olley GS, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. Econometrica 64:1263–1297
Pearl J (2000) Causality. Cambridge University Press, Cambridge
Powell JL (1994) Estimation of semiparametric models. In: Engle R, McFadden D (eds) Handbook of econometrics, vol 4. Elsevier, Amsterdam, pp 2443–2521
Quandt RE (1958) The estimation of the parameters of a linear regression systemobeying two separate regimes. J Am Stat Assoc 53:873–880
Quandt RE (1972) A new approach to estimating switching regressions. J Am Stat Assoc 67:306–310
Roy A (1951) Some thoughts on the distribution of earnings. Oxford Econ Pap 3:135–146
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58

Schennach SM (2004) Estimation of nonlinear models with measurement error. Econometrica 72:33–75
Telser LG (1964) Iterative estimation of a set of linear regression equations. J Am Stat Assoc 59:845–862
Vytlacil EJ (2002) Independence, monotonicity, and latent index models: an equivalence result. Econometrica 70:331–341

Prior Bayes: Rubin's View of Statistics

Herman Rubin
Purdue University, West Lafayette, IN, USA

Introduction

What is statistics? The common way of looking at it is a collection of methods, somehow or other produced, and one should use one of those methods for a given set of data. The typical user who has little more than this comes to a statistician asking for THE answer, as if the data are sufficient to get this without knowing the problem.

No; the first thing is to formulate the problem. One cannot do better than to assume that there is an unknown state of nature, that there is a probability distribution of observations given a state of nature, a set of possible actions, and that each action in each state of nature has (possibly random) consequences.

Following the von Neumann–Morgenstern (1944) axioms for utility, in 1947 (see Rubin 1987a) I was able to show that if one has a self-consistent evaluation of actions in each state of nature, the utility function for an unknown state of nature has to be an integral of the utilities for the states of nature. Another way of looking at this in the discrete case is that one assigns a weight to each result in each state of nature, and should choose the action which produces the best sum; this generalizes to the best value of the integral. This is the prior Bayes approach.

Let us simplify to the usual Bayes model; it can be generalized to include more, and in fact, must be for the infinite parametric (usually called non-parametric) problems encountered.

So the quantity to be minimized is

$$\int \int L(\omega, q(x))d\mu(x|\omega)d\xi(\omega).$$

If this integral is finite, and if, for example, L is positive, the integration can be interchanged and the result written as

$$\int \int L(\omega, q(x))d\phi(\omega|x)dm(x),$$

and if the properties of q for different x are unrestricted, one can (hopefully) use the usual Bayes proceduce of minimizing the inner integral.

But this can require a huge amount of computing power, possibly even exceeding the capacity of the universe,and sufficient assurance that one has a sufficiently good approximation to the loss-prior combination. One use this latter because it is only the product of L and ξ which is relevant. One can try to approximate, but posterior robustness results are hard, and often impossible, to come by.

On the other hand, the prior Bayes approach can show what is and what is not important, and can, in many cases, provide methods which are not much worse than full Bayes methods, or at least the approximations made. One might question how this can be shown, considering that the full Bayes procedure cannot be calculated; however, a smaller problem with one which can be calculated might be shown to come close. Also, one can get procedures which are good with considerable uncertainty about the prior distribution of the parameter.

Results and Examples

Some early approaches, some by those not believing in the Bayes approach, were the empirical Bayes results of Robbins and his followers. Empirical Bayes extends much farther now, and it is in the spirit of prior Bayes, as the performance of the procedures is what is considered. There is a large literature on this, and I will not go into it in any great detail.

Suppose we consider the case of the usual test of a point null, when the distribution is normal. If we assume the prior is symmetric, we need only consider procedures which accept if the mean is close enough to the null, and reject otherwise. If we assume that the prior gives a point mass at the null, and is otherwise given by a smooth density, the prior Bayes risk is

$$\xi(\{0\})P(|X > c|0) + \int Q(\omega)h(\omega)P(|X| \le c|\omega)d\omega,$$

where Q is the loss of incorrect acceptance.

This shows that the tails of the prior distribution are essentially irrelevant if the variance is at all small, so the prior probability of the null is NOT an important consideration. Only the ratio of the probability of the null to the density at the null of the alternative is important. This shows that the large-sample results of Rubin and Sethuraman (1965) can be a good approximation for moderate, or even small, samples. It also shows how the "p-value" should change with the sample size. An expository paper is in preparation.

If the null is not a point null, this is still a good approximation if the width of the null is smaller than the standard deviation of the observations; if it is not, the problem is much harder, and the form of the loss-prior combination under the null becomes of considerable importance. Again, the prior Bayes approach shows where the problem lies, in the behavior of the loss-prior combination near the division point between accepptance and rejection. In "standard" units, a substantial part of the parameter space may be involved.

For another example, consider the problem of estimating an infinite dimensional normal mean with the covariance matrix a multiple of the identity, or the similar problem of estimating a spectral density function. The first problem was considereed by Rubin (1987b), and was corrrected in Hui Xu's thesis in 2008. If one assumes the prior mean square of the *k*th mean, or the corresponding Fourier coefficient, is non-increasing, an empirical Bayes type result can be obtained, and can be shown to be asymptotically optimal in the class of "simple" procedures, which are the ones currently being used with more stringent assumptions, and which are generally not as good. The results are good even if the precise form is not known, while picking a kernel is much more restrictive and generally not even asymptotically optimal.

For the latter problem, obtaining a reasonable prior seems to be extremely difficult, but the simple procedures obtained are likely to be rather good even if one is found. The positive definiteness of the Toeplitz matrix of the covariances is a difficult condition to work with.

I see a major set of applications to non-parametric problems, properly called infinite parametric, problems such as density estimation, testing with infinite dimensional alternatives, etc. With a large number of dimensions, the classical approach does not work well, and the only "simple Bayesian" approaches use priors which look highly constrained to me.

About the Author

"Professor Rubin has contributed in deep and original ways to statistical theory and philosophy. The statistical community has been vastly enriched by his contributions through his own research and through his influence, direct or indirect on the research and thinking of others." Erich Marchard and William Strawdeman, *A festschrift for Herman Rubin*, A. DasGupta (ed.), p. 21. "He is well known for his broad ranging mathematical research interests and for fundamental contributions in Bayesian decision theory, in set theory, in estimations for simultaneous equations, in probability and in asymptotic statistics." Mary Ellen Block, Ibid. p. 408. Professor Rubin is a Fellow of the IMS, and a Fellow of the AAAS (American Association for the Advancement of Science).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Model Selection
►Moderate Deviations
►Statistics: An Overview
►Statistics: Nelder's View

References and Further Reading

Rubin H (1987a) A weak system of axioms for "rational" behavior and the non-separability of utility from prior. Stat Decisions 5:47–58
Rubin H (1987b) Robustness in generalized ridge regression and related topics. Third Valencia Symp Bayesian Stat 3:403–410
Rubin H, Sethuraman J (1965) Bayes risk efficiency. Sankhya A 27:347–356
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton university press, Princeton
Xu H (2008) Some applications of the prior Bayes approach. Unpublished thesis

Probabilistic Network Models

Ove Frank
Professor Emeritus
Stockholm University, Stockholm, Sweden

A network on vertex set V is represented by a function y on the set V^2 of ordered vertex pairs. The function can be univariate or multivariate and its variables can be numerical or categorical. A graph G with vertex set V and edge set E in V^2 is represented by a binary function $y = \{(u, v, y_{uv}) : (u, v) \in V^2\}$ where y_{uv} indicates whether $(u, v) \in E$. If $V = \{1, \ldots, N\}$ and vertices are ordered according to their labels, y can be given as an N by N adjacency matrix $y = (y_{uv})$. Simple undirected graphs have $y_{uv} = y_{vu}$ and $y_{vv} = 0$ for all u and v. Colored graphs have a categorical variable y with the categories labeled by colors. Graphs with more general variables y are called valued graphs or networks. If Y is a random variable with outcomes y representing networks in a specified family of networks, the probability distribution induced on this family is a probabilistic network model and Y is a representation of a random network.

Simple random graphs are defined with uniform distributions or Bernoulli distributions. Uniform models assign

equal probabilities to all graphs in a specified finite family of graphs, such as all graphs of order N and size M, or all connected graphs of order N, or all trees of order N. Bernoulli graphs (Bernoulli digraphs) have edge indicators that are independent Bernoulli variables for all unordered (ordered) vertex pairs. There is an extensive literature on such random graphs, especially on the simplest Bernoulli (p) graph, which has a common edge probability p for all vertex pairs. An extension of a fixed graph G to a Bernoulli (G, α, β) graph is a Bernoulli graph that is obtained by independently removing edges in G with probability α and independently inserting edges in the complement of G with probability β. Such models have been applied to study reliability problems in communication networks. Attempts to model the web have recently contributed to an interest in random graph models with specified degree distributions and random graph processes for very large dynamically changing graphs.

The literature on social networks describes models for finite random digraphs on $V = \{1, \ldots, N\}$ in which dyads (Y_{uv}, Y_{vu}) for $u < v$ are independent and have probabilities that depend on parameters governing in- and out-edges of each vertex and mutual edges of each vertex pair. Special cases of such models with independent dyads are obtained by assuming homogeneity for the parameters of different vertices or different groups of vertices. Extensions to models with dependent dyads include Markov graphs that allow dependence between incident dyads. Other extensions are log-linear models that assume that the log-likelihood function is a linear function of specified network statistics chosen to reflect various properties of interest in the network.

Statistical analysis of network data comprise exploratory tools for selecting appropriate probabilistic network models as well as confirmatory tools for estimating and testing various models. Many of these tools use computer intensive methods.

Applications of probabilistic network models appear in many different areas in which relationships between the units studied are essential for an understanding of their properties and characteristics. The social and behavioral sciences have contributed to the development of many network models for the study of social interaction, friendship, dominance, co-operation and competition. There are applications to criminal networks and co-offending, communication and transportation networks, vaccination programs in epidemiology, information retrieval and organizational systems, particle systems in physics, biometric cell systems. Random graphs and random fields are also theoretically developed in computer science, mathematics, and statistics. There is an exciting interplay between model development and new applications in a variety of important areas.

Many references to the literature on graphs, random graphs, and random networks are provided by the following sources.

About the Author

For biography *see* the entry ►Network Sampling.

Cross References

►Graphical Markov Models
►Network Models in Probability and Statistics
►Network Sampling
►Social Network Analysis
►Uniform Distribution in Statistics

References and Further Reading

Bonato A (2008) A course on the web graph. American Mathematical Society, Providence
Carrington P, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, New York
Diestel R (2005) Graph theory. Springer, Berlin/Heidelberg
Durrett R (2007) Random graph dynamics. Cambridge University Press, New York
Kolaczyk E (2009) Statistical analysis of network data. Springer, New York
Meyers R (ed) (2009) Encyclopedia of complexity and systems science. Springer, New York

Probability on Compact Lie Groups

David Applebaum
Professor, Head
University of Sheffield, Sheffield, UK

Introduction

Probability on groups enables us to study the interaction between chance and symmetry. In this article I'll focus on the case where symmetry is generated by continuous groups, specifically compact Lie groups. This class contains many examples such as the n-torus, special orthogonal groups $SO(n)$ and special unitary groups $SU(n)$ which are important in physics and engineering applications. It is also a very good context to demonstrate the key role played by non-commutative harmonic analysis via group representations. The classic treatise (Heyer 1977) by Heyer gives a systematic mathematical introduction to this topic while

Diaconis (1988) presents a wealth of concrete examples in both probability and statistics.

For motivation, let ρ be a probability measure on the real line. Its characteristic function $\widehat{\rho}$ is the Fourier transform $\widehat{\rho}(u) = \int_{\mathbb{R}} e^{iux}\rho(dx)$ and $\widehat{\rho}$ uniquely determines ρ. Note that the mappings $x \to e^{iux}$ are the irreducible unitary representations of $\mathbb{R}$.

Now let G be a compact Lie group and ρ be a probability measure defined on G. The group law of G will be written multiplicatively. If we are given a probability space $(\Omega, \mathcal{F}, P)$ then ρ might be the law of a G-valued random variable defined on Ω. The *convolution* of two such measures ρ_1 and ρ_2 is the unique probability measure $\rho_1 * \rho_2$ on G such that

$$\int_G f(\sigma)(\rho_1 * \rho_2)(d\sigma) = \int_G \int_G f(\sigma\tau)\rho_1(d\sigma)\rho_2(d\tau),$$

for all continuous functions f defined on G. If X_1 and X_2 are independent G-valued random variables with laws ρ_1 and ρ_2 (respectively), then $\rho_1 * \rho_2$ is the law of X_1X_2.

Characteristic Functions

Let $\widehat{G}$ be the set of all irreducible unitary representations of G. Since G is compact, $\widehat{G}$ is countable. For each $\pi \in \widehat{G}, \sigma \in G, \pi(\sigma)$ is a unitary (square) matrix acting on a finite dimensional complex inner product space V_π having dimension d_π. Every group has the trivial representation δ acting on $\mathbb{C}$ by $\delta(\sigma) = 1$ for all $\sigma \in G$. The *characteristic function* of the probability measure ρ is the matrix-valued function $\widehat{\rho}$ on $\widehat{G}$ defined uniquely by

$$\langle u, \widehat{\rho}(\pi)v\rangle = \int_G \langle u, \pi(\tau)v\rangle\rho(d\tau),$$

for all $u, v \in V_\pi$. $\widehat{\rho}$ has a number of desirable properties (Siebert 1981). It determines ρ uniquely and for all $\pi \in \widehat{G}$:

$$\widehat{\rho_1 * \rho_2}(\pi) = \widehat{\rho}_1(\pi)\widehat{\rho}_2(\pi).$$

In particular $\widehat{\delta} = 1$.

Lo and Ng (1988) considered a family of matrices $(C_\pi, \pi \in \widehat{G})$ and asked when there is a probability measure ρ on G such that $C_\pi = \widehat{\rho}(\pi)$. They found a necessary and sufficient condition to be that $C_\delta = 1$ and that the following non-negativity condition holds: for all families of matrices $(B_\pi, \pi \in \widehat{G})$ where B_π acts on V_π and for which $\sum_{\pi\in S_\pi} d_\pi \mathrm{tr}(\pi(\sigma)B_\pi) \geq 0$ for all $\sigma \in G$ and all finite subsets S_π of V_π we must have $\sum_{\pi\in S_\pi} d_\pi \mathrm{tr}(\pi(\sigma)C_\pi B_\pi) \geq 0$.

Densities

Every compact group has a bi-invariant finite Haar measure which plays the role of Lebesgue measure on $\mathbb{R}^d$ and which is unique up to multiplication by a positive real number. It is convenient to normalise this measure (so it has total mass 1) and denote it by $d\tau$ inside integrals of functions of τ. We say that a probability measure ρ has a *density* f if $\rho(A) = \int_A f(\tau)d\tau$ for all Borel sets A in G. To investigate existence of densities we need the *Peter–Weyl theorem* that the set of functions $\left\{d_\pi^{\frac{1}{2}}\pi_{ij}; 1 \leq i, j \leq d_\pi, \pi \in \widehat{G}\right\}$ are a complete orthonormal basis for $L^2(G, \mathbb{C})$. So any $f \in L^2(G, \mathbb{C})$ can be written

$$f(\sigma) = \sum_{\pi\in\widehat{G}} d_\pi \mathrm{tr}(\pi(\sigma)\widehat{f}(\pi)), \tag{1}$$

where $\widehat{f}(\pi) = \int_G f(\tau)\pi(\tau^{-1})d\tau$ is the Fourier transform. In Applebaum (2008) it was shown that ρ has a square-integrable density f (which then has an expansion as in (1)) if and only if $\sum_{\pi\in\widehat{G}} d_\pi \mathrm{tr}(\widehat{\rho}(\pi)\widehat{\rho}(\pi)^*) < \infty$ where * is the usual matrix adjoint. A sufficient condition for ρ to have a continuous density is that $\sum_{\pi\in\widehat{G}} d_\pi^{\frac{3}{2}} |\mathrm{tr}(\widehat{\rho}(\pi)\widehat{\rho}(\pi)^*)|^{\frac{1}{2}} < \infty$ in which case the series on the right hand side of (0.1) converges absolutely and uniformly (see Proposition 6.6.1 on pp. 117–118 of Faraut [2008]).

Conjugate Invariant Probabilities

Many interesting examples of probability measures are conjugate invariant, i.e., $\rho(\sigma A\sigma^{-1}) = \rho(A)$ for all $\sigma \in G$. In this case there exists $c_\pi \in \mathbb{C}$ such that $\widehat{\rho}(\pi) = c_\pi I_\pi$ where I_π is the identity matrix in V_π (Said et al. 2010). If a density exists it takes the form $f(\sigma) = \sum_{\pi\in\widehat{G}} d_\pi \overline{c_\pi}\chi_\pi(\sigma)$, where $\chi_\pi(\sigma) = \mathrm{tr}(\pi(\sigma))$ is the group character.

Example 1 *Gauss Measure.* Here $c_\pi = e^{\sigma^2\kappa_\pi}$ where $\kappa_\pi \leq 0$ is the eigenvalue of the group Laplacian corresponding to the Casimir operator $\kappa_\pi I_\pi$ on V_π and $\sigma > 0$. For example if $G = SU(2)$ then it can be parametrized by the Euler angles ψ, ϕ and θ, $\widehat{G} = \mathbb{Z}_+$, $\kappa_m = -m(m+2)$ and we have a continuous density depending only on $0 \leq \theta \leq \frac{\pi}{2}$:

$$f(\theta) = \sum_{m=0}^{\infty} (m+1)e^{-\sigma^2 m(m+2)} \frac{\sin((m+1)\theta)}{\sin(\theta)}.$$

Example 2 *Laplace Distribution.* This is a generalization of the double exponential distribution on $\mathbb{R}$ (with equal parameters). In this case $c_\pi = (1 - \beta^2\kappa_\pi)^{-1}$ where $\beta > 0$ and κ_π is as above.

Infinite Divisibility

A probability measure ρ on G is *infinitely divisible* if for each $n \in \mathbb{N}$ there exists a probability measure $\rho^{\frac{1}{n}}$ on G such that the nth convolution power $\left(\rho^{\frac{1}{n}}\right)^{*n} = \rho$. Equivalently $\widehat{\rho}(\pi) = \widehat{\rho^{\frac{1}{n}}}(\pi)^n$ for all $\pi \in \widehat{G}$. If G is connected as well as compact any such ρ can be realised as μ_1 in

a weakly continuous convolution semigroup of probability measures $(\mu_t, t \geq 0)$. For a general Lie group, such an *embedding* may not be possible and the investigation of this question has generated much research over more than 30 years (McCrudden 1998). The structure of convolution semigroups has been intensely analyzed. These give the laws of group-valued ►*Lévy processes*, i.e., processes with stationary and independent increments (Liao 2004). In particular there is a Lévy–Khintchine type formula (originally due to G.A.Hunt) which classifies these in terms of the structure of the infinitesimal generator of the associated Markov semigroup that acts on the space of continuous functions. One of the most important examples is Brownian motion (see ►Brownian Motion and Diffusions) and this has a Gaussian distribution. Another important example is the *compound Poisson process* (see ►Poisson Processes)

$$Y(t) = X_1 X_2 \cdots X_{N(t)} \tag{2}$$

where $(X_n, n \in \mathbb{N})$ is a sequence of i.i.d. random variables having common law ν (say) and $(N(t), t \geq 0)$ is an independent Poisson process of intensity $\lambda > 0$. In this case $\mu_t = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \nu^{*n}$. Note that μ_t does not have a density and it is conjugate invariant if ν is.

Applications

There is intense interest among statisticians and engineers in the *deconvolution problem* on groups. The problem is to estimate the signal density f_X from the observed density f_Y when the former is corrupted by independent noise having density f_ϵ, so the model is $Y = X\epsilon$ and the inverse problem is to untangle $f_Y = f_X * f_\epsilon$. Inverting the characteristic function enables the construction of non-parametric estimators for f_X and optimal rates of convergence are known for these when f_ϵ has certain smoothness properties (Kim and Richards 2001; Koo and Kim 2008). In Said et al. (2010) the authors consider the problem of *decompounding*, i.e., to obtain non-parametric estimates of the density of X_1 in (2) based on i.i.d. observations of a noisy version of Y: $Z(t) = \epsilon Y(t)$, where ϵ is independent of $Y(t)$. This is applied to multiple scattering of waves from complex media by working with the group $SO(3)$ which acts as rotations on the sphere.

About the Author

David Applebaum is Professor of Probability and Statistics and is currently Head of the Department of Probability and Statistics, University of Sheffield UK. He has published over 60 research papers and is the author of two books, *Probability and Information* (second edition), Cambridge University Press (1996, 2008) and *Lévy Processes and Stochastic Calculus* (second edition), Cambridge University Press (2004, 2009). He is joint managing editor of the *Tbilsi Mathematical Journal* and associate editor for *Bernoulli*, *Methodology and Computing in Applied Probability*, *Journal of Stochastic Analysis and Applications* and *Communications on Stochastic Analysis*. He is a member of the Atlantis Press Advisory Board for Probability and Statistics Studies.

Cross References

►Brownian Motion and Diffusions
►Characteristic Functions
►Lévy Processes
►Poisson Processes

References and Further Reading

Applebaum D (2008) Probability measures on compact groups which have square-integrable densities. Bull Lond Math Sci 40:1038–1044
Diaconis P (1988) Group representations in probability and statistics, Lecture Notes – Monograph Series Volume 11. Institute of Mathematical Statistics, Hayward
Faraut J (2008) Analysis on Lie groups. Cambridge University Press, Cambridge
Heyer H (1977) Probability measures on locally compact groups. Springer, Berlin/Heidelberg
Kim PT, Richards DS (2001) Deconvolution density estimators on compact Lie groups. Contemp Math 287:155–171
Koo J-Y, Kim PT (2008) Asymptotic minimax bounds for stochastic deconvolution over groups. IEEE Trans Inf Theory 54:289–298
Liao M (2004) Lévy processes in Lie groups. Cambridge University Press, Cambridge
Lo JT-H, Ng S-K (1988) Characterizing Fourier series representations of probability distributions on compact Lie groups. Siam J Appl Math 48:222–228
McCrudden M (1998) An introduction to the embedding problem for probabilities on locally compact groups. In: Hilgert J, Lawson JD, Neeb K-H, Vinberg EB (eds) Positivity in Lie theory: open problems. Walter de Gruyter, Berlin/New York, pp 147–164
Said S, Lageman C, LeBihan N, Manton JH (2010) Decompounding on compact Lie groups. IEEE Trans Inf Theory 56(6):2766–2777
Siebert E (1981) Fourier analysis and limit theorems for convolution semigroups on a locally compact group. Adv Math 39:111–154

Probability Theory: An Outline

Tamás Rudas
Professor, Head of Department of Statistics, Faculty of Social Sciences
Eötvös Loránd University, Budapest, Hungary

Sources of Uncertainty in Statistics

Statistics is often defined as the science of the methods of data collection and analysis, but from a somewhat more

conceptual perspective, statistics is also the science of methods dealing with uncertainty. The sources of uncertainty in statistics may be divided into two groups. Uncertainty associated with the data one has and uncertainty associated with respect to the mechanism which produced the data. These kinds of uncertainty are often interrelated in practice, yet it is useful to distinguish them.

Uncertainties in Data

One may distinguish two main sources of uncertainty with respect to data. One is related to measurement error, the other to sampling error.

Measurement error is the difference between the actual measurement obtained and the true value of what was measured. This applies to both cases when a numerical measurement taken (typical in the physical, biological, medical, psychological sciences) but also to qualitative observations when subjects are classified (typical in the social and behavioral sciences), except that in the latter case, instead of the numerical value of the error, its lack or presence is considered. In the former case, it is often observed that the errors are approximately distributed as normal, even if very precise and expensive measuring instruments are used. In the latter case, the existence of measurement error, that is misclassification, is often attributed to self-reflection of the human beings observed. In both situations, the lack of understanding of the precise mechanisms behind measurement errors suggest applying a stochastic model assuming that the result of the measurement is the sum of the true value plus a random error.

Sampling error stems from the uncertainty of how our results would differ, if a sample that is different from the actual one were observed. It is usually associated with the entire sample (and not with the individual observations) and is measured as the difference between the estimates obtained from the actual sample, and the census value that could be obtained if the same data collection methods were applied to the entire population. The census value may or may not be equal to the true population parameter of interest. For example, if the measurement error is assumed to be constant, then the census value differs from the population value by this quantity. Usually, the likely size of the sampling error is characterized by the standard deviation (standard error) of the estimates. Under many common random sampling schemes, the distribution of the estimates is normal, and the choice of the standard error, as a characteristic quantity, is well justified.

Uncertainties in Modeling

While uncertainties associated with the data seem to be inherent characteristics, uncertainties related to modeling are more determined by our choice of models, which depends very often on the existing knowledge regarding the research problem at hand. The most common assumption is that the quantity of interest has a specified, though unknown to the researcher, distribution, that may or may not be assumed to belong to some parametric family of distributions. A further possible choice, gaining increasing popularity during the recent decades, is that the distribution of interest belongs to a parametric family, though not with a specified parameter value, rather characterized by a probability distribution (the prior distribution) on the possible parameter values. The former view is adopted in frequentist statistics and the latter view is the Bayesian approach to statistics.

To model uncertainty, frequentist statistics uses frequentist or classical probability theory, while ▶Bayesian statistics often relies on a subjective concept of probability.

Classical and Frequentist Probability

Historically, there are two sources of modern probability theory. One is the theory of gambling, where the main goal was to determine how probable certain outcomes were in a game of chance. These problems could be appropriately handled under the assumptions that all possible outcomes of an experiment (rolling a die, for example) are equally likely and probabilities could be determined as the ratio of the number of outcomes with a certain characteristic, to the total number of outcomes. This interpretation of probability is called classical probability. Questions related to gambling also made important contributions to developing the concepts of Boolean algebra (the algebra of events associated with an experiment), conditional probability and infinite sequences of random variables (which play an important role in the frequentist interpretation of probability see below). The other source of modern probability theory is the analysis of errors associated with a measurement. This led, among others, to the understanding of the central role played by the normal distribution.

It is remarkable, that the main concepts and results of these two apparently very different fields, all may be based on one set of axioms, proposed by Kolmogorov and given in the next section.

The Kolmogorov Axioms

The axioms, summarizing the concepts developed within the classical and frequentist approaches, apply to experiments that may be repeated infinitely many times, where all circumstances of the experiment are supposed to remain constant. An experiment may be identified with its possible outcomes. Certain subsets of outcomes are called events, with he assumption that no outcome (the impossible event) and all the outcomes (the certain event) are

events and countable unions or intersections of events are also events. This means that the set of events associated with an experiment form a sigma-field. Then the Kolmogorov axioms (basic assumptions that are accepted to be true) are the following:

For any event A, its probability

$$P(A) \geq 0$$

For the certain event Ω

$$P(\Omega) = 1$$

For a series $A_i, i = 1, 2, \ldots$ of pairwise disjoint events

$$\Sigma_{i=1,2,\ldots} P(A_i) = P(\Sigma_{i=1,2,\ldots} A_i)$$

The heuristic interpretation is that the probability of an event manifests itself via the relative frequency of this event over long series of repetitions of the experiment. This is why this approach to probability is often called frequentist probability. Indeed, the axioms are true for relative frequencies instead of probabilities.

The Laws of Large Numbers

The link between the heuristic notion of probability and the mathematical theory of probability is established by the result that if $f_n(A)$ denotes the frequency of event A after n repetitions of an experiment, then

$$f_n(A)/n \to P(A),$$

where the convergence $\to$ may be given various interpretations. More generally, if X is a random variable (that is, such a function that $X \in I$ is an event for every interval I), then for the average of n independent observations of X, $\overline{X}_n$,

$$\overline{X}_n \to E(X),$$

where $E(X)$ is the expected value of X. Here, convergence is in probability (weak law) or almost surely (strong law) (See also ►Laws of Large Numbers).

The Central Limit Theorem

This fundamental result explains why the normal distribution plays such a central role of statistics. Many of the statistics are sample averages and for their asymptotic distributions the following result holds. If $V(X)$ denotes the variance of X, then the asymptotic distribution of

$$\frac{\overline{X}_n - E(X)}{\sqrt{V(X)/n}}$$

is standard normal (see also ►Central Limit Theorems).

Subjective Probability

This interpretation of the concept of probability associates it with the strength of trust or belief that a person has in the occurrence of an event. Such beliefs manifest themselves, for example, in betting preferences: out of two events, a rational person would have a betting preference for the one with which he/she associates a larger subjective probability. A fundamental difference between frequentist and subjective probability is that the latter may also be applied to experiments and events that may not be repeated many times. Of course, the subjective probabilities of different individuals may be drastically different from each other and it has been demonstrated repeatedly that the subjective probabilities an individual associates with different events, may not be logically consistent. Bayesian statistics sometimes employs the elicitation of such subjective probabilities to construct a prior distribution.

About the Author

Professor Rudas was the Founding Dean of the Faculty of Social Sciences, Eötvös Loránd University, from 2003 to 2009. He is also an Affiliate Professor in the Department of Statistics, University of Washington, Seattle, and a Recurrent Visiting Professor in the Central European University, Budapest. He is Vice President, European Association of Methodology, 2008–. Professor Rudas has been awarded the Erdei Prize of the Hungarian Sociological Association for the application of statistical methods in sociology (1988), and the Golden Memorial Medal of the Eötvös Loránd University (2009).

Cross References

►Axioms of Probability
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Central Limit Theorems
►Convergence of Random Variables
►Foundations of Probability
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Measure Theory in Probability
►Philosophy of Probability
►Probability, History of
►Statistics and Gambling

References and Further Reading

Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York

Kolmogorov AN (1950) Foundations of the theory of probability. Chelsey, New York (Original work: Grundbegriffe der Wahrscheinlichkeits Rechnung, 1933, Berlin: Springer-Verlag)
Rudas T (ed) (2008) Handbook of probability: theory and applications. Sage, Thousand Oaks

Probability, History of

Jordi Vallverdú
Universitat Autònoma de Barcelona, Barcelona, Spain

Five thousand years ago dice were invented in India (David 1998). This fact implies that their users had at least a common sense approach to the idea of probability. Those dice were not the contemporary cubical standard dice, but fruit stones or animal bones (Dandoy 2006). They must surely have been used for fun and gambling as well as for fortune-telling practices. The worries about the future and the absurd idea that the world was causally guided by supernatural forces led those people to a belief in the explanatory power of rolling dice.

In fact, cosmogonical answers were the first attempt to explain in a causal way the existence of things and beings. The Greek creation myth involved a game of dice between Zeus, Poseidon, and Hades. Also in the classic Hindu book *Mahabharata* (section "Sabha-parva"), we can find the use of dice for gambling. But in both cases there is no theory regarding probability in dice, just their use "for fun."

Later, and beyond myths, Aristotle was the strongest defender of the causal and empirical approach to reality (*Physics*, II, 4–6) although he considered the possibility of chance, especially the problem of the game of dice (*On Heavens*, II, 292a30) and probabilities implied in it. These ideas had nothing to do with those about atomistic chance by Leucippus and Democritus nor Lucrecius' controversial *clinamen*'s theory. Hald (1988) affirms the existence of probabilistic rather than mathematical thought in Classical Antiquity; we can accept that some authors (like Aristotle) were worried about the idea of chance (as well as about the primordial emptiness and other types of conceptual *cul-de-sac*), but they made no formal analysis of it. Later, we can find traces of interest in the moral aspects of gambling with dice in Talmudic (*Babylonian Talmud, Book* 8*: Tract Sanhedrin, chapter* III*, Mishnas* I *to* III) and Rabbinical texts, and we know that in 960, Bishop Wibolf of Cambrai calculated 56 diverse ways of playing with three dice. *De Vetula*, a Latin poem from the thirteenth century, tells us of 216 possibilities. But the first occurrence of combinatorics per se arose from Chinese interest in future prediction through the 64 hexagrams of the *I Ching* (previously eight trigrams derived from four binary combinations of two elemental forces, *yin* and *yang*).

In 1494 Luca Paccioli defined the basic principles of algebra and multiplication tables up to 60 × 60 in his book *Summa de arithmetica, geometria, proportioni e proportionalita*. He posed the first serious statistical problem of two men playing a game called "balla," which is to end when one of them has won six rounds. However, when they stop playing *A* has only won five rounds and *B* three. How should they divide the wager? It would be another 200 years before this problem was solved.

In 1545 Girolamo Cardano wrote the books *Ars magna* (the great art) and *Liber de ludo aleae* (the book on games of chance). This was the first attempt to use mathematics to describe statistics and probability, and accurately described the probabilities of throwing various numbers with dice. Galileo expanded on this by calculating probabilities using two dice. At the same time the quantification of all aspects of daily life (art, music, time, space) between the years 1250 and 1600 made possible the numerical analysis of nature and, consequently, the discovery of the distribution of events and their rules (Crosby 1996).

It was finally Blaise Pascal who refined the theories of statistics and, with Pierre de Fermat, solved the "balla" problem of Paccioli (Devlin 2008). All these paved the way for modern statistics, which essentially began with the use of actuarial tables to determine insurance for merchant ships (Hacking 1984, 1990). Pascal was also the first to apply probability studies to the theory of decision (see his *Pensées*, 233), curiously, in the field of religious decisions. It is in this historical moment that the Latin term "probabilis" acquires its actual meaning, evolving from "worthy of approbation" to "numerical assessment of likelihood on a determined scale" (Moussy 2005).

In 1662, Antoine Arnauld and Pierre Nicole published the influential *La logique ou l'art de penser*, where we can find statistical probabilities. Games and their statistical roots worried people like Cardano, Pascal, Fermat, and Huygens (Weatherford 1982), although all of them were immersed in a strict mechanistic paradigm. Huygens is considered the first scientist interested in scientific probability, and in 1657 he published *De ratiotiniis in aleae ludo*. In 1708 Pierre Raymond de Montmort published his *Essay d'Analyse sur les Jeux de Hazard*, probably the first comprehensive text on probability theory. It was the next step after Pascal's work on combinatorics and its application to the solution of problems on games of chance. Later, De Moivre wrote the influential *De mensura sortis* (1711), and 78 years later, Laplace published

his *Philosophical Essay About Probability*. In the 1730s, Daniel Bernoulli (Jacob Bernoulli's nephew) developed the idea of utility as the mathematical combination of the quantity and perception of risk. Gottfried Leibniz at the beginning of the eighteenth century argued in several of his writings against the idea of chance, defending deterministic theories. According to him, chance was not part of the true nature of reality but the result of our incomplete knowledge. In this sense, probability is the estimation of facts that could be completely known and predicted, not the basic nature of things. Even morality was guided by natural laws, as Immanuel Kant argued in his *Foundations of the Metaphysics of Morals* (1785).

In 1763 an influential paper written by the Reverend Thomas Bayes was published posthumously. Richard Price, who was a friend of his, worked on the results of his efforts to find the solution to the problem of computing a distribution for the parameter of a ▶binomial distribution: *An Essay towards solving a Problem in the Doctrine of Chances*. Proposition 9 in the essay represented the main result of Bayes. Degrees of belief are therein considered as a basis for statistical practice. In a nutshell, Bayes proposed a theorem in which "probability" is defined as an index of subjective confidence, at the same time taking into account the relationships that exist within an array of simple and conditional probabilities. ▶Bayes' theorem is a tool for assessing how probable evidence can make a given hypothesis (Swinburne 2005). So, we can revise predictions in the light of relevant evidence and make a Bayesian inference, based on the assignment of some a priori distribution of a parameter under investigation (Stigler 1990). The classical formula of Bayes' rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where our *posterior* belief $P(A|B)$ is calculated by multiplying our *prior* belief $P(A)$ by the *likelihood* $P(B|A)$ that B will occur if A is true. This classical version of Bayesianism had a long history, beginning with Bayes and continuing through Laplace to Jeffreys, Keynes, and Carnap in the twentieth century. Later, in the 1930s, a new type of Bayesianism appeared, the "subjective Bayesianism" of Ramsey and De Finetti (Ramsey 1931; de Finetti 1937; Savage 1954).

At the end of the nineteenth century, a lot of things were changing in the scientific and philosophical arena. The end of the idea of "causality" and the conflicts about observation lay at the heart of the debate. Gödel attacked Hilbert's axiomatic approach to mathematics and Bertrand Russell, as clever as ever, told us: "The law of causality (. . .) is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm (. . .) The principle "same cause, same effect," which philosophers imagine to be vital to science, is therefore utterly otiose" (Suppes 1970, p. 5). Nevertheless, scientists like Einstein were reluctant to accept the loss of determinism in favor of a purely random Universe; Einstein's words "God does not play dice" are the example of the difficulty of considering the whole world as a world of probabilities, with no inner intentionality, nor moral direction. On the other hand, scientists like Monod (*Chance and Necessity*, 1970) accepted this situation. In both cases, there is a deep consideration of the role of probability and chance in the construction of the philosophical and scientific meaning about reality.

In the 1920s there arose from theworks of Fisher (1922) and Neyman and Pearson (1928) the *classic* statistical paradigm: frequentism. They use the relative frequency concept, that is, you must perform one experiment many times and measure the proportion where you get a positive result. This proportion, if you perform the experiment enough times, is the probability. If Neyman and Pearson wrote their first joint paper and presented their approach as *one among alternatives*, Fisher, with his null hypothesis testing, gave a different message: his statistics was the formal solution of the problem of inductive inference (Gigerenzer 1990, p. 228).

From then on, these two main schools, Bayesian and Frequentist, were fighting each other to demonstrate that theirs was the superior and only valid approach (Vallverdú 2008).

Finally, with the advent of the information era and all the (super)computer scientific simulations, Bayesianism has again achieved a higher status inside the community of experts on probability. Bayesian inference also allows intelligent and real-time monitoring of computational clusters, and its application in belief networks has proved to be a good technique for diagnosis, forecasting, and decision analysis tasks. This fact has contributed to the increasing application of parallel techniques for large Bayesian networks in expert systems (automated causal discovery, AI…) (Korb and Nicholson 2003).

Acknowledgments

This research was supported by the project "El diseño del espacio en entornos de cognición distribuida: plantillas y affordances," MCI [FFI2008-01559/FISO].

About the Author

Jordi Vallverdú is a lecturer professor of philosophy in science and computing at Universitat Autonoma de Barcelona. He holds a Ph.D. in philosophy of science (UAB) and a master's degree in history of sciences (UAB).

After a short research stay as a fellowship researcher at Glaxo-Wellcome Institute for the History of Medicine, London (1997), and a research assistant of Dr. Jasanoff at J.F.K. School of Government, Harvard University (2000), he worked in computing epistemology issues and bioethic and synthetic emotions. He is listed as an EU Biosociety Research Expert and is a member of the E-CAP' Steering (http://www.ia-cap.org/administration.php). He leads a research group, SETE (Synthetic Emotions in Technological Environments), which has published about computational models of synthetic emotions and their implementation into social robotic systems. He is Editor-in-Chief of the *International Journal of Synthetic Emotions* (IJSE) and has edited (and written as an included author) the following books: *Handbook of research on synthetic emotions and sociable robotics: new applications in affective computing and artificial intelligence* (with D. Casacuberta, Information Science Publishing, 2009) and *Thinking machines and the philosophy of computer science: concepts and principles* (Ed., 2010).

Cross References

►Actuarial Methods
►Bayes' Theorem
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Foundations of Probability
►Philosophy of Probability
►Probability Theory: An Outline
►Statistics and Gambling
►Statistics, History of

References and Further Reading

Crosby AW (1996) The measure of reality: quantification in Western Europe 1250–1600. Cambridge University Press, Cambridge
Dandoy JR (2006) Astragali through time. In: Maltby M (ed) Integrating zooarchaeology. Oxbow Books, Oxford, pp 131–137
David FN (1998) Games, gods and gambling, a history of probability and statistical ideas. Dover, Mineola
De Finetti B (1937) La Prevision: Ses Lois Logiques, Ses Sources Subjectives. Annals de l'Institute Henri Poincaré 7:1–68
Devlin K (2008) The unfinished game: Pascal, Fermat, and the seventeenth-century letter that made the world modern. Basic Books, New York
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos trans Roy Soc London Ser A 222:309–368
Gigerenzer G et al (1990) The Empire of chance. How probability changed science and everyday life. Cambridge University Press, Cambridge
Hacking I (1984) The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Cambridge University Press, Cambridge
Hacking I (1990) The taming of chance. Cambridge University Press, Cambridge
Hald A (1988) A history of probability and statistics and their applications before 1750. Wiley, New York
Korb KB, Nicholson AE (2003) Bayesian artificial intelligence. CRC Press, Boca Raton
Moussy C (2005) Probare, probatio, probabilis' dans de vocabulaire de la démonstration. Pallas 69:31–42
Neymanv J, Pearson ES (1928) On the use and interpretation certain test criteria for purposes of statistical inferences. Biometrika 20(A):175–240, 263–294
Ramsey FP (1931) Truth and probability (1926). In: Braithwaite RB (ed) The foundations of mathematics and other logical essays, Ch. VII. Kegan Paul, Trench, Trubner/Harcourt, Brace, London/New York, pp 156–198
Savage LJ (1954) The foundations of statistics. Wiley, New York
Stigler SM (1990) The history of statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge
Suppes P (1970) A probabilistic theory of causality. North-Holland, Helsinki
Swinburne R (ed) (2005) Bayes's theorem. In: Proceedings of the British academy, vol 113. Oxford University Press, London
Vallverdú J (2008) The false Dilemma: Bayesian vs. Frequentist. E – LOGOS Electron J Philos 1–17. http://e-logos.vse.cz/index.php?target=indexyear
Weatherford R (1982) Philosophical foundations of probability theory. Routledge & Kegan Paul, London

Probit Analysis

Tiberiu Postelnicu
Professor Emeritus
Romanian Academy, Bucharest, Romania

Introduction

The idea of probit analysis was originally published in *Science* by Chester Ittner Bliss (1899–1979) in 1934. He was primarily concerned with finding an effective pesticide to control insects that fed on grape leaves. By plotting the response of the insects to various concentrations of pesticides, he could visually see that each pesticide affected the insects at different concentrations, but he did not have a statistical method to compare this difference. The most logical approach would be to fit a regression of the response versus the concentration or dose and compare between the different pesticides. The relationship of response to dose was sigmoid in nature and at that time regression was only used on linear data. Therefore, Bliss developed the idea of transforming the sigmoid dose–response curve to a straight line. When biological responses are plotted against their causal stimuli (or their logarithms) they often form a

sigmoid curve. Sigmoid relationships can be linearized by transformations such as logit, probit, and angular. For most systems the probit (normal sigmoid) and logit (logistic sigmoid) give the most closely fitting result. Logistic methods are useful in epidemiology, but in biological assay work, probit analysis is preferred. David Finney, from the University of Edinburgh, took Bliss' idea and wrote a book entitled *Probit Analysis* in 1947, and since this year it is still the preferred statistical method in understanding dose–response relationships.

Probit

In probability theory and statistics, the *probit function* is the inverse cumulative distribution function (CDF), or *quantile function* associated with the standard normal distribution. It has applications in exploratory statistical graphics and specialized regression modeling of binary response variables. For the standard normal distribution, the CDF is commonly denoted $\Phi(z)$, which is a continuous, monotone-increasing sigmoid function whose domain is the real line and range is $(0,1)$. The probit function gives the inverse computation, generating a value of an $N(0,1)$ random variable, associated with specified cumulative probability. Formally, the probit function is the inverse of $\Phi(z)$, denoted $\Phi^{-1}(p)$. In general, we have $\Phi(\text{probit}(p)) = p$ and $\text{probit}(\Phi(z)) = z$. Bliss proposed transforming the percentage into "*probability unit*" (or "*probit*") and included a table to aid researchers to convert their kill percentages to his probit, which they could then plot against the logarithm of the dose and thereby, it was hoped, obtain a more or less straight line. Such a probit model is still important in toxicology, as well as in other fields. It should be observed that probit methodology, including numerical optimization for fitting of probit functions, was introduced before widespread availability of electronic computing and, therefore, it was convenient to have probits uniformly positive.

Related Topics

The probit function is useful in statistical analysis for diagnosing deviation from normality, according to the method of Q–Q plotting. If a set of data is actually a sample of a normal distribution, a plot of the values against their probit scores will be approximately linear. Specific deviation from normality such as asymmetry, heavy tails, or bimodality can be diagnosed based on the detection of specific deviations from linearity. While the $Q - Q$ plot can be used for comparison with any distribution family (not only the normal), the normal Q–Q plot is a relatively standard exploratory data analysis procedure because the assumption of normality is often a starting point for analysis.

The normal distribution CDF and its inverse are not available in closed form, and computation requires careful use of numerical procedures. However, the functions are widely available in software for statistics and probability modeling, and also in spreadsheets. In computing environments where numerical implementations of the inverse error function are available, the probit function may be obtained as $\text{probit}(p) = \sqrt{2}\text{erf}^{-1}(2p - 1)$. An example is MATLAB, where an "erfinv" function is available and the language MATHEMATICA implements "InverseErf." Other environments directly implement the probit function in the R programming language.

Closely related to the probit function is the *logit function* using the "odds" $p/(1-p)$, where p is the proportional response, i.e., r out of n responded, so there is $p = r/n$ and $\text{logit}(p) = \log \text{odds} = \log(p/(1-p))$. Analogously to the probit model, it is possible to assume that such a quantity is related linearly to a set of predictors, resulting in the logit model, the basis in particular of logistic regression model (see ►Logistic Regression), the most prevalent form of regression analysis for binary response data. In current statistical practice, probit and logit regression models are often handled as cases of the generalized linear model (see ►Generalized Linear Models).

Probit Model

In statistics and related fields, a *probit model* is a specification for a binary response model that employs a probit link function. This model is most often estimated using the standard maximum likelihood procedure; such an estimation is called *probit regression*. A fast method for computing maximum likelihood estimates for probit models was proposed by Ronald Fisher in an Appendix to the article of Bliss in 1935.

Probit analysis is a method of analyzing the relationship between a stimulus (dose) and the quantal (all or nothing) response. Quantitative responses are almost always preferred, but in many situations they are not practical. In these cases, it is only possible to determine if a certain response has occurred. In a typical quantal response experiment, groups of animals are given different doses of a drug. The percent dying at each dose level is recorded. These data may then be analyzed using probit analysis. StatPlus includes two different methods of probit analysis, but the Finney method is the most important and useful. The probit model assumes that the percent response is related to the log dose as the cumulative normal distribution, that is, the log doses may be used as variables to read the percent dying from the cumulative normal. Using the

normal distribution, rather than other probability distributions, influences the predicted response rate at the high and low ends of possible doses, but has little influence near the middle. Much of the comparison of different drugs is done using response rates of 50%. The probit model may be expressed mathematically as follows:

$$P = a + b(\log(\text{Dose})),$$

where P is five plus the inverse normal transform of the response rate (called the probit). The five is added to reduce the possibility of negative probits, a situation that caused confusion when solving the problem by hand.

Suppose the response variable Y is *binary*, that is, it can have only two possible outcomes, which we will denote as 1 and 0. For example, Y may represent presence/absence of a certain condition, success/failure of some device, and answer yes/no on a survey. We also have a vector of regression X, which are assumed to influence the outcome Y. Specifically, we assume that the model takes the form

$$P[Y = 1|X] = \Phi(X'\beta),$$

where P is the probability and Φ is the probit function – the CDF of the standard normal distribution. The parameters β are typically estimated by maximum likelihood. For more complex probit analysis, such as the calculation of relative potencies from several related dose–response curves, consider nonlinear optimization software or specialist dose–response analysis software. The latter is a FORTRAN routine written by David Finney and Ian Craigie from Edinburgh University Computing Center. MLP or GENSTAT can be used for a more general nonlinear model fitting. We must take into account that the standard probit analysis is designed to handle only quantal responses with binomial error distributions. Quantal data, such as the number of subjects responding versus the total number of subjects tested, usually have binomial error distributions. We should not use continuous data, such as percent maximal response, with probit analysis as these data are likely to require regression methods that assume a different error distribution.

Applications

Probit analysis is used to analyze many kinds of dose–response or binomial response experiments in a variety of fields. It is commonly used in toxicology to determine the relative toxicity of chemicals to living organisms. This is done by testing the response of an organism under various concentrations of each of the chemicals in question and then comparing the concentrations at which one encounters a response. The response is always binomial and the relationship between the response and the various concentrations is always sigmoid. Probit analysis acts as a transformation from sigmoid to linear and then runs a regression on the relationship. Once the regression is run, we can use the output of the probit analysis to compare the amount of chemical required to create the same response in each of the various chemicals. There are many points used to compare the differing toxicities of chemicals, but the LC50 (liquids) or LD50 (solids) are the most widely used outcomes of the modern dose–response experiments. The LC50/LD50 represent the concentration (LC50) or dose (LD50) at which 50% of the population responds. It is possible to use probit analysis with various methods such as statistical packages SPSS, SAS, R, or S, but it is good to see the history of the methodology to get a thorough understanding of the material. We must take care that probit analysis assumes that the relationship between number responding (non-percent response) and concentration is normally distributed; if not, logit is preferred.

The properties of the estimates given by probit analysis have been studied also by Ola Hertzberg (1974). The up-and-down technique is the best known among staircase methods for estimating the parameters in quantal response curves (QRC). Some small sample properties of probit analysis are considered and in the estimate distribution the medians are used as a measure of location.

About the Author

Tiberiu Postelnicu (born on June 15, 1930, Campina), received his Ph.D. in Mathematics, University of Bucharest, in 1957. He was Head of Department of Biostatistics, Carol Davila University of Medicine and Pharmacy, Bucharest, and Department of Biometrics, Centre of Mathematical Statistics of the Romanian Academy, Bucharest. He is a member of the International Statistical Institute, Biometric Society, Bernoulli Society for Mathematical Statistics and Probability, Italian Society for Statistics, and the New York Academy of Science. Dr. Postelnicu was a member of the editorial boards of the *Biometrical Journal* and *Journal for Statistical Planning and Inference*. He was awarded the Gheorghe Lazar Prize for Mathematics, Romanian Academy (1972). Currently, Professor Postelnicu is President of the Commission for Biometrics of the Romanian Academy.

Cross References

►Agriculture, Statistics in
►Econometrics
►Generalized Linear Models
►Logistic Regression

►Nonlinear Models
►Normal Distribution, Univariate

References and Further Reading

Bliss CI (1934) The method of probit. Science 79(2037):38–39
Bliss CI (1935) The calculation of the dosage–mortality curve. Ann Appl Biol 22:1, 134–167
Bliss CI (1938) The determination of the dosage-mortality curve from small numbers. Q J Pharmacol 11:192–216
Collett D (1991) Modelling binary data. Chapman & Hall, London
Finney DJ (1947) Probit analysis, 1st edn. Cambridge University Press, Cambridge
Finney DJ, Stevens WL (1948) A table for the calculation of working probits and weights in probit analysis. Biometrika 35(1–2): 191–201
Finney DJ (1971) Probit analysis, 3rd edn. Cambridge University Press, Cambridge
Greenberg BG (1980) Chester I Bliss, 1899–1979. Int Stat Rev 8(1):135–136
Hertzberg JO (1974) On small sample properties of probit analysis. In: Proceedings of the 8th International Biometric Conference. Constantza, Romania, pp 153–162
McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall, London

Promoting, Fostering and Development of Statistics in Developing Countries

Noël H. Fonton[1], Norbert Hounkonnou[2]
[1]Professor, Head of Centre of Biostatistics and Director Laboratory of Modeling of Biological Phenomena University of Abomey-Calavi, Cotonou, Benin Republic
[2]Professor, Chairman of International Chair of Mathematical Physics and Applications (ICMPA UNESCO Chair) Cotonou, Benin Republic

Statistical methods are universal and hence their applicability depends neither on geographical area nor on a people's culture. Promoting and increasing the use of statistics in developing countries can help to find solutions to the needs of their citizens. Developing countries are confronted with endemic poverty that requires implementable solutions for alleviating suffering. Such poverty is a signal call to the world to meet fundamental human needs—food, adequate shelter, access to education and healthcare, protection from violence, and freedom. Statistics and statistical tools, the matching between hypothesis, data collection, and statistical method, are necessary as development strategies in the developing counties are formulated to address these needs.

First, a reliable basis for the implementation of strategies against poverty and achievement of the Millennium Development Goals require good statistics, an essential element of good governance. Therefore, important indicators to inform and monitor development policies are often derived from household surveys, which have become a dominant form of data collection in developing countries. Such surveys are an important source of socioeconomic data. Azouvi (2001) has proposed a low-cost statistical program in four areas: statistical coordination, national accounts, economic and social conjuncture, and dissemination. To increase awareness that good statistics are important for achieving better development results, the Marrakech Action Plan for Statistics (MPS) was developed in 2004. This global plan for improving development statistics was agreed upon at a second round-table for best managing development results (World Bank, 2004). The idea is that better data are needed for better results in order to improve development statistics. One indicator to ensure the application of this MPS is the full participation of developing countries in the 2010 census round. Additionally, funds allocated by the World Bank from the Development Grant Facility and technical assistance from national universities are critical.

Second, national capacity-building in statistics is very limited. Even in developed countries, secondary school students, together with their teachers, rarely see the applicability and the challenge of statistical thinking (Boland, 2002). This situation exists to a greater degree in the university training systems of many developing countries. It is suggested that statistical programs should be reviewed and executed by statisticians and examples of a local nature should be used. This is possible with sufficient number of statisticians in various fields. According to Lo (2009), there are 5 to 10 holders of doctoral degrees in statistics per country, with higher numbers in some countries. There is a low number of statisticians in developing countries due to the lack of master's degree programs in statistics. In sub-Saharan, French-speaking African countries, the "Statistiques pour l'Afrique Francophone et Applications au vivant" (STAFAV) project is being implemented in three parts: a network for master's training in applied statistics at Cotonou (Benin), Saint-Louis (Senegal), and Yaounde (Cameroon); some PhD candidates jointly supervised by scientists from African universities and French universities; and the development of statistical research through an African Network of Mathematical Statistics and Applications (RASMA). STAFAV constitutes a good means for increasing statistical capacity in

developing countries. With the launch of the Statistical Pan African Society (SPAS), more visibility for research and the use of statistics and probability may be achieved via the African Visibility Program in Statistics and Probability. This society is an antidote to the very isolated work of statistics researchers and allows researchers to share experiences and scientific work.

Third, statistics are a tool for solving the problems of developing countries, but the development of research activities is a challenge. Many of these countries are located in tropical areas; therefore, there are major differences between them and other countries due to high biological variability and the probability distribution of the studied phenomena is frequently misunderstood. Control of biological variability requires wide use of probability theory. Because of the disparate populations and subpopulations in the experimental data, the use of one probability distribution should be called into question. Lacking sophisticated statistical methods, development of statistical methods of mix-distributions becomes a challenge. Economic loss, agriculture, water policies, and health (malaria, HIV/AIDS, and recently H1N1) are the major areas for research programs in which statistics have a major role to play. Development of statistical research programs directed at the well-being of local people is necessary.

The sustainability of statistical development is another important issue in the field. Graduate statisticians, when returning to their native countries, often do not have facilities for continuing education, documentation, or statistics software packages. To foster statistics in developing countries, national statistics institutes, universities, and research centers need to increase funds allocated for subscribing to statistical journals, mainly online, and software, with a staff properly trained on those fields. Statisticians must be offered opportunities to attend annual conferences and to do research and/or professional training in a statistical institute outside of their countries.

Contributions of statisticians from developed countries, working or teaching in developing countries, are welcome. Specifically, they can join research teams in developing countries, share experiences with them, help acquire funding, and teach.

Cross References

►African Population Censuses
►Careers in Statistics
►Learning Statistics in a Foreign Language
►National Account Statistics
►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics: Developing Country Perspective
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistics and Climate Change
►Statistics Education

References and Further Reading

Azouvi A (2001) Proposals for a minimum programme for statistics in developing countries. Int Stat Rev 69(2):333–343
Boland PJ (2002) Promoting statistics thinking amongst secondary school students in national context. ICOTS6, p 6
Lo GS (2009) Probability and statistics in Africa. IMS Bull 38(7):8
World Bank (2004) The marrakech action plan for statistics. Second international roundtable on managing for development results. Morocco, p 19

Properties of Estimators

Paul H. Garthwaite
Professor of Statistics
The Open University, Milton Keynes, UK

Estimation is a primary task of statistics and estimators play many roles. Interval estimators, such as confidence intervals or prediction intervals, aim to give a range of plausible values for an unknown quantity. Density estimators aim to approximate a probability distribution. These and other varied roles of estimators are discussed in other sections. Here attention is restricted to point estimation, where the aim is to calculate from data a single value that is a good estimate of an unknown parameter.

We will denote the unknown parameter by θ, which is assumed to be a scalar. In the standard situation there is a statistic T whose value, t, is determined by sample data. T is a random variable and it is referred to as a (point) estimator of θ if t is an estimate of θ. Usually there will be a variety of possible estimators so criteria are needed to separate good estimators from poor ones. There are a number of desirable properties which we would like estimators to possess, though a property will not necessarily identify a unique "best" estimator and rarely will there be an estimator that has all the properties mentioned here. Also, caution must be exercised in using the properties as a reasonable property will occasionally lead to an estimator that is unreasonable.

One property that is generally useful is unbiasedness. T is an unbiased estimator of θ if, for any θ, $E(T) = \theta$. Thus T is unbiased if, on average, it tends neither to be bigger nor smaller than the quantity it estimates, regardless of

the actual value of the quantity. The bias of T is defined to be $E(T) - \theta$. Obviously a parameter can have more than one unbiased estimator. For example, if θ is the mean of a symmetric distribution from which a random sample is taken, then T is an unbiased estimator if it is the mean, median or mid-range of the sample. It is also the case that sometimes a unique unbiased estimator is not sensible. For example, Cox and Hinkley (1974, p. 253) show that if a single observation is taken from a geometric distribution with parameter θ, then there is only one unbiased estimator and its estimate of θ is either 1 (if the observation's value is 1) or 0 (if the observation's value is greater than 1). In most circumstances these are not good estimates.

It is desirable, almost by definition, that the estimate t should be close to θ. Hence the quality of an estimator might be judged by its expected absolute error, $E(|T - \theta|)$, or its mean squared error, $E[(T - \theta)^2]$. The latter is used far more commonly, partly because of its relationship to the mean and variance of T:

$$\text{mean squared error} = \text{variance} + (\text{bias})^2. \qquad (1)$$

If the aim is to find an estimator with small mean squared error (MSE), clearly unbiasedness is desirable, as then the last term in Eq. (1) vanishes. However, unbiasedness is not essential and trading a small amount of bias for a large reduction in variance will reduce the MSE. Perhaps the best known biased estimators are the regression coefficients given by ridge regression (see ▶Ridge and Surrogate Ridge Regressions), which handles multicollinearities in a regression problem by allowing a small amount of bias in the coefficient estimates, thereby reducing the variance of the estimates.

It may seem natural to try to find estimators which minimize MSE, but this is often difficult to do. Moreover, given any estimator, there is usually some value of θ for which that estimator's MSE is greater than the MSE of some other estimator. Hence the existence of an estimator with a *uniformly* minimum MSE is generally in doubt. For example, consider the trivial and rather stupid estimator that ignores the data and chooses some constant θ_0 as the estimator of θ. Should θ actually equal θ_0, then this estimator has an MSE of 0 and other estimators will seldom match it. Thus other estimators will not have a uniformly smaller MSE than this trivial estimator.

Restricting attention to unbiased estimators solves many of the difficulties of working with MSE. The task of minimizing MSE reduces to that of minimizing variance and substantial theory has been developed about minimum variance unbiased estimators (MVUEs). This includes two well-known results, the *Cramér–Rao lower bound* and the ▶*Rao–Blackwell theorem.* The Cramér–Rao lower bound is I_θ^{-1}, where I_θ is the Fisher information about θ. (I_θ is determined from the likelihood for θ.) Subject to certain regularity conditions, the Cramér–Rao lower bound is a lower bound to the variance of any unbiased estimator of θ.

A benefit of the Cramér–Rao lower bound is that it provides a numerical scale-free measure for judging an estimator: the *efficiency* of an unbiased estimator is defined as the ratio of the Cramér–Rao lower bound to the variance of the estimator. Also, an unbiased estimator is said to have the property of being efficient if its variance equals the Cramér–Rao lower bound. Efficient estimators are not uncommon. For example, the sample mean is an efficient estimator of the population mean when sampling is from a normal distribution or a Poisson distribution, and there are many others. By definition, only an MVUE might be efficient.

Sufficiency is a property of a statistic that can lead to good estimators. A statistic S (which may be a vector) is sufficient for θ if it captures all the information about θ that the data contain. For example, the sample variance is sufficient for the population variance when data are a random sample from a normal distribution – hence to make inferences about the population variance we only need to know the sample variance and not the individual data values. The definition of sufficiency is a little more transparent in ▶Bayesian statistics than in classical statistics (though the definitions are equivalent). In the Bayesian approach, S is sufficient for θ if the distribution of θ, given the value of S, is the same as θ's distribution given all the data. i.e. $g(\theta \mid S) = g(\theta \mid \text{data})$, where $g(\,.\,)$ is the p.d.f. of θ. In the classical definition (where θ cannot be considered to have a distribution), S is sufficient for θ if the conditional distribution of the data, given the value of S, does not depend on θ. A sufficient statistic may contain much superfluous information along with the information about θ, so the concept of a *minimal sufficient statistic* is also useful. A statistic is minimal sufficient if it can be expressed as a function of every other sufficient statistic.

The Rao–Blackwell theorem shows the importance of sufficient statistics when seeking unbiased estimators with small variance. It states that if $\widehat{\theta}$ is an unbiased estimator of θ and S is a sufficient statistic, then

1. $T_S = E(\widehat{\theta} \mid S)$ is a function of S alone and is an unbiased estimator of θ.
2. $\text{Var}(T_S) \le \text{var}(\widehat{\theta})$.

The theorem means that we can try to improve on any unbiased estimator by taking its expectation conditional on a sufficient statistic – the resulting estimator will also be unbiased and its variance will be smaller than, or equal

to, the variance of the original estimator. Stronger results hold if a minimal sufficient statistic is also *complete*: S is complete if $E[h(S)]$ cannot equal 0 for all θ unless $h(S) \equiv 0$ almost everywhere (where h is any function). If S is complete there is at most one function of S that is an unbiased estimator of θ. Suppose now, that S is a complete minimal sufficient statistic for θ. An important result is that if $h(S)$ is an unbiased estimator of θ, then $h(S)$ is an MVUE for θ, if an MVUE exists. The consequence is that, when searching for an MVUE, attention can be confined to functions of a complete sufficient statistic.

Turning to asymptotic properties, suppose that data consist of a ►simple random sample of size n and consider the behavior of an estimator T as $n \to \infty$. An almost essential property is that the estimator should be consistent: T is a consistent estimator of θ if T converges to θ in probability as $n \to \infty$. Consistency implies that, as the sample size increases, any bias in T tends to 0 and the variance of T also tends to 0.

Two useful properties that do not relate directly to the accuracy of an estimator are ►*asymptotic normality* and *invariance*. When sample sizes are large, confidence intervals and hypothesis tests are often based on the assumption that the distribution of an estimator is approximately normal. Hence asymptotic normality is a useful property in an estimator, especially if approximate normality holds quite well for modest sample sizes. Invariance of estimators relates to the method of forming them. It is the notion that if we take a transformation of a parameter, then ideally its estimator should transform in the same way. For example, let $\phi = g(\theta)$, where ϕ is a one-to-one function of θ. Then if a method of forming estimators gives t_1 and t_2 as estimates of ϕ and θ, invariance would imply that t_1 necessarily equalled $g(t_2)$. Maximum likelihood estimators are invariant.

We have assumed that the unknown parameter (θ) is a scalar. Concepts such as unbiasedness, sufficiency, consistency, invariance and asymptotic normality extend very naturally to the case where the unknown parameter is a vector. If θ is a vector but an estimate of just one of its components is required, then a vector-form of the Cramer–Rao lower bound yields a minimum variance for any unbiased estimator of the component. Simultaneous estimation of more than one component, however, raises new challenges unless estimating each component separately and combining the estimates is optimal.

While a search for MVUEs has been a focus of one area of statistics, other branches of statistics want different properties in estimators. Robust methods want point estimators that are comparatively insensitive to a few aberrant observations or the odd outlier. Nonparametric methods want to estimate a population mean or variance, say, without making strong assumptions about the population distribution. These branches of statistics do not place paramount importance on unbiasedness or minimal variance, but they nevertheless typically seek estimators with low bias and small variance – it is just that their estimators must also satisfy other requirements. In contrast, Bayesian statistics uses a markedly different framework for choosing estimators. In its basic form the parameters of the sampling model are given a prior distribution, while a loss function specifies the penalty for inaccuracy in estimating a parameter. The task is then to select an estimator or decision rule that will minimize the expected loss, so minimizing expected loss is the property of dominant importance.

Many other sections of this encyclopedia also consider point estimators and point estimation methods. These include sections on nonparametrics, robust estimation, Bayesian methods and decision theory. The focus in this section has been the classical properties of point estimators. Deeper discussion of this topic and proofs of results are given in most advanced textbooks on statistical inference or theoretical statistics, such as Bickel and Doksum (2000), Cox and Hinkley (1974), Garthwaite et al. (2002), and Lehmann and Casella (1998).

About the Author

Paul Garthwaite is Professor of Statistics at the Open University, UK, where he was Head of the Department of Statistics from 2001–2004 and again in 2006. He worked for twenty years in the University of Aberdeen and has held visiting positions at universities in New York State, Minnesota, Brisbane and Sydney. In 1983 he was awarded the L J Savage Prize, a prize now under the auspices of the International Society for Bayesian Analysis. He has published 80 journal papers and is co-author of two books, *Statistical Inference* (Oxford University Press, 2002) and *Uncertain Judgements: Eliciting Expert' Probabilities* (Willey, 2006).

Cross References

►Approximations for Densities of Sufficient Estimators
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Bayesian Statistics
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Cramér–Rao Inequality
►Estimation
►Estimation: An Overview
►Minimum Variance Unbiased
►Rao–Blackwell Theorem

►Sufficient Statistics
►Unbiased Estimators and Their Applications

References and Further Reading

Bickel PJ, Doksum KA (2000) Mathematical statistics: basic ideas and selected topics, 2nd edn. Prentice Hall, London
Cox DR, Hinkley DV (1974) Theoretical statistics. Wiley, New York
Garthwaite PH, Jolliffe IT, Jones B (2002) Statistical inference, 2nd edn. Oxford University Press, Oxford
Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York

Proportions, Inferences, and Comparisons

George A. F. Seber
Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand

A common problem in statistics, and especially in sample surveys, is how to estimate the proportion p $(= 1 - q)$ of people with a given characteristic (e.g., being left-handed) in a population of known size N. If there are M left-handed people in the population, then $p = M/N$. The usual method of estimating p is to take a ►simple random sample (SRS), that is, a random sample without replacement, of size n and count the number of left-handed people, x, in the sample. If the sample is representative, then p can be estimated by the sample proportion $\hat{p} = x/n$. To make inferences about p we need to use the probability distribution of x, namely the Hypergeometric distribution (see ►Hypergeometric Distribution and Its Application in Statistics), a distribution that is difficult to use. From this distribution we can get the mean and variance of x and hence of $\hat{p}$, namely

$$\mu_{\hat{p}} = p \quad \text{and} \quad \sigma^2_{\hat{p}} = r\frac{pq}{n},$$

where $r = (N - n)/(N - 1) \approx 1 - f$, and f is the sampling fraction n/N, which can be ignored if it is less that 0.1 (or better 0.05). Fortunately, for sufficiently large N, M and n, x and $\hat{p}$ are approximately normal so that $z = (\hat{p} - p)/\sigma_{\hat{p}}$ has an approximate standard normal distribution (with mean 0 and variance 1). This approximation will still hold if we replace p by $\hat{p}$ in the denominator of $\sigma_{\hat{p}}$ to get $\hat{\sigma}_{\hat{p}}$ giving us an approximate 95% confidence interval $\hat{p} \pm 1.96\hat{\sigma}_{\hat{p}}$.

Inverse sampling can also be used to estimate p, especially when the characteristic is rare. Random sampling is continued until x units of the given characteristic are selected, n now being random, and Haldane in 1945 gave the estimate $\hat{p}_I = (x - 1)/(n - 1)$. This has variance estimate $\hat{\sigma}^2_{\hat{p}_I} = r_I(\hat{p}_I\hat{q}_I)/(n - 2)$, where $r_I = 1 - (n - 1)/N$ for without replacement and $r_I = 1$ for with replacement, and an approximate 95% confidence interval for p is $\hat{p}_I \pm 1.96\hat{\sigma}_{\hat{p}_I}$ (Salehi and Seber 2001).

Another application of this theory is in the case where M consists of a known number of marked animals released into a population of unknown size, but with f known to be sufficiently small so that we can set $r = 1$. The confidence interval for p can then be rearranged to give a confidence interval for N. This simple idea has lead to a very large literature on capture-recapture (Seber 2002).

Returning to our example relating to left-handed people, when we choose the first person from the population, the probability of getting a left-handed person will be p so that the terms "probability" and "proportion" tend to be used interchangeably in the literature, although they are distinct concepts. They can be brought even closer together if sampling is with replacement for then the probability of getting a left-handed person at each selection remains at p and x now has a ►Binomial distribution, as we have n independent trials with probability of "success" being p. The above formulas for means and variances and confidence interval are still the same except that r is now exactly 1. This is not surprising as we expect sampling with replacement to be a good approximation for sampling without replacement when a small proportion of a population is sampled. Confidence intervals for the Binomial distribution have been studied for many years and a variety of approximations and modifications have been considered, for example, Newcombe (1998a). "Exact" confidence intervals, usually referred to as the Clopper–Pearson intervals, can also be computed using the so-called "tail" probabilities of the binomial distribution, which are related to a ►Beta distribution (cf. Agresti and Coull 1998). We can also use the above theory to test null hypotheses like H_0: $p = p_0$, though such hypotheses apply more to probabilities than proportions.

When it comes to comparing two proportions, there are three different experimental situations that need to be considered. Our example for explaining these relates to voting preferences. Suppose we wish to compare the proportions, say p_i $(i = 1, 2)$, of people voting for a particular candidate in two separate areas and we do so by taking an SRS of size n_i in each area and computing $\hat{p}_i$ for each area. In comparing the areas we will be interested in estimating $\theta = p_1 - p_2$ using $\hat{\theta} = \hat{p}_1 - \hat{p}_2$. As the two estimates are statistically independent, and assuming f can be neglected for each sample, we have

$$\sigma^2_{\hat{\theta}} = \frac{p_1q_1}{n_1} + \frac{p_1q_2}{n_2},$$

which we can estimate by replacing each p_i by its estimate. Assuming the normal approximation is valid for each sample, we have the usual approximate 95% confidence interval $\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}$ for θ. This theory also applies to comparing two Binomial probabilities and, as with a single probability, a number of methods have been proposed (see Newcombe 1998b). Several procedures for testing the null hypothesis H_0: $\theta = 0$ are available including an "exact" test due to Fisher and an approximate Chi-square test with or without a correction for continuity.

A second situation is when we have a single population and the p_is now refer to two different candidates. The estimates $\hat{p}_i$ $(i = 1, 2)$ are no longer independent and we now find, from Scott and Seber (1983), that

$$\sigma^2_{\hat{\theta}} = \frac{1}{n}[p_1 + p_2 - (p_1 - p_2)^2],$$

where n is the sample size. Once again we can replace the unknown p_i by their estimates and, assuming f can be ignored, we can obtain an approximate confidence interval for θ, as before.

A third situation occurs when two proportions from the same populations overlap in some way. Suppose we carry out a sample survey ▶questionnaire of n questions that have answers "Yes" and "No." Considering the first two questions, Q_1 and Q_2, let p_{11} be the proportion of people who say "Yes" to both questions, p_{12} the proportion who say "Yes" to Q_1 and "No" to Q_2, p_{21} the proportion who say "No" to Q_1 and "Yes" to Q_2, and p_{22} the proportion who say "No" to both questions. Then $p_1 = p_{11}+p_{12}$ is the proportion saying "Yes " to Q_1 and $p_2 = p_{11} + p_{21}$ the proportion saying "Yes" to Q_2. We want to estimate $\theta = p_1 - p_2$, as before. If x_{ij} are observed in the category with probability p_{ij} and $x_1 = x_{11} + x_{12}$ and $x_2 = x_{11} + x_{21}$, then, from Wild and Seber (1993),

$$\sigma^2_{\hat{\theta}} = \frac{1}{n}[p_{12} + p_{21} - (p_{12} - p_{21})^2].$$

To estimate the above variance, we would replace each p_{ij} by its estimate $\hat{p}_{ij}$. In many surveys though, only x_1 and x_2 are recorded so that the only parameters we can estimate are the p_i using $\hat{p}_i = x_i/n$. However, we can use these estimates in the following bounds

$$\frac{1}{n}d(1-d) \le \sigma^2_{\hat{\theta}} \le \frac{1}{n}[min(p_1 + p_2, q_1 + q_2) - (p_1 - p_2)^2],$$

where $d = |p_{12} - p_{21}| = |p_1 - p_2|$. Further comments about constructing confidence intervals, testing hypotheses, and dealing with non-responses to the questions are given in the above paper.

For an elementary discussion of some of the above ideas see Wild and Seber (2000).

About the Author

For biography *see* the entry ▶Adaptive Sampling.

Cross References

▶Asymptotic Normality
▶Binomial Distribution
▶Fisher Exact Test
▶Hypergeometric Distribution and Its Application in Statistics
▶Inverse Sampling
▶Statistical Inference
▶Statistical Inference in Ecology
▶Statistical Inference: An Overview

References and Further Reading

Agresti A, Coull BA (1998) Approximate is better than 'exact' for interval estimation of binomial proportions. Am Stat 52:119–126
Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16(2):101–133 (Followed by a discussion by several authors)
Newcombe R (1998a) Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 17(2): 857–872
Newcombe R (1998b) Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 17(2):873–890 (Correction: 1999, 18, 1293)
Salehi MM, Seber GAF (2001) A new proof of Murthy's estimator which applies to sequential sampling. Aust N Z J Stat 43: 901–906
Seber GAF (2002) The estimation of animal abundance, 2nd edn. Blackburn, Caldwell (Reprinted from the 1982 edition).
Scott AJ, Seber GAF (1983) Difference of proportions from the same survey. Am Stat 37(4):319–320
Wild CJ, Seber GAF (1993) Comparing two proportions from the same survey. Am Stat 47(3):178–181 (Correction: 1994, 48(3), 269)
Wild CJ, Seber GAF (2000) Chance encounters: a first course in data analysis and inference. Wiley, New York

Psychiatry, Statistics in

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

Why "Statistics in Psychiatry"? What makes statistics in psychiatry a particularly interesting intellectual challenge? Why is it not merely a sub-discipline of ▶medical statistics such as the application of statistics in rheumatology or statistics in cardiology? It is in the nature of mental illness and of mental health. Mental illness extends beyond

medicine into the realms of the social and behavioral sciences. Similarly, statistics in psychiatry owes as much, or more, to developments in social and behavioral statistics as it does to medical statistics. Statisticians in this area typically user a much wider variety of multivariate statistical methods than do medical statisticians elsewhere. Scientific psychiatry has always taken the problems of measurement much more seriously than appears to be the case in other clinical specialties. This is partly due to the fact that the measurement problems in psychiatry are obviously rather complex, but partly also because the other clinical fields appear to have been a bit backward by comparison. It is also an academic discipline where, at its best, there is fruitful interplay between the ideas typical of the 'medical model' of disease and those coming from the psychometric traditions of, say, educationalists and personality theorists.

"Mental diseases have both psychological, sociological and biological aspects and their study requires a combination of the approaches of the psychologist, the sociologist and the biologist, using the last word rather than physician since the latter must be all three. In *each* of these aspects statistical reasoning plays a part, whether it be in the future planning of hospitals, the classification of the various forms of such illnesses, the study of causation or the evaluation of methods of treatment." (Moran 1969 – my italics)

"The etiology of mental illness inevitably involves complex and inadequately understood interactions between social stressors and genetically and socially determined vulnerabilities – the whole area being overlaid by a thick carpet of measurement and misclassification errors." (Dunn 2000)

Do the varieties of mental illness fall into discrete, theoretically justified, diagnostic categories? Or are the boundaries entirely pragmatic and artificial? Should we be considering dimensions (matters of degree) or categories? Is the borderline between bipolar depression and schizophrenia, for example, real or entirely arbitrary? The same question even applies to the existence of mental illness itself. What distinguishes illness from social deviance and eccentricity? Establishing the validity and utility of psychiatric diagnoses has been, and still is, a major application of statistical thinking involving a whole range of complex multivariate methods (factor analysis being one of the most prominent). Once psychiatrists have created a diagnostic system they also need to be able to demonstrate its reliability. They need to be confident that psychiatrists will consistently allocate patients with a particular profile of symptoms to the same diagnostic group. They need to know that the category "schizophrenia" means the same thing and is used in the same way by all scientific psychiatrists, whether they work in the USA, China or Uganda. Here, again, statistical methods (kappa coefficients, for example) hold the centre stage.

The development of rating scales and the evaluation of the associated measurement errors form the central core of statistics in psychiatry. It is the problem of measurement that makes psychiatry stand apart from the other medical specialties. Scientific studies in all forms of medicine (including psychiatry) need to take account of confounding and selection effects. Demonstration of treatment efficacy for mental illness, like treatment efficacy elsewhere in medicine, always needs the well-designed controlled randomized trial. But measurement and measurement error make psychiatry stand out. Here, the statistician's world has long been populated by latent variable models of all sorts – finite mixture and latent class models, factor analysis and item response models and, of course, structural equations models (SEM) allowing one to investigate associations and - with luck and careful design - causal links between these latent variables.

About the Author

Graham Dunn has been Professor of Biomedical Statistics and Head of the Biostatistics Group at the University of Manchester since 1996. Before that he was Head of the Department of Biostatistics and Computing at the Institute of Psychiatry. Graham's research is primarily focussed on the design and analysis of randomised trials of complex interventions, specialising on the evaluation of cognitive behavioural and other psychological approaches to the treatment of psychosis, depression and other mental health problems. Of particular interest is the design and analysis of multi-centre explanatory trials from which it is possible to test for and estimate the effects of mediation and moderation, and for the effects of dose (sessions attended) and the quality of the therapy provided (including therapist effects). He also has interests in the design and analysis of measurement reliability studies. A key methodological component of both of these fields of applied research is the development and implementation of econometric methods such as the use of instrumental variables. He is author or co-author of seven statistics books, including *Statistical Evaluation of Measurement Errors* (Wiley, 2nd edition 2004), *Statistics in Psychiatry* (Hodder Arnold, 2000) and (with Brian Everitt) *Applied Multivariate Data Analysis* (Wiley, 2nd edition 2001).

Cross References

►Factor Analysis and Latent Variable Modelling
►Kappa Coefficient of Agreement
►Medical Statistics

►Rating Scales
►Structural Equation Models

References and Further Reading

Dunn G (2000) Statistics in psychiatry. Arnold, London
Moran PAP (1969) Statistical methods in psychiatric research (with discussion). J Roy Stat Soc A 132:484–525

Psychological Testing Theory

Vesna Buško
Associate Professor, Faculty of Humanities and Social Sciences
University of Zagreb, Zagreb, Croatia

Testing and assessment of individual differences have been a critical part of the professional work of scientists and practitioners in psychology and related disciplines. It is generally acknowledged that psychological tests, along with the existing conceptualizations of measurements of human potential, are among the most valuable contributions of the behavioral sciences to society. Testing practice is for many reasons an extremely sensitive issue, and is not only a professional but also a public issue. As the decisions based on test results and their interpretations often entail important individual and societal consequences, psychological testing has been the target of substantial public attention and long-standing criticism (AERA, APA, and NCME 2006).

The theory of psychological tests and measurement, or, as typically referred to, test theory or psychometric theory, offers a general framework and a set of techniques for evaluating the development and use of psychological tests. Due to their latent nature, the majority of psychological constructs are typically measured indirectly, i.e., by observing behavior on appropriate tasks or responses to test items. Different test theories have been proposed to provide rationales for behaviorally based measurement.

Classical test theory (CTT) has been the foundation of psychological test development since the turn of the twentieth century (Lord and Novick 1968). It comprises a number of psychometric models and techniques intended to estimate theoretical parameters, including the description of different psychometric properties, such as the derivation of reliability estimates and ways to assess the validity of use of test. This knowledge is crucial if we are to make sound inferences and interpretations from the test scores.

The central notion of CTT is that any observed test score (X) can be decomposed into two additive components: a *true score* (T) and a random *measurement error* term (e). Different models of CTT have been proposed, each defined by specific sets of assumptions that determine the circumstances under which the model may be reasonably applied. Some assumptions are associated with properties of measurement error as random discrepancies between true and observed test scores, whereas others include variations of the assumption that the two tests measure the same attribute. The latter assumption is essential for deducing test reliability, i.e., the ratio of true score variance to observed score variance, from the discrepancy between two measurements of the same attribute in the same person.

CTT and its applications have been criticized for various weaknesses, such as population dependence of its parameters, focus on a single undifferentiated random error, or arbitrary definition of test score variables. Generalizability theory (see Brennan 2001) was developed as an extension of the classical test theory approach, providing a framework for estimating the effects of multiple sources of error or other factors determining test scores. Another generalization of CTT has been put forward within the formulation of the Latent State-Trait Theory (see Steyer 2003). Formal definitions of states and traits have been introduced, and models allowing the separation of persons, situations, and/or interaction effects from measurement error components of the test scores are presented.

A more recent development in psychometric theory, item response theory (IRT), emerged to address some of the limitations of the classical test theory (Embretson and Reise 1999). The core assumption of IRT is that the probability of a person's expected response to an item is the joint function of that person's ability, or her/his position on the latent trait, and one or more parameters characterizing the item. The response probability is presented in the form of an item characteristic curve as a function of the latent trait.

Despite the controversies and criticisms surrounding CTT, and the important advances and challenges in the field of IRT, both classical and modern test theories appear today to be widely used and are complementary in designing and evaluating psychological and educational tests.

Cross References

►Psychology, Statistics in

References and Further Reading

AERA, APA, NCME (2006) Standardi za pedagoško i psihološko testiranje (Standards for Educational and Psychological Testing). Naklada Slap, Jastrebasko

Brennan RL (2001) Generalizability theory. Springer, New York

Embretson SE, Reise SP (1999) Item response theory for psychologists. LEA, Mahwah

Lord FC, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley Publishing Company, Reading

Steyer R (2003) Wahrscheinlichkeit und regression. Springer, Berlin

Psychology, Statistics in

Joseph S. Rossi
Professor, Director of Research at the Cancer Prevention Research Center
University of Rhode Island, Kingston, RI, USA

The use of quantitative methods in psychology is present essentially at its beginning as an independent discipline, and many of the early developers of statistical methods, such as Galton, Pearson, and Yule, are generally considered by psychologists as among the major contributors to the development of psychology itself. In addition, many early psychologists made major contributions to the development of statistical methods, often in the context of psychometric measurement theory and multivariate methods (e.g., Spearman, Thurstone). Among the techniques that psychologists developed or helped to develop during the early part of the twentieth century are the correlation coefficient, the chi-square test, regression analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), ►principal components analysis, and various multivariate procedures. The use of the ►analysis of variance (ANOVA) in psychology did not begin until about 1940 and quickly became widespread.

During the decades of the 1940s and 1950s, a kind of schism arose among psychologists, with experimental psychologists favoring the use of ANOVA techniques and psychologists interested in measurement and individual differences favoring correlation and regression techniques, culminating in Cronbach's famous declaration concerning the "two disciplines" of scientific psychology. That these procedures were both aspects of the general linear model (see ►General Linear Models) and essentially equivalent mathematically did not become widely known among psychologists until about 1968. A similar sort of schism with respect to models of statistical inference has been resolved with a kind of hybrid model that accommodates both the Fisher and Neyman-Pearson approaches, although in this case, most researchers in psychology are completely unaware that such a schism ever existed, and that the models of statistical decision-making espoused in their textbooks and in common everyday use represent a combination of views thought completely antithetical by their original proponents. Bayesian approaches, while not unknown in psychology, remain vastly underutilized.

Statistical methods currently in common use in psychology include: Pearson product-moment correlation coefficient, chi-square test (see ►Chi-Square Tests), *t* test, univariate and multivariate analysis of variance (see ►Analysis of Variance and ►Multivariate Analysis of Variance (MANOVA)) and covariance with associated follow-up procedures (e.g., Tukey test), multiple regression, factor analysis (see ►Factor Analysis and Latent Variable Modelling) and ►principal components analysis, discriminant function analysis, path analysis, and structural equation modeling (see ►Structural Equation Models). Psychologists have been instrumental in the continued development and refinement of many of these procedures, particularly for measurement oriented procedures, such as item response theory, and structural equation modeling techniques, including confirmatory factor analysis, latent growth curve modeling, multiple group structural invariance modeling, and models to detect mediation and moderation effects. There is considerable emphasis on group level data analysis using parametric statistical procedures and the assumptions of univariate and multivariate normality. The use of nonparametric procedures, once fairly common, has declined substantially in recent decades. The use of more modern nonparametric techniques and robust methods is almost unknown among applied researchers.

The Null Hypothesis Significance Testing Controversy

Common to many of the procedures in use in psychology is an emphasis on null hypothesis significance testing (NHST) and concomitant reliance on statistical test *p* values for assessing the merit of scientific hypotheses. Considering its still dominant position, the use of the NHST paradigm in psychology and related disciplines has been subject to numerous criticisms over a surprisingly long period of time, starting at least 70 years ago. Until recently, these criticisms have not gained much traction. Common objections raised against the NHST paradigm include the following:

- The null is not a meaningful hypothesis and is essentially always false.

- Rejection of the null hypothesis provides only weak support for the alternative hypothesis.
- Failure to reject the null hypothesis does not mean that the null can be accepted, so that null results are inconclusive.
- Significance test p values are misleading in that they depend largely on sample size and consequently do not indicate the magnitude or importance of the obtained effect.
- The obtained p value is unrelated to, but frequently confused with both the study alpha (α) level and $1-\alpha$.
- Reliance on p values has led to an overemphasis on the type I error rate and to the neglect of the type II error rate.
- Statistical significance is not the same as scientific or practical significance.
- The NHST approach encourages an emphasis on point estimates of parameter values rather than confidence intervals.
- The use of the $p < 0.05$ criterion for ►statistical significance is arbitrary and has led to dichotomous decision making with regard to the acceptance/rejection of study hypotheses. This has resulted in the phenomenon of "publication bias," which is the tendency for studies that report statistical significance to be published while those that do not are not published, despite the overall quality or merit of the research.
- The dichotomous decision making approach inherent to the NHST paradigm has seriously compromised the ability of researchers to accumulate data and evidence across studies. This has hindered the development of theories in many areas of psychology, since a few negative results tend to be accorded more weight than numerous positive results.

The extent and seriousness of these criticisms has led some to suggest an outright ban on the use of significance testing in psychology. Once inconceivable, this position is receiving serious consideration in numerous journal articles in the most prestigious journals in psychology, has been discussed by recent working groups and task forces on quantitative methods and reporting standards in psychology, and is even the subject of one recent book. Even among those not willing to discard significance testing entirely, there is widespread agreement on a number of alternative approaches that would reduce reliance on p values. These include the use of confidence intervals, effect size indices, ►power analysis, and ►meta-analysis. Confidence intervals (see ►Confidence Interval) provide useful information beyond that supplied by point estimates of parameters and p values. In psychology, the most frequently used has been the 95% confidence interval. Despite its simplicity, confidence intervals are still not widely used and reported or even that well understood by many applied researchers. For example, some recent studies have indicated that even when error bars are shown on graphs, it is not at all clear if authors intended to show standard deviations, standard errors, or confidence intervals.

Measures of effect size have been recommended as supplements to or even as substitutes for reporting p values. These provide a more direct index of the magnitude of study results and are not directly influenced by study sample size. Measures of effect size fall into two broad categories. The most common historically are measures of the proportion of shared variance between two (or more) variables, such as r^2 and R^2. These indices have a long history of use in research employing correlational and regression methods, particularly within the tradition of individual differences research. Within the tradition of experimental psychology, comparable measures have been available for many years but have not seen widespread use until relatively recently. Most commonly used for the t test and ANOVA are η^2 and ω^2, which index the proportion of variance in the dependent variable that is accounted for by the independent variable. Again, under the general linear model (see ►General Linear Models), these indices are essentially equivalent mathematically and retain their separate identities primarily for historical purposes. Multivariate analogs exist but are not well known and not often used.

A second and more recent approach to measuring the magnitude of study outcomes independent of p values is represented by standardized measures of effect size. These were developed by Cohen as early as 1960 but have not seen widespread use until relatively recently. While he developed a wide range of such indices designed to cover numerous study designs and statistical methods, by far the most widely known and used is Cohen's d. An advantage of d is its simplicity:

$$d = (M_1 - M_2)/s_p$$

where M_1 and M_2 represent the means of two independent groups and s_p is the pooled within-groups standard deviation. It is also easily related to proportion of variance measures:

$$\eta^2 = d^2/(d^2 + 4).$$

A disadvantage of d is that is it applicable only to the comparison of two groups. A generalized version applicable to the ANOVA F test is Cohen's f. However, this measure is much less well-known and not often utilized. As a guide to use and interpretation, Cohen developed a simple rubric

for categorizing the magnitudes of effect sizes. For *d*, small, medium, and large effect sizes are defined as 0.20, 0.50, and 0.80, respectively. The equivalent magnitudes for proportion of variance accounted for are 0.01, 0.06, and 0.14, respectively. Cohen recognized these definitions as arbitrary, but subsequent research suggests they hold up well across a broad range of research areas in the "softer" areas of psychology.

Although methods for aggregating data across independent studies have been in use for more than 100 years, a more formal and systematic approach did not begin to take shape until 1976 with the independent work of Rosenthal and Glass, who coined the term meta-analysis. While very controversial at first, and to a lesser extent still, the technique caught on rapidly as an efficient way to summarize quantitatively the results of a large number of studies, thus overcoming the heavy reliance on *p* values used in the more traditional narrative literature review. In principle any outcome measure can be employed; however, in practice meta-analysis relies heavily on the use of effect sizes as the common metric integrated across studies. Many studies employ the Pearson correlation coefficient *r* for this purpose, although Cohen's *d* is without question the most frequently used, primarily due to its simplicity and easy applicability to a wide range of focused two-group comparisons characteristic of many studies in psychology (e.g., control vs treatment, men vs women). It is probably fair to say that the rise of meta-analysis over the past 20–30 years has greatly facilitated and popularized the concept of the effect size in psychology. As a consequence, a great deal of work has been conducted on *d* to investigate its properties as a statistical estimator. This has resulted in substantial advances in meta-analysis as a statistical procedure. Most early meta-analyses employed simple two-group comparisons of effect size across studies using a fixed effects model approach (often implicitly). Most recent applications have emphasized a regression model approach in which numerous study-level variables are quantified and used as predictors of effect size (e.g., subject characteristics, study setting, measures and methods used, study design quality, funding source, publication status and date, author characteristics, etc.). Fixed effects models still predominate, but there is growing recognition that random (or mixed) effects models may be more appropriate in many cases. A considerable array of follow-on procedures have been developed as aids in the interpretation of meta-analysis results (e.g., effect size heterogeneity test *Q*, funnel and forest plots, assessment of publication bias, fail-safe number, power analysis, etc.). When done well, meta-analysis not only summarizes the literature, it identifies gaps and provides clear suggestions for future research.

Current Trends and Future Directions

Quantitative specialists in psychology continue to work on methods and design in a number of areas. These include data descriptive and exploratory procedures and alternatives to parametric methods, such as ▶exploratory data analysis and cluster analysis (see ▶Cluster Analysis: An Introduction), robust methods, and computer-intensive methods. Work focusing on design includes alternatives to randomized designs, methods for field experiments and quasi-experimental designs, and the use of fractional ANOVA designs. Structural equation modeling continues to receive a lot of attention, including work on latent growth curve modeling, latent transition analysis, intensive longitudinal modeling, invariance modeling, multiple group models, multilevel models, hierarchical linear modeling, and models to detect mediation and moderation effects.

Missing data analysis and multiple imputation methods (see ▶Multiple Imputation), especially for longitudinal designs, is also receiving considerable emphasis. The increased interest in longitudinal approaches has not been limited to group designs. The single subject/idiographic approach, also known as the person-specific paradigm, has been the focus of much recent work. This approach focuses on change over time at the individual level, exemplified by time series analysis, intensive longitudinal modeling, dynamic factor analysis, and dynamic cluster analysis.

Psychologists also continue to conduct a great deal of work on meta-analysis and integrative data analysis, particularly on random effects and hierarchical model approaches, as well as on investigations of the properties of various effect size indices as statistical estimators and the application and development of effect size indices for a wider range of study designs. Increased use of meta-analysis by applied researchers is encouraging the use of alternatives to null hypothesis testing, including the specification of non-zero null hypotheses, and the use of alternative hypotheses that predict the magnitude of the expected effect sizes.

About the Author

Joseph S. Rossi, Ph.D., is Professor and Director of the Behavioral Science Ph.D. program in the Department of Psychology, and Director of Research at the Cancer Prevention Research Center, at the University of Rhode Island. He has been principal investigator or co-investigator on more than 50 grants and has published more than 150 papers and chapters. In 1996, the American Psychological Society and the Institute for Scientific Information listed Dr. Rossi 5th in author impact (citations/paper) and 12th in number of citations (*APS Observer*, January

1996, pp. 14–18). In 2006, he was named one of the most highly cited researchers in the world in the fields of Psychology/Psychiatry by Thomson Reuters (http://isihighlycited.com/). He won the University of Rhode Island's Scholarly Excellence Award in 2003, was elected to membership in the Society of Multivariate Experimental Psychology in 1995, and is a fellow of Division 5 of the American Psychological Association. Dr. Rossi is one of the principal developers of the transtheoretical model of health behavior change. His areas of interest include quantitative psychology, health promotion and disease prevention, and expert system development for health behavior change. Dr. Rossi is a member of the University of Rhode Island's Institutional Review Board.

Cross References

►Confidence Interval
►Cross Classified and Multiple Membership Multilevel Models
►Effect Size
►Factor Analysis and Latent Variable Modelling
►Frequentist Hypothesis Testing: A Defense
►Meta-Analysis
►Moderating and Mediating Variables in Psychological Research
►Multidimensional Scaling
►Multidimensional Scaling: An Introduction
►Null-Hypothesis Significance Testing: Misconceptions
►Psychological Testing Theory
►Sociology, Statistics in
►Statistics: Controversies in Practice

References and Further Reading

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) Reporting standards for research in psychology: why do we need them? What might they be? Am Psychol 63:839–851
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, NJ
Cohen J (1990) Things I have learned (so far). Am Psychol 45: 1304–1312
Cohen J (1994) The earth is round ($p < .05$). Am Psychol 49: 997–1003
Cowles M (1989) Statistics in psychology: an historical perspective. Lawrence Erlbaum, Hillsdale, NJ
Cronbach LJ (1957) The two disciplines of scientific psychology. Am Psychol 12:671–684
Cumming G, Finch S (2005) Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60:170–180
Cumming G, Fidler F, Leonard M, Kalinowski P, Christiansen A, Kleinig A, Lo J, McMenamin N, Wilson S (2007) Statistical reform in psychology: is anything changing? Psychol Sci 18: 230–232
Grissom RJ, Kim JJ (2005) Effect sizes for research: a broad practical approach. Lawrence Erlbaum, Hillsdale, NJ
Harlow LL, Mulaik SA, Steiger JH (eds) (1997) What if there were no significance tests? Lawrence Erlbaum, Hillsdale, NJ
Kline RB (2004) Beyond significance testing: Reforming data analysis methods in behavioral research. American Psychological Association, Washington, DC
Lipsey MW, Wilson DB (2000) Practical meta-analysis. Sage, Thousand Oaks, CA
MacKinnon DP (2008) Introduction to statistical mediation analysis. Lawrence Erlbaum, New York
Maxwell SE (2004) The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol Meth 9:147–163
Meehl PE (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. J Consult Clin Psychol 46:806–834
Molenaar P, Campbell CG (2009) The new person-specific paradigm in psychology. Current Directions in Psychological Science 18:112–117
Nickerson RS (2000) Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Meth 5:241–301
Shadish WR, Cook TD (2009) The renaissance of field experimentation in evaluating interventions. Annual Review of Psychology 60:607–629
Wilkinson L, and the Task Force on Statistical Inference (1999) Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 54:594–604

Public Opinion Polls

Christopher Wlezien
Professor of Political Sciences
Temple University, Philadelphia, PA, USA

A public opinion poll is a survey of the views of a sample of people. It is what we use to measure public opinion in the modern day. This was not always true, of course, as nonrandom "straw polls" have been in regular use at least since the early nineteenth century. We thus had information even back then about what the public thought and wanted, though it was not very reliable. The development of probability sampling, and its application by George Gallup, Archibald Crossley, and Elmo Roper, changed things in important ways, as we now had more reliable information about public opinion (see Geer 1996). The explosion of survey data since that time has fueled the growth in research on attitudes and opinion and behavior that continues today.

There are various forms of probability sampling. In simple random sampling respondents are selected purely at random from the population. This is the most basic form of probability sampling and does a good job representing

the population particularly as the sample size increases and sampling error declines. In stratified random sampling, the population is divided into strata, e.g., racial or ethnic groups, and respondents are selected randomly from within the strata. This approach helps reduce sampling error across groups, which can result from simple random sampling. Traditionally, most survey organizations have relied on ▶cluster sampling. Here the population is divided into geographic clusters, and the survey researcher draws a sample of these clusters and then samples randomly from within them. This is particularly useful when respondents are geographically disbursed. Survey organizations using any of these methods traditionally have relied on face-to-face interviews.

The technology of public opinion polling has changed quite dramatically over time. The invention of random digit dialing (RDD) had an especially significant impact, as interviewing could be done over the telephone based on lists of randomly-generated phone numbers. The more recent introduction of internet polling is having a similar impact. These developments have clear and increasing advantages in cost and speed, and have made it much easier to conduct polls. Witness the growth in the number of pre-election trial-heat polls in presidential election years in the United States (US) (Wlezien and Erikson 2002). Consider also that National Annenberg Election Survey (NAES) conducted over 100,000 telephone interviews during the 2000 presidential election campaign, with similar numbers in 2004 and 2008. In the same election years Knowledge Networks' conducted repeated interviews with 29,000 individuals via the internet. Similar developments can be seen in other countries, including Canada and the United Kingdom. These numbers would be almost inconceivable using face-to-face interviews.

The developments also come with disadvantages. To begin with, there is coverage error. Not everyone has a telephone, and the number relying solely on a cell phone – which poses special challenges for telephone surveys – is growing. Fewer have access to the internet and we cannot randomly e-mail them (note that this precludes calculations of sampling error, and thus confidence intervals. Internet polls do have a number of advantages for scholarly research, however (Clarke et al. 2008)). Even among those we can reach, nonresponse is a problem. The issue here is that respondents who select out of surveys may not be representative of the population. Survey organizations and scholars have long relied on weighting devices to address coverage and nonresponse error (besides sampling error and coverage and nonresponse error, all polls are subject to measurement error, which reflects flaws in the survey instrument itself, including question wording, order, interviewer training and other things. For a treatment of the different forms of survey error, see Weissberg (2005)). In recent years more complicated approaches have begun to be used, including "propensity scores" (see, e.g., Terhanian 2008).

A recent analysis of polling methods in the 2008 US presidential election campaign suggested that the survey mode had little effect on poll estimates (AAPOR 2009). The extent to which the weighting fixes used by survey organizations succeed is the subject of ongoing research – see, e.g., work on internet polls by Malhotra and Krosnick (2007) and Sanders et al. (2007). It is of special importance given the appeal of internet surveys owing to the speed with which they can be conducted and their comparatively low cost.

Despite the difficulties, polls have performed very well. Pre-election polls have proved very accurate at predicting the final vote, particularly at the end of the campaign (Traugott 2005). They also predict well earlier on, though it is not an identity relation, e.g., in US presidential elections, early leads tend to decline by Election Day (Wlezien and Erikson 2002). Pre-election polls in recent election years have provided more information about the election outcome than highly-touted election prediction markets (Erikson and Wlezien 2008). Polls also tell quite a lot about public opinion (see, e.g., Stimson 1991; Page and Shapiro 1992; Erikson and Tedin 2009). Policymakers now have reliable information about the preferences of those with a stake in the policies they make. It also appears to make a difference to what they actually do (Geer 1996).

Acknowledgments

I thank my colleague Michael G. Hagen for very helpful comments.

About the Author

Dr. Christopher Wlezien is Professor, Department of Political Science, Temple University, Philadelphia. While at Oxford University, he co-founded the Spring School in Quantitative Methods for Social Research. He is an elected member of the International Statistical Institute (2006), and holds or has held visiting positions at Columbia University, the European University Institute, Institute Empresa, Juan March Institute, McGill University, L'Institut d'Etudes Politiques de Paris, and the University of Manchester. He has authored or co-authored many papers and books, including *Degrees of Democracy* (Cambridge, 2010), in which he develops and tests a thermostatic model of public opinion and policy. Currently, he is founding co-editor of the *Journal of Elections, Public Opinion and Parties* and Associate editor of *Public Opinion Quarterly*.

Cross References

►Margin of Error
►Nonresponse in Surveys
►Questionnaire
►Social Statistics
►Sociology, Statistics in
►Telephone Sampling: Frames and Selection Techniques

References and Further Reading

American Association for Public Opinion Research (2009) An evaluation of the methodology of the 2008 pre-election primary polls. Available at http://aapor.org/uploads/AAPOR_Rept_FINAL-Rev-4-13-09.pdf

Clarke HD, Sanders D, Stewart MC, Whiteley P (2008) Internet surveys and national election studies: a symposium. Journal of Elections, Public Opinion and Parties 18:327–330

Erikson RS, Tedin KL (2009) American public opinion. Longman, New York

Erikson RS, Wlezien C (2008) Are political markets really superior to polls as election predictions? Public Opin Quart 72:190–215

Geer J (1996) From tea leaves to public opinion polls. Columbia University Press, New York

Malhotra N, Krosnick JA (2007) The effect of survey mode on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with non-probability samples. Polit Anal 15:286–323

Page B, Shapiro RY (1992) The rational public. University of Chicago Press, Chicago

Sanders D, Clarke HD, Stewart MC, Whiteley P (2007) Does mode matter for modeling political choice: evidence from the british election study. Polit Anal 15:257–285

Stimson JN (1991) Public opinion in american: moods, cycles and swings. Westview Press, Boulder, CO

Terhanian G (2008) Changing times, changing modes: the future of public opinion polling. Journal of Elections, Public Opinion and Parties 18:331–342

Traugott MW (2005) The accuracy of the pre-election polls in the 2004 presidential election. Public Opin Quart 69:642–654

Weissberg H (2005) The total survey error approach. University of Chicago Press, Chicago

Wlezien C, Erikson RS (2002) The timeline of presidential election campaigns. J Polit 64(4):969–993

P–Values

Raymond Hubbard
Thomas F. Sheehan Distinguished Professor of Marketing
Drake University, Des Moines, IA, USA

The origin of the *p*-value is credited to Karl Pearson (1900), who introduced it in connection with his chi-square test (see ►Chi-Square Tests). However, it was Sir Ronald Fisher who popularized significance tests and *p*-values in the multiple editions of his hugely influential books *Statistical Methods for Research Workers* and *The Design of Experiments*, first published in 1925 and 1935, respectively. Fisher used divergencies in the data to reject the null hypothesis by calculating the probability of the data on a true null hypothesis, or $\Pr(x|H_0)$. More formally, $p = \Pr(T(X) \geq T(x)|H_0)$. The *p*-value is the probability of getting a test statistic $T(X)$ larger than or equal to the observed result, $T(x)$, as well as more extreme ones, assuming a true null hypothesis, H_0, of no effect or relationship. Thus, the *p*-value is an index of the (im)plausibility of the actual observations (together with more extreme, unobserved ones) if the null is true, and is a random variable whose distribution is uniform over the interval $[0,1]$.

The reasoning is that if the data are viewed as being rare or extremely unlikely under H_0, this constitutes *inductive evidence* against the null hypothesis. Fisher (1966, p. 13) immortalized a *p*-value of 0.05 for rejecting the null: "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard." Consequently, values like $p < 0.01$, $p < 0.001$, and so on, are said to furnish even stronger evidence against H_0. So Fisher considered *p*-values to play an important epistemological role (Hubbard and Bayarri 2003).

Moreover, Fisher (1959, p. 43) saw the *p*-value as an *objective* measure for judging the (im)plausibility of H_0:

> "…the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders."

Researchers across the world have enthusiastically adopted *p*-values as a "scientific" and "objective" criterion for certifying knowledge claims.

Unfortunately, the *p*-value is neither an objective nor very useful measure of evidence in statistical significance testing (see Hubbard and Lindsay 2008, and the references therein). In particular, *p*-values exaggerate the evidence against H_0. Because of this, the validity of much published research with comparatively small (including 0.05) *p*-values must be called into question.

Of great concern, though obviously no fault of the index itself, members of the research community insist on investing *p*-values with capabilities they do not possess (for critiques of this, see, among others, Carver 1978; Nickerson 2000). Some common misconceptions regarding the *p*-value are that it denotes an objective measure of:

- The probability of the null hypothesis being true
- The probability (in the sense of $1 - p$) of the alternative hypothesis being true
- The probability (again, in the sense of $1 - p$) that the results will replicate
- The magnitude of an effect
- The substantive or practical significance of a result
- The Type I error rate
- The generalizability of a result

Despite its ubiquity, the p-value is of very limited use. Indeed, I agree with Nelder's (1999, p. 261) assertion that the most important task in developing a helpful statistical science is "to demolish the *P*-value culture."

About the Author

Dr. Raymond Hubbard is Thomas F. Sheehan Distinguished Professor of Marketing in the College of Business and Public Administration, Drake University, Des Moines, Iowa, USA. He has served as the Chair of the Marketing Department (1988–1989; 1992–1994; 2000–2003). He is a member of the American Marketing Association, Academy of Marketing Science, and the Association for Consumer Research. He has authored or coauthored over 50 journal articles, many of them methodological in nature. He is presently working on a book (with R. Murray Lindsay), tentatively titled "From Significant Difference to Significant Sameness: A Proposal for a Paradigm Shift in Managerial and Social Science Research."

Cross References

►Bayesian P-Values
►Effect Size
►False Discovery Rate
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Psychology, Statistics in
►P-Values, Combining of
►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

Carver RP (1978) The case against statistical significance testing. Harvard Educ Rev 48:378–399
Fisher RA (1959) Statistical methods and scientific inference, 2nd edn. Oliver and Boyd, Edinburgh. Revised
Fisher RA (1966) The design of experiments, 8th edn. Oliver and Boyd, Edinburgh
Hubbard R, Bayarri MJ (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing (with comments). Am Stat 57:171–182
Hubbard R, Lindsay RM (2008) Why *P*-values are not a useful measure of evidence in statistical significance testing. Theory Psychol 18:69–88
Nelder JA (1999) From statistics to statistical science (with comments). Stat 48:257–269
Nickerson RS (2000) Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Meth 5: 241–301
Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Philos Mag J Sci 50:157–175

P-Values, Combining of

Dinis Pestana
Professor, Faculty of Sciences
Universidade de Lisboa, DEIO, and CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal

Let us assume that the p-values p_k are known for testing H_{0k} versus H_{Ak}, $k = 1, \ldots, n$, in n independent studies on some common issue, and our aim is to achieve a decision on the overall question H_0^*: all the H_{0k} are true *versus* H_A^*: some of the H_{Ak} are true. As there are many different ways in which H_0^* can be false, selecting an appropriate test is in general unfeasible. On the other hand, combining the available p_k's so that $T(p_1, \ldots, p_n)$ is the observed value of a random variable whose sampling distribution under H_0^* is known is a simple issue, since under H_0^*, $\boldsymbol{p}$ is the observed value of a random sample $\boldsymbol{P} = (P_1, \ldots P_n)$ from a $Uniform(0,1)$ population. In fact, several different sensible combined testing procedures are often used.

A rational combined procedure should of course be *monotone*, in the sense that if one set of p-values $\boldsymbol{p} = (p_1, \ldots, p_n)$ leads to rejection of the overall null hypothesis H_0^*, any set of componentwise smaller p-values $\boldsymbol{p}' = (p_1', \ldots, p_n')$, $p_k' \leq p_k$, $k = 1, \ldots, n$, must also reject H_0^*.

Tippett (1931) used the fact that $P_{1:n} = \min\{P_1, \ldots, P_n\} \underset{|_{H_0^*}}{\sim} Beta(1,n)$ to reject H_0^* if the minimum observed p-value $p_{1:n} < 1-(1-\alpha)^{1/n}$. This *Tippett's minimum method* is a special case of *Wilkinson's method* (Wilkinson, 1951), advising rejection of H_0^* when some low rank order statistic $p_{k:n} < c$; as $P_{k:n} \sim Beta(k, n+1-k)$, to reject H_0^* at level α the cut-of-point c is the solution of $\int_0^c u^{k-1}(1-u)^{n-k}\mathrm{d}u = \alpha\, B(k, n+1-k)$.

The exact distribution of $\overline{P}_n = \frac{1}{n}\sum_{k=1}^n P_k$ is cumbersome, but for large n an approximation based on the central limit theorem (see ▶Central Limit Theorems) can be used to perform an overall test on H_0^* vs. H_A^*. On the other hand, the probability density function of the ▶geometric mean $G_n = \left(\prod_{k=1}^n P_k\right)^{\frac{1}{n}}$ of n independent uniform random variables is readily computed, $f_{G_n}(x) = \dfrac{n\,(-n\,x\,\ln\,x)^{n-1}}{\Gamma(n)}\mathrm{I}_{(0,1)}(x)$, leading to a more powerful test; see, however, the discussion below on publication bias.

Another way of constructing combined p-values is to use additive properties of simple functions of uniform random variables. Fisher (1932) used the fact that $P_k \sim Uniform(0,1) \Longrightarrow -2\ln P_k \sim \chi^2_2$, and therefore, $-2\sum_{k=1}^n \ln P_k \underset{|_{H_0^*}}{\sim} \chi^2_{2n}$. Then H_0^* is rejected at the significance level α if the $-2\sum_{k=1}^n \ln p_k > \chi^2_{2n,1-\alpha}$. Stouffer et al. (1949) used as test statistic $\sum_{k=1}^n \dfrac{\Phi^{-1}(P_k)}{\sqrt{n}} \underset{|_{H_0^*}}{\sim} Gaussian(0,1)$, where Φ^{-1} denotes the inverse of the distribution function of the standard Gaussian, rejecting H_0^* at level α if $\left|\sum_{k=1}^n \frac{\Phi^{-1}(P_k)}{\sqrt{n}}\right| > z_{1-\alpha}$.

Another simple transformation of uniform random variables P_k is the logit transformation, $\ln\frac{P_k}{1-P_k} \sim Logistic(0,1)$. As $\sum_{k=1}^n \dfrac{\ln\frac{P_k}{1-P_k}}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4)}}} \approx t_{5n+4}$, reject H_0^* at the significance level α if $-\sum_{k=1}^n \dfrac{\ln\frac{p_k}{1-p_k}}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4}}} > t_{5n+4,\,1-\alpha}$.

Birnbaum (1954) has shown that every monotone combined test procedure is *admissible*, i.e., provides a most powerful test against some alternative hypothesis for combining some collection of tests, and is therefore optimal for some combined testing situation whose goal is to harmonize eventually conflicting evidence, or to pool inconclusive evidence. In the context of social sciences Mosteller and Bush (1954) recommend Stouffer's method, but Littel and Folks (1971, 1973) have shown that under mild conditions Fisher's method is optimal for combining independent tests.

As in many other techniques used in ▶meta-analysis, publication bias can easily lead to erroneous conclusions. In fact, the set of available p-values comes only from studies considered worth publishing because the observed p-values were small, seeming to point out significant results. Thus the assumption that the p_k's are observations from independent $Uniform(0,1)$ random variables is questionable, since in general they are in fact a set of low order statistics, given that p-values greater than 0.05 have not been recorded. For instance, $\mathbb{E}\left(G_n^k\right) = \left(\frac{1}{1+\frac{k}{n}}\right)^n \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-k}$, and in particular $\mathbb{E}\left(G_n\right) = \left(\frac{n}{n+1}\right)^n \underset{n\to\infty}{\downarrow} \frac{1}{\mathrm{e}} \approx 0.3679$, the standard deviation decreases to zero, the skewness steadily decreases after a maximum 0.2645 for $n = 5$, and the kurtosis increases from -0.8541 (for $n = 2$) towards 0. Whenever $p_{n:n}$ falls below the critical rejection point, this test will lead to the rejection of H_0^*, but $p_{n:n}$ smaller than the critical point (for $n \geq 14$, the expected value of G_n is greater than 0.36 and the standard deviation is smaller than 0.1) is what should be expected as a consequence of publication bias.

Another important issue: H_A^* states that some of the H_{Ak} are true, and so a meta-decision on H_0^* implicitly assumes that some of the P_k may have non-uniform distribution, cf. Hartung et al. (2008, pp. 81–84) and Kulinskaya et al. (2008, pp. 117–119), and references therein, on the promising concepts of generalized and of random p-values. Gomes et al. (2009) investigated the effect of augmenting the available set of p-values with uniform and with non uniform pseudo-p-values, using results such as: Let X_{m_1} and X_{m_2} be independent random variables, X_m denoting a random variable with probability density function $f_m(x) = \left(mx + \frac{2-m}{2}\right)\mathrm{I}_{(0,1)}(x)$, $m \in [-2,2]$, i.e., a convex mixture of uniform and $Beta(1,2)$ (if $m \in [-2,0)$), thus favoring pseudo-p-values near 0 the sharper the slope is, the slope $m = 0$ corresponding to standard uniform, or of uniform and $Beta(2,1)$ (if $m \in (0,2]$), in this case favoring the occurrence of p-values near 1. Then $\min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right)$ is a member of the same family – more precisely $X_{\frac{m_1 m_2}{6}}$. In particular, if either $m_1 = 0$ or $m_2 = 0$, then $\min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right)$ can be used to generate a new set of uniform random variables, which moreover are independent of the ones used to generate them.

Extensive simulation, namely with computationally augmented samples of p-values (Gomes et al. 2009;

Brilhante et al. 2010) led to the conclusion that in what concerns decreasing power, and increasing number of unreported cases needed to reverse the overall conclusion of a meta-analysis, the methods of combining *p*-values rank as follows:

1. Arithmetic mean
2. ►Geometric mean
3. Chi-square transformation (Fisher's method)
4. Logistic transformation
5. Gaussian transformation (Stouffer's method)
6. Selected order statistics (Wilkinson's method)
7. Minimum (Tippett's method)

About the Author

Professor Dinis Pestana has been President of the Department of Statistics and Operations Research, University of Lisbon, for two consecutive terms (1986–1989), President of the Faculty of Sciences of Lisbon extension in Madeira (1985–1988) before the University of Madeira has been founded, President of the Center of Statistics and Applications, Lisbon University, for three consecutive terms (1981–1987). He supervised the Ph. D. studies of many students that played a leading role in the development of Statistics at the Portuguese universities, and is co-author of 40 papers published in international journals or as chapters of books, and many other papers, namely explaining Statistics to the layman. He launched the annual meetings of the Portuguese Statistical Society, and had a leading role on the local organization of international events, such as the 2001 European Meeting of Statisticians in Funchal.

Cross References

►Bayesian P-Values
►Meta-Analysis
►P-Values

References and Further Reading

Birnbaum A (1954) Combining independent tests of significance. J Amer Stat Assoc 49:559–575

Brilhante MF, Pestana D, Sequeira F (2010) Combining *p*-values and random *p*-values. In: Luzar-Stiffler V, Jarec I, Bekic Z (eds) Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, 515–520

Fisher RA (1932) Statistical methods for research workers, 4th edn. Oliver and Boyd, London

Gomes MI, Pestana D, Sequeira F, Mendonça, S, Velosa S (2009) Uniformity of offsprings from uniform and non-uniform parents. In: Luzar-Stiffler V, Jarec I, Bekic Z (eds) Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, pp 243–248

Hartung J, Knapp G, Sinha BK (2008) Statistical meta-analysis with applications. Wiley, New York

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis. a guide to calibrating and combining statistical evidence. Wiley, Chichester

Littel RC, Folks LJ (1971, 1973) Asymptotic optimality of Fisher's method of combining independent tests, I and II. J Am Stat Assoc 66:802–806, 68:193–194

Mosteller F, Bush R (1954) Selected quantitative techniques In: Lidsey G (ed) Handbook of social psychology: theory and methods, vol I. Addison-Wesley, Cambridge

Stouffer SA, Schuman EA, DeVinney LC, Star S, Williams RM (1949) The American Soldier, vol I: Adjustment during army life. Princeton University Press, Princeton

Tippett LHC (1931) The methods of statistics. Williams and Norgate, London

Wilkinson B (1951) A statistical consideration in psychological research. Psychol Bull 48:156–158

Pyramid Schemes

Robert T. Smythe
Professor
Oregon State University, Corvallis, OR, USA

A pyramid scheme is a business model in which payment is made primarily for enrolling other people into the scheme. Some schemes involve a legitimate business venture, but in others no product or services are delivered. A typical pyramid scheme combines a plausible business opportunity (such as a dealership) with a recruiting operation that promises substantial rewards. A recruited individual makes an initial payment, and can earn money by recruiting others who also make a payment; the recruiter receives part of these receipts, and a cut of future payments as the new recruits go on to recruit others. In reality, because of the geometrical progression of (hypothetical) recruits, few participants in a pyramid scheme will be able to recruit enough others to recover their initial investment, let alone make a profit, because the pool of potential recruits is rapidly exhausted.

Although they are illegal in many countries, pyramid schemes have existed for over a century. As recently as November 2008, riots broke out in several towns in Colombia after the collapse of several pyramid schemes, and in 2006 Ireland launched a website to better educate consumers to pyramid fraud after a series of schemes were perpetrated in Cork and Galway.

Perhaps the best-known type of pyramid scheme is a *chain letter*, which often does not involve even a fictitious product. A chain letter may contain k names; purchasers of the letter invest $\$2x$, with $\$x$ paid to the name at the top of the letter and $\$x$ to the seller of the letter. The purchaser deletes the name at the top of the list, adds his own at the bottom, and sells the letter to new recruits. The promoter's pitch is that if the purchaser, and each subsequent recruit for k−1 stages, sells just two letters, there will be 2^{k-1} people selling 2^k letters featuring the purchaser's name at the top of the list, so that the participant would net $\$2^k x$ from the venture. Many variants of this basic "get rich quick" scheme have been, and continue to be, promoted.

A structure that can be used to model many pyramid schemes is that of recursive trees. A tree with n vertices labeled $1, 2, \ldots, n$ is a *recursive tree* if node 1 is distinguished as the *root*, and for each j with $2 \le j \le n$, the labels of the vertices in the unique path from the root to node j form an increasing sequence. The special case of *random* or *uniform* recursive trees, in which all trees in the set of trees of given order n are equally probable, has been extensively analyzed (cf. Smythe and Mahmoud (1994), for example); however, most pyramid schemes or chain letters in practice have restrictions making their probability models non-uniform. The number of places where the next node may join the tree is then a random variable, unlike the uniform case. This complicates the analysis considerably (and may account for the relative sparsity of mathematical analysis of the properties of pyramid schemes).

Bhattacharya and Gastwirth (1983) analyze a chain letter scheme allowing reentry, in which each purchaser may sell only two letters, unless he purchases a new letter to re-enter the chain. In terms of recursive trees, this means that a node of the tree is *saturated* once it has two offspring nodes, and no further nodes can attach to it. It is further assumed that at each stage, participants who have not yet sold two letters all have an equal chance to make the next sale, i.e., all unsaturated nodes of the recursive tree have an equal chance of being the "parent" of the next node to be added. If L_n denotes the number of leaves of the recursive tree (nodes with no offspring) at stage n under this growth rule, L_n/n corresponds to the proportion of "shutouts" (those receiving no revenue) in this chain letter scheme. The analysis of Bhattacharya and Gastwirth sets up a nonhomogeneous Markov chain model and derives a diffusion approximation for large n. They find that L_n/n converges to 0.382 in this model and that the (centered and scaled) number of shutouts has a normally distributed limit. Mahmoud (1994) considers the height h_n of the tree of order n in this same "random pyramid" scheme and show that it converges with probability 1 to 3.98912; the proof involves embedding the discrete-time growth process of the pyramid in a continuous time birth-and-death process. Mahmoud notes that a similar analysis could be carried out for schemes permitting the sale of m letters, provided that the probabilistic behavior of the total number of shutouts could be derived (as it was in the binary case by Bhattacharya and Gastwirth).

Gastwirth (1977) and Gastwirth and Bhattacharya (1984) analyze another variant of pyramid schemes, known as a quota scheme. This places a limit on the maximum number of participants, so that the scheme corresponds to a recursive tree of some fixed size n. This scheme derives from a case in a Connecticut court (Naruk 1975) in which people bought dealerships in a "Golden Book of Values," then were paid to recruit other dealers. In this scheme, each participant receives a commission from all of his descendants; thus for the jth participant, the size of the branch of the tree rooted at j determines his profit. If S_j denotes the size of this branch, and j/n converges to a limit θ, Gastwirth and Bhattacharya showed that the distribution of S_j converges to the geometric law

$$P(S_j = i+1) = \theta(1-\theta)^i \quad \text{for } i = 0, 1, 2, \ldots$$

(It was later shown (Mahmoud and Smythe (1991)) that if j is fixed, the limiting distribution of S_j/n is $Beta(1, j-1)$.) Calculations made by Gastwirth and Bhattacharya show that, for example, when n is fixed at 270, the 135th entry has probability only about 0.15 of recruiting two or more new entrants, and probability 0.03 of three or more recruits. Gastwirth (1977) shows that for large n, the expected proportion of all participants who are able to recruit at least r persons is 2^{-r}.

Other variants of pyramid schemes include the "8-Ball Model" and the "2-Up System" (http://www.mathmotivation.com/money/pyramid-scheme.html). In the eight-Ball model, the participant again recruits two new entrants, but does not receive any payment until two further levels have been successfully recruited. Thus a person at any level in the scheme would theoretically receive $2^3 = 8$ times his "participation fee," providing incentive to help those in lower levels succeed. In the two-Up scheme, the income from a participant's first two recruits goes to the individual who recruited the participant; if the participant succeeds in recruiting three or more new entrants, the income received from these goes to the participant, along with the income from the first two sales made by each subsequent recruit. This scheme creates considerable incentive to pursue the potentially lucrative third recruit. For both of these schemes, it is easily calculated that when the pool of prospective recruits is exhausted, the majority of the participants in the scheme end up losing money.

About the Author

Robert Smythe is Professor of Statistics at Oregon State University. He was Chair of the Department of Statistics for ten years and previously chaired the Department of Statistics at George Washington University for eight years. He has written in many areas of probability and statistics, and is a Fellow of the American Statistical Association and a Fellow of the Institute for Mathematical Statistics.

Cross References

►Beta Distribution
►Geometric and Negative Binomial Distributions
►Markov Chains

References and Further Reading

Bhattacharya P, Gastwirth J (1983) A non-homogeneous Markov model of a chain letter scheme. In: Rizvi M, Rustagi J, Siegmund D (eds) Recent advances in statistics: papers in honor of Herman Chernoff. Academic, New York, pp 143–174

Gastwirth J (1977) A probability model of a pyramid scheme. Am Stat 31:79–82

Gastwirth J, Bhattacharya P (1984) Two probability models of pyramids or chain letter schemes demonstrating that their promotional claims are unreliable. Oper Res 32:527–536

Mahmoud H (1994) A strong law for the height of random binary pyramids. Ann Appl Probab 4:923–932

Mahmoud H, Smythe RT (1991) On the distribution of leaves in rooted subtrees of recursive trees. Ann Prob 1: 406–418

http://www.mathmotivation.com/money/pyramid-scheme.html

Naruk H (1975) Memorandum of decision: State of Connecticut versus Bull Investment Group, 32 Conn. Sup. 279

Smythe RT, Mahmoud H (1994) A survey of recursive trees. Theor Probab Math Stat 51:1–27