C

Calibration

Christos P. Kitsos
Professor and Head
Technological Educational Institute of Athens, Athens, Greece

Various methods and different (linear or not, simple linear, or multivariate) models have been adopted in industry to address the calibration problem. In practice, most of the models attempt to deal with the simple linear calibration technique, mostly applied in chemical applications, especially when some instruments are to be calibrated (examples include pH meters, NIR instruments, and establishing calibration graphs in chromatography).

The early work of Shukla (1972) put forward the problem on the real statistical dimensions, and even early on it was realized that when a non-linear model describes the phenomenon (Schwartz 1978), a linear approximation is eventually adopted. But even so, in the end we come to a nonlinear function to be estimated as best as possible (Kitsos and Muller 1995). When the variance of the measurement is due to many sources of variability, different techniques are used. Statistical calibration has been reviewed by Osborn (1991), who provides a list of pertinent references; when a robust approach might be appropriate, see Kitsos and Muller (1995). Certainly, to consider the variance constant and to follow a statistical quality control method (see ▶Statistical Quality Control), Hochberg and Marom (1983) might be helpful, but not in all cases. For the multivariate case, see the compact book of Brown (1993), Brereton (2000), and for an application Oman and Wax (1984). Moreover, different methods have been used on the development of the calibration problem like cross-validation (see Clark 1980).

Next we briefly introduce the statistical problem and the optimal design approach is adopted in the sequence to tackle the problem.

Consider the simple regression model with

$$n = E(y|u) = \theta_0 + \theta_1 u_1 \quad u_1 \in U = [-1, 1]$$

where U is the design space, which can always be transformed to $[-1, 1]$. Moreover, the involved error is assumed to be from the normal distribution with mean zero and variance $\sigma^2 > 0$.

The aim is to estimate the value of $u_1 = u_0$ given $n = C$, i.e.,

$$u_0 = (C - \theta_0)/\theta_1$$

which is a nonlinear function of the involved linear parameters, as we have already emphasized above.

The most well-known competitive estimators of u_0 when y_0 is provided are the so-called "*classical predictor*"

$$C(u_0) = \bar{x} + \frac{S_{xx}}{S_{xy}}(\dot{y}_0 - \bar{y})$$

and the "*inverse predictor*"

$$I(u_0) = \bar{u} + \frac{S_{xy}}{S_{yy}}(\dot{y}_0 - \bar{y})$$

with:

$$S_{\mathrm{tr}} = \sum (t_i - \bar{t})(r_i - \bar{r})$$

were by $\dot{y}_0$ we mean the average of the possible k observations taken at the prediction stage (or experimental condition) and $\bar{y}$ as usually the average of the collected values.

The comparison of $C(u_0)$ and $I(u_0)$ is based on the values of the sample size n and the proportion $|\sigma/\theta_1|$ under the assumption that x_0 belongs to the experimenter area.

One of the merits of $C(u_0)$ is that when the usual normal assumption for the errors is imposed, the classical predictor is the maximum likelihood estimator. Moreover, $C(u_0)$ is a consistent estimator while $I(u_0)$ is inconsistent. The $I(u_0)$ estimation is criticized as it provides a confidence interval that might be the whole real line or, in the best case, two disjoint semi-infinite intervals. When $|\sigma/\theta_1|$ is small the asymptotic mse (mean square error) of $C(u_0)$ is smaller than with the use of $I(u_0)$, when x_0 does not lie in the neighborhood of $\bar{u}$.

The main difficulty is the construction of confidence intervals, as the variance of u_0 does not exist. This provides food for thought for an optimal design approach for the calibration problem. To face these difficulties the

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

optimal experimental approach is adopted (see *Optimum Experimental Designs*, also see Kitsos 2002).

For the one-stage design we might use of the criterion function Φ, either D-optimality for (θ_0, θ_1) or c-optimality for estimating u_0. The D-optimal design is of interest because its effectiveness can be investigated, as measured by the c-optimality criterion. Under c-optimality, thanks to Elfving's theorem, locally optimal two-point design can be constructed geometrically. The criterion that experimenters like to use is

$$\min Var(\hat{u}_0).$$

Different approaches have been adopted for this crucial problem: Bayesian, see Chaloner and Verdinelli (1995), Hunter and Lamboy (1981); structural inference, see Kalotay (1971). There is a criticism that structural inference is eventually Bayesian, but this is beyond the scope of this discussion.

When suitable priors for u_0 are chosen the calibrative density functions come from the non-central Student with mean $\mathrm{Ba}(u_0)$ as

$$\mathrm{Ba}(u_0) = \bar{u} + \frac{S_{yy}}{r}(\dot{y}_0 - \bar{y})$$

where $r = S_{yy} + \sum_j^k (y_{0_j} - \bar{y})^2$. When $k = 1$ the Bayesian estimator coincides with the inverse, namely $\mathrm{Ba}(u_0) = I(u_0)$.

The structural approach forms the simple linear model as a "structural model" and obtains a rather complicated model, which, again, with $k = 1$, coincides with the inverse regression.

The nonlinear calibration has attracted classical and Bayesian approaches, both based on the Taylor expansion of the nonlinear model. Therefore, calibration is based on the linear approach of the nonlinear model.

About the Author

Dr. Christos Kitsos is a Professor and Head, Department of Mathematics, of the Technological Educational Institute of Athens, Greece. His first degree is in mathematics, Athens University, his MA degree from the Math and Stat of University of New Brunswick, Canada, while his PhD is from the Glasgow University, UK. He is a member of ISI (and ISI-Committee of Risk Analysis), IEEE and was elected member of IASC, were he is also a member. He has organized a number of statistical conferences in Greece, and has founded the series of International Conference on Cancer Risk Assessment (ICCRA). He has published 88 papers in journals and proceedings volumes, participated in 80 conferences, and published 12 books in Greek (8 are textbooks), and is a coauthor of 5 international books as editor. Professor Kitsos has been the national representative at EUROSTAT and OECD for educational statistics.

Cross References

►Chemometrics
►Measurement Error Models
►Optimal Regression Design
►Optimum Experimental Design

References and Further Reading

Brereton GR (2000) Introduction to multivariate calibration in analytical chemistry. Analyst 125(11):2125–2154

Brown JP (1993) Measurement, regression and calibration. Oxford Science Publication, Oxford Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:273–304

Clark RM (1980) Calibration, cross-validation and carbon-14. II. J Roy Stat Soc, Ser A 143:177–194

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

Hochberg Y, Marom I (1983) On improved calibrations of unknowns in a system of quality-controlled assays. Biometrics 39:97–108

Hunter WG, Lamboy WFA (1981) Bayesian analysis of the linear calibration problem. Technometrics 23:323–338

Kalotay AJ (1971) Structural solution to the linear calibration problem. Technometrics 13:761–769

Kanatani K (1992) Statistical analysis of focal-length calibration using vanishing points. IEEE Trans Robotic Autom 8:767–775

Kitsos CP (1992) Quasi-sequential procedures for the calibration problem. In Dodge Y, Whittaker J (eds) COMPSTAT 1992, vol 2. Physica-Verlag, Heidelberg, pp 227–231

Kitsos CP (2002) The simple linear calibration problem as an optimal experimental design. Commun Stat – Theory Meth 31:1167–1177

Kitsos CP, Muller Ch (1995) Robust linear calibration. Statistics 27:93–106

Oman SD, Wax Y (1984) Estimating fetal age by ultrasound measurements: an example of multivariate calibration. Biometrics 40:947–960

Osborne C (1991) Statistical calibration: a review. Int Stat Rev 59:309–336

Schwartz LM (1978) Statistical uncertainties of analyses by calibration of counting measurements. Anal Chem 50:980–985

Shukla GK (1972) On the problem of calibration. Technometrics 14:547–553

Canonical Analysis and Measures of Association

Jacques Dauxois[1], Guy Martial Nkiet[2]
[1]Professor
Institut de Mathématiques de Toulouse, Toulouse, France
[2]Professor
Université des Sciences et Techniques de Masuku, Franceville, Gabon

Introduction

The Bravais–Pearson linear correlation coefficient and the Sarmanov maximal coefficient are well known statistical

tools that permit to measure, respectively, correlation (also called linear dependence) and stochastic dependence of two suitable random variables X_1 and X_2 defined on a probability space $(\Omega, \mathcal{A}, P)$. Since these coefficients just are the first canonical coefficients obtained from linear and nonlinear canonical analysis, respectively, it is relevant to improve them by using all the canonical coefficients. In order to give an unified framework for these notions, we introduce the canonical analysis (CA) of two closed subspaces H_1 and H_2 of a Hilbert space H. Then, a class of measures of association that admits the aforementioned coefficients as particular cases can be constructed.

Canonical Analysis

Let H be a separable real Hilbert space with inner product and related norm denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ respectively, and H_1 and H_2 be two closed subspaces of H. Then we have the following definition that comes from Dauxois and Pousse (1975).

Definition 1 *The canonical analysis (CA) of H_1 and H_2 is any triple*

$$\left(\{\rho_i\}_{i\in I_0}, \{f\}_{i\in I_1}, \{g\}_{i\in I_2}\right),$$

with $I_\ell \subset \mathbb{N}^*$ for $\ell \in \{0,1,2\}$, that satisfies:

1. *The system $\{f\}_{i\in I_1}$ (resp. $\{g\}_{i\in I_2}$) is an orthonormal basis of H_1 (resp. H_2)*
2. *$\rho_1 = \langle f_1, g_1 \rangle = \sup\{\langle f, g\rangle; (f,g) \in H_1 \times H_2,\ \|f\| = \|g\| = 1\}$*
3. *For any $i \in I_0$ such that $i \geq 2$, one has:*

$$\rho_i = \langle f_i, g_i \rangle = \sup\left\{\langle f, g\rangle; (f,g) \in F_i^{\perp} \times G_i^{\perp}, \|f\| = \|g\| = 1\right\}$$

where $F_i = span\{f_1, \cdots, f_{i-1}\}$ and $G_i = span\{g_1, \cdots, g_{i-1}\}$.

Conditions for existence of canonical analysis have been investigated in the aforementioned work. More precisely, denoting by Π_E the orthogonal projector onto the closed subspace E of H, a sufficient condition is the compacity of $T_1 = \Pi_{H_1}\Pi_{H_2|H_1}$ or, equivalently, that of $T_2 = \Pi_{H_2}\Pi_{H_1|H_2}$. In this case, we say that we have a compact CA, and the following proposition holds:

Proposition 1 *Consider a compact CA $\left(\{\rho_i\}_{i\in I_0}, \{f\}_{i\in I_1}, \{g\}_{i\in I_2}\right)$, of H_1 and H_2, where the ρ_i's are arranged in nonincreasing order. Then:*

1. *$\left\{\rho_i^2\right\}_{i\in I_0}$ is the noincreasing sequence of eigenvalues of T_1 and T_2 and, for any $i \in I_0$, one has $0 \leq \rho_i \leq 1$.*
2. *$\{f\}_{i\in I_1}$ (resp. $\{g\}_{i\in I_2}$) is an orthonormal basis of H_1 (resp. H_2) such that, for any $i \in I_0$, f_i (resp. g_i) is an eigenvector of T_1 (resp. T_2) associated with ρ_i^2.*
3. *$\forall (i,j) \in (I_0)^2$, $\langle f_i, g_j \rangle = \delta_{ij}\rho_i$, $\Pi_{H_1} f_i = \rho_i g_i$, $\Pi_{H_2} g_i = \rho_i f_i$.*
4. *$\{f\}_{i\in I_1 - I_0}$ (resp. $\{g\}_{i\in I_2 - I_0}$) is an orthonormal basis of $\ker(T_1) = H_1 \cap H_2^{\perp}$ (resp. $\ker(T_2) = H_2 \cap H_1^{\perp}$).*

Remark 1

1. The ρ_i's are termed the *canonical coefficients*. They permit to study the relative postions of each of the preceding subspace with respect to the other. For instance, the nullity of all these coefficients is equivalent to the orthogonality of H_1 and H_2, and if one of these subspaces is included into the other these coefficients are all equal to 1. Note that, in this later case there does not exist a compact CA when the considered subspaces are infinite-dimensional ones. Nevertheless, it is possible to find a triple having the same properties than a compact CA. Such a triple can be given by $\left(\mathbb{1}, (f_i)_{i\in\mathbb{N}^*}, (g_i)_{i\in\mathbb{N}^*}\right)$, where $\mathbb{1}$ is the numerical sequence with all terms equal to 1, $(f_i)_{i\in\mathbb{N}^*}$ is a orthonormal basis of H_1 and $(g_i)_{i\in\mathbb{N}^*}$ is the previous system possibly completed with an orthonormal basis of $\ker T_2 = H_2 \cap H_1^{\perp}$ so as to obtain an orthonormal basis of H_2.
2. From the previous notion of CA it is possible to define a canonical analysis of two subspaces H_1 and H_2 relatively to a third one H_3. It is just the CA of the subspaces $H_{1\cdot 3} := (H_1 \oplus H_3) \cap H_3^{\perp}$ and $H_{2\cdot 3} := (H_2 \oplus H_3) \cap H_3^{\perp}$. This CA leads to interesting properties given in Dauxois et al. (2004a), and is useful in statistics for studying conditional independence between random vectors (see, e.g., Dauxois et al. [2004b]).
3. When $X_1 = \left(X_1^1, \cdots, X_1^{p_1}\right)^T$ and $X_2 = \left(X_2^1, \cdots, X_2^{p_2}\right)^T$ are two random vectors such that any X_i^j belongs to $L^2(P)$, their *Linear Canonical Analysis* (LCA) is the CA of H_1 and H_2 where $H_i = \text{span}\left(X_i^1, \cdots, X_i^{p_i}\right)$. The spectral analysis of T_1 is equivalent to that of $V_1^{-1}V_{12}V_2^{-1}V_{21}$, where V_i (resp. V_{12}; resp. V_{21}) denotes the covariance (resp. cross-covariance) operator of X_i (resp. X_1 and X_2; resp. X_2 and X_1). So, it is just the CA of random vectors introduced by Hotelling (1936). The first canonical coefficient is the Bravais–Pearson linear correlation coefficient.
4. When X_1 and X_2 are arbitrary random variables, their *Nonlinear Canonical Analysis* (NLCA) is the CA of H_1 and H_2 where H_i is the closed subspace of $L^2(P)$ consisting in random variables of the form $\varphi(X_i)$, where φ is a measurable function valued into $\mathbb{R}$. In this case, the first canonical coefficient just is the Sarmanov maximal coefficient.

Measures of Association

Let $\mathcal{C}$ be the set of pairs (H_1, H_2) of closed subspaces of a Hilbert space, having a compact CA or being infinite-dimensional and such that $H_1 \subset H_2$ or $H_2 \subset H_1$. We consider an equivalence relation $\simeq$ defined on $\mathcal{C}$, such that $(H_1, H_2) \simeq (E_1, E_2)$ if there exists a pair (I_1, I_2) of isometries satisfying: $I_1(H_1) = E_1$, $I_2(H_2) = E_2$ and $\forall (x,y) \in H_1 \times H_2$, $< I_1(x), I_2(y) >_E = < x, y >_H$, where

H (resp. E) denotes the separable real Hilbert space which contains H_1 and H_2 (resp. E_1 and E_2). We also consider a preordering relation $\preceq$ on $\mathcal{C}$, such that $(H_1, H_2) \preceq (E_1, E_2)$ if there exists a pair $\left(E_1', E_2'\right)$ of closed subspaces satisfying: $E_1' \subset E_1$, $E_2' \subset E_2$ and $(H_1, H_2) \simeq \left(E_1', E_2'\right)$.

Definition 2 *A measure of association r between Hilbertian subspaces is any map from a subset $\mathcal{C}_r$ of $\mathcal{C}$ into $[0,1]$ such that the following conditions are satisfied:*

$$
\begin{aligned}
r(H_1, H_2) &= r(H_2, H_1);\\
H_1 \perp H_2 &\Leftrightarrow r(H_1, H_2) = 0;\\
H_1 \subset H_2 \quad \text{or} \quad H_1 \subset H_2 &\Rightarrow r(H_1, H_2) = 1;\\
(H_1, H_2) \simeq (E_1, E_2) &\Rightarrow r(H_1, H_2) = r(E_1, E_2);\\
(H_1, H_2) \preceq (E_1, E_2) &\Rightarrow r(H_1, H_2) \leq r(E_1, E_2).
\end{aligned}
$$

Remark 2

1. When $X_1 = \left(X_1^1, \cdots, X_1^{p_1}\right)^T$ and $X_2 = \left(X_2^1, \cdots, X_2^{p_2}\right)^T$ are two random vectors such that any X_i^j belongs to $L^2(P)$, we obtain a measure of linear dependence between X_1 and X_2 by putting $r(X_1, X_2) := r(H_1, H_2)$ with $H_i = \text{span}\left(X_i^1, \cdots, X_i^{p_i}\right)$. Indeed, from second axiom given above, $r(X_1, X_2) = 0$ if and only if X_1 and X_2 are uncorrelated, that is $V_{12} = 0$. From the third one, it is seen that if there exists a linear map A such that $X_1 = AX_2$ then $r(X_1, X_2) = 1$.
2. When X_1 and X_2 are arbitrary random variables, considering $H_i = \{\varphi(X_i) / \mathbb{E}(\varphi(X_i)^2) < +\infty\}$, a measure of stochastic dependence of X_1 and X_2 is obtained by putting $r(X_1, X_2) := r(H_1, H_2)$. In this case, the above axioms are closed to the conditions proposed by Rényi (1959) for good measures of dependence. In particular, the second axiom gives the equivalence between the independence of X_1 and X_2 and the nullity of $r(X_1, X_2)$, and from the third axiom it is seen that for any one to one and bimeasurable functions f and g, one has $r(f(X_1), g(X_2)) = r(X_1, X_2)$.

A class of measures of association can be built by using symmetric non decreasing functions of canonical coefficients. In what follows, $\mathcal{P}(\mathbb{N}^*)$ denotes the set of permutations of $\mathbb{N}^*$. For $\sigma \in \mathcal{P}(\mathbb{N}^*)$ and $x = (x_n)_n \in c_0$, we put $x_\sigma = \left(x_{\sigma(n)}\right)_n$ and $|x| = (|x_n|)_n$. We denote by c_0 the space of numerical sequences $x = (x_n)_n$ such that $\lim_{n\to\infty} x_n = 0$.

Definition 3 *A symmetric nondecreasing function (sndf) is a map Φ from a subset c_Φ of c_0 to $\mathbb{R}_+$ satisfying:*

1. *For all $x \in c_\Phi$ and $\sigma \in \mathcal{P}(\mathbb{N}^*)$, one has $x_\sigma \in c_\Phi$ and $\Phi(|x_\sigma|) = \Phi(x)$.*
2. *For all $(x, y) \in (c_\Phi)^2$, if $\forall n$, $|x_n| \leq |y_n|$, then $\Phi(x) \leq \Phi(y)$.*
3. *There exists a nondecreasing function f_Φ from $\mathbb{R}$ to $\mathbb{R}$ such that : $f_\Phi(0) = 0$; $\forall u \in \mathbb{R}$, $(u, 0, \cdots) \in c_\Phi$ and $\Phi(u, 0, \cdots) = f_\Phi(|u|)$.*

We denote by Ψ the map from $\mathcal{C}$ to $c_0 \cup \{\mathbb{1}\}$ such that $\Psi(H_1, H_2)$ is the noincreasing sequence of canonical coefficients of H_1 and H_2. Then we have:

Proposition 2 *Let Φ be a sndf with definition domain c_Φ, and such that $\Phi(\mathbb{1}) = 1$. Then, the map $r_\Phi = \Phi \circ \Psi$ is a measure of association defined on the subset $\mathcal{C}_\Phi = \{(H_1, H_2) \in \mathcal{C};\ \Psi(H_1, H_2) \in c_\Phi \cup \{\mathbb{1}\}\}$.*

This proposition means that a measure of association between two subspaces is obtained as a function of the related nonincreasing sequence of canonical coefficients through a sndf. Some examples of such measures are:

$$
\begin{aligned}
r_1(H_1, H_2) &= 1 - \exp\left(-\sum_{i=1}^{+\infty} \rho_i^2,\ r_{2,p}(H_1, H_2)\right)\\
&= \sqrt{\frac{\sum_{i=1}^{+\infty} \rho_i^{2p}}{1 + \sum_{i=1}^{+\infty} \rho_i^{2p}}}\ (p \in \mathbb{N}^*),\ r_3(H_1, H_2)\\
&= \max_{i\geq 1} |\rho_i| = \rho_1.
\end{aligned}
$$

On the one hand, this class of measures of association contains all the measures built by using LCA of random vectors (see Cramer and Nicewander (1979), Lin (1987), Dauxois and Nkiet (1997b)). On the other hand, when H_1 and H_2 are the subspaces considered in the second assertion of Remark 2, r_3 just is the Sarmanov maximal coefficient. In this case, estimation of the coefficients from NLCA and, therefore, the related measures of associations can be obtained from approximation based on step functions or B-spline functions, and from sampling. Using this approach, a class of independence tests that admits the chi-squared test of independence as particular case, have been be proposed (see Dauxois and Nkiet (1998)).

Cross References

►Canonical Correlation Analysis
►Multivariate Data Analysis: An Overview

References and Further Reading

Cramer EM, Nicewander WA (1979) Some symmetric invariant measures of multivariate association. Psychometrika 41:347–352

Dauxois J, Nkiet GM (1997a) Canonical analysis of Euclidean subspaces and its applications. Linear Algebra Appl 264:355–388

Dauxois J, Nkiet GM (1997b) Testing for the lack of a linear relationship. Statistics 30:1–23

Dauxois J, Nkiet GM (1998) Nonlinear canonical analysis and independence tests. Ann Stat 26:1254–1278

Dauxois J, Pousse A (1975) Une extension de l'analyse canonique. Quelques Applications. Ann Inst Henri Poincaré XI:355–379

Dauxois J, Nkiet GM, Romain Y (2004a) Canonical analysis relative to a closed subspace. Linear Algebra Appl 388:119–145

Dauxois J, Nkiet GM, Romain Y (2004b) Linear relative canonical analysis, asymptotic study and some applications. Ann Inst Stat Math 56:279–304

Hotelling H (1936) Relations between two sets of variables. Biometrika 28:321–377

Lin PE (1987) Measures of association between vectors. Commun Stat Theory Meth 16:321–338

Rényi A (1959) On measures of dependence. Acta Math Acad Sci Hung 10:57–71

Canonical Correlation Analysis

Tenko Raykov
Professor of Measurement and Quantitative Methods
Michigan State University, East Lansing, MI, USA

Introduction

Canonical correlation analysis (CCA) is one of the most general multivariate statistical analysis methods (see ►Multivariate Statistical Analysis). To introduce CCA, consider two sets of variables, denoted A and B for ease of reference (e.g., Raykov and Marcoulides 2008). Let A consist of p members collected in the vector $\underline{\mathbf{x}}$, and let B consist of q members placed in the vector $\underline{\mathbf{y}}$ $(p > 1, q > 1)$. In an application setting, the variables in either set may or may not be considered response variables (dependent or outcome measures) or alternatively independent variables (predictors, explanatory variables). As an example, A may consist of variables that have to do with socioeconomic status (e.g., income, education, job prestige, etc.), while B may comprise cognitive functioning related variables (e.g., verbal ability, spatial ability, intelligence, etc.).

Consider the correlation matrix R of all variables in A and B taken together, which has $(p + q) \cdot (p + q - 1)/2$ non-duplicated (non-redundant) correlations. Obviously, even for relatively small p and q, there are many non-duplicated elements of R. CCA deals with reducing this potentially quite large number of correlations to a more manageable group of interrelationship indices that represent the ways in which variables in A covary with variables in B, i.e., the interrelationships among these two sets of variables. More specifically, the purpose of CCA is to obtain a small number of derived variables (measures) from those in A on the one hand, and from those in B on the other, which show high correlations across the two sets (e.g., Johnson and Wichern 2002). That is, a main goal of CCA is to "summarize" the correlations between variables in set A and those in set B into a much smaller number of corresponding linear combinations of them, which in a sense are representative of those correlations. With this feature, CCA can be used as a method for (1) examining independence of two sets of variables (viz. A and B), (2) data reduction, and (3) preliminary analyses for a series of subsequent statistical applications.

Achieving this goal is made feasible through the following steps (cf. Raykov and Marcoulides 2008). First, a linear combination Z_1 of the variables $\underline{\mathbf{x}}$ in A is sought, as is a linear combination W_1 of the variables $\underline{\mathbf{y}}$ in B, such that their correlation $\rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$ is the highest possible across all choices of combination weights for W_1 and Z_1 (see next section for further details). Call Z_1 and W_1 the *first pair of canonical variates*, and $\rho_{1,1}$ the *first canonical correlation*. In the next step, another linear combination Z_2 of variables in A is found and a linear combination W_2 of variables in B, with the following property: their correlation $\rho_{2,2} = \mathrm{Corr}(Z_2, W_2)$ is the highest possible under the assumption of Z_2 and W_2 being uncorrelated with the variables in the first combination pair, Z_1 and W_1. Z_2 and W_2 are referred to as the *second pair of canonical variates*, and $\rho_{2,2}$ as the *second canonical correlation*. This process can be continued until t pairs of canonical variates are obtained, where $t = \min(p, q)$ being the smaller of p and q. While in many applications t may be fairly large, it is oftentimes the case that only up to the first two or three pairs of canonical variates are really informative (see following section). If all canonical correlations are then uniformly weak and close to zero, A and B can be considered largely (linearly) unrelated. Otherwise, one could claim that there is some (linear) interrelationship between variables in A with those in B. Individual scores on the canonical variates can next be computed and used as values on new variables in subsequent analyses. These scores may be attractive then, since they capture the essence of the cross-set variable interrelationships.

Procedure

To begin a CCA, two linear combinations $Z_1 = \underline{\mathbf{a}}_1' \underline{\mathbf{x}}$ and $W_1 = \underline{\mathbf{b}}_1' \underline{\mathbf{y}}$ are correspondingly sought from the variables

in A and in B, such that $\rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$ is at its maximal possible value across all possible choices of $\underline{\mathbf{a}}_1$ and $\underline{\mathbf{b}}_1$. Consider the covariance matrix S of the entire set of $p + q$ variables in A and B:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where S_{11} is the covariance matrix of the p variables in A, S_{22} that of the q variables in B, S_{21} that of the q variables in B with the p in A, and S_{12} denotes the covariance matrix of the p variables in A with the q measures in B. It can be shown (e.g., Johnson and Wichern 2002) that this maximum correlation $\rho_{1,1}$ will be achieved if the following holds:

1. $\underline{\mathbf{a}}_1$ is taken as the (generalized) eigenvector pertaining to the largest solution ρ^2 of the equation $\left|S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right| = 0$, where $|.|$ denotes determinant, that is, $\underline{\mathbf{a}}_1$ fulfils the equation $\left(S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right)\underline{\mathbf{a}}_1 = \underline{0}$, with ρ^2 being the largest solution of the former equation.
2. $\underline{\mathbf{b}}_1$ is the (generalized) eigenvector pertaining to the largest root of the equation $\left|S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right| = 0$, that is, $\underline{\mathbf{b}}_1$ fulfils the equation $\left(S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right)\underline{\mathbf{b}}_1 = \underline{0}$, with the largest π^2 satisfying the former equation.

The solutions of the two involved determinantal equations are identical, that is, $\rho^2 = \pi^2$, and the positive square root of the largest of them equals $\rho_{(1)} = \pi_{(1)} = \rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$, the maximal possible correlation between a linear combination of variables in A with a linear combination of those in B. Then $Z_1 = \underline{\mathbf{a}}_1' \underline{\mathbf{x}}$ and $W_1 = \underline{\mathbf{b}}_1' \underline{\mathbf{y}}$ represent the first canonical variate pair, with this maximal correlation, $\mathrm{Corr}(Z_1, W_1)$, being the first canonical correlation.

As a next step, the second canonical variate pair is furnished as a linear combination of the variables in A, using the eigenvector pertaining to the second largest solution of $\left|S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right| = 0$ on the one hand, and a linear combination of the B variables using the second largest solution of $\left|S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right| = 0$ on the other hand; then their correlation is the second canonical correlation. One continues in the same manner until $t = \min(p, q)$ canonical variate pairs are obtained; the corresponding canonical correlations are calculated as their interrelationship indices (correlations). From the construction of the canonical variates follows that they are uncorrelated with one another:

$$\begin{aligned} \mathrm{Cov}(Z_i, Z_j) &= \mathrm{Cov}(W_i, W_j) = \mathrm{Cov}(Z_i, W_j) \\ &= 0 \text{ (for all } i \neq j;\ i, j = 1, \ldots, t). \end{aligned}$$

Interpretation

Even though there are $t = \min(p, q)$ canonical variate pairs and canonical correlations, oftentimes in applications not all are important for understanding the relationships among variables in A and B. Statistical tests are available which help evaluate the importance of canonical variate pairs and aid a researcher in finding out how many pairs could be retained for further analysis. The tests assume multivariate normality and examine hypotheses of canonical correlations being 0 in a given population. The first test evaluates the null hypothesis that all canonical correlations are 0. If this hypothesis is rejected, at least the first canonical variate pair is of relevance when trying to understand the interrelationship between the variables in A and B; more specifically, at least the first canonical correlation is not zero in the population. Then the second test examines the null hypothesis that apart from the first canonical correlation, all remaining ones are 0; and so on. If the first tested hypothesis is not rejected, it can be concluded that A and B are (linearly) unrelated to one another.

After completing these tests, and in case at least the first canonical correlation is significant, the next question may well be how to interpret the canonical variates. To this end, one can use the correlations of each canonical variate with variables within its pertinent set. That is, when trying to interpret Z_1, one can look at its correlations with the variables in A. Similarly, when trying to interpret W_1, one can examine its correlations with the variables B; and so on for the subsequent canonical variate pairs and their members. The principle to follow thereby, is to interpret each canonical variate as representing the common features of initial variables correlated highly with that variate. Furthermore, for a given canonical correlation $\rho_i = \pi_i$, its square ρ_i^2 can be interpreted as a squared multiple correlation coefficient for the regression relating the ith canonical variate for any of the sets A or B, with the variables of the other set (B or A, respectively; $i = 1, \ldots, t$). With this in mind, ρ_i^2 can be viewed as proportion shared variance between A and B, as captured by the ith canonical variate pair ($i = 1, \ldots, t$); the square of the first canonical correlation is interpretable as a measure of "set overlap."

Similarly to principal components and factors, canonical variates can be used to obtain individual subject scores on them. They can be used in subsequent analyses, e.g., as scores on explanatory variables. Like principal components, the units of a canonical variate may not be meaningful. It is stressed that canonical variates are not latent variables, but instead share the same observed status as manifest (recorded) variables, since they are linear combinations of the latter.

Relationship to Discriminant Function Analysis

It can be shown (Tatsuoka 1971) that with $k > 2$ groups discriminant function analysis (DFA) is identical to CCA using additionally defined variables $D_1, D_2, \ldots, D_{k-1}$ as comprising set A, while the original explanatory (predictor) variables, say $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_p)'$, are treated as set B. These ▶dummy variables $D_1, D_2, \ldots, D_{k-1}$ are defined in exactly the same way they would be for purposes of regression analysis with categorical predictors. If one then performs a CCA with these sets A and B, the results will be identical to those obtained with a DFA on the original variables $\underline{\mathbf{x}}$. Specifically, the first canonical variate for B will equal the first discriminant function; the second canonical variate for B will equal the second discriminant function, etc. The test for significance of the canonical correlations is then a test for significance of discriminant functions, and the number of significant such functions and of canonical correlations is the same. Further, each consecutive eigenvalue for the discriminant criterion, v_i, is related to a corresponding generalized eigenvalue (determinantal equation root) ρ_i : $v_i = \frac{\rho_i^2}{1-\rho_i^2}$ ($i = 1, 2, \ldots, r$; Johnson and Wichern 2002). Testing the significance of discriminant functions is thus equivalent to testing significance of canonical correlations.

Generality of Canonical Correlation Analysis

CCA is a very general multivariate statistical method that unifies a number of analytic approaches. The canonical correlation concept generalizes the notion of bivariate correlation that is a special case of the former for $p = q = 1$ variables. The multiple correlation coefficient of main relevance in regression analysis is also a special case of canonical correlation, which is obtained when the set A consists of $p = 1$ variable – the response measure – and the set B consists of q variables that are the predictors (explanatory variables) in the pertinent regression model. The multiple correlation coefficient is then identical to the first canonical correlation. Third, since various uni- and multivariate ANOVA designs can be obtained as appropriate special cases of regression analysis, these designs and corresponding ANOVAs can be seen as special cases of canonical correlation analysis. Also, as indicated in the preceding section, discriminant function analysis is a special case of CCA as well. (Since DFA is a "reverse" MANOVA – e.g., Raykov and Marcoulides 2008 – one can alternatively see the latter also as a special case of CCA.) Hence, canonical correlation analysis is a very general multivariate analysis method, which subsumes a number of others that are widely used in statistical applications.

About the Author

Tenko Raykov is a Professor of Measurement and Quantitative Methods at Michigan State University. He received his Ph.D. in Mathematical Psychology from Humbolt University in Berlin. He is an editorial board member of the *Structural Equation Modeling*, *Multivariate Behavioral Research*, *Psychological Methods* and the *British Journal of Mathematical and Statistical Psychology*. He is a coauthor (with G.A. Marcoulides) of the text *A First Course in Structural Equation Modeling* (Lawrence Erlbaum Associates, 2006), *An Introduction to Applied Multivariate Analysis* (Routledge 2008), and *Introhhduction to Psychometric Theory* (Routledge 2010).

Cross References

▶Analysis of Multivariate Agricultural Data
▶Canonical Analysis and Measures of Association
▶Discriminant Analysis: An Overview
▶Eigenvalue, Eigenvector and Eigenspace
▶Multivariate Data Analysis: An Overview
▶Multivariate Statistical Analysis

References and Further Reading

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River

Raykov T, Marcoulides GA (2008) An introduction to applied multivariate analysis. Taylor & Francis, New York

Tatsuoka MM (1971) Multivariate analysis: techniques for educational and psychological research. Wiley, New York

Careers in Statistics

Daniel R. Jeske[1], Janet Myhre[2]
[1]Professor and Chair
University of California-Riverside, Riverside, CA, USA
[2]MARC and Professor Emeritus
Claremont McKenna College, Claremont, CA, USA

Statistics has changed over the last decades from being a discipline that primarily studied ways to characterize randomness and variation to a discipline that emphasizes the importance of data in the explanation of phenomenon and in problem solving. While statisticians routinely use mathematics and computer programming languages as key

tools in their work, they usually also function as an important data-driven decision maker within their application domain. Consequently, a statistician must have a genuine curiosity about the subject domain they work within, and furthermore, must have strong collaborative and communication skills in order successfully interact with the many colleagues they will encounter and rely on for information.

As the world becomes more quantitative through the data revolution, more professions and businesses depend on data and on the understanding and analyses of these data. Data are not simply numbers. Data contain information that needs to be understood and interpreted. As a result, statisticians are much more than bean counters or number crunchers. They possess skills to find needles in haystacks and to separate noise from signal. They are able to translate a problem or question into a framework that enables data collection and data analysis to provide meaningful insights that can lead to practical conclusions.

Loosely speaking there is a spectrum of statisticians that ranges from very applied on one end to very theoretical on the other end. Applied statisticians skillfully select and implement an appropriate statistical methodology to solve a problem. They are a statistician who has a problem and is looking for a solution. Theoretical statisticians are interested in trying to expand the toolkit of applied statisticians by generalizing or creating new methods that are capable of solving new problems or solving existing problems more efficiently. They are statisticians who might get motivated by a problem someone else encountered in practice, but who then abstract the problem as much as possible so that their solution has as broad an impact as possible. Most statisticians are not planted firmly on either end of this spectrum, but instead find themselves moving around and adapting to the particular challenge they are facing.

Another way to loosely categorize statisticians is in terms of industrial (or government) versus academic statisticians. Academic statisticians are primarily involved with innovative research and the teaching of statistics classes. Aside from Statistics departments, there are many alternative departments for academic statisticians including Mathematics, Economics, Business, Sociology and Psychology. Research goals for an academic statistician vary with their interests, and also depend on their emphasis toward either applied or theoretical research. In addition, the University at which they work can emphasize either a teaching or research mission that further dictates the quantity and type of research they engage in. In any case, it is a primary responsibility of an academic statistician to publish papers in leading statistics journals to advance the field. Teaching responsibilities can include introductory Statistics for undergraduate non-majors, core statistical theory and methods classes for Statistics majors and in many cases advanced graduate-level Statistics classes for students pursuing an MS and/or PhD degree in Statistics.

Industrial statisticians are frequently focused on problems that have some bearing on the company's business. In some large companies there may be a fundamental research group that operates more like an academic environment, but in recent years the number and size of these groups are diminishing as corporations are more squarely focused on their bottom lines. Industrial statisticians are expected to assimilate the company culture and add value to the projects they work on that goes well beyond the contributions that their statistical skills alone enable. They might, for example, become project managers and even technical managers where their organizational, motivational, and leadership skills become important assets to the company.

Many statisticians engage in statistical consulting, either as their primary vocation or as a part-time endeavor. Academic statisticians, for example, often have opportunities to lend their data analysis and quantitative problem solving skills to government and industry clients, and can contribute to litigation cases as an expert consultant or even an expert witness. Consultants must have exceptionally strong communication skills to be able to translate the interpretation of their findings into the language of the client. In the same way, they have to be able to elicit information from their clients that will ensure the efficacy of their data analyses. Industrial statisticians often function as internal consultants to the company they work for. This is particularly true in large companies where there can be a group of statisticians that serve as a shared central resource for the entire company.

The following alphabetical list is meant to provide an appreciation of the diversity of fields where statisticians are gainfully employed: Agriculture, Archaeology, Astronomy, Biology, Chemistry, Computer Science, Demography, Economics, Ecology, Education, Engineering, Epidemiology, Finance, Forestry, Genetics, Health Sciences, Insurance, Law, Manufacturing, Medicine, National Defense, Pharmacology, Physics, Psychology, Public Health, Safety, Sociology, Sports, Telecommunications, and Zoology. To be more specific, consider the following brief descriptions of work and employment of statisticians in the following fields:

Medicine

Florence Nightingale was not only a historic figure because of what she brought to the profession of nursing, but she was also a pioneering figure in the use of statistics. Statistical work in medicine involves designing studies

and analyzing their data to determine if new (or existing) drugs, medical procedures and medical devices are safe and effective. Statisticians find careers at pharmaceutical companies, medical research centers and governmental agencies concerned with drugs, public health and medicine.

Ecology

Research laboratories, commercial firms and government environmental agencies employ statisticians to evaluate the environmental impact of air, water and soil pollutants. Statisticians also work with government lawyers to analyze the impact (false positive or false negative) of proposed federal or state pollution tests and regulations.

Market Research

Statisticians analyze consumer demand for products and services, analyze the effectiveness of various types of advertising, and analyze the economic risk of satisfying consumer demand for products and services.

Manufacturing

The success of manufacturing industries such as aerospace, electronics, automobile, chemical or other product producing industries depends, at least in part, on the efficiency of production and the quality and reliability of their products. Statistical techniques and models are used for predicting inventory needs, improving production flow, quality control, reliability prediction and improvement, and development of product warranty plans. The Deming Prize, named after the prolific statistician W. Edwards Deming, was established in 1950 and is annually awarded to companies that make major advances in quality improvement. The Malcolm Baldridge National Quality Award, named after Malcolm Baldridge who served as the United States Secretary of Commerce under President Regan, was established in 1988 and is annually awarded to U.S. organizations for performance excellence.

Actuarial Sciences

Actuarial statisticians use Mathematics and Statistics to assess the risk of insurance and financial portfolios. Statistical methods are used, for example, to determine a wide variety of appropriate insurance premiums (e.g., homeowner, life, automobile, flood, etc.) and to manage investment and pension funds.

Safety

Statisticians are employed by many businesses and agencies to model safety concerns and to estimate and predict the probability of occurrence of these safety concerns. Nuclear power plants, national defense agencies and airlines are just a few of the businesses that statistically analyze safety risks.

Telecommunications

The reliability of voice and data networks is paramount to a telecommunication company's revenue and their brand name image. Statisticians work collaboratively with engineers to model alternative design architectures and choose the most cost-effective design that minimizes customer-perceived downtime. Statisticians working in telecommunication companies frequently shift into new technology areas to keep up with the vastly changing landscape of high-tech companies.

The authors have found that their careers in statistics involve work that is usually very interesting, often involves new ideas and learning experiences, and can definitely bring value to problem solving.

For more information on careers in statistics consult www.amstat.org or e-mail asainfo@amstat.org.

About the Authors

Dr. Daniel R. Jeske is a Professor and Chair, Department of Statistics, University of California – Riverside, CA, and is the first Director of the Statistical Consulting Collaboratory at UCR. He has published over 45 journal articles and over 35 refereed conference papers. Prior to his academic career, he was a Distinguished Member of Technical Staff and a Technical Manager at AT&T Bell Laboratories. Concurrent with that, he was a part-time Visiting Lecturer in the Department of Statistics at Rutgers University. Currently, he is an Associate Editor for The *American Statistician*.

Dr. Janet Myhre is President and Founder of Mathematical Analysis Research Corporation of Claremont, California. She is Professor Emeritus, Honorary Alumna, and Founder of the Reed Institute of Applied Statistics at Claremont McKenna College. She is a Fellow of the American Statistical Association and and has served as an Associate Editor of *Technometrics* and as Chair of the Committee on Careers in Statistics of the American Statistical Association.

Cross References

►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics in Advancing Quantitative Education
►Statistics Education

Case-Control Studies

Alastair Scott, Chris Wild
Professors
The University of Auckland, Auckland, New Zealand

Introduction

The basic aim of a case-control study is to investigate the association between a disease (or some other condition of interest) and potential risk factors by drawing separate samples of "cases" (people with the disease, say) and "controls" (people at risk of developing the disease). Let Y denote a binary response variable which can take values $Y = 1$, corresponding to a case, or $Y = 0$, corresponding to a control, and let $\boldsymbol{x}$ be a vector of explanatory variables or covariates. Our aim is to fit a binary regression model to explain the probabilistic behavior of Y as a function of the observed values of the explanatory variables recorded in $\boldsymbol{x}$. We focus particularly on the logistic regression model (see ►Logistic Regression),

$$\begin{aligned}\text{logit}\{\text{pr}(Y \mid \boldsymbol{x};\boldsymbol{\beta})\} &= \log\left\{\frac{\text{pr}(Y=1 \mid \boldsymbol{x};\boldsymbol{\beta})}{\text{pr}(Y=0 \mid \boldsymbol{x};\boldsymbol{\beta})}\right\} \\ &= \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1, \qquad (1)\end{aligned}$$

since this makes the analysis of case-control data particularly simple and is the model of choice in most applications.

In principle, the most straightforward way of obtaining data from which to build regression models for $\text{pr}(Y \mid \boldsymbol{x})$ would be to use a prospective sampling design. Here covariate information is ascertained for a cohort of individuals who are then tracked through time until the end of the study when whether they have contracted the disease ($Y = 1$) or not ($Y = 0$) is recorded. With prospective sampling designs, observation proceeds from covariates (explanatory variables) to response, corresponding to the logic underlying the modelling. With case-control sampling, the order is reversed, with data collection proceeding from response to covariates. The parameter $\boldsymbol{\beta}_1$ in Model (1) can still be estimated, however. Consider the simplest situation of a single binary covariate taking values $x = 0$ or $x = 1$. Using Bayes Theorem, Cornfield (1951) showed that the prospective odds ratio, $\frac{\text{pr}(Y=1|x=1)}{\text{pr}(Y=0|x=1)} / \frac{\text{pr}(Y=1|x=0)}{\text{pr}(Y=0|x=0)}$, can be expressed as $\frac{\text{pr}(x=1|Y=1)}{\text{pr}(x=0|Y=1)} / \frac{\text{pr}(x=1|Y=0)}{\text{pr}(x=0|Y=0)}$, which only involves quantities that can be estimated directly from case-control data. Cornfield also pointed out that the relative risk, $\text{pr}(Y = 1 \mid x = 1)/\text{pr}(Y = 1 \mid x = 0)$, which is usually of more interest, is approximated well by the odds ratio if the disease is rare. If the overall probability of a case can be estimated from other sources, then this can be combined with the relative risk to give estimates of the absolute risk of disease for exposed ($x = 1$) and non-exposed ($x = 0$) groups. All this extends immediately to general $\boldsymbol{\beta}_1$, all of whose components represent individual log odds ratios.

Types of Case-Control Studies

There are two broad types of case-control study, population-based and matched, corresponding to two different ways of controlling for confounding variables. In the simplest form of population-based sampling sampling, random samples are drawn independently from the case- and control-subpopulations of a real, finite target population or cohort. Covariate information, $\boldsymbol{x}$, is then ascertained for sampled individuals. Fitting logistic model (1) is particularly simple here. Following earlier work for discrete covariates, Prentice and Pyke (1979) showed that we can get valid inferences about $\boldsymbol{\beta}_1$ by running the case-control data through a standard logistic regression program designed for prospective data. The intercept β_0, which is needed if we want to estimate the absolute risk for given values of the covariates, is completely confounded with the relative sampling rates of cases and controls but can be recovered using additional information such as the finite population totals of cases and controls.

Prentice and Pyke extended this to stratified case-control sampling, where the target population is first split into strata on the basis of variables known for the whole population and separate case-control samples are drawn from each stratum. Again we get valid inferences about all the other coefficients by running the data through a prospective logistic regression program, **provided** that we introduce a separate intercept for each stratum into our model. Otherwise standard logistic programs need to be modified slightly to produce valid inferences (Scott andh Wild 1997).

In designing a population-based study, it is important to make sure that the controls really are drawn from the same population, using the same protocols, as the cases. Increasingly, controls are selected using modern sample survey techniques, involving multi-stage sampling and varying selection probabilities, to help ensure this. The modifications needed to handle these complications are surveyed in Scott and Wild (2009).

In a matched case-control study, each case is individually matched with one or more controls. This could be regarded as an limiting case of a stratified study with the

strata so finely defined that each stratum includes only a single case. If we introduce an extra intercept for each matched set, then we can no longer use a simple logistic program since the plethora of parameters will lead to inconsistent parameter estimates. Instead we need to carry out a conditional analysis. More specifically, suppose that there are M controls in the *jth* matched set and model (1) is replaced by $\text{logit}\{\text{pr}(Y \mid \boldsymbol{x};\boldsymbol{\beta})\} = \beta_{0j} + \boldsymbol{x}^T\boldsymbol{\beta}_1$ for these observations. Then the conditional probability that the covariates $\boldsymbol{x}_{j0}$ are those of the case and $(\boldsymbol{x}_{j1},\ldots,\boldsymbol{x}_{jM})$ are those of the M controls, given the set of $M+1$ covariates can be expressed in the form $exp\left(\boldsymbol{x}_{j0}^T\boldsymbol{\beta}_1\right) / \sum_{m=0}^{M+1} exp\left(\boldsymbol{x}_{jm}^T\boldsymbol{\beta}_1\right)$, which does not involve the intercept terms. Inferences about $\boldsymbol{\beta}_1$ can then be made from the conditional likelihood obtained when we combine these terms over all matched sets. With pair matching ($M = 1$), this likelihood is identical to a simple logistic regression on the difference between the paired covariates.

More sophisticated designs, including incidence density sampling, nested case-control studies and case-cohort studies, that can handle complications such as time-varying covariates and ►survival data are discussed in other chapters in this volume.

Discussion

Case-control sampling is a cost-reduction device. If we could afford to collect data on the whole finite population or cohort, then we would do so. There many practical difficulties that need to be overcome to run a successful case-control study; a good account of these is given in Breslow (2005). Despite this, the case-control study in its various forms is one of the most common designs in health research. In fact, Breslow and Day (1980) described such studies as "perhaps the dominant form of analytical research in epidemiology" and since that time the rate of appearance of papers reporting on case-control studies has gone up by a factor of more than 20. These designs are also used in other fields, sometimes under other names. In econometrics, for example, the descriptor "choice-based" is used (see Manski and McFadden (1981)).

There are several reasons for the popularity of case-control studies. The first is the simplicity of the logistic analysis outlined above. The other two reasons concern efficiency: time efficiency and statistical efficiency. The former comes from being able to use historical information immediately rather than having to follow individuals through time and then wait to observe an outcome as in a prospective study. The first chapter of Breslow and Day (1980) has a good discussion of the attendant risks. The gain in statistical efficiency can be huge. For example, suppose that we have a condition that affects only 1 individual in 20 on average and we wish to investigate the effect of an exposure that affects 50% of people. In this situation a case-control study with equal numbers of cases and controls has the same power for detecting a small increase in risk as a prospective study with approximately five times as many subjects. If the condition affects only one individual in 100 then the prospective study would need 25 times as many subjects!

About the Authors

Professor Scott is Past President of the New Zealand Statistical Society (1989–1990). He is of one of New Zealand's foremost mathematical scientists. He was founding head of the University of Auckland's Department of Statistics (and previously Head of the Department of Mathematics and Statistics). He is an elected member of the International Statistical Institute, and one of 12 honorary life members of the New Zealand Statistical Association. His 1981 paper with JNK Rao, published in the *Journal of American Statistical Association*, was selected as one of 19 landmark papers in the history of survey sampling for the 2001 centenary volume of the International Association of Survey Statisticians.

Professor Wild is, with Professor Scott, the only statistician in New Zealand that has been elected a Fellow of the American Statistical Association. He is a Past President of the International Association for Statistics Education (2003–2005). He is currently Editor of the *International Statistical Review*.

Cross References

►Medical Statistics
►Statistical Methods in Epidemiology

References and Further Reading

Breslow NE (1996) Statistics in epidemiology: the case-control study. J Am Stat Assoc 91:14–28
Breslow NE (2005) Case-control studies. In: Aherns W, Pigeot I (eds) Handbook of epidemiology. Springer, New York, pp 287–319
Breslow NE, Day NE (1980) The analysis of case-control studies. International Agency for Research on Cancer, Lyon
Cornfield J (1951) A method of estimating comparitive rates from clinical data. J Natl Cancer Inst 11:1269–1275
Manski CF, McFadden D (eds) (1981) Structural analysis of discrete data with econometric applications. Wiley, New York. Models and case-control studies. Biometrika 66:403–411
Scott AJ, Wild CJ (1997) Fitting regression models to case-control data by maximum likelihood. Biometrika 84:57–71
Scott AJ, Wild CJ (2009) Population-based case control studies. In: Pfefferman D, Rao CR (eds) Ch 38 in Handbook of statistics 29: sample surveys. Elsevier, Amsterdam, pp 1009–1031

Categorical Data Analysis

Alan Agresti[1], Maria Kateri[2]
[1]Distinguished Professor Emeritus
University of Florida, Gainesville, FL, USA
[2]Associate Professor
University of Ioannina, Ionnina, Greece

Introduction

A categorical variable consists of a set of non-overlapping categories. Categorical data are counts for those categories. The measurement scale is *ordinal* if the categories exhibit a natural ordering, such as opinion variables with categories from "strongly disagree" to "strongly agree." The measurement scale is *nominal* if there is no ordering. The types of possible analysis depend on the measurement scale.

When the subjects measured are cross-classified on two or more categorical variables, the table of counts for the various combinations of categories is a *contingency table*. The information in a contingency table can be summarized and further analyzed through appropriate *measures of association and models*. A standard reference on association measures is Goodman and Kruskal (1979).

Most studies distinguish between one or more *response variables* and a set of *explanatory variables*. When the main focus is on the association and interaction structure among a set of response variables, such as whether two variables are conditionally independent given values for the other variables, *log-linear models* are useful. More commonly, research questions focus on effects of explanatory variables on a categorical response variable. *Logistic regression models* (see ▶Logistic Regression) are then of particular interest. For *binary* (success-failure) response variables, they describe the *logit,* which is $\log[P(Y=1)/P(Y=2)]$, using

$$\log[P(Y=1)/P(Y=2)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where Y is the binary response variable and $x_1, \ldots, x_p$ the set of the explanatory variables. For a nominal response Y with J categories, the model simultaneously describes

$$\log[P(Y=1)/P(Y=J)],$$
$$\log[P(Y=2)/P(Y=J)], \ldots, \log[P(Y=J-1)/P(Y=J)].$$

For ordinal responses, a popular model uses explanatory variables to predict a logit defined in terms of a cumulative probability (McCullagh 1980),

$$\log[P(Y \le j)/P(Y > j)],\ j = 1, 2, \ldots, J-1.$$

For categorical data, the binomial (see ▶Binomial Distribution) and multinomial distributions (see ▶Multinomial Distribution) play the central role that the normal does for quantitative data. Models for categorical data assuming the binomial or multinomial were unified with standard regression and ▶analysis of variance (*ANOVA*) models for quantitative data assuming normality through the introduction by Nelder and Wedderburn (1972) of the *generalized linear model* (*GLM*, see ▶Generalized Linear Models). This very wide class of models can incorporate data assumed to come from any of a variety of standard distributions (such as the normal, binomial, and Poisson). The GLM relates a function of the mean (such as the log or logit of the mean) to explanatory variables with a linear predictor.

Contingency Tables

Two categorical variables are *independent* if the probability of response in any particular category of one variable is the same for each category of the other variable. The most well-known result on two-way contingency tables is the test of the null hypothesis of independence, introduced by Karl Pearson in 1900. If X and Y are two categorical variables with I and J categories respectively, then their cross-classification leads to a $I \times J$ table of observed frequencies $\mathbf{n} = (n_{ij})$. Under this hypothesis, the expected cell frequencies equal $m_{ij} = n\pi_{i\cdot}\pi_{\cdot j}, i = 1, \ldots, I, j = 1, \ldots, J$, where n is the total sample size $\left(n = \sum_{i,j} n_{ij}\right)$ and $\pi_{i\cdot}(\pi_{\cdot j})$ is the ith row (jth column) marginal of the underlying probabilities matrix $\boldsymbol{\pi} = (\pi_{ij})$. Then the corresponding maximum likelihood (ML) estimates equal $\hat{m}_{ij} = np_{i\cdot}p_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n}$, where p_{ij} denotes the sample proportion in cell (i,j). The hypothesis of independence is tested through Pearson's chi-squared statistic

$$\chi^2 = \frac{\sum_{i,j}(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \qquad (1)$$

The P-value is the right-tail probability above the observed χ^2 value. The distribution of χ^2 under the null hypothesis is approximated by a $\chi^2_{(I-1)(J-1)}$, provided that the individual expected cell frequencies are not too small. When a contingency table has ordered row or column categories (ordinal variables), specialized methods can take advantage of that ordering.

More generally, models can be formulated that are more complex than independence, and expected frequencies m_{ij} can be estimated under the constraint that the model holds. If $\hat{m}_{ij}$ are the corresponding maximum likelihood estimates, then, to test the hypothesis that the model holds, we can use the Pearson statistic (1) or the statistic that results from the standard statistical approach of

conducting a *likelihood-ratio test,* which has test statistic

$$G^2 = 2\sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \quad (2)$$

Independence between the classification variables X and Y (i.e., $m_{ij} = n\pi_{i\cdot}\pi_{\cdot j}$, for all i and j) can be expressed in terms of a log-linear model as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \quad i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The more general model that allows association between the variables is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J. \quad (3)$$

Log-linear models describe the way the categorical variables and their association influence the count in each cell of the contingency table. They can be considered as a discrete analogue of *ANOVA*. The two-factor interaction terms relate to odds ratios describing the association.

Associations can be modeled through simpler *association models*. The simplest such model, the *linear-by-linear association model,* is relevant when both classification variables are ordinal. It replaces the interaction term λ_{ij}^{XY} by the product $\phi\mu_i\nu_j$, where μ_i and ν_j are known scores assigned to the row and column categories respectively. This model is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j, \quad i = 1, \ldots, I,\ j = 1, \ldots, J. \quad (4)$$

More general models treat one or both sets of scores as parameters.

The special case of square $I \times I$ contingency tables with the same categories for the rows and the columns occurs with matched-pairs data. For example, such tables occur in the study of *rater agreement* and in the analysis of social mobility. A condition of particular interest for such data is *marginal homogeneity,* that $\pi_{i\cdot} = \pi_{\cdot i}, i = 1, \ldots, I$. For the 2×2 case of binary matched pairs, the test comparing the margins using the chi-squared statistic $(n_{12} - n_{21})^2/(n_{12} + n_{21})$ is called *McNemar's test*.

The models for two-way tables extend to higher dimensions. The various models available vary in terms of the complexity of the association and interaction structure.

Inference and Software

Standard statistical packages, such as *SAS*, *R*, and *SPSS*, are well suited for analyzing categorical data, mainly using maximum likelihood for inference. For *SAS*, a variety of codes are presented and discussed in the Appendix of Agresti (2002), and see also Stokes et al. (2000). For *R*, see the on-line manual of Thompson (2008). Bayesian analysis of categorical data can be carried out through WinBUGS (http://wlww.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

The standard reference on log-linear models is Bishop et al. (1975). For logistic regression, Hosmer and Lemeshow (2000) is popular. A more comprehensive book dealing with categorical data analysis using various types of models and analyses is Agresti (2002), with Agresti (2010) focusing on ordinal data.

About the Author

Professor Agresti is recipient of the first Herman Callaert Leadership Award in Statistical Education and Dissemination, Hasselt University, Diepenbeek, Belgium (2004), and Statistician of the Year award, Chicago chapter of American Statistical Association (2002–2003). He received an honorary doctorate from De Montfort University in the U.K. in 1999. He has presented invited lectures and short courses for universities and companies in about 30 countries. Professor Agresti is author or coauthor of five textbooks, including the internationally respected text "*Categorical Data Analysis*." Professor Kateri won the Myrto Lefkopoulou award for her Ph.D. thesis in Greece and has since published extensively on methods for categorical data.

Cross References

►Algebraic Statistics
►Association Measures for Nominal Categorical Variables
►Chi-Square Test: Analysis of Contingency Tables
►Data Analysis
►Exact Inference for Categorical Data
►Generalized Linear Models
►Logistic Regression
►Variation for Categorical Variables

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Agresti A (2010) Analysis of ordinal categorical data, 2nd edn. Wiley, New York

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Goodman LA, Kruskal WH (1979) Measures of association for cross classifications. Springer, New York

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

McCullagh P (1980) Regression models for ordinal data (with discussion). J R Stat Soc B 42:109–142

Nelder J, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Stokes ME, Davis CS, Koch GG (2000) Categorical data analysis using the SAS system, 2nd edn. SAS Institute, Cary

Thompson LA (2008) R (and S-PLUS) manual to accompany Agresti's Categorical data analysis (2002), 2nd edn. https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf

Causal Diagrams

Sander Greenland[1], Judea Pearl[2]
[1]Professor
University of California-Los Angeles, Los Angeles, CA, USA
[2]Professor, Director of Cognitive Systems Laboratory
University of California-Los Angeles, Los Angeles, CA, USA

From their inception, causal systems models (more commonly known as structural-equations models) have been accompanied by graphical representations or path diagrams that provide compact summaries of qualitative assumptions made by the models. These diagrams can be reinterpreted as probability models, enabling use of graph theory in probabilistic inference, and allowing easy deduction of independence conditions implied by the assumptions. They can also be used as a formal tool for causal inference, such as predicting the effects of external interventions. Given that the diagram is correct, one can see whether the causal effects of interest (target effects, or causal estimands) can be estimated from available data, or what additional observations are needed to validly estimate those effects. One can also see how to represent the effects as familiar standardized effect measures. The present article gives an overview of: (1) components of causal graph theory; (2) probability interpretations of graphical models; and (3) methodologic implications of the causal and probability structures encoded in the graph, such as sources of bias and the data needed for their control.

Introduction

From their inception in the early twentieth century, causal models (more commonly known as structural-equations models) were accompanied by graphical representations or path diagrams that provided compact summaries of qualitative assumptions made by the models. Figure 1 provides a graph that would correspond to any system of five equations encoding these assumptions:

1. independence of A and B
2. direct dependence of C on A and B
3. direct dependence of E on A and C
4. direct dependence of F on C and
5. direct dependence of D on B, C, and E

The interpretation of "direct dependence" was kept rather informal and usually conveyed by causal intuition, for example, that the entire influence of A on F is "mediated" by C.

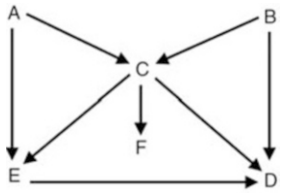
By the 1980s it was recognized that these diagrams could be reinterpreted formally as probability models, enabling use of graph theory in probabilistic inference, and allowing easy deduction of independence conditions implied by the assumptions (Pearl 1988). By the 1990s it was further recognized that these diagrams could also be used as tools for guiding causal and counterfactual inference (Pearl 1995, 2000; Pearl and Robins 1995; Spirtes et al. 2001) and for illustrating sources of bias and their remedy in empirical research (Greenland et al. 1999; Greenland 2000, 2003; Robins 2001; Greenland and Brumback 2002; Cole and Hernán 2002; Hernán et al. 2002; Jewell 2004; Pearl 2009; Glymour and Greenland 2008). Given that the graph is correct, one can see whether the causal effects of interest (target effects, or causal estimands) can be estimated from available data, or what additional observations are needed to validly estimate those effects. One can also see how to represent the effects as familiar standardized effect measures.

The present article gives an overview of: (1) components of causal graph theory; (2) probability interpretations of graphical models; and (3) methodologic implications of the causal and probability structures encoded in the graph, such as sources of bias and the data needed for their control. *See* ▶Causation and Causal Inference for discussion of definitions of causation and statistical models for causal inference.

Graphical Models and Causal Diagrams

Basics of Graph Theory

As befitting a well developed mathematical topic, graph theory has an extensive terminology that, once mastered, provides access to a number of elegant results which may be used to model any system of relations. The term *dependence* in a graph, usually represented by connectivity, may refer to mathematical, causal, or statistical dependencies. The connectives joining variables in the graph are called *arcs, edge,* or *links,* and the variables are also called *nodes* or *vertices.* Two variables connected by an arc are *adjacent* or *neighbors* and arcs that meet at a variable are also adjacent. If the arc is an arrow, the tail (starting) variable is the *parent* and the head (ending) variable is the *child*. In causal diagrams, an arrow represents a "direct effect" of the parent on the child, although this effect is direct only relative



Causal Diagrams. Fig. 1 $E \leftarrow C \rightarrow D$ is open, $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ is closed

to a certain level of abstraction, in that the graph omits any variables that might mediate the effect.

A variable that has no parent (such as A and B in Fig. 1) is *exogenous* or *external*, or a *root* or *source* node, and is determined only by forces outside of the graph; otherwise it is *endogenous* or *internal*. A variable with no children (such as D in Fig. 1) is a *sink* or *terminal node*. The set of all parents of a variable X (all variables at the tail of an arrow pointing into X) is denoted $\text{pa}[X]$; in Fig. 1, $\text{pa}[D] = \{B, C, E\}$.

A *path* or *chain* is a sequence of adjacent arcs. A *directed path* is a path traced out entirely along arrows tail-to-head. If there is a directed path from X to Y, X is an *ancestor* of Y and Y is a *descendant* of X. In causal diagrams, directed paths represent causal pathways from the starting variable to the ending variable; a variable is thus often called a cause of its descendants and an effect of its ancestors. In a *directed* graph the only arcs are arrows, and in an *acyclic* graph there is no feedback loop (directed path from a variable back to itself). Therefore, a *directed acyclic graph* or DAG is a graph with only arrows for edges and no feedback loops (i.e., no variable is its own ancestor or its own descendant).

A variable *intercepts* a path if it is in the path (but not at the ends); similarly, a set of variables S intercepts a path if it contains any variable intercepting the path. Variables that intercept directed paths are *intermediates* or *mediators* on the pathway. A variable is a *collider* on the path if the path enters and leaves the variable via arrowheads (a term suggested by the collision of the arrows at the variable). Note that being a collider is relative to a path; for example in Fig. 1, C is a collider on the path $A \rightarrow C \leftarrow B \rightarrow D$ and a noncollider on the path $A \rightarrow C \rightarrow D$. Nonetheless, it is common to refer to a variable as a collider if it is a collider along any path (i.e., if it has more than one parent). A path is *open* or *unblocked* at noncolliders and *closed* or *blocked* at colliders; hence a path with no collider (like $E \leftarrow C \leftarrow B \rightarrow D$) is *open* or *active*, while a path with a collider (like $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$) is closed or inactive.

Some authors use a bidirectional arc (two-headed arrow, $\leftrightarrow$) to represent the assumption that two variables share ancestors that are not shown in the graph; $A \leftrightarrow B$ then means that there is an unspecified variable U with directed paths to both A and B (e.g., $A \leftarrow U \rightarrow B$).

Interpretations of Graphs

Depending on assumptions used in its construction, graphical relations may be given three distinct levels of interpretation: probabilistic, causal, and functional. We now briefly describe these levels, providing further details in later sections.

The probabilistic interpretation requires the weakest set of assumptions. It treats the diagram as a carrier of conditional independencies constraints on the joint distribution of the variables in the graph. To serve in this capacity, the parents $\text{pa}[X]$ of each variable X in the diagram are chosen so as to render X independent of all its nondescendants, given $\text{pa}[X]$. When this condition holds, we say that the the diagram is *compatible* with the joint distribution. In Fig. 1, for example, variable E is assumed to be independent of its nondescendants $\{B, D, F\}$ given its parents $\text{pa}[E] = \{A, C\}$. We will see that compatibility implies many additional independencies (e.g., E and F are independent given C) that could be read from the diagram by tracing its paths. In real-life problems, compatibility arises if each parent–child family $\{X, \text{pa}[X]\}$ represents a stochastic process by which nature determines the probability of the child X as a function of the parents $\text{pa}[X]$, independently of values previously assigned to variables other than the parents.

To use diagrams for causal inference, we must assume that the direction of the arrows correspond to the structure of the causal processes generating the data. More specifically, the graph becomes a *causal diagram* if it encodes the assumption that for each parent–child family, the conditional probability $\Pr(x|\text{pa}[X])$ would remain the same regardless of whether interventions take place on variables not involving $\{X, \text{pa}[X]\}$, even if they are ancestors or descendants of X. In Fig. 1, for example, the conditional probability $P(C|A, B)$ is assumed to remain invariant under manipulation of the consequences of C, i.e., E, F or D. A causal DAG represents a complete causal structure, in that all sources of causal dependence are explained by causal links; in particular, it is assumed that all common (shared) causes of variables in the graph are also in the graph, so that all exogenous variables (root nodes) are causally independent (although they may be unobserved).

If we assume further that the arrows represent functional relationships, namely processes by which nature assigns a definite value to each internal node, the diagram can then be used to process counterfactual information and display independencies among potential outcomes

(including counterfactual variables) (Pearl 1995, Chap. 7). We will describe such *structural diagrams* and potential outcomes below.

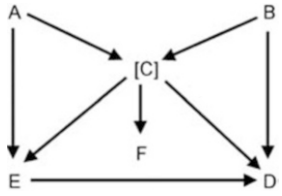
Control: Manipulation Versus Conditioning

The word "control" is used throughout science, but with a variety of meanings that are important to distinguish. In experimental research, to control a variable C usually means to manipulate or set its value. In observational studies, however, to control C (or more precisely, to control for C) more often means to condition on C, usually by stratifying on C or by entering C in a regression model. The two processes are very different physically and have very different representations and implications (Pearl 1995; Greenland et al. 1999).

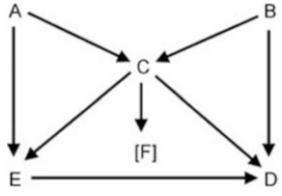
If a variable X is influenced by a researcher, a realistic causal diagram would need an ancestor R of X to represent this influence. In the classical experimental case in which the researcher alone determines X, R and X would be identical. In human trials, however, R more often represents just an *intention* to treat (with the assigned level of X), leaving X to be influenced by other factors that affect compliance with the assigned treatment R. In either case, R might be affected by other variables in the graph. For example, if the researcher uses age to determine assignments (an age-biased allocation), age would be a parent of R. Ordinarily however R would be exogenous, as when R represents a randomized allocation.

In contrast, by definition in an observational study there is no such variable R representing the researcher's influence on X. Conditioning is often used as a substitute for experimental control, in the hopes that with sufficient conditioning, X will be independent of uncontrolled influences. Conditioning on a variable C closes open paths that pass through C. However, if C is a collider, conditioning on C opens paths that were blocked by C or by an ancestral collider A. In particular, conditioning on a variable may open a path even if it is not on the path, as with F in Figs. 1 and 3.

To illustrate conditioning in a graph, we will redraw the graph to surround conditioned variables with square brackets (conditioned variables are often circled instead). We may now graphically determine the status of paths after conditioning by regarding the path open at colliders that are bracketed or have bracketed descendants, open at unbracketed noncolliders, and closed elsewhere. Figure 2 shows Fig. 1 after conditioning on C, from which we see that the $E-D$ paths $E \leftarrow C \leftarrow B \rightarrow D$ and $E \leftarrow A \rightarrow C \rightarrow D$ have been blocked, but the path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ has been opened. Were we to condition on F but not C, no open path would be blocked, but the path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ would again be opened.



Causal Diagrams. Fig. 2 Conditional on C, $E \leftarrow C \rightarrow D$ is closed but $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ is open



Causal Diagrams. Fig. 3 Conditional on F, $E \leftarrow C \rightarrow D$ and $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ are both open

The opening of paths at conditioned colliders reflect the fact that we should expect two unconditionally independent causes A and B become dependent if we condition on their consequences, which in Fig. 1 are C and F. To illustrate, suppose A and B are binary indicators (i.e., equal to 1 or 0), marginally independent, and $C = A + B$. Then among persons with $C = 1$, some will have $A = 1$, $B = 0$ and some will have $A = 0$, $B = 1$ (because other combinations produce $C \neq 1$). Thus when $C = 1$, A and B will exhibit perfect negative dependence: $A = 1 - B$ for all persons with $C = 1$.

The distinction between manipulation and conditioning is brought to the fore when considering the notion of "holding a variable constant." Conditioning on a variable X means that we choose to narrow the scope of discussion to those situations only where X attains a given value, regardless of how that value is attained. Manipulating X means that we physically intervene and set X to a given value, say $X = x$. The difference can be profound. For example, in cancer screening, conditioning on the absence of lighters and matches in the home lowers dramatically the probability of finding lung cancer, because restricting our attention to those who do not have these tools for smoking is tantamount to examining nonsmokers. In contrast, removing lighters and matches from people's homes during the screening will not lower the probability of finding lung cancer, since any lung cancers present will be unaffected by this act. Likewise, conditional on a low barometer reading we will have a lower probability of rain than

unconditionally, but setting the barometer to a low reading (e.g., by pushing its needle down) will have no effect on the weather.

Graphical Representation of Manipulation

One way of representing manipulation in the graph is to simulate the act of setting X to a constant, or the immediate implications of that act. If prior to intervention the probability of X is influenced by its parents via $P(x|\text{pa}[X])$, such influence no longer exists under an intervention that is made without reference to the parents or other variables. In that case, physically setting X at x dislodges X from the influence of its parents and subjects it to a new influence that keeps its value at $X = x$ regardless of the values taken by other variables. This can be represented by cutting out all arrows pointing to X and thus creating a new graph, in which X is an exogenous (root) node, while keeping the rest of the graph (with its associated conditional probabilities) intact. For example, setting C to a constant in Fig. 1, will render E and D independent, because all $E - D$ paths will be blocked by such intervention, including $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$, even though the latter path would be opened by conditioning on C. On the other hand, manipulating F but not C would leave all $E - D$ paths intact, and the $E - D$ association will therefore remain unaltered.

Assuming the graph is correct, graphical representation of interventions by deleting arrows enables us to compute post-intervention distributions from pre-intervention distributions (Pearl 1995, 2001, 2009; Spirtes et al. 2001; Lauritzen 2001) for a wide variety of interventions, including those that have side effects or that are conditioned upon other variables in the graph (Pearl 1995, pp. 105, 113). Nonetheless, "holding X constant" does not always correspond to a physically feasible manipulation, not even conceptually. Consider systolic blood pressure (SBP) as a cause of stroke (Y). It is easy to "hold SBP constant" in the sense of conditioning on each of its observed values. But what does it mean to "hold SBP constant" in the manipulative sense? There is only one condition under which SBP is constant: Death, when it stays put at zero. Otherwise, SBP is fluctuating constantly in some strictly positive range in response to posture, activity, and so on. Furthermore, no one knows how to influence SBP except by interventions R which have side effects on stroke (directed paths from R to Y that do not pass through SBP). Yet these side effects vary dramatically with intervention (e.g., there are vast differences between exercise versus medication side effects).

On the other hand, consider the problem of estimating the causal effect of SBP on the rate of blood flow in a given blood vessel. At this physiological level of discussion we can talk about the effect on blood flow of "changing SBP from level s to level s'," without specifying any mechanism for executing that change. We know from basic physics that the blood flow in a vessel depends on blood pressure, vessel diameter, blood viscosity, and so on; and we can ask what the blood flow would be if the blood pressure were to change from s to s' while the other factors remained at their ambient values. Comparing the results from conditioning on SBP $= s$ versus conditioning on SBP $= s'$ would not give us the desired answer because these conditioning events would entail different distributions for the causes (ancestors) of SBP, some of which might also affect those determinants of flow which we wish held constant when comparing.

We may thus conclude that there are contexts in which it makes no practical sense to speak of "holding X constant" via manipulation. In these contexts, manipulation of a given variable X can only be represented realistically by an additional node R representing an actual intervention, which may have side effects other than those intended or desired. On the other hand, such an R node will be redundant if X itself is amenable to direct manipulation. For such an X, manipulation can be represented by removing the arrows ending in X which correspond to effects overridden by the manipulation (Pearl 1995, 2000, 2009; Spirtes et al. 2001; Lauritzen 2001). When X is completely randomized or held constant physically, this corresponds to removing all arrows into X.

The phrase "holding X constant" may also be meaningful when X is not directly manipulable. In these cases, we may still be able to estimate a causal effect of X if we can find an instrumental variable Z (a variable that is associated with X but not with any uncontrolled confounding variable U, and Z has no effect on Y except through X). Although the operational meaning of these effects is not immediately apparent when direct manipulation of X free of side effects is not conceivable, estimation of these effects can help judge proposed interventions that affect Y via effects on X.

Separation

The intuition of closing and opening paths by conditioning is captured by the concept of "separation" which will be defined next. We say that a path is *blocked* by a set S if the path contains either an arrow-emitting node that is in S, or a collider that is outside S and has no descendant in S.

Two variables (or sets of variables) in the graph are *d-separated* (or just separated) by a set S if, after conditioning on S, there is no open path between them. Thus S d-separates X from Y if S blocks all paths from X to Y. In Fig. 1, $\{A, C\}$ d-separates E from B, but $\{C\}$ does

not (because conditioning on C alone results in Fig. 2, in which E and B are connected via the open path A). In a causal DAG, pa$[X]$ d-separates X from every variable that is not affected by X (i.e., not a descendant of X). This feature of DAGs is sometimes called the "Markov condition," expressed by saying the parents of a variable "screen off" the variable from everything but its effects. Thus in Fig. 1 pa$[E] = \{A, C\}$, which d-separates E from B but not from D.

In a probability graph, d-separation of X and Y by S implies that X and Y are independent given S in any distribution compatible with graph. In a causal diagram, d-separation of X and Y by S implies additionally that manipulation of X will not alter the distribution of Y if the variables in S are held constant physically (assuming this can be done). More generally, the distribution of Y will remain unaltered by manipulation of X if we can hold constant physically a set S that intercepts all directed paths from X to Y, even if S does not d-separate X and Y. This is so because only descendants of X can be affected by manipulation of X. In sharp contrast, conditioning on X may change the probabilities of X's ancestors; hence the stronger condition of d-separation by S is required to insure that conditioning on X does not alter the distribution of Y given S.

Statistical Interpretations and Applications

Earlier we defined the notion of compatibility between a joint probability distribution for the variables in a graph and the graph itself. It can be shown that compatibility is logically equivalent to requiring that two sets of variables are independent given S whenever S separates them in the graph. Moreover these conditional independencies constitute the *only* testable implications of a causal model specified by the diagram (Pearl 1988, p. 120). Thus, given compatibility, two sets of variables will be independent in the distribution if there is no open path between them in the graph.

Many special results follow for distributions compatible with a DAG. For example, if in a DAG, X is not an ancestor of any variable in a set T, then T and X will be independent given pa$[X]$. A distribution compatible with a DAG thus can be reduced to a product of factors $\Pr(x|\text{pa}[X])$, with one factor for each variable X in the DAG; this is sometimes called the "Markov factorization" for the DAG. When X is a treatment, this condition implies the probability of treatment is fully determined by the parents of X, pa$[X]$. Algorithms are available for constructing DAGs that are compatible with a given distribution (Pearl 1988, pp. 119–121).

Some of the most important constraints imposed by a graphical model on a compatible distribution correspond to the independencies implied by absence of open paths; e.g., absence of an open path from A to B in Fig. 1 constrains A and B to be marginally independent (i.e., independent if no stratification is done). Nonetheless, the converse does not hold; i.e., presence of an open path allows but does not imply dependency. Independence may arise through cancellation of dependencies; as a consequence even adjacent variables may be marginally independent; e.g., in Fig. 1, A and E could be marginally independent if the dependencies through paths $A \rightarrow E$ and $A \rightarrow C \rightarrow E$ cancelled each other. The assumption of faithfulness, discussed below, is designed to exclude such possibilities.

Bias and Confounding

Usually, the usage of terms like "bias," "confounding" and related concepts refer to dependencies that reflect more than just the effect under study. To capture these notions in a causal graph, we say that an open path between X and Y is a *biasing path* if it is not a directed path. The association of X with Y is then *unbiased* for the effect of X on Y if the only open paths from X to Y are the directed paths. Similarly, the dependence of Y on X is *unbiased given* S if, after conditioning on S, the open paths between X and Y are exactly (only and all) the directed paths in the starting graph. In such a case we say S is sufficient to block bias in the $X - Y$ dependence, and is minimally sufficient if no proper subset of S is sufficient.

Informally, confounding is a source of bias arising from causes of Y that are associated with but not affected by X (*see* ►Confounding). Thus we say an open nondirected path from X to Y is a *confounding path* if it ends with an arrow into Y. Variables that intercept confounding paths between X and Y are *confounders*. If a confounding path is present, we say *confounding* is present and that the dependence of Y on X is *confounded*. If no confounding path is present we say the dependence is *unconfounded*, in which case the only open paths from X to Y through a parent of Y are directed paths. Similarly, the dependence of Y on X is *unconfounded given* S if, after conditioning on S, the only open paths between X and Y through a parent of Y are directed paths.

An unconfounded dependency may still be biased due to nondirected open paths that do not end in an arrow into Y. These paths can be created when one conditions on a descendant of both X and Y, or a descendant of a variable intercepting a directed path from X to Y (Pearl 2000, p. 339). The resulting bias is called *Berksonian bias*, after its discoverer Joseph Berkson (Rothman et al. 2008). Most epidemiologists call this type of bias "selection bias" (Rothman et al. 2008) while computer scientists refer to

it as "explaining away" (Pearl 1988). Nonetheless, some writers (especially in econometrics) use "selection bias" to refer to confounding, while others call any bias created by conditioning "selection bias".

Consider a set of variables S that contains no effect (descendant) of X or Y. S is *sufficient* to block confounding if the dependence of Y on X is unconfounded given S. "No confounding" thus corresponds to sufficiency of the empty set. A sufficient S is called *minimally sufficient* to block confounding if no proper subset of S is sufficient. The initial exclusion from S of descendants of X or Y in these definitions arises first, because conditioning on X-descendants can easily block directed (causal) paths that are part of the effect of interest, and second, because conditioning on X or Y descendants can unblock paths that are not part of the $X - Y$ effect, and thus create new bias.

These considerations lead to a graphical criterion called the *back-door criterion* which identifies sets S that are sufficient to block bias in the $X - Y$ dependence (Pearl 1995, 2000). A *back-door path* from X to Y is a path that begins with a parent of X (i.e., leaves X from a "back door") and ends at Y. A set S then satisfies the back-door criterion with respect to X and Y if S contains no descendant of X and there are no open back-door paths from X to Y after conditioning on S.

In a unconditional DAG, the following properties hold (Pearl 1995, 2000; Spirtes et al. 2001; Glymour and Greenland 2008):

1. All biasing paths are back-door paths.
2. The dependence of Y on X is unbiased whenever there are no open back-door paths from X to Y.
3. If X is exogenous, the dependence of any Y on X is unbiased.
4. All confounders are ancestors of either x or of y.
5. A back-door path is open if and only if it contains a common ancestor of X and Y.
6. If S satisfies the back-door criterion, then S is sufficient to block $X - Y$ confounding.

These conditions do not extend to conditional DAGs like Fig. 2. Also, although $\mathrm{pa}[X]$ always satisfies the back-door criterion and hence is sufficient in a DAG, it may be far from minimal sufficient. For example, there is no confounding and hence no need for conditioning whenever X separates $\mathrm{pa}[X]$ from Y (i.e., whenever the only open paths from $\mathrm{pa}[X]$ to Y are through X).

As a final caution, we note that the biases dealt with by the above concepts are only confounding and selection biases. To describe biases due to measurement error and model-form misspecification, further nodes representing mismeasured or misspecifed variables must be introduced (Glymour and Greenland 2008).

Estimation of Causal Effects

Suppose now we are interested in the effect of X on Y in a causal DAG, and we assume a probability model compatible with the DAG. Then, given a sufficient set S, the only source of association between X and Y within strata of S will be the directed paths from X to Y. Hence the *net effect* of $X = x_1$ vs. $X = x_0$ on Y when $S = s$ is defined as $\Pr(y|x_1, s) - \Pr(y|x_0, s)$, the difference in risks of $Y = y$ at $X = x_1$ and $X = x_0$. Alternatively one may use another effect measure such as the risk ratio $\Pr(y|x_1, s)/\Pr(y|x_0, s)$. A *standardized effect* is a difference or ratio of weighted averages of these stratum-specific $\Pr(y|x, s)$ over S, using a common weighting distribution. The latter definition can be generalized to include intermediate variables in S by allowing the weighting distribution to causally depend on X. Furthermore, given a set Z of intermediates along all directed paths from X to Y and identification of the $X - Z$ and $Z - Y$ effects, one can produce formulas for the $X - Y$ effect as a function of the $X - Z$ and $Z - Y$ effects ("front-door adjustment" (Pearl 1995, 2000)).

The above form of standardized effect is identical to the forms derived under other types of causal models, such as potential-outcome models (*see* ▶Causation and Causal Inference). In those models, the outcome Y of each unit is replaced by a vector of outcomes Y containing components Y_x, where Y_x represents the outcome when $X = x$ is the treatment given. When S is sufficient, some authors (Pearl 2000) go so far as to identify the $\Pr(y|x, s)$ with the distribution of potential outcomes Y_x given S, thereby creating a *structural model* for the potential outcomes. If the graph is based on functional rather than probabilistic relationships between parents and children, this identification can also model unit-based counterfactuals $Y_x(u)$ for any pair (X, Y), where u is a unit index or a vector of exogeneous variables characterizing the units.

There have been objections to this identification on the grounds that not all variables in the graph can be manipulated, and that potential-outcome models do not apply to nonmanipulable variables. The objection loses force when X is an intervention variable, however. In that case, sufficiency of a set S implies that the marginal potential-outcome distribution $\Pr(Y_x = y)$ equals $\sum_s \Pr(y|x, s)\Pr(s)$, which is the risk of $Y = y$ given $X = x$ standardized to the S distribution.

In fact, sufficiency of S implies the stronger condition of *strong ignorability* given S, which says that X and the vector Y of potential outcomes are independent given S. In particular, strong ignorability given S follows if S

satisfies the back-door criterion, or if X is randomized given S. Nonetheless, for the equation $\Pr(Y_x = y) = \sum_s \Pr(y|x,s)\Pr(s)$ it suffices that X be independent of each component potential outcome Y_x given S, a condition sometimes called weak ignorability given S.

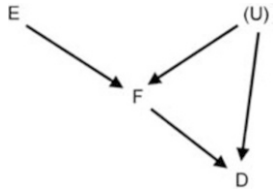
Identification of Effects and Biases

To check sufficiency and identify minimally sufficient sets of variables given a graph of the causal structure, one need only see whether the open paths from X to Y after conditioning are exactly the directed paths from X to Y in the starting graph. Mental effort may then be shifted to evaluating the reasonableness of the causal independencies encoded by the graph, some of which are reflected in conditional independence relations. This property of graphical analysis facilitates the articulation of necessary background knowledge for estimating effects, and eases teaching of algebraically difficult identification conditions.

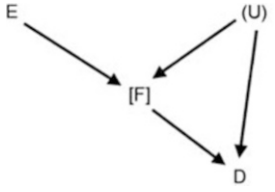
As an example, spurious sample associations may arise if each variable affects selection into the study, even if those selection effects are independent. This phenomenon is a special case of the collider-stratification effect illustrated earlier. Its presence is easily seen by starting with a DAG that includes a selection indicator $F = 1$ for those selected, 0 otherwise, as well as the study variables, then noting that we are always forced to examine associations within the $F = 1$ stratum (i.e., by definition, our observations stratify on selection). Thus, if selection (F) is affected by multiple causal pathways, we should expect selection to create or alter associations among the variables.

Figure 4 displays a situation common in randomized trials, in which the net effect of E on D is unconfounded, despite the presence of an uncontrolled cause U of D. Unfortunately, a common practice in health and social sciences is to stratify on (or otherwise adjust for) an intermediate variable F between a cause E and effect D, and then claim that the estimated (F-residual) association represents that portion of the effect of E on D not mediated through F. In Fig. 4 this would be a claim that, upon stratifying on the collider F, the $E-D$ association represents the direct effect of E on D. Figure 5 however shows the graph conditional on F, in which we see that there is now an open path from E to D through U, and hence the residual $E-D$ association is confounded for the direct effect of E on D.

The $E-D$ confounding by U in Fig. 5 can be seen as arising from the confounding of the $F-D$ association by U in Fig. 4. In a similar fashion, conditioning on C in Fig. 1 opens the confounding path through A, C, and B as seen in Fig. 2; this path can be seen as arising from the confounding of the $C-E$ association by A and the $C-D$ association by B in Fig. 1. In both examples, further stratification on either A or B blocks the created path and thus removes the new confounding.



Causal Diagrams. Fig. 4 $E \to F \to D$ is open, $E \to F \leftarrow U \to D$ is closed



Causal Diagrams. Fig. 5 Conditional on F, $E \to F \to D$ is closed but $E \to F \leftarrow U \to D$ is open

Bias from conditioning on a collider or its descendant has been called "collider bias" (Greenland 2003; Glymour and Greenland 2008). Starting from a DAG, there are two distinct forms of this bias: Confounding induced in the conditional graph (Figs. 2, 3, and 5), and Berksonian bias from conditioning on an effect of X and Y. Both biases can in principle be removed by further conditioning on certain variables along the biasing paths from X to Y in the conditional graph. Nonetheless, the starting DAG will always display ancestors of X or Y that, if known, could be used remove confounding; in contrast, no variable need appear or even exist that could be used to remove Berksonian bias.

Figure 4 also provides a schematic for estimating the $F-D$ effect, as in randomized trials in which E represents assignment to or encouragement toward treatment F. In this case E acts as an *instrumental variable* (or instrument), a variable associated with F such that every open path from E to D includes an arrow pointing into F (Pearl 2000; Greenland 2000; Glymour and Greenland 2008). Although the $F-D$ effect is not generally estimable, using the instrument E one can put bounds on confounding of the $F-D$ association, or use additional assumptions that render the effect of F on D estimable.

Questions of Discovery

While deriving statistical implications of graphical models is uncontroversial, algorithms that claim to discover causal (graphical) structures from observational data have been

subject to strong criticism (Freedman and Humphreys 1999; Robins and Wasserman 1999). A key assumption in certain "discovery" algorithms is a converse of compatibility called *faithfulness* Spirtes et al. 2001. A compatible distribution is *faithful* to the graph (or *stable* Pearl (2000)) if for all X, Y, and S, X and Y are independent given S **only** when S separates X and Y (i.e., the distribution contains no independencies other than those implied by graphical separation). Faithfulness implies that minimal sufficient sets in the graph will also be minimal for consistent estimation of effects. Nonetheless, there are real examples of near cancellation (e.g., when confounding obscures a real effect), which make faithfulness questionable as a routine assumption. Fortunately, faithfulness is not needed for the uses of graphical models discussed here.

Whether or not one assumes faithfulness, the generality of graphical models is purchased with limitations on their informativeness. Causal diagrams show whether the effects can be estimated from the given information, and can be extended to indicate effect direction when that is monotone VanderWeele and Robins 2010;. Nonetheless, the nonparametric nature of the graphs implies that parametric concepts like effect-measure modification (heterogeneity of arbitrary effect measures) cannot be displayed by the basic graphical theory. Similarly, the graphs may imply that several distinct conditionings are minimal sufficient (e.g., both $\{A, C\}$ and $\{B, C\}$ are sufficient for the *ED* effect in Fig. 1), but offer no further guidance on which to use. Open paths may suggest the presence of an association, but that association may be negligible even if nonzero. Because association transmitted by an open path may become attenuated as the length of the path increases, there is often good reason to expect certain phenomena (such as the conditional $E - D$ confounding shown in Figs. 2, 3 and 5) to be small in practical terms.

Further Readings

Full technical details of causal diagrams and their relation to causal inference can be found in the books by Pearl (2000) and Spirtes et al. (2001). A compact survey is given in Pearl (2009). Less technical reviews geared toward health scientists include Greenland et al. (2002), Greenland and Brumback (2008), and Glymour and Greenland (1999).

About the Authors

For Dr. Greenland's biography *see* the entry ►Confounding and Confounder Control.

Dr. Pearl is Professor of Computer Science at the University of California in Los Angeles, and Director of UCLA's Cognitive Systems Laboratory. He is one of the pioneers of Bayesian networks and the probabilistic approach to artificial intelligence. He is a member National Academy of Engineering (1995) and Corresponding Member, Spanish Academy of Engineering (2002). Professor Pearl was awarded the Lakatos Award, London School of Economics and Political Science for "an outstanding contribution to the philosophy of science" (2001); the ACM Allen Newell Award for "groundbreaking contributions that have changed the scientific world beyond computer science and engineering. Dr. Pearl made seminal contributions to the field of artificial intelligence that extend to philosophy, psychology, medicine, statistics, econometrics, epidemiology and social science."; the Benjamin Franklin Medal in Computers and Cognitive Science for "creating the first general algorithms for computing and reasoning with uncertain evidence, allowing computers to uncover associations and causal connections hidden within millions of observations. His work has had a profound impact on artificial intelligence and statistics, and on the application of these fields to a wide range of problems in science and engineering" (2008). He has written three fundamental books in artificial intelligence: *Heuristics: Intelligent Search Strategies for Computer Problem Solving* (Addison-Wesley, 1984), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan-Kaufmann, 1988) and *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2000). Profeshhhhsor Pearl holds two Honorary Doctorates.

Cross References

►Causation and Causal Inference
►Confounding and Confounder Control
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Rubin Causal Model
►Structural Equation Models

References and Further Reading

Cole S, Hernán MA (2002) Fallibility in estimating direct effects. Int J Epidemiol 31:163–165

Freedman DA, Humphreys P (1999) Are there algorithms that discover causal structure? Synthese 121:29–54

Glymour MM, Greenland S (2008) Causal diagrams. Ch. 12. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729 (Erratum: 2000, 29, 1102)

Greenland S (2003) Quantifying biases in causal models: classical confounding versus collider-stratification bias. Epidemiology 14:300–306

Greenland S, Brumback BA (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48
Hernán MA, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation. Am J Epidemiol 155:176–184
Jewell NP (2004) Statistics for epidemiology. Chapman and Hall/CRC Press, Boca Raton, Sect. 8.3
Lauritzen SL (2001) Causal inference from graphical models. In: Cox DR, Kluppelberg C (eds) Complex stochastic systems. Chapman and Hall/CRC Press, Boca Raton, pp 63–107
Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo
Pearl J (1995) Causal diagrams for empirical research (with discussion). Biometrika 82:669–710
Pearl J (2000) Causality. Cambridge University Press, New York. 2nd edition, 2009
Pearl J (2009) Causal inference in statistics: an overview. Statist Surv 3:96–146
Pearl J, Robins JM (1995) Probabilistic evaluation of sequential plans from causal models with hidden variables. In: Proceedings of the eleventh conference annual conference on uncertainty in artificial intelligence (UAI-95). Morgan Kaufmann, San Francisco, pp 444–453
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Wasserman L (1999) On the impossibility of inferring causation from association without background knowledge. In: Glymour C, Cooper G (eds) Computation, causation, and discovery. AAAI Press/The MIT Press, Menlo Park/Cambridge, pp 305–321
Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott, Philadelphia
Spirtes P, Glymour C, Scheines R (2001) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge
VanderWeele, TJ, Robins JM (2010) Signed directed acyclic graphs for causal inference. J R Stat Soc Ser B 72(1):111–127

Causation and Causal Inference

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

In the health sciences, definitions of cause and effect have not been tightly bound with methods for studying causation. Indeed,many approaches to causal inference require no definition, leaving users to imagine causality however they prefer. As Sir Austin Bradford Hill said in his famous article on causation: "I have no wish…to embark upon a philosophical discussion of the meaning of 'causation'" (Hill 1965). Without a formal definition of causation, an association is distinguished as causal only by having been identified as such based on external and largely subject-matter considerations, such as those Hill put forth.

Nonetheless, beneath most treatments of causation in the health sciences, one may discern a class of definitions built around the ideas of counterfactuals or potential outcomes. These ideas have a very long history and form the foundation of most current statistical methods for causal inference. Thus, the present article will begin with these definitions and the methods they entail. It will then turn methods that explicitly presume no definition of causation but rather begin with an idea of what a causal association should look like (perhaps derived from subject-matter judgments, including consideration of possible counterfactuals), and employ statistical methods to estimate those associations.

Counterfactuals and Potential Outcomes

Skeptical that induction in general and causal inference in particular could be given a sound logical basis, David Hume nonetheless captured the foundation of the potential-outcome approach when he wrote

> We may define a cause to be an object, followed by another, … where, if the first object had not been, the second had never existed.
>
> (Hume 1748, p. 115)

A key aspect of this view of causation is its *counterfactual* element: It refers to how a certain outcome event (the "second object," or effect) would not have occurred if, *contrary to fact*, an earlier event (the "first object," or cause) had not occurred. In this regard, it is no different from conventional statistics, which refers to samples that might have occurred, but did not. This counterfactual view of causation was adopted by numerous philosophers and scientists after Hume (e.g., Mill 1843; Fisher 1918; Cox 1958; Simon and Rescher 1966; MacMahon and Pugh 1967; Stalnaker 1968; Lewis 1973).

The development of this view into a statistical theory with methods for causal inference is recounted by Rubin (1990), Greenland et al. (1999), Greenland (2004), and Pearl (2009). The earliest such theories were developed in the 1920s by Fisher, Neyman, and others for the analysis of randomized experiments and are today widely recognized under the heading of *potential-outcome models* of causation (also known in engineering as *destructive-testing* models). Suppose we wish to study the effect of an intervention variable X on a subsequent outcome variable Y defined on an observational unit or a population; for example, X could be the daily dose regimen for a drug in a clinical trial, and Y could be survival time. Given X has potential values $x_1, \ldots, x_J$ (e.g., drug doses), we suppose that there is a list of *potential outcomes* $\mathbf{y} = (y(x_1), \ldots, y(x_J))'$ such that if $X = x_j$ then $Y = y(x_j)$. The list $\mathbf{y}$ thus exhibits

the correspondence between treatments, interventions, or actions (the X values) and outcomes or responses (the Y values) for the unit, and so is sometimes called a *response schedule* (Berk 2004). A simpler and common notation has $\mathbf{y} = (y_1, \ldots, y_J)'$, with Y_j denoting the random variable "outcome when treated with $X = x_j$."

Under this model, assignment of a unit to a treatment level x_j is a choice of which potential outcome $y(x_j)$ from the list $\mathbf{y}$ to attempt to observe. It is ordinarily assumed that the assignments made for other units do not affect the outcomes of another unit, although there are extensions of the model to include between-unit interactions, as in contagious outcomes (Halloran and Struchiner 1995). Regardless of the X assignment, the remaining potential outcomes are treated as existing pre-treatment covariates on which data are missing (Rubin 1978, 1991). Because at most one of the J potential outcomes is observed per unit, the remaining potential outcomes can be viewed as missing data, and causal inference can thus be seen as a special case of inference with missing data.

To say that intervention x_i causally affects Y relative to intervention x_j means that $y(x_i) \neq y(x_j)$, i.e., X "matters" for Y for the unit. The *sharp* (or strong) null hypothesis is that $y(x)$ is constant over x within units. This hypothesis states that changing X would not affect the Y of any unit, i.e., $y(x_i) = y(x_j)$ for every unit and every x_i and x_j; it forms the basis of exact ►permutation tests such as ►Fisher's exact test (Greenland 1991). The effect of intervention x_i relative to x_j on a unit may be measured by the difference in potential outcomes $y(x_i) - y(x_j)$. If the outcome is strictly positive (like life expectancy or mortality risk), it could instead be measured by the ratio $y(x_i)/y(x_j)$.

Because we never observe two potential outcomes on a unit, we can only estimate population averages of the potential outcomes. We do this by observing average outcomes in differently exposed groups and substituting those observations for the average potential outcomes in the group of interest – a perilous process whenever the observed exposure groups are atypical of the population of interest with respect to other risk factors for the outcome (Maldonado and Greenland 2002) (see Confounding and Confounder Control).

A more subtle problem is that only for difference measures will the population effect (the difference of average potential outcomes) equal the population average effect (the average difference of potential outcomes). Hence the average of the differences $y(x_i) - y(x_j)$ in the population is often called the *average causal effect* (ACE) (Angrist et al. 1996). For some popular measures of effect, such as rate ratios and odds ratios, the population effect may not even equal any average of individual effects (Greenland 1987, 1996; Greenland et al. 1999).

The theory extends to probabilistic outcomes by replacing the $y(x_j)$ by probability functions $p_j(y)$ (Greenland 1987; Robins 1988; Greenland et al. 1999). The theory also extends to continuous X by allowing the potential-outcome list $\mathbf{y}$ to contain the potential outcome $y(x)$ or $p_x(y)$ for every possible value x of X. Both extensions are embodied in Pearl's notation for intervention effects, in which $p_x(y)$ becomes $P(Y = y|\text{set}[X = x])$ or $P(Y = y|\text{do}[X = x])$ (Pearl 1995, 2009). Finally, the theory extends to complex longitudinal data structures by allowing the treatments to be different event histories or processes (Robins 1987, 1997).

From Randomized to Observational Inference

Potential outcomes were developed part of a design-based strategy for causal inference in which ►randomization provided the foundation for inference. Indeed, before the 1980s, the model was often referred to as "the randomization model," even though the causal concepts within it do not hinge on randomization (e.g., Wilk 1955; Copas 1973). It thus seems that the early strong linkage of potential outcomes to randomized designs deflected consideration of the model for observational research. In the 1960s, however, a number of philosophers used counterfactuals to build general foundations for causal analysis (e.g., Simon and Rescher 1966; Stalnaker 1968; Lewis 1973). Similar informal ideas can be found among epidemiologists of the era (e.g., MacMahon and Pugh 1967), and conceptual models subsuming counterfactuals began to appear shortly thereafter (e.g., Miettinen 1972; Rothman 1976; Hamilton 1979).

The didactic value of these models was quickly apparent in the clarification they brought to ideas of strength of effect, synergy, and antagonism (MacMahon and Pugh 1967; Rothman 1976; see also Rothman et al. 2008, Chaps. 2 and 5). Most importantly, the models make clear distinctions between causal and statistical relations: Causal relations refer to relations of treatments to potential outcomes *within* treated units, whereas statistical relations refer associations of treatments with actual outcomes *across* units (Rothman et al. 2008, Chap. 4). Consequently, the models have aided in distinguishing confounding from collapsibility (Greenland and Robins 1986; Greenland et al. 1999), synergy from statistical interaction (Greenland and Poole 1988), and causation probabilities from attributable fractions (Greenland et al. 1999; Greenland and Robins 2000).

The conceptual clarification also stimulated development of statistical methods for observational studies. Rubin (1974, 1978) and his colleagues extended statistical machinery based on potential outcomes from

the experimental setting to observational data analysis, leading, for example, to propensity-scoring and inverse-probability-of-treatment methods for confounder adjustment (Rosenbaum 2002; Hirano et al. 2003), as well as new insights into analysis of trials with noncompliance (Angrist et al. 1996) and separation of direct and indirect effects (Robins and Greenland 1992, 1994; Frangakis and Rubin 2002; Kaufman et al. 2004). In many cases, such insights have led to methodologic refinements and better-informed choices among existing methods. In the longitudinal-data setting, potential-outcome modeling has led to entirely new methodologies for analysis of time-varying covariates and outcomes, including g-estimation and marginal structural modeling (Robins 1987, 1998; Robins et al. 1992, 1999, 2000).

A serious caution arises, however, when it is not clear that the counterfactual values for X (treatments other than the actual one) represent physical possibilities or even unambiguous states of nature. A classic example is gender (biological sex). Although people speak freely of gender (male vs. female) as cause of heart disease, given a particular man, it is not clear what it would mean for that man to have been a woman instead. Do we mean that the man cross-dressed and lived with a female identity his entire life? Or that he received a sex-change operation after birth? Or that the zygote from which he developed had its male chromosome replaced by a female chromosome?

Potential-outcome models bring to light such ambiguities in everyday causal language but do not resolve them (Greenland 2005a; Hernán 2005). Some authors appear to insist that use of the models be restricted to situations in which ambiguities are resolved, so that X must represent an intervention variable, i.e., a precise choice among treatment actions or decisions (Holland 1986). Many applications do not meet this restriction, however, and some go so far as to confuse outcomes (Y) with treatments (X), which can lead to nonsense results. Examples include estimates of mortality after "cause removal," e.g., removal of all lung-cancer deaths. Sensible interpretation of any effect estimate requires asking what intervention on a unit could have given the unit a value of X (here, lung-cancer death) other than the one that was observed, and what the side effects that intervention would have. One cannot remove all lung-cancer deaths by smoking cessation. A treatment with a 100% cure rate might do so but need not guarantee the same subsequent lifespan as if the cancer never occurred. If such questions cannot be given at least a speculative answer, the estimates of the impact of cause removal cannot be expected to provide valid information for intervention and policy purposes (Greenland 2005a).

More sweeping criticisms of potential-outcome models are given by Dawid (2000), for example, that the distribution of the full potential-outcome vector $\mathbf{Y}$ (i.e., the joint distribution of the $Y(x_1), \ldots, Y(x_J)$) cannot be nonparametrically identified by randomized experiments. Nonetheless, as the discussants point out, the practical implication of these criticisms are not clear, because the marginal distributions of the separate potential outcomes $Y(x_j)$ are nonparametrically identifiable, and known mechanisms of action may lead to identification of their joint distribution as well.

Canonical Inference

Before the extension of potential outcomes to observational inference, the only systematic approach to causal inference in epidemiology was the informal comparison of observations to characteristics expected of causal relations. Perhaps, the most widely cited of such approach is based on Hill's considerations (Hill 1965), which are discussed critically in numerous sources (e.g., Koepsell and Weiss 2003; Phillips and Goodman 2004; Rothman et al. 2008, Chap. 2) as well as by Hill himself.

The canonical approach usually leaves terms like "cause" and "effect" as undefined concepts around which the self-evident canons are built, much like axioms are built around concepts like "set" and "is an element of" in mathematics. Only proper temporal sequence (cause must precede effect) is a necessary condition for a cause–effect relation to hold. The remaining considerations are more akin to diagnostic symptoms or signs of causation – that is, they are properties an association is assumed more likely to exhibit if it is causal than if it is not (Hill 1965; Susser 1988, 1991). Furthermore, some of these properties (like specificity and dose response) apply only under specific causal models (Weiss 1981, 2002).

Thus, the canonical approach makes causal inference more closely resemble clinical judgment than experimental science, although experimental evidence is listed among the considerations (Hill 1965; Rothman et al. 2008, Chap. 2; Susser 1991). Some of the considerations (such as temporal sequence, association, dose response or predicted gradient, and specificity) are empirical signs and thus subject to conventional statistical analysis. Others (such as plausibility) refer to prior belief, and thus (as with disease symptoms) require elicitation, the same process used to construct priors for Bayesian analysis.

The canonical approach is widely accepted in health sciences, subject to many variations in detail. Nonetheless, it has been criticized for its incompleteness and informality, and the consequent poor fit it affords to the deductive

or mathematical approaches familiar to classic science and statistics (Rothman et al. 2008, Chap. 2). Although there have been some interesting attempts to reinforce or reinterpret certain canons as empirical predictions of causal hypotheses (e.g., Susser 1988; Weed 1986; Weiss 1981, 2002; Rosenbaum 2002), there is no generally accepted mapping of the entire canonical approach into a coherent statistical methodology; one simply uses standard statistical techniques to test whether empirical canons are violated. For example, if the causal hypothesis linking X to Y predicts a strictly increasing trend in Y with X, a test of this statistical prediction may serve as a statistical criterion for determining whether the hypothesis fails the dose-response canon. Such usage falls squarely in the falsificationist/frequentist tradition of the twentieth-century statistics, but leaves unanswered most of the policy questions that drive causal research; this gap led to the development of methodologic modeling or *bias analysis*.

Bias Analysis

In the second half of the twentieth-century, a more rigorous approach to observational studies emerged in the wake of major policy controversies such as those concerning cigarette smoking and lung cancer (e.g., Cornfield et al. 1959). This approach begins with the idea that, conditional on some sufficient set of confounders Z, there is a population association or relation between X and Y that is the target of inference. In other words, the Z-stratified associations are presumed to accurately reflect the effect of X on Y in that population stratum, however "effect" may be defined. Estimates of this presumably causal association are then the effect estimates.

Observational and analytic shortcomings bias or distort these estimates: Units may be selected for observation in a nonrandom fashion; stratifying on additional unmeasured covariates U may be essential for the X-Y association to approximate a causal effect; inappropriate covariates may be entered into the analysis; components of X or Y or Z may not be adequately measured; and so on. In methodologic modeling or *bias analysis*, one models these shortcomings. In effect, one attempts to model the design and execution of the study, including features (such as selection biases and measurement errors) beyond investigator control. The process is thus a natural extension to observational studies of the design-based paradigm in experimental and survey statistics. For further details, see BIAS MODELING or the overviews by Greenland (2005b, 2009).

Structural Equations and Causal Diagrams

Paralleling the development of potential-outcome models, an entirely different approach causal analysis arose in observational research in economics and related fields. Like methodologic modeling, this *structural-equations* approach does not begin with a formal definition of cause and effect, but instead develops models to reflect assumed causal associations, from which empirical (and hence testable) associations may be derived. Like most of statistics before the 1980s, structural-equations methods were largely limited to normal linear models to derive statistical inferences. Because these models bear no resemblance to typical epidemiologic data, this limitation may in part explain the near absence of structural equations from epidemiology, despite their ubiquity in social-science methodology. From their inception, however, causal system models have been accompanied by graphical representations or path diagrams that provided compact summaries of qualitative assumptions made by the structural model; see ►Causal Diagrams for a review.

Conclusion

Different approaches to causal inference represent separate historical streams rather than distinct methodologies, and can be blended in various ways. The result of any modeling exercise is simply one more input to informal judgments about causal relations, which may be guided by canonical considerations. Insights and innovations in any approach can thus benefit the entire process of causal inference, especially when that process is seen as part of a larger context. Other traditions or approaches (some perhaps yet to be imagined) may contribute to the process. It thus seems safe to say that no one approach or blend is a complete solution to the problem of causal inference, and that the topic remains one rich with open problems and opportunities for innovation.

Acknowledgments

Some of the above material is adapted from Greenland (2004).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Bias Analysis
►Causal Diagrams
►Complier-Average Causal Effect (CACE) Estimation

►Confounding and Confounder Control
►Event History Analysis
►Forecasting Principles
►Interaction
►Misuse of Statistics
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Rubin Causal Model
►Simpson's Paradox
►Spurious Correlation
►Structural Equation Models

References and Further Reading

Angrist J, Imbens G, Rubin DB (1996) Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91:444–472

Berk R (2003) Regression analysis: a constructive critique. Sage Publications, Thousand Oaks

Copas JG (1973) Randomization models for matched and unmatched 2 × 2 tables. Biometrika 60:267–276

Cornfield J, Haenszel W, Hammond WC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Nat Cancer Inst 22:173–203

Cox DR (1958) The planning of experiments. Wiley, New York

Dawid P (2000) Causal inference without counterfactuals (with discussion). J Am Stat Assoc 95:407–448

Fisher RA (1918) The causes of human variability. Eugenics Rev 10:213–220

Frangakis C, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58:21–29

Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analysis. Am J Epidemiol 125:761–768

Greenland S (1991) On the logical justification of conditional tests for two-by-two contingency tables. Am Stat 45:248–251

Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501

Greenland S (1998) Induction versus Popper: substance versus semantics. Int J Epidemiol 27:543–548

Greenland S (1999) The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem. Am J Publ Health 89: 1166–1169

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729 (Erratum: Int J Epidemiol 29:1102)

Greenland S (2003) Quantifying biases in causal models: classical confounding versus collider-stratification bias. Epidemiology 14:300–306

Greenland S (2004) An overview of methods for causal inference from observational studies. In: Gelman A, Meng XL (eds) Applied Bayesian modeling and causal inference from an incomplete-data perspective. Wiley, New York, pp 3–13

Greenland S (2005a) Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). Emerg Themes in Epidemiol (online journal), http://www.ete-online.com/content/2/1/5

Greenland S (2005b) Multiple-bias modeling for observational studies (with discussion). J Roy Stat Soc, ser A 168:267–308

Greenland S (2009) Relaxation penalties and priors for plausible modeling of nonidentified bias sources. Stat Sci 24:195–210

Greenland S, Brumback BA (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Poole C (1988) Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 14:125–129

Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419

Greenland S, Robins JM (2000) Epidemiology, justice, and the probability of causation. Jurimetrics 40:321–340

Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14:29–46

Halloran ME, Struchiner CJ (1995) Causal inference for infectious diseases. Epidemiology 6:142–151

Hamilton MA (1979) Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. Am J Epidemiol 109:362–379

Hernán MA (2005) Hypothetical interventions to define causal effects — afterthought or prerequisite? Am J Epidemiol 162: 618–620

Hill AB (1965) The environment and disease: association or causation? Proc Roy Soc Med 58:295–300

Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189

Hume D (1748) An enquiry concerning human understanding. 1988 reprint by Open Court Press, LaSalle

Jewell NP (2004) Statistics for epidemiology. Chapman & Hall/CRC Press, Boca Raton

Kaufman JS, MacLehose R, Kaufman S (2004) A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. Epidemiol Perspect Innov (online journal) 1:4

Lewis DK (1973) Causation. J Philos 70:556–567

MacMahon B, Pugh TF (1967) Causes and entities of disease. In: Clark DW, MacMahon B (eds) Preventive medicine. Little Brown, Boston, pp 11–18

Maldonado G, Greenland S (2002) Estimating causal effects (with discussion). Int J Epidemiol 31:421–438

Miettinen OS (1972) Standardization of risk ratios. Am J Epidemiol 96:383–388

Mill JS (1843) A system of logic, ratiocinative and inductive. 1956 reprint by Longman & Greens, London

Morrison AS (1985) Screening in chronic disease. Oxford, New York

Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710

Pearl J (2009) Causality, 2nd edn. Cambridge University Press, New York

Phillips CV, Goodman K (2004) The missed lessons of Sir Austin Bradford Hill. Epidemiologic Perspectives Innovations (online journal), http://www.epi-perspectives.com/content/1/1/3

Robins JM (1987) A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Dis 40 (suppl 2):139s–161s

Robins JM (1988) Confidence intervals for causal parameters. Stat Med 7:773–785

Robins JM (1989) The control of confounding by intermediate variables. Stat Med 8:679–701

Robins JM (1997) Causal inference from complex longitudinal data. In: Berkane M (ed) Latent variable modeling and applications to

causality. Lecture notes in statistics (120). Springer, New York, pp 69–117
Robins JM (1998) Structural nested failure time models. In: Armitage P, Colton T (eds) The encyclopedia of biostatistics. Wiley, New York, pp 4372–4389
Robins JM (1999) Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry DA (eds) Statistical models in epidemiology. Springer, New York, pp 95–134
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Greenland S (1989) The probability of causation under a stochastic model for individual risks. Biometrics 46: 1125–1138
Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143–155
Robins JM, Greenland S (1994) Adjusting for differential rates of PCP prophylaxis in high-versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc 89: 737–749
Robins JM, Blevins D, Ritter G, Wulfson M (1992) G-estimation of the effect of prophylaxis therapy for Pnemocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3:319–336 (Errata: Epidemiology 4:189)
Robins JM, Greenland S, Hu FC (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). J Am Stat Assoc 94: 687–712
Robins JM, Hernán MA, Brumback BA (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560
Rosenbaum P (2002) Observational studies, 2nd edn. Springer, New York
Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592
Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott, Philadelphia
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol 66: 688–701
Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58
Rubin DB (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci 5: 472–480
Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47:1213–1234
Simon HA, Rescher N (1966) Cause and counterfactual. Philos Sci 33:323–340
Stalnaker RC (1968) A theory of conditionals. In: Rescher N (ed) Studies in logical theory. Blackwell, Oxford
Stone (1993) The assumptions on which causal inference rest. J Roy Stat Soc, ser B 55:455–466
Susser M (1988) Falsification, verification and causal inference in epidemiology: reconsideration in light of Sir Karl Popper's philosophy. In: Rothman KJ (ed) Causal inference. Epidemiology Resources, Boston, pp 33–57
Susser M (1991) What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 133: 635–648
Weed DL (1986) On the logic of causal inference. Am J Epidemiol 123:965–979
Weiss NS (1981) Inferring causal relationships: elaboration of the criterion of "dose-response." Am J Epidemiol 113:487–490
Weiss NS (2002) Can "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology 13:6–8
Welch BL (1937) On the z-test in randomized blocks and Latin squares. Biometrika 29:21–52
Wilk MB (1955) The randomization analysis of a generalized randomized block design. Biometrika 42:70–79

Censoring Methodology

Ng Hon Keung Tony
Associate Professor
Southern Methodist University, Dallas, TX, USA

Basic Concepts on Censored Data

In industrial and clinical experiments, there are many situations in which units (or subjects) are lost or removed from experimentation before the event of interest occurs. The experimenter may not always obtain complete information on the time to the event of interest for all experimental units or subjects. Data obtained from such experiments are called *censored data*. Censoring is one of the distinguishing features of lifetime data. Censoring can be either unintentional due to accidental breakage or an individual under study drops out or intentional in which the removal of units or subjects is pre-planned, or both. Censoring restricts our ability to observe the time-to-event and it is a source of difficulty in statistical analysis.

Censoring can occur at either end (single censoring) or at both ends (double censoring). If the event of interest is only known to be occured before a certain time, it is called *left censoring*. The term "left censored" implies that the event of interest is to the left of the observed time point. The most common case of censoring is *right censoring*, in which the exact time to the event of interest is not observed and it is only known to be occured after a certain time. Different types of right censoring schemes are discussed in the subsequent section. For *interval censoring*, the event of interest is only known to be occurred in a given time interval. This type of data frequently comes from experiments where the items under test are not constantly monitored, for example, the patients in a clinical trial have periodic follow-up and events of interest occur in between two consecutive follow-ups. Note that left censoring is a special case of interval censoring where the starting time for the interval is zero.

For life-testing experiments where the event of interest is the failure of the item on test, two common reasons for pre-planned censoring are saving the total time on test and reducing the cost associated with the experiment because failure implies unit's destruction which can be costly. When budget and/or facility constraints are in place, suitable censoring scheme can be used to control the time spent and the cost of the experiment. Nevertheless, censored data usually will reduce the efficient of statistical inference compare to complete data. Therefore, it is desirable to develop censoring scheme which can balance between (i) total time spent for the experiment; (ii) number of units used in the experiment; and (iii) the efficient of statistical inference based on the results of the experiment.

Different Types of Censoring Schemes

Suppose n units are placed on a life-testing experiment. Further, suppose $X_1, X_2, \cdots, X_n$ denote the lifetimes of these n units taken from a population with lifetime distribution function $F(x;\theta)$ and density function $f(x;\theta)$, where θ is an unknown parameter(s) of interest. Let $X_{1:n} \le \cdots \le X_{n:n}$ denote the corresponding ordered lifetimes observed from the life-test. Some commonly used censoring schemes are discussed in the following.

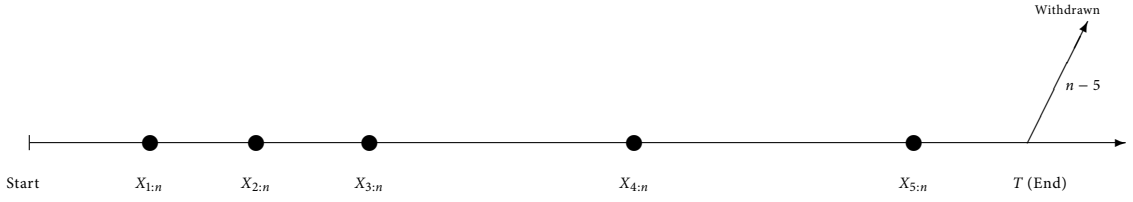
Type-I Censoring

Suppose it is planned that the life-testing experiment will be terminated at a pre-fixed time T. Then, only the failures until time T will be observed. The data obtained from such a restrained life-test will be referred to as a *Type-I censored sample*. It is also called time-censoring since the experimental time is fixed. Note that the number of failures observed here is random and, in fact, has a *Binomial*$(n, F(T;\theta))$ distribution. Figure 1 shows a schematic representation of a Type-I censored life-test with $m = 7$. Inferential procedures based on Type-I censored samples have been discussed extensively in the literature; see, for example, Cohen (1991) and Balakrishnan and Cohen (1991).

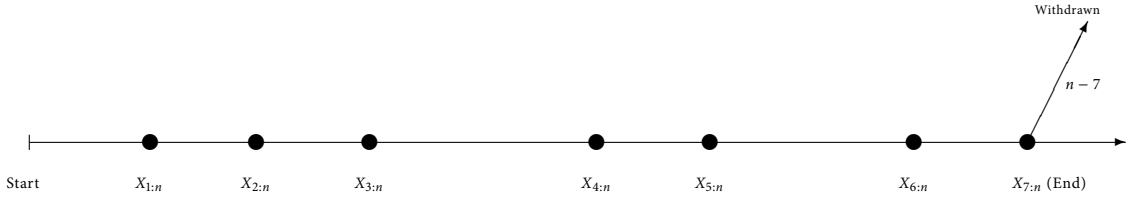
Type-I censoring scheme has the advantage that the experimental time is controlled to be at most T while it has the disadvantage that the effective sample size can turn out to be a very small number (even equal to zero) so that usual statistical inference procedures will not be applicable or they will have low efficiency.

Type-II Censoring

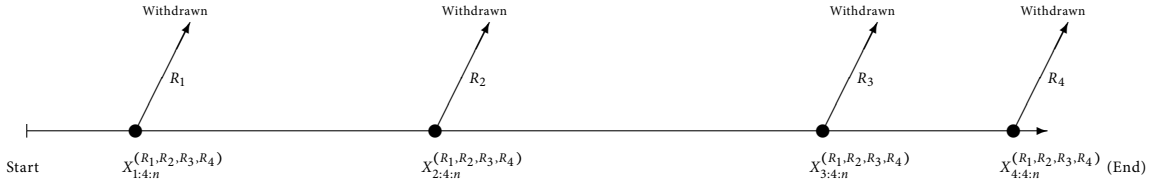
Suppose it is planned that the life-testing experiment will be terminated as soon as the mth (where m is pre-fixed) failure is observed. Then, only the first m failures out of n units under test will be observed. The data obtained from such a restrained life-test will be referred to as a *Type-II censored sample*. In contrast to Type-I censoring, the number of failures observed is fixed (viz., m) while the duration of the experiment is random (viz., $X_{m:n}$). Figure 2 shows a schematic representation of a Type-II censored life-test with m = 7. Inferential procedures based on Type-II censored samples have been discussed extensively in the



Censoring Methodology. Fig. 1 Schematic representation of a Type-I censored life-test



Censoring Methodology. Fig. 2 Schematic representation of a Type-II censored life-test



Censoring Methodology. Fig. 3 Schematic representation of a progressively Type-II censored life-test

literature; see, for example, Nelson (1982), Cohen (1991), and Balakrishnan and Cohen (1991).

Type-II censoring scheme has the advantage that the number of observed failures is fixed to be m which ensure reasonable information is available for statistical inference. However, it has the disadvantage that the experimental time is random and it can be large.

Progressive Censoring

Both the conventional Type-I and Type-II censoring schemes do not have the flexibility of allowing removal of units at points other than the terminal point of the experiment. This restricts our ability to observe extreme failures which may lead to inefficient statistical inference if we are interested in the behavior of the upper tail of the lifetime distribution. For this reason, a more general censoring scheme called *progressive censoring* has been introduced. The censored life-testing experiments described above can be extended to situations wherein censoring occurs in multiple stages. Data arising from such life-tests are referred to as *progressively censored data*. Naturally, progressive censoring can be introduced in both Type-I and Type-II forms.

For example, a progressive Type-II censored life-testing experiment will be carried out in the following manner. Prior to the experiment, a number $m < n$ is determined and the censoring scheme $(R_1, R_2, \ldots, R_m)$ with $R_j > 0$ and $\sum_{j=1}^{m} R_j + m = n$ is specified. During the experiment, j-th failure is observed and immediately after the failure, R_j functioning items are removed from the test. We denote the m completely observed (ordered) lifetimes by $X_{j:m:n}^{(R_1,R_2,\ldots,R_m)}$, $j = 1, 2, \ldots, m$, which are the observed progressively Type-II right censored sample. Figure 3 shows a schematic representation of a progressively Type-II censored life-test with $m = 4$. Notice that the conventional Type-II censoring scheme is a special case of a progressive Type-II censoring scheme when $R_i = 0$, for $i = 1, \ldots, m-1$ and $R_m = n-m$. Similarly, progressive Type-I censoring scheme can be introduced in a similar manner. Inferential procedures based on progressively Type-II censored samples have been discussed in the literature; see, for example, Balakrishnan and Aggarwala (2000) and Balakrishnan (2007) for excellent reviews on the literatures on this topic.

Hybrid Censoring

As mentioned previously, both Type-I and Type-II censoring schemes have some shortcomings. To keep away from these shortcomings, hybrid censoring schemes combining Type-I and Type-II censoring schemes have been proposed. Specifically, if the experiment is terminated at $T^* = \min\{X_{m:n}, T\}$, where m and T are pre-fixed prior to the experiment, then the censoring scheme is called *Type-I hybrid censoring scheme*; if the experiment is terminated at $T^* = \max\{X_{m:n}, T\}$, then the censoring scheme is called *Type-II hybrid censoring scheme*. We can see that both Type-I and Type-II hybrid censoring schemes try to balance between the advantages and disadvantages of conventional Type-I and Type-II censoring schemes. Hybrid censoring schemes has been studied extensively in the literature, one may refer Epstein (1954), Draper and Guttman (1987), Gupta and Kundu (1998), and Childs et al. (2003, 2008), Kundu (2007) for details. In recent years, the idea of hybrid censoring has been generalized to progressive censoring, for discussions on different types of hybrid progressive censoring schemes, see, for example, Kundu and Joarder (2006), Banerjee and Kundu (2008), Ng et al. (2009) and Lin et al. (2009).

About the Author

H. K. T. Ng is an Associate Professor in the Department of Statistical Science at Southern Methodist University. He received his Ph.D. degree in Mathematics (2002) from McMaster University, Hamilton, Canada. He is an elected member of International Statistical Institute (ISI) and an elected senior member of Institute of Electrical and Electronics Engineers (IEEE). He is currently an Associate Editor of Communications in Statistics.

Cross References

►Astrostatistics
►Event History Analysis

►Nonparametric Estimation Based on Incomplete Observations
►Ordered Statistical Data: Recent Developments
►Step-Stress Accelerated Life Tests
►Survival Data

References and Further Reading

Balakrishnan N (2007) Progressive censoring methodology: an appraisal (with discussions). Test 16:211–296
Balakrishnan N, Aggarwala R (2000) Progressive censoring: theory, methods and applications. Birkhäuser, Boston
Balakrishnan N, Cohen AC (1991) Order statistics and inference: estimation methods. Academic, San Diego
Banerjee A, Kundu D (2008) Inference based on Type-II hybrid censored data from a Weibull distribution. IEEE Trans Reliab 57:369–378
Childs A, Chandrasekar B, Balakrishnan N, Kundu D (2003) Exact likelihood inference based on Type-I and Type-II hybrid censored samples from the exponential distribution. Ann Inst Stat Math 55:319–330
Childs A, Chandrasekar B, Balakrishnan N (2008) Exact likelihood inference for an exponential parameter under progressive hybrid censoring schemes. In: Vonta F, Nikulin M, Limnios N, Huber-Carol C (eds) Statistical models and methods for biomedical and technical systems. Birkhäuser, Boston, pp 323–334
Cohen AC (1991) Truncated and censored samples: theory and applications. Marcel Dekker, New York
Draper N, Guttman I (1987) Bayesian analysis of hybrid life tests with exponential failure times. Ann Inst Stat Math 39:219–225
Epstein B (1954) Truncated life tests in the exponential case. Ann Math Stat 25:555–564
Gupta RD, Kundu D (1998) Hybrid censoring schemes with exponential failure distribution. Commun Stat Theory Meth 27:3065–3083
Kundu D (2007) On hybrid censoring Weibull distribution. J Stat Plan Infer 137:2127–2142
Kundu D, Joarder A (2006) Analysis of Type-II progressively hybrid censored data. Comput Stat Data Anal 50:2509–2528
Nelson W (1982) Applied life data analysis. Wiley, New York
Ng HKT, Kundu D, Chan PS (2009). Statistical analysis of exponential lifetimes under an adaptive Type-II progressive censoring scheme, Naval Research Logistics 56:687–698

Census

Margo J. Anderson
Professor of History and Urban Studies
University of Wisconsin–Milwaukee, Milwaukee, WI, USA

Introduction

A census usually refers to a complete count by a national government of the population, with the population further defined by demographic, social or economic characteristics, for example, age, sex, ethnic background, marital status, and income. National governments also conduct other types of censuses, particularly of economic activity. An economic census collects information on the number and characteristics of farms, factories, mines, or businesses.

Most countries of the world conduct population censuses at regular intervals. By comparing the results of successive censuses, analysts can see whether the population is growing, stable, or declining, both in the country as a whole and in particular geographic regions. They can also identify general trends in the characteristics of the population. Because censuses aim to count the entire population of a country, they are very expensive and elaborate administrative operations and thus are conducted relatively infrequently. The United States and the United Kingdom, for example, conduct a population census every 10 years (a *decennial* census), and Canada conducts one every 5 years (a *quinquennial* census). Economic censuses are generally conducted on a different schedule from the population census.

Censuses of population usually try to count everyone in the country as of a fixed date, often known as Census Day. Generally, governments collect the information by sending a ►questionnaire in the mail or a census taker to every household or residential address in the country. The recipients are instructed to complete the questionnaire and send it back to the government, which processes the answers. Trained interviewers visit households that do not respond to the questionnaire and individuals without mail service, such as the homeless or those living in remote areas.

History

Censuses have been taken since ancient times by emperors and kings trying to assess the size and strength of their realms. These early censuses were conducted sporadically, generally to levy taxes or for military conscription. Clay tablet fragments from ancient Babylon indicate that a census was taken there as early as 3800 BCE to estimate forthcoming tax revenues. The ancient Chinese, Hebrews, Egyptians, and Greeks also conducted censuses. However, enumerations did not take place at regular intervals until the Romans began to count of the population in the Republic and later the empire. Among the Romans the census was usually a count of the male population and assessment of property value. It was used mainly for drafting men into military service and for taxing property.

After the fall of the Roman Empire in the fifth century CE, census taking disappeared for several hundred years in the West. The small feudal communities of the Middle

Ages had neither the mechanisms nor the need for censuses. However, in 1086 William the Conqueror ordered the compilation of the census-like Domesday Book, a record of English landowners and their holdings. From the data given in this survey, which was made to determine revenues due to the king, historians have reconstructed the social and economic conditions of the times.

The modern census dates from the seventeenth century, when European powers wanted to determine the success of their overseas colonies. Thus the British crown and the British Board of Trade ordered repeated counts of the colonial American population in the seventeenth and eighteenth centuries, starting in the 1620s in Virginia. The first true census in modern times was taken in New France, France's North American empire, beginning in 1665. The rise of democratic governments resulted in a new feature of the census process: The 1790 census of the United States was the first to have its Constitution require a census and periodic reapportionment of its House of Representatives on the basis of the decennial census results. Sweden began to conduct censuses in the mid-eighteenth century, and England and Wales instituted a regular decennial census in 1801. During the nineteenth century and the first half of the twentieth century, the practice of census taking spread throughout the world. India conducted its first national census in 1871, under British rule. China's first modern census, in 1953, counted 583 million people.

The United Nations encourages all countries to conduct a population count through a census or population registration system. It also promotes adoption of uniform standards and census procedures. The United Nations Statistical Office compiles reports on worldwide population.

Uses of Census Information

Governments use census information in almost all aspects of public policy. In some countries, the population census is used to determine the number of representatives each area within the country is legally entitled to elect to the national legislature. The Constitution of the United States, for example, provides that seats in the House of Representatives should be apportioned to the states according to the number of their inhabitants. Each decade, Congress uses the population count to determine how many seats each state should have in the House and in the electoral college, the body that nominally elects the president and vice president of the United States. This process is known as *reapportionment.* States frequently use population census figures as a basis for allocating delegates to the state legislatures and for redrawing district boundaries for seats in the House, in state legislatures, and in local legislative districts. In Canada, census population data are similarly used to apportion seats among the provinces and territories in the House of Commons and to draw electoral districts.

Governments at all levels – such as cities, counties, provinces, and states – find population census information of great value in planning public services because the census tells how many people of each age live in different areas. These governments use census data to determine how many children an educational system must serve, to allocate funds for public buildings such as schools and libraries, and to plan public transportation systems. They can also determine the best locations for new roads, bridges, police departments, fire departments, and services for the elderly.

Besides governments, many others use census data. Private businesses analyze population and economic census data to determine where to locate new factories, shopping malls, or banks; to decide where to advertise particular products; or to compare their own production or sales against the rest of their industry. Community organizations use census information to develop social service programs and child-care centers. Censuses make a huge variety of general statistical information about society available to researchers, journalists, educators, and the general public.

Conducting a Census

Most nations create a permanent national statistical agency to take the census. In the United States, the Bureau of the Census (Census Bureau), an agency of the Department of Commerce, conducts the national population census and most economic censuses. In Canada, the Census Division of Statistics Canada is responsible for taking censuses.

Conducting a census involves four major stages. First, the census agency plans for the census and determines what information it will collect. Next, it collects the information by mailing questionnaires and conducting personal interviews. Then the agency processes and analyzes the data. Finally, the agency publishes the results to make them available to the public and other government agencies.

Planning the Census

Census agencies must begin planning for a census years in advance. One of the most important tasks is to determine what questions will appear on the census questionnaire. Census agencies usually undertake a lengthy public review process to determine the questions to be asked. They conduct public meetings, consider letters and requests from the general public, and consult with other government agencies and special advisory committees. In the United States, census questions must be approved by Congress and

the Office of Management and Budget. In Canada, questions must be approved by the governor-general on the recommendations of the Cabinet.

The questions included on census forms vary from nation to nation depending on the country's particular political and social history and current conditions. Most censuses request basic demographic information, such as the person's name, age, sex, educational background, occupation, and marital status. Many censuses also include questions about a person's race, ethnic or national origin, and religion. Further questions may ask the person's place of birth; relationship to the head of the household; citizenship status; the individual's or the family's income; the type of dwelling the household occupies; and the language spoken in the household.

Questions that are routine in one nation may be seen as quite controversial in another, depending on the history of the country. The United States census does not ask about religious affiliation because such a question is considered a violation of the First Amendment right to freedom of religion or an invasion of privacy. Other nations, such as India, do collect such information. Questions on the number of children born to a woman were quite controversial in China in recent years because of government efforts to limit families to having only one child. In the United States, asking a question on income was considered controversial in 1940 when it was first asked. It is no longer considered as objectionable. Questions change in response to public debate about the state of society. For example, Americans wanted to know which households had radios in 1930, and the census introduced questions on housing quality in 1940. Canadians have recently begun to ask census questions on disability status and on the unpaid work done in the home.

Besides determining the content of the census, census agencies must make many other preparations. Staffing is a major concern for census agencies because censuses in most countries require a huge number of temporary workers to collect and process data. Consequently, census agencies must begin recruiting and training workers months or years in advance. For example, the U.S. Census Bureau had to fill 850,000 temporary, short-term positions to conduct the 2000 census. In order to hire and retain enough staff, it had to recruit nearly three million job applicants. The majority of temporary workers are hired to go door-to-door to interview households that do not respond to the census questionnaire. In some countries, government employees at a local level, such as schoolteachers, are asked to help conduct the count.

Prior to any census, a census agency must develop an accurate list of addresses and maps to ensure that everyone is counted. The U.S. Census Bureau obtains addresses primarily from the United States Postal Service and from previous census address lists. It also works closely with state, local, and tribal governments to compile accurate lists. Finally, census agencies often conduct an extensive marketing campaign before Census Day to remind the general population about the importance of responding to the census. This campaign may involve paid advertising, distributing materials by direct mail, promotional events, and encouraging media coverage of the census.

Collecting the Information

Until relatively recently, population censuses were taken exclusively through personal interviews. The government sent *enumerators* (interviewers) to each household in the country. The enumerators asked the head of the household questions about each member of the household and entered the person's responses on the census questionnaire. The enumerator then returned the responses to the government. Today, many censuses are conducted primarily through *self-enumeration,* which means that people complete their own census questionnaire. Self-enumeration reduces the cost of a census to the government because fewer enumerators are needed to conduct interviews. In addition, the procedure provides greater privacy to the public and generally improves the accuracy of responses, because household members can take more time to think over the questions and consult their personal records.

Nevertheless, census operations still require hiring very large numbers of temporary enumerators to conduct address canvassing in advance of a mail census and to retrieve forms from non responding households and check on vacant units. Other nations continue to conduct censuses partially or totally through direct enumeration. Some, such as Turkey, require people to stay home on Census Day to await the census taker.

Census agencies make a special effort to count people who may not receive a questionnaire by mail or who have no permanent address. For example, the U.S. Census Bureau sends census takers to interview people at homeless shelters, soup kitchens, mobile food vans, campgrounds, fairs, and carnivals. It consults with experts to find migrant and seasonal farmworkers. Finally, the agency distributes census questionnaires to people living in group quarters, such as college dormitories, nursing homes, hospitals, prisons and jails, halfway houses, youth hostels, convents and monasteries, and women's shelters.

The level of detail on the complete count census varies by country, particularly after the development of probability survey techniques in the 1940s. In the United States, for example, until the 2010 census, most households received a "short form," a brief set of questions on basic

characteristics such as name, age, sex, racial or ethnic background, marital status, and relationship to the household head. But from the mid-twentieth century until 2000, a smaller sample of households received the "long form," with many additional detailed questions. These included questions about the individual's educational background, income, occupation, language knowledge, veteran status, and disability status as well as housing-related questions about the value of the individual's home, the number of rooms and bedrooms in it, and the year the structure was built. These "long form" questions have been collected in the American Community Survey since the early 2000s, and thus are no longer asked on the U.S. Census in 2010.

Processing and Analysis of Data

For most of the 19th century in the United States and Canada, census data were tabulated and compiled by hand, without the aid of machines. Manual processing was very slow, and some figures were obsolete by the time they were published. The invention of mechanical tabulating devices in the late nineteenth century made processing of the data much faster and improved the accuracy of the results. For example, in 2010, the U.S. Census Bureau will scan the data from 100 + million paper questionnaires, and capture the responses using optical character recognition software. Once in electronic form, the data can be analyzed and turned into statistics. Unreadable or ambiguous responses are checked by census clerks and manually keyed into the computer.

Publication of Results

U.S. and Canadian censuses publish only general statistical information and keep individual responses confidential. By law, the U.S. Census Bureau and Statistics Canada are prohibited from releasing individual responses to any other government agency or to any individual or business. Census workers in both countries must swear under oath that they will keep individual responses confidential. Employees who violate this policy face a monetary fine and possible prison term. If an individual's personal data were not kept confidential, people might refuse to participate in the census for fear that their personal information would be made public or used by the government to track their activities. In the United States, individual census responses are stored at the National Archives. After 72 years, the original forms are declassified and opened to the public. These original responses are frequently used by people researching the history of their families or constructing genealogies. In Canada, census responses from 1906 and later are stored at Statistics Canada. Microfilmed records of census responses from 1911 and earlier are stored at the National Archives of Canada; these are the only individual census responses currently available for public use.

Until the 1980s, census agencies published their results in large volumes of numeric tables – sometimes numbering in the hundreds of volumes. Today, the majority of census data is distributed electronically, both in tabulated form, and through anonymized public use microdata samples.

Problems in Census Taking and Issues for the Future

Censuses provide important information about the population of a country. But they can become embroiled in political or social controversy simply by reporting information. Complaints about the census generally involve concerns about the accuracy of the count, the propriety of particular questions, and the uses to which the data are put.

All censuses contain errors of various kinds. Some people and addresses are missed. People may misunderstand a question or fail to answer all the questions. Census officials have developed elaborate procedures to catch and correct errors as the data are collected, but some errors remain. For example, the 1990 U.S. census missed 8.4 million people and mistakenly counted 4.4 million people, according to Census Bureau estimates. The latter figure included people counted more than once, fictitious people listed on forms, and fabrications by enumerators. Such errors undermine the credibility of the census as a mechanism for allocating seats in legislative bodies and government funds.

In recent years, developments in statistical analysis have made it possible to measure the accuracy of censuses. Census results may be compared with population information from other sources, such as the records of births, deaths, and marriages in vital statistics. Census officials can also determine the level of accuracy of the count by conducting a second, sample count called a *post-enumeration survey* or *post-censal survey*. In this technique, census staff knock on the door of each housing unit in selected blocks around the country, regardless of whether the housing unit was on the master address list. The staff member determines whether the household was counted in the census. By comparing the results from this survey with the census records, census officials can estimate how many people from each geographic region were missed in the original census count. Some nations, such as Canada and Australia, have begun to adjust the census results for omissions and other errors.

Concerns about the confidentiality of the census represent another source of data error. Censuses require public understanding, support, and cooperation to be successful. Concerns about government interference with private life

can prevent people from cooperating with what is essentially a voluntary counting process. People may be suspicious of giving information to a government agency or may object that particular census questions invade their privacy. When public trust is lacking, people may not participate. In some nations, past census controversies have led to the elimination of the national census. During World War II (1939–1945), for example, the German Nazi forces occupying The Netherlands used the country's census records and population registration data to identify Jews for detention, removal, and extermination. This use ultimately undermined the legitimacy of the census after World War II. In The Netherlands, the legacy of the Nazi era was one of the major justifications to end census taking. The Netherlands took its last regular census in 1971 and now collects population information through other mechanisms.

Many nations are currently exploring alternatives to or major modernizations of the traditional population census. France, for example, has recently implemented a continuous measurement population counting system. The United States is exploring the use of administrative records and electronic methods of data collection to replace the mail enumeration in 2020.

About the Author

Dr. Margo J. Anderson is Professor of History and Urban Studies, University of Wisconsin–Milwaukee. She specializes in the social history of the United States in the nineteenth and twentieth centuries. She was Chair, History Department (1992–1995). She was a member of the National Academy of Sciences' Panel on Census Requirements for the Year 2000 and Beyond. Dr. Anderson was Vice President (2005), and President, Social Science History Association (2006). She is a Fellow, American Statistical Association (ASA) (1998), and was Chair, Social Statistics Section of ASA (1998). Currently, she is ASA Chair, Committee on Committees. She has authored and coauthored numerous papers and several books including, *The American Census: A Social History* (New Haven: Yale University Press, 1988), and *Who Counts? The Politics of Census-Taking in Contemporary America* (with Stephen E. Fienberg, Russell Sage, 1999, revised and updated 2001), named as one of Choice Magazine's Outstanding Academic Books of 2000. She was Editor-in-Chief of the *Encyclopedia of the U.S. Census* (Washington, D.C.: CQ Press, 2000). Professor Anderson is widely regarded as the leading scholar on the history of the U.S. census, and has made distinguished contributions to research in American social science.

Cross References

►African Population Censuses
►Demography
►Economic Statistics
►Federal Statistics in the United States, Some Challenges
►Population Projections
►Sample Survey Methods
►Simple Random Sample
►Small Area Estimation
►Statistical Publications, History of

References and Further Reading

Anderson M (1988) The American census: a social history. Yale University Press, New Haven

Anderson M, Fienberg SE (2001) Who counts? The politics of census taking in contemporary America, rev edn. Russell Sage Foundation, New York

Desrosieres A. La Politique des Grands Nombres: Histoire de la Raison Statistique. Edition La Découverte, Paris (1998, The politics of large numbers: a history of statistical reasoning. Harvard University Press, Cambridge)

Minnesota Population Center (2010) Integrated public use microdata series. http://ipums.org.

U.K. Office of National Statistics (2001) 200 Years of the Census. http://www.statistics.gov.uk/census2001/bicentenary/pdfs/200years.pdf

U.N. Statistics Division (2010) 2010 World Population and Housing Census Programme. http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/more.htm

Ventresca M (1996) When states count: institutional and political dynamics in modern census establishment, 1800–1993. PhD dissertation, Stanford University, Stanford

Worton DA (1997) The Dominion bureau of statistics: a history of Canada's central statistics office and its antecedents: 1841–1972. McGill-Queens University Press, Kingston

Central Limit Theorems

Julio M. Singer
Professor, Head
Universidade de São Paulo, São Paulo, Brazil

Introduction

One of the objectives of statistical inference is to draw conclusions about some parameter, like the mean or the variance of a (possibly conceptual) population of interest based

on the information obtained in a sample conveniently selected therefrom. For practical purposes, estimates of these parameters must be coupled with statistical properties and except in the most simple cases, exact properties are difficult to obtain and one must rely on approximations. It is quite natural to expect estimators to be consistent, but it is even more important that their (usually mathematically complex) exact sampling distribution be adequately approximated by a simpler one, such as the normal or the χ^2 distribution, for which tables or computational algorithms are available. Here we are not concerned with the convergence of the actual sequence of statistics $\{T_n\}$ to some constant or random variable T as $n \to \infty$, but with the convergence of the corresponding distribution functions $\{G_n\}$ to some specific distribution function F. This is known as weak convergence and for simplicity, we write $T_n \xrightarrow{D} F$. Although this is the weakest mode of stochastic convergence, it is very important for statistical applications, since the related limiting distribution function F may generally be employed in the construction of approximate confidence intervals for and significance tests about the parameters of interest. In this context, central limit theorems (CLT) are used to show that statistics expressed as sums of the underlying random variables, conveniently standardized, are asymptotically normally distributed, i.e., converge weakly to the normal distribution. They may be proved under different assumptions regarding the original distributions.

The simplest CLT states that the (sampling) distribution of the sample mean of independent and identically distributed (*i.i.d*) random variables with finite second moments may be approximated by a normal distribution. Although the limiting distribution is continuous, the underlying distribution may even be discrete. CLT are also available for independent, but not identically distributed (e.g., with different means and variances) underlying random variables, provided some (relatively mild) assumptions hold for their moments. The Liapounov CLT and the Lindeberg-Feller CLT are useful examples. Further extensions cover cases of dependent random underlying variables; in particular, the Hájek-Šidak CLT is extremely useful in regression analysis, where as the sample size increases, the response variables form a triangular array in which for each row (i.e., for given n), they are independent but this is not true among rows (i.e., for different values of n). Extensions to cover cases where the underlying random variables have more complex (e.g., martingale-type) dependence structures are also available. When dealing with partial sum or empirical distributional processes, we must go beyond the finite-dimensional case and assume some *compactness* conditions to obtain suitable results, wherein the so-called *weak invariance principles* play an important role.

Different Versions of the Central Limit Theorem

We now present (without proofs) the most commonly used versions of the CLT. Details and a list of related references may be obtained in Sen et al. (2010).

Theorem 1 *(Classical CLT) Let* $\{X_k, k \geq 1\}$ *be a sequence of i.i.d. random variables such that*

1. $\mathbb{E}(X_k) = \mu$.
2. $\mathbb{V}ar(X_k) = \sigma^2 < \infty$.

Also, let $Z_n = (T_n - n\mu)/(\sigma\sqrt{n})$ *where* $T_n = \sum_{k=1}^n X_k$. *Then,* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.

In practice, this result implies that for large n, the distribution of the sample mean $\overline{X}_n = T_n/n$ may be approximated by a normal distribution with mean μ and variance σ^2/n. An interesting special case occurs when the underlying variables X_k have Bernoulli distributions with probability of success π. Here the expected value and the variance of X_k are π and $\pi(1-\pi)$, respectively. It follows that the large-sample distribution of the sample proportion, $p_n = T_n/n$ may be approximated by a $\mathcal{N}[\pi, \pi(1-\pi)/n]$ distribution. This result is known as the *De Moivre–Laplace CLT*.

An extension of Theorem 1 to cover the case of sums of independent, but not identically distributed random variables requires additional assumptions on the moments of the underling distributions. In this direction, we consider the following result.

Theorem 2 *(Liapounov CLT) Let* $\{X_k, k \geq 1\}$ *be a sequence of independent random variables such that*

1. $\mathbb{E}(X_k) = \mu_k$.
2. $\nu_{2+\delta}^{(k)} = \mathbb{E}(|X_k - \mu_k|^{2+\delta}) < \infty, \quad k \geq 1$ *for some* $0 < \delta \leq 1$.

Also let $T_n = \sum_{k=1}^n X_k$, $\mathbb{V}ar(X_k) = \sigma_k^2$, $\tau_n^2 = \sum_{k=1}^n \sigma_k^2$, $Z_n = (T_n - \sum_{k=1}^n \mu_k)/\tau_n$ *and* $\rho_n = \tau_n^{-(2+\delta)} \sum_{k=1}^n \nu_{2+\delta}^{(k)}$. *Then, if* $\lim_{n\to\infty} \rho_n = 0$, *it follows that* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.

This as well as other versions of the CLT may also be extended to the multivariate case by referring to the *Cramér-Wold Theorem*, which essentially states that the asymptotic distribution of the multivariate statistic under

investigation may be obtained by showing that every linear combination of its components follows an asymptotic normal distribution. Given a sequence $\{\boldsymbol{X}_n, n \geq 1\}$ of random vectors in $\mathbb{R}^p$, with mean vectors $\boldsymbol{\mu}_n$ and covariance matrices $\boldsymbol{\Sigma}_n$, $n \geq 1$, to show that $n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$, one generally proceeds according to the following strategy:

1. Use one of the univariate CLT to show that for every fixed $\boldsymbol{\lambda} \in \mathbb{R}^p$, $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}(0, \gamma^2)$ with $\gamma^2 = \lim_{n\to\infty} n^{-1}\boldsymbol{\lambda}'\left(\sum_{i=1}^{n}\boldsymbol{\Sigma}_i\right)\boldsymbol{\lambda}$.
2. Use the Cramér-Wold Theorem to complete the proof.

As an example we have:

Theorem 3 *(Multivariate version of the Liapounov CLT) Let $\{\boldsymbol{X}_n, n \geq 1\}$ be a sequence of random vectors in $\mathbb{R}^p$ with mean vectors $\boldsymbol{\mu}_n$ and finite covariance matrices $\boldsymbol{\Sigma}_n, n \geq 1$, such that $\max_{1\leq i\leq n}\max_{1\leq j\leq p}\mathbb{E}(|X_{ij} - \mu_{ij}|^{2+\delta}) < \infty$ for some $0 < \delta < 1$, and $\boldsymbol{\Sigma} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$ exists. Then $n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$.*

In the original formulation, Liapounov used $\delta = 1$, but even the existence of $\nu_{2+\delta}^{(k)}, 0 < \delta \leq 1$ is not a necessary condition, as we may see from the following theorem.

Theorem 4 *(Lindeberg-Feller CLT) Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables satisfying*

1. $\mathbb{E}(X_k) = \mu_k$.
2. $\mathbb{V}ar(X_k) = \sigma_k^2 < \infty$.

Also, let $T_n = \sum_{k=1}^{n}X_k$, $\tau_n^2 = \sum_{k=1}^{n}\sigma_k^2$ and $Z_n = \sum_{k=1}^{n}Y_{nk}$ where $Y_{nk} = (X_k - \mu_k)/\tau_n$ and consider the following additional conditions:

1. *Uniform asymptotic negligibility (UAN):* $\max_{1\leq k\leq n}\left(\sigma_k^2/\tau_n^2\right) \to 0$ *as* $n \to \infty$.
2. *Asymptotic normality:* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.
3. *Lindeberg-Feller (uniform integrability):*

$$\forall \varepsilon > 0, \frac{1}{\tau_n^2}\sum_{k=1}^{n}\mathbb{E}\left[(X_k - \mu_k)^2 I(|X_k - \mu_k| > \varepsilon\tau_n)\right] \to 0 \text{ as } n \to \infty,$$

where $I(A)$ denotes the indicator function.

Then, (A) and (B) hold simultaneously if and only if (C) holds.

Condition (A) implies that the random variables Y_{nk} are *infinitesimal*, i.e., that $\max_{1\leq k\leq n}P(|Y_{nk}| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$, or, in other words, that the random variables $Y_{nk}, 1 \leq k \leq n$, are uniformly in k, asymptotically in n, negligible.

When the underlying random variables under consideration are bounded, i.e., when $P(a \leq X_k \leq b) = 1$ for some finite scalars $a < b$, it follows that a necessary and sufficient condition for $Z_n \xrightarrow{D} \mathcal{N}(0,1)$ is that $\tau_n \to \infty$ as $n \to \infty$.

Up to this point we have devoted attention to the weak convergence of sequences of statistics $\{T_n, n \geq 1\}$ constructed from independent underlying random variables $X_1, X_2, \ldots$. We consider now some extensions of the CLT where such restriction may be relaxed. The first of such extensions holds for sequences of (possibly dependent) random variables which may be structured as a *double array* of the form

$$\begin{pmatrix} X_{11}, & X_{12}, & \cdots, & X_{1k_1} \\ X_{21}, & X_{22}, & \cdots, & X_{2k_2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}, & X_{n2}, & \cdots, & X_{nk_n} \end{pmatrix}$$

where the X_{nk} are row-wise independent. The case where $k_n = n$, $n \geq 1$, is usually termed a *triangular array* of random variables. This result is very useful in the field of ▶order statistics.

Theorem 5 *(Double array CLT) Let the random variables $\{Y_{nk}, 1 \leq k \leq k_n, n \geq 1\}$ where $k_n \to \infty$ as $n \to \infty$ be such that for each n, $\{Y_{nk}, 1 \leq k \leq k_n\}$ are independent. Then*

1. *$\{Y_{nk}, 1 \leq k \leq k_n, n \geq 1\}$ is an infinitesimal system of random variables, i.e., satisfies the UAN condition.*
2. *$Z_n = \sum_{k=1}^{k_n}Y_{nk} \xrightarrow{D} \mathcal{N}(0,1)$.*

hold simultaneously, if and only if, for every $\varepsilon > 0$, as $n \to \infty$ the following two conditions hold

1. $\sum_{k=1}^{k_n}P(|Y_{nk}| > \varepsilon) \to 0$.
2. $$\sum_{k=1}^{k_n}\left\{\int_{\{|y|\leq\varepsilon\}} y^2 dP(Y_{nk} \leq y) - \left[\int_{\{|y|\leq\varepsilon\}} y dP(Y_{nk} \leq x)\right]^2\right\} \to 1.$$

Linear regression and related models pose special problems since the underlying random variables are not identically distributed and in many cases, the exact functional form of their distributions is not completely specified. Least-squares methods (see ▶Least Squares) are attractive under these conditions, since they may be employed in a rather general setup. In this context, the following CLT specifies sufficient conditions on the explanatory variables such that the distributions of the least squares estimators of the regression parameters may be approximated by normal distributions.

Theorem 6 *(Hájek-Šidak CLT) Let* $\{Y_n, n \geq 1\}$ *be a sequence of i.i.d. random variables with mean* μ *and finite variance* σ^2*; let* $\{\boldsymbol{x}_n, n \geq 1\}$ *be a sequence of real vectors* $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nn})'$*. Then if Noether's condition holds, i.e., if*

$$\max_{1 \leq i \leq n} \left[x_{ni}^2 \Big/ \sum_{i=1}^{n} x_{ni}^2 \right] \to 0 \quad as \quad n \to \infty,$$

holds, it follows that

$$Z_n = \left[\sum_{i=1}^{n} x_{ni}(Y_{ni} - \mu) \right] \Big/ \left[\sigma^2 \sum_{i=1}^{n} x_{ni}^2 \right]^{1/2} \xrightarrow{D} \mathcal{N}(0,1).$$

As an illustration, consider the simple linear regression model (see ▶Simple Linear Regression)

$$y_{ni} = \alpha + \beta x_{ni} + e_{ni}, \; i = 1, \ldots, n,$$

where y_{ni} and x_{ni} represent observations of the response and explanatory variables, respectively, α and β are the parameters of interest and the e_{ni} correspond to uncorrelated random errors with mean 0 and variance σ^2. The least squares estimators of β and α are respectively $\widehat{\beta}_n = \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)(y_{ni} - \bar{y}_n) / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$ and $\widehat{\alpha}_n = \bar{y}_n - \widehat{\beta}_n \bar{x}_n$ where $\bar{x}_n$ and $\bar{y}_n$ correspond to the sample means of the explanatory and response variables. Irrespectively of the form of underlying distribution of e_{ni}, we may use standard results to show that $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are unbiased and have variances given by $\sigma^2 \left[\sum_{i=1}^{n} x_{ni}^2 / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2 \right]$ and $\sigma^2 \left[\sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2 \right]^{-1}$, respectively. Furthermore, the covariance between $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ is $-\sigma^2 \bar{x}_n / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$. When the underlying distribution of e_{ni} is normal, we may show that $(\widehat{\alpha}_n, \widehat{\beta}_n)$ follows a bivariate normal distribution. If Noether's condition holds and both $\bar{x}_n$ and $n^{-1} \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$ converge to finite constants as $n \to \infty$, we may use the Hájek-Šidak CLT and the Cramér-Wold Theorem to conclude that the same bivariate normal distribution specified above serves as an approximation of the true distribution of $(\widehat{\alpha}_n, \widehat{\beta}_n)$, whatever the form of the distribution of e_{ni}, provided that n is sufficiently large.

The results may also be generalized to cover alternative estimators obtained by means of generalized and weighted least-squares procedures as well as via robust M-estimation procedures. They may also be extended to generalized linear and nonlinear models. Details may be obtained in Sen et al. (2010), for example.

It is still possible to relax further the independence assumption on the underlying random variables. The following theorems constitute examples of CLT for dependent random variables having a martingale (or reverse martingale) structure. For further details, the reader is referred to Loynes (1970), Brown (1971), Dvoretzky (1971), or McLeish (1974).

Theorem 7 *(Martingale CLT) Consider a sequence* $\{X_k, k \geq 1\}$ *of random variables satisfying*

1. $\mathbb{E}(X_k) = 0$.
2. $\mathbb{E}\left(X_k^2\right) = \sigma_k^2 < \infty$.
3. $\mathbb{E}\{X_k | X_1, \ldots, X_{k-1}\} = 0$, $X_0 = 0$.

Also let $T_n = \sum_{k=1}^{n} X_k$, $\tau_n^2 = \sum_{k=1}^{n} \sigma_k^2$, $v_k^2 = \mathbb{E}\left(X_k^2 | X_1, \ldots, X_{k-1}\right)$ *and* $w_n^2 = \sum_{k=1}^{n} v_k^2$*. If*

1. $w_n^2 / \tau_n^2 \xrightarrow{P} 1$ *as* $n \to \infty$.
2. $\forall \varepsilon > 0$, $\tau_n^{-2} \sum_{k=1}^{n} \mathbb{E}\left[X_k^2 I(|X_k| > \varepsilon \tau_n) \right] \to 0$ *as* $n \to \infty$ *(Lindeberg-Feller condition),*

then the sequence $\{X_k, k \geq 1\}$ *is infinitesimal and* $Z_n = T_n / \tau_n \xrightarrow{D} \mathcal{N}(0,1)$.

Note that the terms v_k^2 are random variables since they depend on $X_1, \ldots, X_{k-1}$; condition (A) essentially states that all the information about the variability in the X_k is contained in $X_1, \ldots, X_{k-1}$. Also note that $\{T_n, n \geq 1\}$ is a *zero mean martingale* (See also ▶Martingale Central Limit Theorem.)

Theorem 8 *(Reverse Martingale CLT) Consider a sequence* $\{T_k, k \geq 1\}$ *of random variables such that*

$$\mathbb{E}(T_n | T_{n+1}, T_{n+2}, \ldots) = T_{n+1} \quad and \quad \mathbb{E}(T_n) = 0,$$

i.e., $\{T_k, k \geq 1\}$ *is a zero mean reverse martingale. Assume that* $\mathbb{E}\left(T_n^2\right) < \infty$ *and let* $Y_k = T_k - T_{k+1}, k \geq 1$, $v_k^2 = \mathbb{E}\left(Y_k^2 | T_{k+1}, T_{k+2}, \ldots\right)$ *and* $w_n^2 = \sum_{k=n}^{\infty} v_k^2$*. If*

1. $w_n^2 / \mathbb{E}\left(w_n^2\right) \xrightarrow{a.s.} 1$.
2. $w_n^{-2} \sum_{k=n}^{\infty} \mathbb{E}\left[Y_k^2 I(|Y_k| > \varepsilon w_n) \middle| T_{k+1}, T_{k+2}, \ldots \right] \xrightarrow{P} 0$, $\varepsilon > 0$ *or* $w_n^{-2} \sum_{k=n}^{\infty} Y_k^2 \xrightarrow{a.s.} 1$,

it follows that $T_n / \sqrt{\mathbb{E}\left(w_n^2\right)} \xrightarrow{D} \mathcal{N}(0,1)$.

Rates of Convergence to Normality

In the general context discussed above, a question of both theoretical and practical interest concerns the speed with which the convergence to the limiting normal distribution takes place. Although there are no simple answers to this question, the following result may be useful.

Theorem 9 *(Berry-Esséen) Let* $\{X_n, n \geq 1\}$ *be a sequence of i.i.d. random variables with* $\mathbb{E}(X_1) = \mu$, $\mathbb{V}ar(X_1) = \sigma^2$ *and suppose that* $\mathbb{E}(|X_1 - \mu|^{2+\delta}) = \nu_{2+\delta} < \infty$ *for*

some $0 < \delta \leq 1$. *Also let* $T_n = \sum_{i=1}^{n} X_i$ *and* $F_{(n)}(x) = P\left[(T_n - n\mu)/\left(\sigma\sqrt{n}\right) \leq x\right]$, $x \in \mathbb{R}$. *Then there exist a constant* C *such that*

$$\Delta_n = \sup_{x\in\mathbb{R}} |F_{(n)}(x) - \Phi(x)| \leq C\frac{\nu_{2+\delta} n^{-\delta/2}}{\sigma^{2+\delta}}$$

where Φ *denotes the standard normal distribution function.*

The reader is referred to Feller (1971) for details. Berry (1941) proved the result for $\delta = 1$ and Esséen (1956) showed that $C \geq 0.4097$. Although the exact value of the constant C is not known, many authors have proposed upper bounds. In particular, van Beeck (1972) showed that $C \leq 0.7975$ and more recently, Shevtsova (2007) concluded that $C \leq 0.7056$. The usefulness of the theorem, however, is limited, since the rates of convergence attained are not very sharp.

Alternatively, the rates of convergence of the sequence of distribution functions $F_{(n)}$ to Φ or of the density functions $f_{(n)}$ (when they exist) to φ (the density function of the standard normal distribution) may be assessed by *Gram-Charlier* or *Edgeworth expansions* as discussed in Cramér (1946), for example. Although this second approach might offer a better insight to the problem of evaluating the rate of convergence to normality than that provided by the former, it requires the knowledge of the moments of the parent distribution and, thus, is less useful in practical applications.

Convergence of Moments

Given that weak convergence has been established, a question of interest is whether the moments (e.g., mean and variance) of the statistics under investigation converge to the moments of the limiting distribution. Although the answer is negative in general, an important theorem, due to Cramér, indicates conditions under which the result is true. The reader is referred to Sen et al. (2010) for details.

Asymptotic Distributions of Statistics not Expressible as Sums of Random Variables

The *Slustky theorem* is a handy tool to prove weak convergence of statistics that may be expressed as the sum, product or ratio of two terms, the first known to converge weakly to some distribution and the second known to converge in probability to some constant. As an example, consider independent and identically distributed random variables $Y_1, \ldots, Y_n$ with mean μ and variance σ^2. Since the corresponding sample standard deviation S converges in probability to σ and the distribution of $\overline{Y}$ may be approximated by a $\mathcal{N}(\mu, \sigma^2/n)$ distribution, we may apply Slutsky's theorem to show that the large-sample distribution of $\sqrt{n}\,\overline{Y}/S = \left(\sqrt{n}\,\overline{Y}/\sigma\right) \times (\sigma/S)$ may be approximated by a $\mathcal{N}(\mu, 1)$ distribution. This allows us to construct approximate confidence intervals for and tests of hypotheses about μ using the standard normal distribution. A similar approach may be employed to the Bernoulli example by noting that p_n is a consistent estimator of π.

An important application of Slutky's Theorem relates to statistics that can be decomposed as a sum of a term for which some CLT holds and a term that converges in probability to 0. Assume, for example, that the variables Y_i have a finite fourth central moment γ and write the sample variance as

$$S^2 = [n/(n-1)]\left\{n^{-1}\sum_{i=1}^{n}[(Y_i - \mu)^2 - \sigma^2/n] + \left[\sigma^2 - \sum_{i=1}^{n}(\overline{Y} - \mu)^2\right]\right\}.$$

Since the first term within the $\{\}$ brackets is known to converge weakly to a normal distribution by the CLT and the second term converges in probability to 0, we conclude that the distribution of S^2 may be approximated by a $\mathcal{N}(\sigma^2, \gamma/n)$ distribution. This is the basis of the projection results suggested by Hoeffding (1948) and extensively explored by Jurečkova and Sen (1996) to obtain large-sample properties of ▶*U*-statistics as well as of more general classes of estimators.

Another convenient technique to obtain the asymptotic distributions of many (smooth) functions of asymptotically normal statistics is the *Delta-method*: if g is a locally differentiable function of a statistic T_n whose distribution may be approximated (for large samples) by a $\mathcal{N}(\mu, \tau^2)$ distribution, then the distribution of the statistic $g(T_n)$ may be approximated by a $\mathcal{N}\{g(\mu), [g'(\mu)]^2\tau^2\}$ distribution, where $g'(\mu)$ denotes the first derivative of g computed at μ. Suppose that we are interested in estimating the odds of a failed versus pass response, i.e., $\pi/(1-\pi)$ in an exam based on a sample of n students. A straightforward application of the De Moivre Laplace CLT may be used to show that the estimator of π, namely, k/n, where k is the number of students that failed the exam, follows an approximate $\mathcal{N}[\pi, \pi(1-\pi)/n]$ distribution. Taking $g(x) = x/(1-x)$, we may use the Delta-method to show that the distribution of the sample odds $k/(n-k)$ may be approximated by a $\mathcal{N}\{\pi/(1-\pi), \pi/[n(1-\pi)^3]\}$ distribution. This type of result has further applications in variance-stabilizing transformations used in cases (as the above example) where the variance of the original statistic depends on the parameter it is set to estimate.

For some important cases, like the Pearson χ^2-statistic or more general quadratic forms $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\mu}) = (\mathbf{Y} - \boldsymbol{\mu})^t \boldsymbol{A}(\mathbf{Y} - \boldsymbol{\mu})$ where $\mathbf{Y}$ is a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ and $\mathbf{A}$ is a p-dimensional square matrix of full rank, the (multivariate) Delta-method may not be employed because the derivative of Q computed at $\boldsymbol{\mu}$ is null. If $\mathbf{A}$ converges to an inverse of $\mathbf{V}$, a useful result known as the *Cochran theorem*, states that the distribution of Q may be approximated by a χ^2 instead of a normal distribution. In fact, the theorem holds even if $\mathbf{A}$ is not of full rank, but converges to a generalized inverse of $\mathbf{V}$. This is important for applications in categorical data.

The CLT also does not hold for extreme order statistics like the sample minimum or maximum; depending on some regularity conditions on the underlying random variables, the distribution of such statistics, conveniently normalized, may be approximated by one of three types of distributions, namely the extreme value distributions of the first, second or third type, which, in this context, are the only possible limiting distributions as shown by Gnedenko (1943).

Central Limit Theorems for Stochastic Processes

Empirical distribution functions and ▶order statistics have important applications in nonparametric regression models, resampling methods like the ▶*jackknife* and *bootstrap* (see ▶Bootstrap Methods), sequential testing as well as in Survival and Reliability analysis. In particular it serves as the basis for the well known goodness-of-fit *Kolmogorov-Smirnov* and *Cramér-von Mises statistics* and for *L*- and *R-estimators* like *trimmed* or *Winsorized means*. Given the sample observations $Y_1, \ldots, Y_n$ assumed to follow some distribution function F and a real number y, the empirical distribution function is defined as

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \le y)$$

where $I(Y_i \le y)$ is an indicator function assuming the value 1 if $Y_i \le y$ and 0, otherwise. It is intimately related to the order statistics, $Y_{n:1} \le Y_{n:2} \le \ldots \le Y_{n:n}$ where $Y_{n:1}$ is the smallest among $Y_1, \ldots, Y_n$, $Y_{n:2}$ is the second smallest and so on. For each fixed sample, F_n is a distribution function when considered as a function of y. For every fixed y, when considered as a function of $Y_1, \ldots, Y_n$, $F_n(y)$ is a random variable; in this context, since the $I(Y_i \le y), i = 1, \ldots, n$, are independent and identically distributed zero-one valued random variables, we may apply the classical CLT to conclude that for each fixed y the distribution of $F_n(y)$ may be approximated by a $\mathcal{N}\{F(y), F(y)[1-F(y)]/n\}$ distribution provided that n is sufficiently large. In fact, using standard asymptotic results, we may show that given any finite number m of points $y_1, \ldots, y_m$, the distribution function of the vector $[F_n(y_1), \ldots, F_n(y_m)]$ may be approximated by a multivariate normal distribution function. This property is known as *convergence of finite-dimensional distributions*.

On the other hand, $F_n - F = \{F_n(y) - F(y) : y \in \mathbb{R}\}$ is a random function defined on the set of real numbers, and, hence, to study its various properties we may need more than the results considered so far. Note that as the sample size n increases, so does the cardinality of the set of order statistics used to define the empirical distribution function and we may not be able to approximate this n-dimensional joint distribution by an m-dimensional one unless some further *tightness* or *compactness* conditions are imposed on the underlying distributions. This is the basis of the weak invariance principles necessary to show the convergence of empirical and other ▶stochastic processes to *Brownian bridge* or *Brownian motion* processes. An outline of the rationale underlying these results follows.

Let $t = F(y)$ and $W_n^0(t) = \sqrt{n}[G_n(t) - t]$, $t \in (0,1)$ where $G_n(t) = F_n[F^{-1}(t)] = F_n(y)$ with $F^{-1}(x) = \inf\{y : F(y) > x\}$, so that $\{W_n^0(t),\ t \in (0,1)\}$ is a stochastic process with $\mathbb{E}[W_n^0(t)] = 0$ and $\mathbb{E}[W_n^0(s)W_n^0(t)] = \min(s,t) - st$, $0 \le s,t \le 1$. Using the multivariate version of the CLT we may show that as $n \to \infty$, for all $m \ge 1$, given $0 \le t_1 \le \ldots \le t_m \le 1$, the vector $\boldsymbol{W}_{nm}^0 = [W_n^0(t_1), \ldots, W_n^0(t_m)] \xrightarrow{D} [W^0(t_1), \ldots, W^0(t_m)] = \boldsymbol{W}_m^0$ where $\boldsymbol{W}_m^0$ follows a $\mathcal{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_m)$ distribution with $\boldsymbol{\Gamma}_m$ denoting a positive definite matrix with elements $\min(t_i, t_j) - t_i t_j$, $i,j = 1, \ldots, m$.

Now, define a stochastic process $\{Z(t),\ t \in (0,1)\}$ with independent and homogeneous increments such that, for every $0 \le s < t \le 1$, the difference $Z(t) - Z(s)$ follows a $\mathcal{N}(0, t-s)$ distribution. Then, it follows that $\mathbb{E}[Z(s)Z(t)] = \min(s,t)$. This process is known as a *standard Brownian motion* or *standard Wiener process*. Furthermore, letting $W^0(t) = Z(t) - tZ(1)$, $0 \le t \le 1$, it follows that $\{W^0(t),\ t \in (0,1)\}$ is also a Gaussian stochastic process such that $\mathbb{E}[W^0(t)] = 0$ and $\mathbb{E}[W^0(s)W^0(t)] = \min(s,t) - st$, $0 \le s,t \le 1$. Then for all $m \ge 1$, given $0 \le t_1 \le \ldots \le t_m \le 1$, the vector $\boldsymbol{W}_m^0 = [W^0(t_1), \ldots, W^0(t_m)]$ also follows a $\mathcal{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_m)$ distribution. Since $W^0(0) = W^0(1) = 0$ with probability 1, this process is called a *tied down Wiener process* or *Brownian bridge*.

Using *the Kolmogorov maximal inequality*, we may show that $\{W_n^0(t),\ t \in (0,1)\}$ is *tight* and referring to standard results in weak convergence of probability measures, we may conclude that $\{W_n^0(t),\ t \in (0,1)\} \xrightarrow{D}$

$\{W^0(t),\ t \in (0,1)\}$. Details and extensions to statistical functionals may be obtained in Jurečkova and Sen (1996) among others.

About the Author

Dr. Julio da Motta Singer is Professor of Biostatistics and Head, Centre for Applied Statistics, University of São Paulo, Brazil. He has obtained his Ph.D. at the University of North Carolina in 1983 (advisor P.K. Sen). He has coauthored over 80 refereed papers and several books including, *Large Sample Methods in Statistics: An Introduction with Applications* (with P.K. Sen, Chapman and Hall, 1993), and *From finite sample to asymptotic methods in Statistics* (with P.K. Sen and A.C. Pedroso-de-Lima, Cambridge University Press, 2010).

Cross References

►Almost Sure Convergence of Random Variables
►Approximations to Distributions
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Empirical Processes
►Limit Theorems of Probability Theory
►Martingale Central Limit Theorem
►Normal Distribution, Univariate
►Probability Theory: An Outline

References and Further Reading

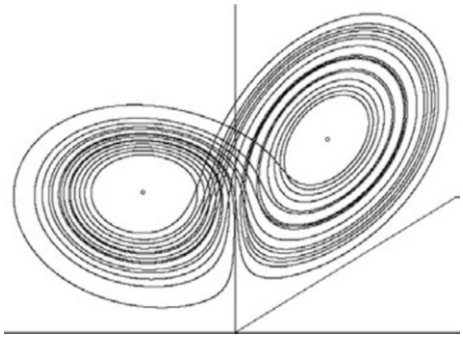
Billingsley P (1968) Convergence of probability measures. Wiley, New York
Berry AC (1941) The accuracy of the Gaussian approximation to the sum of independent variates. Trans Am Math Soc 49:122–136
Brown BM (1971) Martingale central limit theorems. Ann Math Stat 42:59–66
Chow YS, Teicher H (1978) Probability theory: independence, interchangeability, martingales Springer, New York
Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton
Dvoretzky A (1971) Asymptotic normality for sums of dependent random variables. In: Proceedings of the sixth berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley, vol 2, pp 513–535
Esséen CG (1971) A moment inequality with an application to the central limit theorem. Skandinavisk Aktuarietidskrift 39: 160–170
Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London
Gnedenko BV (1943) Sur la distribution limite du terme maximum d'une série aléatoire. Ann Math 44:423–453
Hoeffding W (1948) A class of statistics with asymptotically normal distributions. Ann Math Stat 19:293–325
Jurečková J, Sen PK (1996) Robust statistical procedures. Wiley, New York
Lehmann EL (2004) Elements of large sample theory. Springer, New York
Loynes RM (1970) An invariance principle for reversed martingales. Proc Am Math Soc 25:56–64
McLeish DL (1974) Dependent central limit theorems and invariance principles. Ann Probab 2:620–628
Reiss RD (1989) Approximate distributions of order statistics with applications to nonparametric statistics. Springer, New York
Sen PK (1981) Sequential nonparametrics: invariance principles and statistical inference. Wiley, New York
Sen PK, Singer JM, Pedroso-de-Lima AC (2010) From finite sample to asymptotic methods in statistics. Cambridge University Press, Cambridge
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York
Shevtsova IG (1974) Sharpening the upper bound of the absolute constant in the Berry–Esséen inequality. Theory Probab Appl 51:549–533
van Beeck P (1972) An application of Fourier methods to the problem of sharpening the Berry–Esséen inequality. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 23: 187–197
van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, New York

Chaotic Modelling

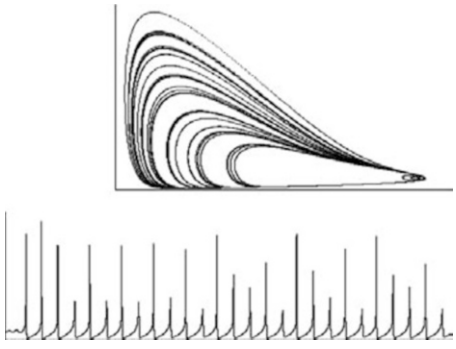
Christos H. Skiadas
Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece

Chaotic modeling is a term used to express the representation of the state of a system or a process by using chaotic models or tools developed in the chaotic literature and the related scientific context. In the following we present the main elements of the chaotic modeling including chaotic terms, differential and difference equations and main theorems (Skiadas 2009).

Chaos is a relatively new science mainly developed during last decades with the use of computers and supercomputers. It touches almost all the scientific fields. However, the basic elements can be found at the end of the nineteenth century and the attempts to solve the famous three-body problem by Henri Poincaré (1890). Although he succeeded to solve only the special case when the three bodies move in the same plane, he could explore the main characteristics of the general three-body problem and to see the unpredictability of the resulting paths in space. In other words he could realize the main characteristic of



Chaotic Modelling. Fig. 1 The Lorenz attractor (xyz view)



Chaotic Modelling. Fig. 2 Autocatalytic attractor and chaotic oscillations

a chaotic process that very small changes in initial conditions have significant impact to the future states of a system.

This was verified by Edwin Lorenz in 1963 with his work on modeling the atmospheric changes. He reduced the Navier-Stokes equations, used to express fluid flows, to a system of three nonlinear coupling differential equations and performed simulations in a computer trying to model the weather changes. He surprisingly found that the system was very sensitive to small changes of initial conditions thus making the forecasts of the future weather unpredictable. Famous are the forms of his simulated paths that look like a butterfly with open wings. The three-dimensional model which he proposed has the form (σ, r and b are parameters):

$$\dot{x} = -\sigma x + \sigma y,\ \dot{y} = -xz + rx - y,\ \dot{z} = xy - bz.$$

The famous Lorenz attractor also known as the butterfly attractor is illustrated in Fig. 1.

Several years later Rössler (1976) proposed a simpler three-dimensional model including only one nonlinear term thus verifying the assumption that a set of simple differential equations with only one nonlinear term may express chaotic behavior. The Rössler system is the following (e, f and m are parameters):

$$\dot{x} = -y - z,\ \dot{y} = x - ez,\ \dot{z} = f + xz - mz.$$

It can be verified that the number of chaotic parameters is equal to the number of the equations.

Chemical chaotic oscillations where observed by Belousov (1959) and later on by Zhabotinsky (1964) when they where working with chemical autocatalytic reactions. The Nobel Prize in chemistry (1977) was awarded to Prigogine for his work on dynamics of dissipative systems (see Prigogine 1961) including the mathematical representation of autocatalytic reactions. A simple autocatalytic reaction is expressed by the following set of three differential equations:

$$\dot{x} = \left(\frac{1}{1+k} + m\right)(k+z) - xy^2 - x,\ \dot{y} = \frac{xy^2 + x - y}{e}, \dot{z} = y - z$$

This model is illustrated in Fig. 2; the parameters set are: $e = 0.013, k = 2.5, m = 0.017$.

The use of computing power gave rise to the exploration of chaos in astronomy and astrophysics. A paper that influenced much the future developments of the chaotic applications was due to Hénon and Heiles in 1964. They had predicted chaos in Hamiltonian systems that could apply to astronomy and astrophysics. Few years before George Contopoulos (1958) had also found chaotic behavior when he explored the paths of stars in a galaxy. That it was most important was that they had shown that the computer experiments had much more to show than simply verify the results coming from the mathematical formulations. Hidden and unexplored scientific fields would emerge by the use of computers.

It was found that chaos could emerge from a system of three or more differential equations with at list one nonlinear term. This comes from the Poincaré–Bendixson theorem which states that a two dimensional system of nonlinear equations may have a regular behavior.

Another theorem is the famous *KAM* theorem from the initials of the names of Kolmogorov, Arnold and Moser. This theorem applies to dynamical systems and may explain the stability or not of these systems to small perturbations. It is interesting that the chaotic forms could be quite stable as it happens for vortex and tornados.

However, the main scientific discovery on chaos came only in 1978 by Michel Feigenbaum when he found that the simple logistic map could produce a chaotic sequence. Feigenbaum tried a difference equation instead of the differential equations that where used in the previous works on chaos. That is different is that chaos can emerge even

from only one difference equation with at list one nonlinear term. This is because a difference equation defines a recurrence scheme which is a set of numerous equations in which every equation uses the outcomes from the preceding one. The complexity resulting from a nonlinear difference equation is large and it can be measured with a power law of the number of iterations.

In the logistic model a mapping into itself is defined by the difference equation and gives rise to period doubling bifurcations and chaos for a specific range of the chaotic parameter. The logistic map is of the form: $x_{n+1} = bx_n(1 - x_n)$, where b is the chaotic parameter and x_n is the chaotic function (see a (x_{n+1}, x_n) diagram of the Logistic model in Fig. 3; $b = 2.9$).

For the logistic map as also for other maps there exists the bifurcation diagram. This is a diagram, usually two dimensional, defining the bifurcation points with respect to the chaotic parameter or parameters (see Fig. 4).

The chaotic modeling has also to do with *strange attractors* by means forms in space that have a great detail and complexity. These forms can arise in nature and also can be simulated from chaotic equations. A very interesting future of a chaotic attractor is that for a variety of initial conditions the chaotic system leads the final results or solutions to a specific area, the strange or chaotic attractor.

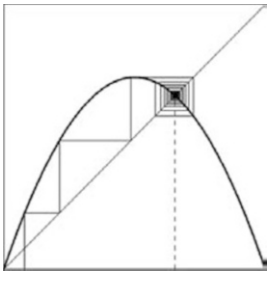
Chaos may also arise from a set of two or more difference equations with at least one nonlinear term. The most popular model is the Hénon (1976) model given by:

$$x_{n+1} = y_n + 1 - ax_n^2,\ y_{n+1} = bx_n.$$

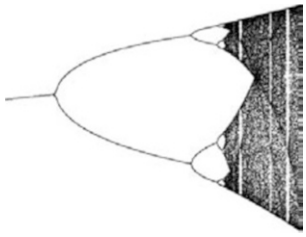
The Jacobian determinant of this model is:

$$\det J = \begin{vmatrix} \frac{\partial x_{n+1}}{\partial x_n} & \frac{\partial y_{n+1}}{\partial x_n} \\ \frac{\partial x_{n+1}}{\partial y_n} & \frac{\partial y_{n+1}}{\partial y_n} \end{vmatrix} = -b.$$

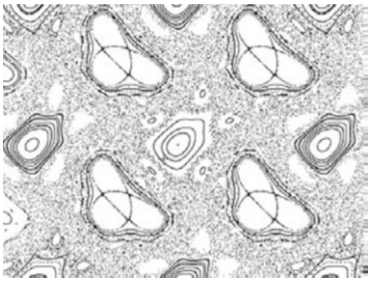
The system is stable for $0 < b < 1$. When $b = 1$ the system is area preserving, but it is unstable.



Chaotic Modelling. Fig. 3 The logistic model



Chaotic Modelling. Fig. 4 The bifurcation diagram



Chaotic Modelling. Fig. 5 A carpet-like form

An alternative of the Hénon map is provided by the following cosine model:
$x_{n+1} = by_n + 2a\cos(x_n) - 2a + 1,\ y_{n+1} = x_n$. This map provides a carpet-like form (see Fig. 5) for $b = -1$ and $a = -0.6$.

Very many cases in nature have to do with delays. This mathematically can be modeled by a delay differential or difference equation. Simpler is to use difference equations to express delay cases. An example is the transformation of the previous Hénon map to the corresponding delay difference equation of the form:

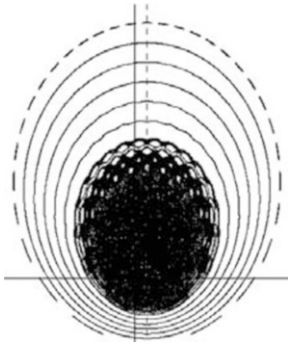
$$x_{n+1} = bx_{n-1} + 1 - ax_n^2.$$

This delay differential equation has the same properties of the Hénon map. In general modeling delays leads to differential or difference equations which produce oscillations and may produce chaos for appropriate selection of the parameters. One of the first proposed chaotic models including delays is the famous Mackey-Glass (1977) model regarding oscillation and chaos in physiological control systems.

Ikeda found his famous attractor in 1979 (see Fig. 6; parameters $a = 1, b = 0.83, c = 0.4$ and $d = 6$) when he was experimenting on the light transmitted by a ring cavity system. The equations' set is:

$$x_{n+1} = a + b(x_n\cos(\varphi_n) - y_n\sin(\varphi_n)),$$
$$y_{n+1} = b(x_n\sin(\varphi_n) + y_n\cos(\varphi_n)),$$

Chaotic Modelling. Fig. 6 The Ikeda attractor



Chaotic Modelling. Fig. 7 Chaotic rotating forms

where the rotation angle is: $\varphi_n = c - \dfrac{d}{1 + x_n^2 + y_n^2}$

The last form of difference equations express a rotation-translation phenomenon and can give very interesting chaotic forms (see Skiadas 2009). Figure 7 illustrates such a case where the rotation angle follows an inverse low regarding the distance r from the origin: $\varphi_n = \dfrac{c}{\sqrt{x_n^2 + y_n^2}} = \dfrac{c}{r}$. A chaotic bulge is located in the central part followed by elliptic trajectories in the outer part (the parameters are: $a = 0.6$ and $b = c = 1$).

Other interesting aspects of chaotic modeling are found in numerous publications regarding control of chaos with applications in various fields.

Chaotic mixing and chaotic advection have also studied with chaotic models as well as economic and social systems.

About the Author

Christos H. Skiadas is the Founder and Director of the Data Analysis and Forecasting Laboratory, Technical University of Crete (TUC). He was Vice-Rector of the Technical University of Crete (1997–1999), Chairman of the Production Engineering and Management Department, TUC (1995–1997) and participated in many committees and served as a consultant in various firms while he directed several scientific and applied projects. He was a visiting fellow in the University of Exeter, UK (1986) and Université Libre de Bruxelles, Belgium (1993–1994). He has been the guest-editor of the journals *Communications in Statistics*, *Methodology and Computing in Applied Probability* (MCAP) and *Applied Stochastic Models and Data Analysis*. He is a member of the Committee and Secretary of ASMDA International Society and was the Chair or co-chair of the International Conferences: 6th ASMDA 1993, 12th ASMDA 2007. As a member of the Greek Statistical Institute he organised and co-chaired two Annual National Conferences of the Institute (11th Chania 1998 and 22nd Chania 2009). Professor Skiadas contributions mainly are directed to the modeling and simulation of innovation diffusion and application of the findings in various fields as finance and insurance, energy consumption, forecasting and modeling. The related publications include more than 90 papers and 10 books, including *Chaotic Modelling and Simulation: Analysis of Chaotic Models, Attractors and Forms* (with Charilaos Skiadas, Chapman and Hall/CRC Press, 2009).

Cross References

►Stochastic Modeling, Recent Advances in

References and Further Reading

Belousov BP (1959) Периодически действующая реакция и ее механизм. [A periodic reaction and its mechanism]. Сборник рефератов по радиационной медицине (*Compilation of Abstracts on Radiation Medicine*) 147:145

Contopoulos G (1958) On the vertical motions of stars in a galaxy. *Stockholm Ann* 20(5):20

Hénon M (1976) A two-dimensional mapping with a strange attractor. *Commun Math Phys* 50:69–77

Hénon M, Heiles C (1964) The applicability of the third integral of motion: some numerical experiments. *Astron J* 69:73–79

Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. *J Stat Phys* 19:25–52

Ikeda K (1979) Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Opt Commun* 30:257–261

Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141

Mackey MC, Glass L (1977) Oscillation and chaos in physiological control systems. *Science* 197:287–289

Poincaré H (1890) Sur les équations de la dynamique et le problème de trois corps. *Acta Math* 13:1–270

Prigogine I (1961) *Thermodynamics of irreversible processes*, 2nd edn. Interscience, New York

Rössler OE (1976) An equation for continuous chaos. *Phys Lett A* 57:397–398

Skiadas CH, Skiadas C (2009) *Chaotic modeling and simulation: analysis of chaotic models attractors and forms.* Taylor & Francis/CRC Press, London

Zhabotinsky AM (1964) Периодический процесс окисления малоновой кислоты растворе (исследование кинетики реакции Белоусова). [Periodic processes of malonic acid oxidation in a liquid phase.] Биофизика [*Biofizika*] 9:306–311

Characteristic Functions

Miljenko Huzak
University of Zagreb, Zagreb, Croatia

Characteristic functions play an outstanding role in the theory of probability and mathematical statistics (Ushakov 1999). The characteristic function (c.f.) of a probability distribution function (d.f.) is the Fourier–Stieltjes transform of the d.f. More precisely, if F is a probability d.f. on d-dimensional real space R^d ($d \geq 1$), then its c.f. is a complex function $\phi : \mathrm{R}^d \to \mathrm{C}$ such that for any $\mathbf{t} = (t_1, \ldots, t_d) \in \mathrm{R}^d$,

$$\begin{aligned}\phi(\mathbf{t}) &= \int_{\mathrm{R}^d} e^{i\sum_{j=1}^d t_j x_j}\, dF(x_1, \ldots, x_d) := \\ &= \int_{\mathrm{R}^d} \cos\left(\sum_{j=1}^d t_j x_j\right) dF(x_1, \ldots, x_d) \\ &\quad + i \int_{\mathrm{R}^d} \sin\left(\sum_{j=1}^d t_j x_j\right) dF(x_1, \ldots, x_d),\end{aligned}$$

where the integrals are Lebesgue–Stieltjes integrals with respect to d.f. F.

If $\mathbf{X} = (X_1, \ldots, X_d)$ is a d-dimensional random vector, then c.f. $\phi = \phi_{\mathbf{X}}$ associated to $\mathbf{X}$ is the c.f. of its d.f. $F = F_{\mathbf{X}}$. Hence

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\sum_{j=1}^d t_j X_j}\right], \quad \mathbf{t} = (t_1, \ldots, t_d) \in \mathrm{R}^d. \qquad (1)$$

Particularly, c.f. $\phi = \phi_X : \mathrm{R} \to \mathrm{C}$ of a random variable (r.v.) X is equal to

$$\phi(t) = \mathbb{E}[e^{itX}], \ t \in \mathrm{R}.$$

Examples of c.f.s of some r.v.s are in Table 1.

C.f.s have many good properties (see Table 2). One of the most important properties of c.f.s is that there is a one-to-one correspondence between d.f.s and their c.f.s, which is a consequence of the *Lévy inversion formula* (see Chow and Teicher 1988 or Feller 1971). Since it is usually simpler to manipulate with c.f.s than with corresponding d.f.s,

Characteristic Functions. Table 1 Characteristic functions of some univariate probability distributions

Distribution	Density $f(x)$	c.f. $\phi(t)$
Degenerate at c		e^{itc}
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$(pe^{it} + 1 - p)^n$
Poisson	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\exp\{\lambda(e^{it} - 1)\}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$e^{i\mu t - \sigma^2 t^2/2}$
Symmetric uniform over $(-\theta, \theta)$	$\frac{1}{2\theta}$	$\frac{\sin \theta t}{\theta t}$
Gamma	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$(1 - it\beta)^{-\alpha}$
Cauchy	$\frac{\alpha}{\pi(\alpha^2 + x^2)}$	$e^{-\alpha\lvert t\rvert}$

Characteristic Functions. Table 2 List of properties of characteristic functions $\phi_{\mathbf{X}}(\mathbf{t})$ given by (1) (the list follows one from Ferguson (1996))

(1)	$\phi_{\mathbf{X}}(\mathbf{t})$ exists for all $\mathbf{t} \in \mathrm{R}^d$ and is continuous.
(2)	$\phi_{\mathbf{X}}(\emptyset) = 1$ and $\lvert\phi_{\mathbf{X}}(\mathbf{t})\rvert \leq 1$ for all $\mathbf{t} \in \mathrm{R}^d$.
(3)	For a scalar a, $\phi_{a\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{X}}(a\mathbf{t})$.
(4)	For a matrix A and a vector $\mathbf{c}$, $\phi_{A\mathbf{X}+\mathbf{c}}(\mathbf{t}) = e^{i\mathbf{t}^\tau \mathbf{c}} \cdot \phi_{\mathbf{X}}(A^\tau \mathbf{t})$.
(5)	For $\mathbf{X}$ and $\mathbf{Y}$ independent, $\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t})$.
(6)	If $\mathbb{E}\lvert\mathbf{X}\rvert < \infty$, $\dot{\phi}_{\mathbf{X}}(\mathbf{t})$ exists and is continuous and $\dot{\phi}_{\mathbf{X}}(\emptyset) = i\,\mathbb{E}\,\mathbf{X}^\tau$.
(7)	If $\mathbb{E}[\lvert\mathbf{X}\rvert^2] < \infty$, $\ddot{\phi}_{\mathbf{X}}(\mathbf{t})$ exists and is continuous and $\ddot{\phi}_{\mathbf{X}}(\emptyset) = -\mathbb{E}[\mathbf{X}\mathbf{X}^\tau]$.
(8)	If $P(\mathbf{X} = \mathbf{c}) = 1$ for a vector $\mathbf{c}$, $\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}^\tau \mathbf{c}}$.
(9)	If $\mathbf{X}$ is normal r. vec. with $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{X}$ and $\mathrm{cov}(\mathbf{X}) = \Sigma$, $\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left\{i\mathbf{t}^\tau \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\tau \Sigma \mathbf{t}\right\}$.

this property makes c.f.s useful in proving many theorems on probability distributions. For example, it can be proved that the components of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ are independent r.v.s if and only if

$$\begin{aligned}&(\forall t_1, \ldots, t_d \in \mathrm{R}) \ \ \phi_{\mathbf{X}}(t_1, \ldots, t_d) \\ &= \phi_{X_1}(t_1) \cdot \phi_{X_2}(t_2) \ldots \phi_{X_d}(t_d).\end{aligned}$$

Moreover, since for any independent r.v.s $X_1, X_2, \ldots, X_n$, c.f. of their sum $S_n = X_1 + \ldots + X_n$ is equal to the product of their c.f.s, to obtain the d.f. of S_n, it is usually easier

to find the c.f. of their sum and to apply the Lévy inversion formula than to find the convolution of their d.f.s.

Another very important property of c.f.s comes from the *continuity theorem* (see Chow and Teicher 1988 or Feller 1971): r.v.s X_n, $n \geq 1$, with corresponding c.f.s ϕ_n, $n \geq 1$, converge in law to a r.v. X with c.f. ϕ if and only if c.f.s ϕ_n, $n \geq 1$, converge to ϕ pointwise. For example, this property makes proving ►central limit theorems easier if not only possible.

C.f.s have been important tools in developing theories of infinite divisible and particularly stable distributions (e.g., see Feller 1971; Chow and Teicher 1988).

Cross References

►Probability on Compact Lie Groups

References and Further Reading

Chow YS, Teicher H (1988) Probability theory, independence, interchangeabilty, martingales. 2nd edn. Springer-Verlag, New York

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

Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London

Lukacs E (1970) Characteristic functions. 2nd edn. Griffin, London

Ushakov NG (1999) Selected topics in characteristic functions. Brill Academic Publishers, Leiden

Chebyshev's Inequality

Gerold Alsmeyer
Professor
Institut für Mathematische Statistik, Münster, Germany

Chebyshev's inequality is one of the most common inequalities used in probability theory to bound the tail probabilities of a random variable X having finite variance $\sigma^2 = \mathrm{Var}X$. It states that

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{for all } t>0, \tag{1}$$

where $\mu = \mathbb{E}X$ denotes the mean of X. Of course, the given bound is of use only if t is bigger than the standard deviation σ. Instead of proving (1) we will give a proof of the more general *Markov's inequality* which states that for any nondecreasing function $g : [0,\infty) \to [0,\infty)$ and any nonnegative random variable Y

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}g(Y)}{g(t)} \quad \text{for all } t>0. \tag{2}$$

Indeed, choosing $Y = |X-\mu|$ and $g(x) = x^2$ gives (1). The proof of Markov's inequality is very easy: For any $t > 0$,

$$\mathbb{P}(Y \geq t) = \int 1_{\{Y\geq t\}}\, d\mathbb{P} \leq \int_{\{Y\geq t\}} \frac{g(Y)}{g(t)}\, d\mathbb{P} \leq \frac{\mathbb{E}g(X)}{g(t)}.$$

Plainly, (1) provides us with the same bound $\sigma^2 t^{-2}$ for the one-sided tail probability $\mathbb{P}(X-\mu > t)$, but in this case an improvement is obtained by the following consideration: For any $c \geq 0$, we infer from Markov's inequality with $g(x) = x^2$

$$\mathbb{P}(X-\mu \geq t) = \mathbb{P}(X-\mu+c \geq t+c) \leq \frac{\mathbb{E}(X-\mu+c)^2}{(t+c)^2} = \frac{\sigma^2+c^2}{(t+c)^2}.$$

The right-hand side becomes minimal at $c = \sigma^2/t$ giving the one-sided tail bound

$$\mathbb{P}(X-\mu > t) \leq \frac{\sigma^2}{\sigma^2+t^2} \quad \text{for all } t>0, \tag{3}$$

sometimes called *Cantelli's inequality*.

Although Chebyshev's inequality may produce only a rather crude bound its advantage lies in the fact that it applies to any random variable with finite variance. Moreover, within the class of all such random variables the bound is indeed tight because, if X has a symmetric distribution on $\{-a, 0, a\}$ with $\mathbb{P}(X = \pm a) = 1/(2a^2)$ and $\mathbb{P}(X = 0) = 1 - 1/a^2$ for some $a > 1$, then $\mu = 0$, $\sigma^2 = 1$ and

$$\mathbb{P}(|X| \geq a) = \mathbb{P}(|X| = a) = \frac{1}{a^2},$$

which means that equality holds in (1) for $t = a$.

On the other hand, tighter bounds can be obtained when imposing additional conditions on the considered distributions. On such example is the following *Vysočanskiĭ-Petunīn inequality* for random variables X with an unimodal distribution:

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{4\sigma^2}{9t^2} \quad \text{for all } t > \sqrt{3/8}\,\sigma, \tag{4}$$

This improves (1) by a factor 4/9 for sufficiently large t.

One of the most common applications of Chebyshev's inequality is the weak law of large numbers (WLLN). Suppose we are given a sequence $(S_n)_{n\geq 1}$ of real-valued random variables with independent increments $X_1, X_2, \ldots$ such that $\mu_n := \mathbb{E}X_n$ and $\sigma_n^2 := \mathrm{Var}X_n$ are finite for all $n \geq 1$. Defining

$$m_n := \mathbb{E}S_n = \sum_{k=1}^{n} \mu_k \quad \text{and} \quad s_n^2 := \mathrm{Var}S_n = \sum_{k=1}^{n} \sigma_k^2$$

and assuming *Markov's condition*

$$\lim_{n\to\infty} \frac{s_n^2}{n^2} = 0 \tag{5}$$

we infer by making use of (1) that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n - m_n}{n}\right| \geq \epsilon\right) \leq \frac{s_n^2}{\epsilon^2 n^2} \to 0 \quad \text{as } n \to \infty$$

and therefore

$$\frac{S_n - m_n}{n} \to 0 \quad \text{in probability.} \tag{WLLN}$$

This result applies particularly to the case of i.i.d. $X_1, X_2, \ldots$ Then $m_n = n\mu$ and $s_n^2 = n\sigma^2$ where $\mu := \mathbb{E}X_1$ and $\sigma^2 := \mathrm{Var}X_1$. In this case, Chebyshev's inequality further gives, for all $\epsilon, \beta > 0$, that

$$\sum_{n\geq 1} \mathbb{P}\left(\left|\frac{S_n - n\mu}{n}\right| \geq \epsilon \log^\beta n\right) \leq \sum_{n\geq 1} \frac{\sigma^2}{\epsilon^2 n \log^{2\beta} n} < \infty$$

and thus, by invoking the Borel-Cantelli lemma (see ►Borel–Cantelli Lemma and Its Generalizations),

$$\frac{S_n - n\mu}{n \log^\beta n} \to 0 \quad \text{a.s. for all } \beta > 0 \tag{6}$$

This is not quite the strong law of large numbers ($\beta = 0$) but gets close to it. In fact, in order for this to derive, a stronger variant of Chebyshev's inequality, called *Kolmogorov's inequality*, may be employed which states that

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k - m_k| \geq t\right) \leq \frac{s_n^2}{t^2} \quad \text{for all } t > 0$$

under the same independence assumptions stated above for the WLLN. Notice the similarity to Chebyshev's inequality in that only $S_n - m_n$ has been replaced with $\max_{1\leq k\leq n}(S_k - m_k)$ while retaining the bound.

Let us finally note that, if X has mean μ, median m and finite variance σ^2, then the one-sided version (3) of Chebyshev's inequality shows that

$$\mathbb{P}(X - \mu \geq \sigma) \leq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X - \mu \leq -\sigma) \leq \frac{1}{2},$$

in other words, the median of X is always with one standard deviation of it mean.

Bibliographical notes: (1) dates back to Chebyshev's original work (Chebyshev 1867), but is nowadays found in any standard textbook on probability theory, like (Feller 1971). The latter contains also a proof of the one-sided version (3) which differs from the one given here. (4) for unimodal distributions is taken from Vysočanskiĭ and Petunĭn (1979), see also Sellke and Sellke (1997). For multivariate extensions of Chebyshev's inequality see Olkin and Pratt (1958) and Monhor (2007).

About the Author

Dr. Alsmeyer is Professor at the Department of Mathematics and Computer Science of the University of Münster and was the Chairman from 2000 till 2008. He has written more than 50 research articles and one book. He has supervised eight Ph.D. students.

Cross References

►Borel–Cantelli Lemma and Its Generalizations
►Expected Value
►Laws of Large Numbers
►Standard Deviation

References and Further Reading

Chebyshev P (1867) Des valeurs moyennes. Liouville's J Math Pure Appl 12:177–184
Feller W (1971) An introduction to probability theory and its applications, vol II, 2nd edn. Wiley, New York
Monhor D (2007) A Chebyshev inequality for multivariate normal distribution. Prob Eng Inform Sci 21(2):289–300
Olkin I, Pratt JW (1958) A multivariate Tchebycheff inequality. Ann Math Stat 29:226–234
Sellke TM, Sellke SH (1997) Chebyshev inequalities for unimodal distributions. Am Stat 51(1):34–40
Vysočanskiĭ DF, Petunĭn JĬ (1979) Proof of the 3σ rule for unimodal distributions. Teor Veroyatnost i Mat Stat 21:23–35

Chemometrics

Rolf Sundberg
Professor of Mathematical Statistics
Stockholm University, Stockholm, Sweden

The role of statistics in chemistry is over a century old, going back to the Guinness brewery chemist and experimenter Gosset, more well-known under the pseudonym "Student." For his applications, he was in need of small-sample statistical methods. Until the 1970s, chemistry methods and instruments were typically univariate, but in that decade analytical chemistry and some other branches of chemistry had to start handling data of multivariate character. For example, instead of measuring light intensity at only a single selected wavelength, instruments became available that could measure intensities at several different wavelengths at the same time. The instrumental capacity rapidly increased, and the multivariate spectral dimension soon exceeded the number of chemical samples analysed (the "$n < p$" problem). In parallel, other chemists worked with Quantitative Structure–Activity Relationships (QSAR), where they tried to explain and predict biological activity or similar properties of a molecule from

a large number of structural physical-chemical characteristics of the molecule, but having an empirical data set of only a moderate number of different molecules. Generally, as soon as multivariate data are of high dimension, we must expect near collinearities among the variables, and when $n < p$, there are necessarily exact collinearities. These were some of the problems faced, long before statisticians got used to $n < p$ in genomics, proteomics etc. This was the birth of the field of chemometrics, a name coined by Svante Wold to characterize these research and application activities.

A standard definition of *Chemometrics* would be of type "The development and use of mathematical and statistical methods for applications in chemistry," with more weight on statistical than mathematical. Another characterization, formulated by Wold, is that the aim of chemometrics is to provide methods for

- How to get chemically relevant information out of measured chemical data
- How to get it into data
- How to represent and display this information

and that in order to achieve this, chemometrics is heavily dependent on statistics, mathematics and computer science. The first task is much concerned with analysis of dependencies and relationships (regression, calibration, discrimination, etc.) within a multivariate framework, because complex chemical systems are by necessity multidimensional. The second task is largely represented by experimental design, both classical and newer, where chemometrics has contributed the idea of design in latent factors (principal variates). For representation of high-dimensional data, projection on a low-dimensional latent variable space is the principal tool. Using diagrams in latent factors from PCA or other dimension-reducing methods is also a way of displaying the information found.

Another type of definition, often quoted, is that "Chemometrics is what chemometricians do." This is not only to laugh at. A vital part of chemometrics is connected with chemistry, but the methods developed might be and are applied in quite different fields, where the data analysis problems are similar, such as metabolomics, food science, sensometrics, and image analysis. This could motivate to distinguish chemometrics and chemometric methods, where the latter could as well be described as statistical methods originally inspired by problems in chemistry.

A statistician's look at the contents of *Journal of Chemometrics* for the period 2003–2005 (150 papers) showed that regression and calibration dominated, covering a third of the contents. Much of this was on regularized regression methods, such as PCR (Principal Components Regression) and PLSR (Partial Least Squares Regression). Other statistical areas represented were multiway methods (where each observation is a matrix or an even higher-dimensional array, see Smilde et al. 2004), classification (discrimination and clustering), multivariate statistical process control, and occasionally other areas, for example experimental design, wavelets, genetic algorithms.

A difference between chemometrics and biometrics (►biostatistics) is that chemometricians are mostly chemists by principal education, with more or less of additional statistical education, whereas biometricians are typically statisticians by education. This has had several consequences for chemometrics. Statistical methods are sometimes reinvented. Metods are sometimes proposed without a theoretical underpinning. The popular method of partial least squares (see ►Partial Least Squares Regression Versus Other Methods) is a good such example, nowadays relatively well understood, but proposed and advocated as a computational algorithm, that was widely regarded with suspicion among statisticians. Thus there is often a role for theoretical statistical studies to achieve a deeper understanding of the chemometric methods and their properties, not least to reveal how various suggested methods relate to each others.

About the Author

Professor Sundberg is Past President of the Swedish Statistical Association (1979–1981). He is an Elected member of the International Statistical Institute (1984). He was an Associate editor for *Scandinavian Journal of Statistics* (1987–1994). In 2004 he was awarded (with Marie Linder) the Kowalski prize for best theoretical paper in *Journal of Chemometrics* (2002–2003).

Cross References

►Sensometrics

References and Further Reading

Regression and calibration

Brown PJ (1993) Measurement, regression, and calibration. Oxford University Press, Oxford

Martens H, Næs T (1989) Multivariate calibration. Wiley, Chichester

Sundberg R (1999) Multivariate calibration – direct and indirect regression methodology (with discussion). Scand J Stat 26: 161–207

Other fields of chemometrics

Carlson R, Carlson JE (2005) Design and optimization in organic synthesis, 2nd edn. Elsevier, Amsterdam

Martens H, Martens M (2001) Multivariate analysis of quality. An introduction. Wiley, Chichester

Smilde A, Bro R, Geladi P (2004) Multi-way analysis with applications in the chemical sciences. Wiley, Chichester

Two journals are devoted to chemometrics, started in 1986/87

Journal of Chemometrics. John Wiley & Sons.

Chemometrics and Intelligent Laboratory Systems. Elsevier.

There are several introductions to chemometrics written for chemists, not listed here. Not all of them are satisfactory in their more statistical parts.

Chernoff Bound

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

The Chernoff Bound, due to Herman Rubin, states that if $\overline{X}$ is the average of n independent observations on a random variable X with mean $\mu < a$ then, for all t,

$$P(\overline{X} > a) \le \left[E(e^{t(X-a)})\right]^n.$$

The proof which follows shortly is a simple application of the Markov inequality that states that for a positive random variable $Y, P(Y \ge b) \le E(Y)/b$, for $b > 0$. The Chernoff bound was a step in the early development of the important field "Large Deviation Theory." It became prominent among computer scientists because of its usefulness in Information Theory.

The Markov inequality is derived from the fact that for $b > 0$,

$$E(Y) = \int ydF(Y) \ge \int_b^\infty ydF(y) \ge bP(Y \ge b)$$

where F is the cumulative distribution of Y.

We observe that

$$E(e^{nt(\overline{X}-a)}) = \left[E(e^{t(X-a)})\right]^n$$

and hence $P(e^{nt(\overline{X}-a)} \ge 1)$ is less than or equal to the bound. This implies the Chernoff bound for $t > 0$. For $t \le 0$ the inequality is automatically satisfied because the bound is at least one. That follows form the fact that the ►moment generating function $M(t) = E(e^{tZ})$ is convex with $M(0) = 1, M'(0) = E(Z)$ and $E(X - a) < 0$.

The prominence of the bound is due to a natural inclination to extend beyond its proper range of applicability. The Central Limit Theorem (see ►Central Limit Theorems), for which an informal statement is that $\overline{X}$ is approximately normally distributed with mean $\mu = E(X)$ and variance σ^2/n where σ is the standard deviation of X. For large deviations (see ►Large Deviations and Applications), or many standard deviations from the mean, the theorem implies that the probability of exceeding a would approach zero, but a naive interpretation would state that this probability would be approximately $\exp(-na^2/2)(2\pi na)^{-1/2}$ and could be seriously wrong.

In 1951, for a special problem of testing a simple hypothesis versus a simple alternative using a statistic of the form $\overline{X}$ where X could take on a few integer values, I realized that the normal approximation was inappropriate. I derived (Chernoff 1952), for $a > E(X)$,

$$n^{-1} \log P(\overline{X} > a) \to \inf_t E(e^{t(X-a)})$$

which was, as far as I know, the first application of Large Deviation Theory to Statistical Inference. This result was used to define a measure of information useful for experimental design and to show that the Kullback-Leibler information numbers (Kullback and Leibler 1951; Chernoff 1956) measure the exponential rate at which one error probability approaches zero when the other is held constant.

At the time I was informed of Cramér's (1938) earlier elegant derivation of more encompassing results, using exponentially tilted distributions. Cramér dealt with deviations which were not limited to those of order square root of n standard deviations, but required a condition that excluded the case which I needed, where the range of X was a multiple of the integers. Blackwell and Hodges (1959) later dealt with that case.

One of my colleagues, Herman Rubin, claimed that he could derive my results more simply, and when I challenged him, he produced the upper bound that I included in my manuscript. At the time the proof seemed so trivial, that I did not mention that it was his. I made two serious errors. First, the inequality is stronger than the upper limit implied by my result, and therefore deserves mention of authorship even though the derivation is simple. Second, because I was primarily interested in the exponential rate at which the probability approached zero, it did not occur to me that this trivially derived bound could become prominent.

About the Author

For biography *see* the entry ►Chernoff–Savage Theorem.

Cross References

►Central Limit Theorems
►Chebyshev's Inequality
►Kullback-Leibler Divergence
►Large Deviations and Applications

References and Further Reading

Blackwell D, Hodges JL (1959) The probability in the extreme tail of a convolution. Ann Math Stat 30:1113–1120

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 23:495–507

Chernoff H (1956) Large-sample theory: parametric case. Ann Math Stat 27:1–22

Cramér H (1938) Sus un nouveau théorème-limite de la théorie des probabilités. Acutalités Scientifiques et Industrielles, 736, Paris

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Chernoff Faces

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

The graphical representation of two dimensional variables is rather straightforward. Three dimensional variables presents more of a challenge, but dealing with higher dimensions is much more difficult. Two methods using profiles and stars suffers from a confusion of which variable is represented when the dimensionality is greater than six.

The method called "Chernoff Faces" (Chernoff 1973) involves a computer program which draws a caricature of a face when given 18 numbers between 0 and 1. These numbers correspond to features of the face. Thus one may represent the length of the nose, another curvature of the mouth, and a third the size of the eyes. If we have 12 dimensional data, we can adjoin 6 constants to get points in 18 dimensional space, each represented by a face. As the point moves in 18 dimensional space the face changes.

The method was developed in response to a problem in cluster analysis (see ►Cluster Analysis: An Introduction). There are many methods proposed to do clustering. It seems that an appropriate method should depend on the nature of the data, which is difficult to comprehend without visualization. The grouping of faces which look alike serves as a preliminary method of clustering and of recognizing which features are important in the clustering.

In the two original applications of the method, the scientists involved claimed that the implementation was lucky because the features which were most important were represented respectively by the size of the eyes and the shape of the face, both of which are prominent features. I claimed that it did not matter which features were selected for the various variables and challenged the scientists to select an alternative choice of features for the variables to degrade the effect of the faces. Their candidate choices had little degradation effect.

To test the conjecture that the choice of variables would have no effect, Rizvi and I carried out an experiment (Chernoff and Rizvi 1975). Of course it is clear that the conjecture cannot be absolutely sound, since the position of the pupils in the eyes cannot be detected if the eyes are small and other features interact similarly. However we set up an experiment where subjects were supposed to cluster 36 faces into two groups of approximately 18 each. The faces were generated from two six dimensional ►multivariate normal distributions with means δ units apart, in Mahalanobis distance, and identity covariance matrix. These data were then subjected to a linear transformation to an 18-dimensional space, and 12 feature selections were made at random. The subjects were given three clustering problem. For the first δ was so large that there was no problem recognizing the clusters. That was a practice problem to train the students in the experiment. For the other two problems two choices of δ were made to establish greater difficulty in separating the two distributions. The result of this experiment was that when the error rate in clustering varies from 8% to 22%, the typical random permutations could change the error rate by a proportion which decreases from 45% to 18%.

Originally, Faces were designed to serve to understand which variables were important and which interacted with each other. Once such relations are understood, analytic methods could be used to probe further. In many applications, Faces could also be used to comprehend data where the roles of the various factors were well understood. For example, in business applications, a smiling face could indicate that some aspect of the business was doing well. With training of the users, such applications could be useful in providing a quick and easy comprehension of a moderately complicated system. For example, one could use a face to represent the many meters an airplane pilot watches, so that he could be alerted when the face begins to look strange. The method of stars could also serve such a function.

Jacob (1978) used faces to represent five particular scales of the Minnesota Multiphasic Personality Inventory (MMPI). The scales represented Hypochondriasis, Depression, Paranoia, Schizophrenia and Hypomania. Realizing that training a psychologist to recognize a smiling face as belonging to a depressed patient would be difficult, he developed an innovative approach to selecting features for the five scales. He presented a random selection of faces

to some psychologists and asked them to rate these faces on the MMPI scales. Then he used regression techniques to decide how the numerical values of an MMPI scale should be translated into features of the face, so that the face presented to a psychologist would resemble that of a person with those scaled values. This would facilitate the process of training psychologists to interpret the faces.

The method of Faces handles many dimensions well. For more than 18 variables, one could use a pair of faces. It does not deal so well with a large number of faces unless we have a time series in which they appear in succession. In that case they can be used to detect changes in characteristics of important complicated systems.

It seems that face recognition among humans is handled by a different part of the brain than that handling other geometrical data and humans are sensitive to very small changes in faces. Also, it seems that cartoons and caricatures of faces are better remembered than realistic representations.

Before the computer revolution, graphical representations, such as nomograms, could be used to substitute for accurate calculations. The Faces are unlikely to be useful for calculation purposes.

About the Author

For biography *see* the entry ▶Chernoff–Savage Theorem.

Cross References

▶Cluster Analysis: An Introduction

References and Further Reading

Chernoff H (1973) The use of faces to represent points in k-dimensional space graphically. J Am Stat Assoc 68:301–308

Chernoff H, Rizvi MH (1975) Effect on classification error of random permutations of features in representing multivariate data by faces. J Am Stat Assoc 70:548–554

Jacob RJK (1978) Facial representation of multivariate data. In: Wang PCC (ed) Graphical representation of multivariate data. Academic, New York, pp 143–168

Chernoff-Savage Theorem

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

Hodges and Lehmann (1956) conjectured in 1956 that the nonparametric competitor to the t-test, the Fisher-Yates-Terry-Hoeffding or c_1 test (Terry 1952), was as efficient as the t-test for normal alternatives and more efficient for nonnormal alternatives.

To be more precise, we assume that we have two large samples, of sizes m and n with $N = m + n$, from two distributions which are the same except for a translation parameter which differs by an amount δ. To test the hypothesis that $\delta = 0$ against one sided alternatives, we use a test statistic of the form

$$T_N = m^{-1} \sum_{i=1}^{N} E_{Ni} z_{Ni}$$

where z_{Ni} is one or zero depending on whether the ith smallest of the N observations is from the first or the second sample. For example the Wilcoxon test is of the above form with $E_{Ni} = i/N$. It was more convenient to represent the test in the form

$$T_N = \int_{-\infty}^{\infty} J_N[H_N(x)]dF_m(x).$$

where F_m and G_n are the two sample cdf's, $\lambda_N = m/N$ and $H_N = \lambda_N F_m + (1 - \lambda_N)G_n$. These two forms are equivalent when $E_{Ni} = J_N(i/N)$.

The proof of the conjecture required two arguments. One was the ▶asymptotic normality of T when $\delta \neq 0$. The Chernoff-Savage theorem (Chernoff and Savage 1958) establishes the asymptotic normality, under appropriate regularity conditions on J_N, satisfied by c_1, using an argument where F_m and G_n are approximated by continuous time ▶Gaussian Processes, and the errors due to the approximation are shown to be relatively small.

The second argument required a variational result using the Pitman measure of local efficacy of the test of $\delta = 0$, which may be calculated as a function of the underlying distribution. For distributions with variance 1, the efficiency of the test relative to the t-test is minimized with a value of 1 for the normal distribution. It follows that the c_1 test is as efficient as the t-test for normal translation alternatives and more efficient for nonnormal translation alternatives.

About the Author

Dr. Herman Chernoff (born in New York City on July 1, 1923) is Professor Emeritus of Statistics at Harvard University and Emeritus Professor at M.I.T. He received a PhD in Applied Mathematics at Brown University in 1948 under the supervision of Abraham Wald (at Columbia University). Dr. Chernoff worked for the Cowles Commission at the University of Chicago and then spent three years in the Mathematics Department at the University of Illinois before joining the Department of Statistics at Stanford University in 1952, where he remained for 22 years.

He moved to M.I.T. in 1974, where he founded the Statistics Center. Since 1985 he has been in the Department of Statistics at Harvard. He retired from Harvard in 1997. Professor Chernoff was President of the Institute of Mathematical Statistics (1967–1968) and is an Elected member of both the American Academy of Arts and Sciences and the National Academy of Sciences. He has been honored for his contributions in many ways. He is a recipient of the Townsend Harris Medal and Samuel S. Wilks Medal "for outstanding research in large sample theory and sequential analysis, for extensive service to scholarly societies and on government panels, for effectiveness and popularity as a teacher, and for his continuing impact on the theory of statistics and its applications in diverse disciplines" (1987). He was named Statistician of the Year, Boston Chapter of the ASA (1991). He holds four honorary doctorates. Professor Chernoff is the co-author, with Lincoln Moses, of a classic text, now a Dover Reprint, entitled *Elementary Decision Theory*. He is also the author of the SIAM monograph 8 entitled *Sequential Analysis and Optimal Design*. The book *Recent Advances in Statistics* (MH Rizvi, J Rustagi and D Siegmund (Eds.), Academic Press, New York, 1983) published in honor of his 60th birthday in 1983 contained papers in the fields where his influence as a researcher and teacher has been strong: design and sequential analysis, optimization and control, nonparametrics, large sample theory and statistical graphics.

Cross References

►Asymptotic Normality
►Asymptotic Relative Efficiency in Testing
►Gaussian Processes
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

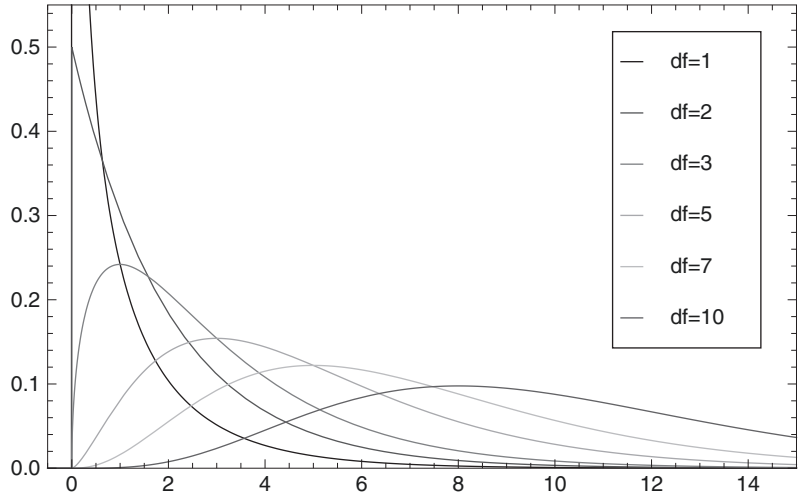
References and Further Reading

Chernoff H, Savage IR (1958) Asymptotic normality and efficiency of certain nonparametric test statistics. Ann Math Stat 29:972–994
Hodges JL, Lehmann EL (1956) The efficiency of some nonparametric competitors to the *t*-test. Ann Math Stat 27:321–325
Terry ME (1952) Some rank order tests which are most powerful against specific parametric alternatives. Ann Math Stat 23: 346–366

Chi-Square Distribution

Miljenko Huzak
University of Zagreb, Zagreb, Croatia

The chi-square distribution is one of the most important continuous probability distributions with many uses in statistical theory and inference. According to O. Sheynin (1971), Ernst Karl Abbe obtained it in 1863, Maxwell formulated it for three degrees of freedom in 1860, and Boltzman discovered the general expression in 1881. Lancaster (1966) ascertained that Bienaymé derived it as early as in 1838. However, their derivations "had no impact on the progress of the mainstream statistics" (R. L. Plackett 1983, p. 68)



Chi-Square Distribution. Fig. 1 Densities of χ^2 -distributions with 1, 2, 3, 5, 7, and 10 degrees of freedom (df)

since chi-square is not only a distribution, but also a statistic and a test procedure, all of which arrived simultaneously in the seminal paper written by Karl Pearson in 1900.

Let $n \geq 1$ be a positive integer. We say that a random variable (r.v.) has χ^2 (*chi-square*, χ is pronounced ki as in kind) *distribution with n degrees of freedom* (d.f.) if it is absolutely continuous with respect to the Lebesgue measure with density:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \Gamma\left(\frac{n}{2}\right)^{-1} 2^{-n/2} x^{n/2-1} e^{-x/2} & \text{if } x > 0 \end{cases}$$

where Γ denotes the Gamma function.

Figure 1 shows some of the densities.

Hence, the χ^2-distribution (with n d.f.) is equal to the Γ - distribution with the parameters $(n/2, 2)$, that is, with the mean and variance equal to n and $2n$ respectively.

The χ^2-distribution is closely connected with the normal distribution. It turns out that the sample variance S^2 of a random sample from a normally distributed population has, up to the constant, the χ^2 – sample distribution. More precisely, if $X_1, \ldots, X_n$ are independent and identically distributed normal r.v.s with the population variance σ^2, then

$$\frac{n-1}{\sigma^2} \cdot S^2 = \frac{1}{\sigma^2}\left((X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right) = \frac{1}{\sigma^2}\left(X_1^2 + X_2^2 + \cdots + X_n^2 - n\overline{X}^2\right)$$

is a χ^2-distributed r.v. with $n - 1$ d.f. (see e.g., Shorack 2000). This is a consequence of a more general property of the normality (Feller 1971). For example, let $\mathbf{X}$ be an n-dimensional standard normal vector, that is, a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ such that its components $X_1, \ldots, X_n$ are independent and normally distributed with mean and variance equal to 0 and 1 respectively. Then the square of the Euclidean norm of $\mathbf{X}$, $|\mathbf{X}|^2 = X_1^2 + \ldots + X_n^2$, is χ^2-distributed with n d.f. If means of the components of $\mathbf{X}$ are non-zero, then $|\mathbf{X}|^2$ has *non-central* χ^2-distribution with n d.f. and *non-centrality* parameter equal to the square of the mean of $\mathbf{X}$. In this generality, χ^2-distribution is the *central* χ^2-distribution, that is, a χ^2-distribution with non-centrality parameter equal to 0.

In statistics, many test statistics have a χ^2 or asymptotic χ^2-distribution. For example, goodness of fit χ^2-tests are based on the so-called Pearson's χ^2-statistics or general χ^2-statistics that have, under appropriate null-hypothesis, asymptotic χ^2-distributions; The Friedman test statistic and likelihood ratio tests are also based on asymptotically χ^2-distributed test statistic (see Ferguson 1996). Generally, appropriately normalized quadratic forms of normal (and asymptotic normal) statistics have χ^2 (and asymptotic χ^2) distributions.

Non-central χ^2-distributions are used for calculating the power function of tests based on quadratic forms of normal or asymptotic normal statistics.

Cross References

►Categorical Data Analysis
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests
►Continuity Correction
►Gamma Distribution
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

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

Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London

Lancaster HO (1966) Forerunners of the Pearson χ^2. Aust J Stat 8: 117–126

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 reasonably be supposed to have arisen from random sampling. Philos Mag 5(1):157–175

Plackett RL (1983) Karl Pearson and the Chi-squared test. Int Stat Rev 51(1):59–72

Sheynin OB (1971) Studies in the history of probability and statistics. XXV. On the history of some statistical laws of distribution. Biometrika 58(1):234–236

Shorack GR (2000) Probability for statisticians, Springer-Verlag, New York

Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements

Vassiliy Voinov[1], Mikhail Nikulin[2]
[1]Professor
Kazakhstan Institute of Management, Economics and Strategic Research, Almaty, Kazakhstan
[2]Professor
University of Victor Segalen, Bordeaux, France

The famous chi-squared goodness-of-fit test was discovered by Karl Pearson in 1900. If the partition of a sample space is such that observations are grouped over r disjoined

intervals Δ_i, and denoting ν_i observed frequencies and $np_i(\theta)$ expected that correspond to a multinomial scheme, the Pearson's sum is written

$$\chi^2 = X_n^2(\theta) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)} = \mathbf{V}^T(\theta)\mathbf{V}(\theta), \quad (1)$$

where $\mathbf{V}(\theta)$ is a vector with components $v_i(\theta)=(\nu_i - np_i(\theta))(np_i(\theta))^{-1/2}$, $i=1,\ldots,r$. If the number of observations $n \to \infty$, the statistic (1) for a simple null hypothesis, specifying the true value of θ, will follow chi-squared probability distribution with $r-1$ degrees of freedom.

Until 1934, Pearson believed that the limit distribution of his chi-squared statistic would be the same if unknown parameters of the null hypothesis were replaced by estimates based on a sample (Stigler (2008), p. 266). Stigler noted that this major error of Pearson "has left a positive and lasting impression upon the statistical world." It would be better to rephrase this sentence as follows: "has left a positive (because it inspired the further development of the theory of chi-squared test) and lasting 'negative' impression". Fisher (1924) clearly showed that the number of degrees of freedom of the Pearson's test must be reduced by the number of parameters estimated by a sample. The Fisher's result is true if and only if parameters are estimated by grouped data (minimizing Pearson's chi-squared sum, using multinomial maximum likelihood estimates (MLEs) for grouped data, or by any other asymptotically equivalent procedure).

Nowadays, the Pearson's test with unknown parameters replaced by grouped data estimates $\hat{\theta}_n$ is known as Pearson-Fisher test $X_n^2(\hat{\theta}_n)$. Chernoff and Lehmann (1954) showed that replacing unknown parameters in (1) by their maximum likelihood estimates based on non-grouped data would dramatically change the limit distribution. In this case, it will follow a distribution that in general depends on unknown parameters and, hence, cannot be used for testing. What is difficult to understand for those who apply chi-squared tests is that an estimate is a realization of a random variable with its own probability distribution and that a particular estimate can be too far from the actual unknown value of a parameter or parameters. This misunderstanding is rather typical for those who apply both parametric and non-parametric tests for compound hypotheses.

Roy (1956) extended Chernoff and Lehmann's result to the case of random grouping intervals. Molinari (1977) derived the limit distribution of Pearson's sum if moment type estimates (MMEs) based on raw data are used. Like the case of MLEs it depends on unknown parameters. Thus, a problem of deriving a test statistic, where limiting distribution will not depend on parameters, is aroused. Dahiya and Gurland (1972) showed that for location and scale families with properly chosen random cells, the limit distribution of Pearson's sum may not depend on unknown parameters but on the null hypothesis. Being distribution-free, such tests can be used in practice, but for each specific null distribution one has to evaluate corresponding critical values. So, two ways of constructing distribution-free Pearson's type tests are to use proper estimates of unknown parameters (e.g., based on grouped data), or to use specially constructed grouping intervals. Another possible way is to modify the Pearson's sum such that its limit probability distribution would not depend on unknowns. Nikulin (1973), using a very general theoretical approach (nowadays known as Wald's method (see Moore 1977)), solved the problem in full for any continuous probability distribution if one will use random cells based on predetermined probabilities to fall into a cell with random boundaries depending on efficient estimates (MLEs or best asymptotically normal (BAN) estimates) of unknown parameters. Rao and Robson (1974), using a much less general heuristic approach, confirmed the result of Nikulin for a particular case of exponential family of distributions. Formally their result fully coincides with that of Nikulin (1973)

$$Y1_n^2(\hat{\theta}_n) = X_n^2(\hat{\theta}_n) + \mathbf{V}^T(\hat{\theta}_n)\mathbf{B}(\mathbf{J} - \mathbf{J}_g)^{-1}\mathbf{B}^T\mathbf{V}(\hat{\theta}_n), \quad (2)$$

where $\mathbf{J}$ and $\mathbf{J}_g = \mathbf{B}^T\mathbf{B}$ are Fisher information matrices for non-grouped and grouped data correspondingly, and $\mathbf{B}$ is a matrix with elements $b_{ij} = \frac{1}{\sqrt{p_i(\theta)}}\frac{\partial p_i(\theta)}{\partial \theta_j}$, $i = 1,\ldots,r$, $j = 1,\ldots,s$. The statistic (2) can be presented also as (Moore and Spruill (1975))

$$Y1_n^2(\hat{\theta}_n) = \mathbf{V}^T(\hat{\theta}_n)(\mathbf{I} - \mathbf{B}\mathbf{J}^{-1}\mathbf{B}^T)^{-1}\mathbf{V}(\hat{\theta}_n). \quad (3)$$

The statistic (2) or (3), suggested first by Nikulin (1973a) for testing the normality, will be referred to subsequently as Nikulin-Rao-Robson (NRR) test. Nikulin (1973) assumed that only asymptotically efficient estimates of unknown parameters (e.g., MLEs based on non-grouped data or BAN estimates) are used for testing. Singh (1987), Spruill (1976), and Lemeshko et al. (2001) showed that the NRR test is asymptotically optimal in some sense. This optimality is not surprising because the second term of (2) depends on the difference between Fisher's matrices for grouped and non-grouped data that possibly takes the information lost in full (Voinov (2006)). Dzhaparidze and Nikulin (1992) generalized Fisher's idea to improve any $\sqrt{n}$- consistent estimator to make it asymptotically as efficient as

MLE. This gives the following way of chi-squared test modification: improve an estimator first and then use the NRR statistic. Since this way is not simple computationally, it is worth considering other modifications. At this point it is important to note that the NRR test is very suitable for describing censored data (Habib and Thomas (1986)).

Dzhaparidze and Nikulin (1974) proposed a modification of the standard Pearson's statistic valid for any square root of n consistent estimate $\tilde{\theta}_n$ of an unknown parameter $U_n^2(\tilde{\theta}_n) = \mathbf{V}^T(\tilde{\theta}_n)\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}(\tilde{\theta}_n)$. This test (the DN test), like the asymptotically equivalent Pearson-Fisher one, is not powerful for equiprobable cells (McCulloch (1985), Voinov et al. (2009)) but it can be rather powerful if an alternative hypothesis is specified and one uses the Neyman-Pearson classes for data grouping. Having generalized the idea of Dzhaparidze and Nikulin (1974), Singh (1987) suggested a generalization of the RRN test (3) valid for any $\sqrt{n}$ - consistent estimator $\tilde{\theta}_n$ of an unknown parameter $Q_s^2(\tilde{\theta}_n) = \mathbf{V}_*^T(\tilde{\theta}_n)(\mathbf{I} - \mathbf{B}\mathbf{J}^{-1}\mathbf{B}^T)^{-1}\mathbf{V}_*(\tilde{\theta}_n)$, where $\mathbf{V}_*(\tilde{\theta}_n) = \mathbf{V}(\tilde{\theta}_n) - \mathbf{B}\mathbf{J}^{-1}\mathbf{W}(\tilde{\theta}_n)$, and $\mathbf{W}(\tilde{\theta}_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln f(X_i,\theta)}{\partial\theta}|_{\theta=\tilde{\theta}_n}$.

A unified large-sample theory of general chi-squared statistics for tests of fit was developed by Moore and Spruill (1975). Moore (1977), based upon Wald's approach, formulated a general recipe for constructing modified chi-squared tests for any square root of n consistent estimator that actually is a generalization of Nikulin's idea. He was first to show that a resulting Wald's quadratic form does not depend on the way of limit covariance matrix of generalized frequencies inverting.

Hsuan and Robson (1976) showed that a modified statistic will not be the same as (3) in the case of moment type estimates (MMEs) of unknown parameters. They succeeded in deriving the limit covariance matrix for generalized frequencies and proved the theorem that a corresponding Wald's quadratic form will follow in the limit the chi-squared distribution. Hsuan and Robson provided the test statistic explicitly for the exponential family of distributions, when MMEs coincide with MLEs, thus confirming the already known result of Nikulin (1973). Hsuan and Robson have not derived the general modified test based on MMEs $\tilde{\theta}_n$ explicitly. This was done later by Mirvaliev (2001). Taking into account the input of Hsuan and Robson, and Mirvaliev, this test will be referred to subsequently as the Hsuan-Robson-Mirvaliev (HRM) statistic

$$Y2_n^2(\tilde{\theta}_n) = X_n^2(\tilde{\theta}_n) + R_n^2(\tilde{\theta}_n) - Q_n^2(\tilde{\theta}_n). \qquad (4)$$

Explicit expressions for quadratic forms $R_n^2(\tilde{\theta}_n)$ and $Q_n^2(\tilde{\theta}_n)$ are given, e.g., in Voinov et al. (2009). The approach, based on Wald's transformation, was also used by Bol'shev and Mirvaliev (1978), Nikulin and Voinov (1989), Voinov and Nikulin (1994), and by Chichagov (2006) for minimum variance unbiased estimators (MVUEs).

It is important to mention two types of decompositions of classical and modified chi-squared tests. The first way decomposes a modified test on a sum of the classical Pearson's test and a correcting term that makes the test chi-squared distributed being distribution free in the limit (Nikulin (1973)). A much more important decomposition was first suggested by McCulloch (1985) (see also Mirvaliev (2001)). This is a decomposition of a test on a sum of the DN statistic and an additional quadratic form being asymptotically independent on the DN statistic. Denoting $W_n^2(\theta) = \mathbf{V}^T(\theta)\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}(\theta)$ and $P_n^2(\theta) = \mathbf{V}^T(\theta)\mathbf{B}(\mathbf{J} - \mathbf{J}_g)^{-1}\mathbf{B}^T\mathbf{V}(\theta)$ the decomposition of the NRR statistic (2) in case of MLEs will be $Y1_n^2(\hat{\theta}_n) = U_n^2(\hat{\theta}_n) + \left(W_n^2(\hat{\theta}) + P_n^2(\hat{\theta}_n)\right)$, where $U_n^2(\hat{\theta}_n)$ is asymptotically independent on $\left(W_n^2(\hat{\theta}) + P_n^2(\hat{\theta}_n)\right)$, and on $W_n^2(\hat{\theta})$. The decomposition of the HRM statistic (4) is $Y2_n^2(\tilde{\theta}_n) = U_n^2(\tilde{\theta}_n) + \left(W_n^2(\tilde{\theta}) + R_n^2(\tilde{\theta}_n) - Q_n^2(\tilde{\theta}_n)\right)$, where $U_n^2(\tilde{\theta}_n)$ is asymptotically independent on $\left(W_n^2(\tilde{\theta}) + R_n^2(\tilde{\theta}_n) - Q_n^2(\tilde{\theta}_n)\right)$, but is asymptotically correlated with $W_n^2(\tilde{\theta})$.

The decomposition of a modified chi-squared test on a sum of the DN statistic and an additional term is of importance because the DN test based on non-grouped data is asymptotically equivalent to the Pearson-Fisher's (PF) statistic for grouped data. Hence, that additional term takes into account the Fisher's information lost due to grouping. Later it was shown (Voinov et al. (2009)) that the DN part, like the PF test, is (for equiprobable cells, for example) insensitive to some alternative hypothesis in case of equiprobable cells (fixed or random) and would be sensitive to it for, e.g., non-equiprobable two Neyman-Pearson classes. For equiprobable cells this suggests using the difference between the modified statistic and the DN part that will be the most powerful statistic in case of equiprobable cells (McCulloch (1985), Voinov et al. (2009)). It became clear that the way of sample space partitioning essentially influences power of a test.

Ronald Fisher (1925) was the first to note that "in some cases it is possible to separate the contributions to χ^2 made by the individual degrees of freedom, and so to test the separate components of a discrepancy." Cochran (1954) wrote "that the usual χ^2 tests are often insensitive, and do not indicate significant results when the null hypothesis is actually false" and suggested to "use a single degree of freedom, or a group of degrees of freedom, from the total χ^2," to obtain more powerful and appropriate test. The

problem of implementing the idea of Fisher and Cochran was that decompositions of Pearson's sum and modified test statistics were not known at that time. Anderson (1994) (see also Boero et al. (2004)) was possibly the first who to decompose the Pearson's χ^2 for a simple null hypothesis into a sum of independent χ^2_1 random variables in case of equiprobable grouping cells. A parametric decomposition of Pearson's χ^2 in case of non-equiprobable cells based on ideas of Mirvaliev (2001) was obtained by Voinov et al. (2008) in an explicit form. At the same time Voinov et al. (2008) presented parametric decompositions of NRR and HRM statistics. Voinov (2010) and Voinov and Pya (2010) introduced vector-valued goodness-of-fit tests that, in some cases, can provide a gain in power for specified alternatives.

About the Authors

Vassiliy Voinov is a Professor of the Operations Management and Information Systems Department, Kazakhstan Institute of Management, Economics and Strategic Research (KIMEP). He received his engineering diploma at Tomsk State University; Candidate of Science degree in Kazakh Academy of Science; Doctor of Science degree in Joint Institute for Nuclear Research, Dubna, Moscow region; Professor degree in Kazakh State Polytechnic University, and also in 1998 received a Professor's degree in KIMEP. He has professional experience as an invited professor in statistics at the University Victor Segalen Bordeaux, France. Vassiliy Voinov participated in many international conferences and has more than 120 research papers and books, including: *Unbiased Estimators and Their Applications*, Volume 1: *Univariate Case*, and Volume 2: *Multivariate Case* (with M. Nikulin, Kluwer Academic Publishers: Dordrecht, 1993 and 1996).

Mikhail S. Nikulin (Nikouline) is Professor of Statistics at the University Victor Segalen, Bordeaux 2. He earned his doctorate in the Theory of Probability and Mathematical Statistics from the Steklov Mathematical Institute in Moscow in 1973, under supervision of Professor L. N. Bol'shev. He was Dean of the the Faculty l'UFR "Sciences and Modélisation", the University Victor Segalen (1996–2001) and Head of the Laboratory EA 2961 "Mathematical Statistics and its Applications" (1999–2007). Professor Nikulin has (co-)authored over 250 papers, 28 edited volumes and 13 books, including *A Guide to Chi-Squared Testing* (with P.E. Greenwood, Wiley, 2004), *Probability, Statistics and Modelling in Public Health* (with D. Commenges, Springer, 2005), and was the editor of the volume *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis and Finance* (with N. Balakrishnan, W. Kahle, N. Limnios and C. Huber-Carol, Birkhäuser, 2009). His name is associated with several statistics terms: Dzhaparidze-Nikulin statistic, Rao-Robson-Nikulin statistic, Bagdonavičius-Nikulin model, and Bagdonavičius-Nikulin estimator.

Cross References

►Chi-Square Distribution
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests

References and Further Reading

Anderson G (1994) Simple tests of distributional form. J Economet 62:265–276
Boero G, Smith J, Wallis KF (2004) The sensitivity of chi-squared goodness of-fit tests to the partitioning of data. Economet Rev 23:341–370
Bol'shev LN, Mirvaliev M (1978) Chi-square goodness-of-fit test for the Poisson, binomial, and negative binomial distributions. Theory Probab Appl 23:481–494 (in Russian)
Chernoff H, Lehmann EL (1954) The use of maximum likelihood estimates in tests for goodness of fit. Ann Math Stat 25: 579–589
Chichagov VV (2006) Unbiased estimates and chi-squared statistic for one-parameter exponential family. In: Statistical methods of estimation and hypotheses testing, vol 19. Perm State University, Perm, Russia, pp 78–89
Cohran G (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10:417–451
Dahiya RC, Gurland J (1972) Pearson chi-squared test of fit with random intervals. Biometrica 59:147–153
Dzhaparidze KO, Nikulin MS (1974) On a modification of the standard statistic of Pearson. Theory Probab Appl 19: 851–853
Dzhaparidze KO, Nikulin MS (1992) On evaluation of statistics of chi-square type tests. In: Problem of the theory of probability distributions, vol 12. Nauka, St. Petersburg, pp. 59–90
Fisher RA (1924) The condition under which χ^2 measures the discrepancy between observation and hypothesis. J R Stat Soc 87:442–450
Fisher RA (1925) Partition of χ^2 into its components. In: Statistical methods for research workers. Oliver and Boyd, Edinburg
Habib MG, Thomas DR (1986) Chi-square goodness-of-fit tests for randomly censored data. Ann Stat 14:759–765
Hsuan TA, Robson DS (1976) The goodness-of-fit tests with moment type estimators. Commun Stat Theory Meth A5:1509–1519
Lemeshko BYu, Postovalov SN, Chimitiva EV (2001) On the distribution and power of Nikulin's chi-squared test. Ind Lab 67:52–58 (in Russian)
McCulloch CE (1985) Relationships among some chi-squared goodness of fit statistics. Commun Stat Theory Meth 14:593–603
Mirvaliev M (2001) An investigation of generalized chi-squared type statistics. Academy of Sciences of the Republic of Uzbekistan, Tashkent, Doctoral dissertation
Molinari L (1977) Distribution of the chi-squared test in non-standard situations. Biometrika 64:115–121
Moore DS (1977) Generalized inverses, Wald's method and the construction of chisquared tests of fit. J Am Stat Assoc 72: 131–137

Moore DS, Spruill MC (1975) Unified large-sample theory of general chisquared statistics for tests of fit. Ann Stat 3:599–616
Nikulin MS (1973a) Chi-square test for continuous distributions. Theory Probab Appl 18:638–639
Nikulin MS (1973b) Chi-square test for continuous distributions with shift and scale parameters. Theory Probab Appl 18: 559–568
Nikulin MS, Voinov VG (1989) A chi-square goodness-of-fit test for exponential distributions of the first order. Springer-Verlag Lect Notes Math 1412:239–258
Rao KC, Robson DS (1974) A chi-squared statistic for goodness-of-fit tests within the exponential family. Commun Stat 3: 1139–1153
Roy AR (1956) On χ^2 statistics with variable intervals. Technical report N1, Stanford University, Statistics Department
Singh AC (1987) On the optimality and a generalization of Rao–Robson's statistic. Commun Stat Theory Meth 16, 3255–3273
Spruill MC (1976) A comparison of chi-square goodness-of-fit tests based on approximate Bahadur slope. Ann Stat 2:237–284
Stigler SM (2008) Karl Pearson's theoretical errors and the advances they inspired. Stat Sci 23:261–171
Voinov V (2006) On optimality of the Rao–Robson–Nikulin test. Ind Lab 72:65–70
Voinov V (2010) A decomposition of Pearson–Fisher and Dzhaparidze–Nikulin statistics and some ideas for a more powerful test construction. Commun Stat Theory Meth 39(4):667–677
Voinov V, Pya N (2010) A note on vector-valued goodness-of-fit tests. Commun Stat 39(3):452–459
Voinov V, Nikulin MS, Pya N (2008) Independently distributed in the limit components of some chi-squared tests. In: Skiadas CH (ed) Recent advances in stochastic modelling and data analysis. World Scientific, New Jersey
Voinov V, Pya N, Alloyarova R (2009) A comparative study of some modified chi-squared tests. Commun Stat Simulat Comput 38:355–367

Chi-Square Test: Analysis of Contingency Tables

David C. Howell
Professor Emeritus
University of Vermont, Burlington, VT, USA

The term "chi-square" refers both to a statistical distribution and to a hypothesis testing procedure that produces a statistic that is approximately distributed as the ▶chi-square distribution. In this entry the term is used in its second sense.

Pearson's Chi-Square

The original chi-square test, often known as Pearson's chi-square, dates from papers by Karl Pearson in the earlier 1900s. The test serves both as a "goodness-of-fit" test, where the data are categorized along one dimension, and as a test for the more common "contingency table," in which categorization is across two or more dimensions. Voinov and Nikulin, this volume, discuss the controversy over the correct form for the goodness of fit test. This entry will focus on the lack of agreement about tests on contingency tables.

In 2000 the Vermont State legislature approved a bill authorizing civil unions. The vote can be broken down by gender to produce the following table, with the expected frequencies given in parentheses. The expected frequencies are computed as $R_i \times C_j/N$, where R_i and C_j represent row and column marginal totals and N is the grand total.

	Vote		
	Yes	No	Total
Women	35 (28.83)	9 (15.17)	44
Men	60 (66.17)	41 (34.83)	101
Total	95	50	145

The standard Pearson chi-square statistic is defined as

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(35 - 28.83)^2}{28.83} + \cdots + \frac{(41 - 34.83)^2}{34.83} = 5.50$$

where i and j index the rows and columns of the table. (For the goodness-of-fit test we simply drop the subscript j.) The resulting test statistic from the formula on the left is approximately distributed as χ^2 on $(r-1)(c-1)$ degrees of freedom. The probability of $\chi^2 \geq 5.50$ on 1 df = 0.019, so we can reject the null hypothesis that voting behavior is independent of gender. (Pearson originally misidentified the degrees of freedom, Fisher corrected him, though Pearson long refused to recognize the error, and Pearson and Fisher were enemies for the rest of their lives.)

Likelihood Ratio Chi-Square

Pearson's chi-square statistic is not the only chi-square test that we have. The likelihood ratio chi-square builds on the likelihood of the data under the null hypothesis relative to the maximum likelihood. It is defined as

$$G^2 = 2\sum O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) = 2\left[35\ln\left(\frac{35}{28.83}\right) + 9\ln\left(\frac{9}{15.17}\right) + 60\ln\left(\frac{60}{66.17}\right) + 41\ln\left(\frac{41}{34.83}\right)\right] = 5.81$$

This result is slightly larger than the Pearson chi-square of 5.50. One advantage of the likelihood ratio chi-square is that G^2 for a large dimensional table can be neatly decomposed into smaller components. This cannot be done exactly with Pearson's chi-square, and G^2 is the usual statistic for log-linear analyses. As sample sizes increase the two chi-square statistics converge.

Small Expected Frequencies

Probably no one would object to the use of the Pearson or likelihood ratio chi-square tests for our example. However, the chi-square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When we have very small expected frequencies, the possible values of the chi-square statistic are quite discrete. For example, for a table with only four observations in each row and column, the only possible values of chi-square are 8, 2, and 0. It should be clear that a continuous chi-square distribution is not a good match for a discrete distribution having only three values. The general rule is that the smallest expected frequency should be at least five. However Cochran (1952), who is generally considered the source of this rule, acknowledged that the number "5" seems to be chosen arbitrarily.

Yates proposed a correction to the formula for chi-square to bring it more in line with the true probability. However, given modern computing alternatives, Yates' correction is much less necessary and should be replaced by more exact methods.

For situations in which we do not satisfy Cochran's rule about small expected frequencies, the obvious question concerns what we should do instead. This is an issue over which there has been considerable debate. One of the most common alternatives is Fisher's Exact Test (see below), but even that is controversial for many designs.

Alternative Research Designs

There are at least four different research designs that will lead to data forming a contingency table. One design assumes that all marginal totals are fixed. Fisher's famous "tea-tasting" study had four cups of tea with milk added first and four with milk added second (row totals are fixed). The taster had to assign four cups to each guessed order of pouring, fixing the column totals. The underlying probability model is hypergeometric, and Fisher's exact test (1934) is ideally suited to this design and gives an exact probability. This test is reported by most software for 2×2 tables, though it is not restricted to the 2×2 case.

Alternatively we could fix only one set of marginals, as in our earlier example. Every replication of that experiment would include 44 women and 101 men, although the vote totals could vary. This design is exactly equivalent to comparing the proportion of "yes" votes for men and women, and it is based on the ▶binomial distribution. The square of a z-test on proportions would be exactly equal to the resulting chi-square statistic. One alternative analysis for this design would be to generate all possible tables with those row marginals and compute the percentage of obtained chi-square statistics that are as extreme as the statistic obtained from the actual data. Alternatively, some authorities recommend the use of a mid-p value, which sums the probability of all tables less likely than the one we obtained, plus half of the probability of the table we actually obtained.

For a different design, suppose that we had asked 145 Vermont citizens to record their opinion on civil unions. In this case neither the Gender nor Vote totals would be fixed, only the total sample size. The underlying probability model would be multinomial. Pearson's chi-square test would be appropriate, but a more exact test would be obtained by taking all possible tables (or, more likely, a very large number of randomly generated tables) with 145 observations and calculating chi-square for each. Again the probability value would be the proportion of tables with more extreme outcomes than the actual table. And, again, we could compute a mid-p probability instead.

Finally, suppose that we went into college classrooms and asked the students to vote. In this case not even the total sample size is fixed. The underlying probability model here is Poisson.

Computer scripts written in R are available for each model with a fixed total sample size at http://www.uvm.edu/~dhowell/StatPages/chi-square-alternatives.html

Summary

Based on a large number of studies of the analysis of contingency tables, the current recommendation would be to continue to use the standard Pearson chi-square test whenever the expected cell frequencies are sufficiently large. There seems to be no problem defining large as "at least 5." With small expected frequencies ▶Fisher's Exact Test seems to perform well regardless of the sampling plan, but ▶randomization tests adapted for the actual research design, as described above, will give a somewhat more exact solution. Recently Campbell (2007) carried out a very large sampling study on 2×2 tables comparing different chi-square statistics under different sample sizes and different underlying designs. He found that across all sampling designs, a statistic suggested by Karl Pearson's son Egon Pearson worked best in most situations. The statistic is defined as $\chi^2 \frac{N}{N-1}$. (For the justification for that adjustment see Campbell's paper.) Campbell found that as

long as the smallest expected frequency was at least one, the adjusted chi-square held the Type I error rate at very nearly α. When the smallest expected frequency fell below 1, Fisher's Exact Test did best.

About the Author

David Howell is a Professor Emeritus (since 2002), and former chair of the Psychology department at the University of Vermont (1987–1992) and (2000–2002). Professor Howell's primary area of research is in statistics and experimental methods. He has authored well known texts: *Statistical Methods for Psychology* (Wadsworth Publishing, 7*th* ed., 2009), *Fundamental Statistics for Behavioral Sciences* (Wadsworth Publishing, 7*th* ed., 2010), and is a coauthor (with Brian Everitt) of a four volume *Encyclopedia of Statistics in Behavior Science* (Wiley & Sons, 2005).

Cross References

►Chi-Square Distribution
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Tests

References and Further Reading

Campbell I (2007) Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Stat Med 26:3661–3675

Cochran WG (1952) The χ^2 test of goodness of fit. Ann Math Stat 25:315–345

Fisher RA (1934) The logic of inductive inference. J R Stat Soc 98: 39–54

Chi-Square Tests

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

The χ^2 statistic was developed by Karl Pearson (1900) as a means to compare an obtained distribution of scores with a theoretical distribution of scores. While it is sometimes still employed as a univariate goodness of fit test, other statistics, such as the ►Kolmogorov–Smirnov test and, where the theoretical distribution is normal, the Shapiro–Wilk test, are now more often used for that purpose.

The chi-square statistic on n degrees of freedom is defined as

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 = \sum \frac{(Y-\mu)^2}{\sigma^2},$$

where z_i is normally distributed with mean zero and standard deviation one (Winkler and Hayes 1975, pp. 375–380). If one were repeatedly to draw samples of one Y score from a normally distributed population, transform that score to a standard z score, and then square that z score, the resulting distribution of squared z scores would be a χ^2 distribution on one degree of freedom. If one were repeatedly to draw samples of three scores, standardize, square, and sum them, the resulting distribution would be χ^2 on three degrees of freedom. Because the χ^2 statistic is so closely related to the normal distribution, it is also closely related to other statistics that are related to the normal distribution, such as t and F.

One simple application of the χ^2 statistic is to test the null hypothesis that the variance of a population has a specified value (Winkler and Hayes 1975, pp. 453–455; Wuensch 2009). From the definition of the sample variance, $s^2 = \frac{\sum(Y-M)^2}{N-1}$, where Y is a score, M is the sample mean, and N is the sample size, the corrected sum of squares $\sum(Y-M)^2 = (N-1)s^2$. Substituting this expression for $\sum(Y-\mu)^2$ in the defining formula yields $\chi^2 = \frac{(N-1)s^2}{\sigma^2}$. To test the hypothesis that an observed sample came from a population with a particular variance, one simply divides the sample sum of squares, $(N-1)s^2$, by the hypothesized variance. The resulting χ^2 is evaluated on $N-1$ degrees of freedom, with a two-tailed p value for nondirectional hypotheses and a one-tailed p for directional hypotheses.

One can also compute a confidence interval for the population variance (Winkler and Hayes 1975, pp. 383–385; Wuensch 2009). For a $100(1-\alpha)\%$ confidence interval for the population variance, compute:

$$\frac{(N-1)s^2}{b} \quad \text{and} \quad \frac{(N-1)s^2}{a}$$

where a and b are the $\alpha/2$ and $(1-\alpha/2)$ fractiles of the chi square distribution on $(N-1)df$. It should be noted that these procedures are not very robust to their assumption that the population is normally distributed.

When one states that he or she has conducted a "chi-square test," that test is most often a "one-way chi-square test" or a "two-way chi-square test" (Howell 2010, pp. 141–151). The one-way test is a univariate goodness of fit test. For each of k groups one has an observed frequency (O) and a theoretical frequency (E), the latter being derived from the theoretical model being tested. The test

statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ on $k-1$ degrees of freedom. The appropriate p value is one-tailed, upper-tailed, for nondirectional hypotheses. When $k = 2$, one should make a "correction for continuity":

$$\chi^2 = \sum \frac{(|O-E| - .5)^2}{E}.$$

The two-way chi-square test is employed to test the null hypothesis that two categorical variables are independent of one another. The data may be represented as an $r \times c$ contingency table, where r is the number of rows (levels of one categorical variable) and c is the number of columns (levels of the other categorical variable). For each cell in this table two frequencies are obtained, the observed frequency (O) and the expected frequency (E). The expected frequencies are those which would be expected given the marginal frequencies if the row variable and the column variable were independent of each other. These expected frequencies are easily calculated from the multiplication rule of probability under the assumption of independence. For each cell, the expected frequency is $(R_i C_j / N)$, where R_i is the marginal total for all cells in the same row, C_j is the marginal total for all cells in the same column, and N is the total sample size. The χ^2 is computed exactly as with the one-way chi-square and is evaluated on $(r-1)(c-1)$ degrees of freedom, with an upper-tailed p value for nondirectional hypotheses. Although statistical software often provides a χ^2 with a correction for continuity when there are only two rows and two columns, almost always the uncorrected χ^2 is more appropriate (Camilli and Hopkins 1978).

It is not unusual to see the two-way chi-square inappropriately employed (Howell 2010, pp. 152–153). Most often this is a result of having counted some observations more than once or having not counted some observations at all. Each case should be counted once and only one. Statistical software will often provide a warning if one or more of the cells has a low expected frequency. The primary consequence of low expected frequencies is low power. Even with quite small expected frequencies, actual Type I error rates do not deviate much from the nominal level of alpha (Camilli and Hopkins 1978).

The results of a two-way chi-square test are commonly accompanied by an estimate of the magnitude of the association between the two categorical variables. When the contingency table is 2 × 2, an odds ratio and/or the phi coefficient (Pearson r between the two dichotomous variables) may be useful. With larger contingency tables Cramer's phi statistic may be useful.

The chi-square statistic is also employed in many other statistical procedures, only a few of which will be mentioned here. The Cochran-Mantel-Haenszel χ^2 is employed to test the hypothesis that there is no relationship between rows and columns when you average across two or more levels of a third variable. The Breslow-Day χ^2 is employed to test the hypothesis that the odds ratios do not differ across levels of a third variable. Likelihood ratio chi-square is employed in the log-linear analysis of multidimensional contingency tables, where it can be employed to test the difference between two models, where one is nested within the other. Likewise, in ►logistic regression, chi-square can be employed to test the effect of removing one or more of the predictors from the model. In discriminant function analysis, chi-square may be employed to approximate the p value associated with the obtained value of Wilks' lambda. A chi-square statistic can be employed to test the null hypothesis that k Pearson correlation coefficients are identical. Chi-square is also used to approximate the p value in the Kruskal-Wallis ANOVA and the Friedman ANOVA. Many more uses of the chi-square statistic could be cited.

About the Author

Dr. Karl L. Wuensch is a Professor in the Department or Psychology, East Carolina University, Greenville, NC, U.S.A. He has authored or coauthored 77 scholarly articles and chapters in books. Professor Wuensch has received several teaching awards, including the Board of Governors Award for Excellence in Teaching, the most prestigious teaching award in the University of North Carolina system.

Cross References

►Chi-Square Distribution
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables

References and Further Reading

Camilli G, Hopkins KD (1978) Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. Psychol Bull 85:163–167

Howell DC (2010) Statistical methods for psychology, 7th edn. Cengage Wadsworth, Belmont

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. The London, Edinburgh and Dublin Phil Mag J Sci Ser 5 50:157–175. Retrieved from www.economics.soton.ac.uk/staff/aldrich/1900.pdf

Winkler RL, Hays WL (1975) Statistics: probability, inference, and decision, 2nd edn. Holt Rinehart & Winston, New York

Wuensch KL (2009) Common univariate and bivariate applications of the chi-square distribution. Retrieved from http://core.ecu.edu/psyc/wuenschk/docs30/Chi-square.doc

Clinical Trials, History of

Toshimitsu Hamasaki
Associate Professor
Osaka University Graduate School of Medicine, Osaka, Japan

In recent years, in addition to advances in methodology, the number of clinical trials conducted and published has greatly increased. Clinical trials, in particular, *blinded, randomized, controlled* comparative clinical trials, are widely recognized as the most scientific and reliable method for evaluating the effectiveness of therapies and promoting a culture of evidence-based medicine (Tukey 1977; Byar et al. 1976; Zelen 1979; Cowan 1981; Byar 1991; Royall 1991; Smith 1998).

The first modern clinical trial is generally considered to be the treatment of pulmonary tuberculosis with streptomycin conducted by the UK Medical Research Council (MRC) and published in *British Medical Journal* in 1948 (MRC 1948; Pocock 1984; Ederer 2005; Day 2006). However, there is still some controversy surrounding this claim as some authors refer to the study with the common cold vaccine conducted by Diehl et al. (1938) as the first modern trial (Hart 1972, 1996; Gill 1996). The design of the streptomycin trial included blinding, ►randomization, and control groups as fundamental elements of the clinical trial. The trial included a total of 107 patients from seven centers, who were assigned to either "streptomycin and bed-rest" (S case) or "bed-rest" (C case) groups, by a process involving a statistical series based on random sampling numbers drawn up for each sex and each center and sealed envelopes. The efficacy of streptomycin was evaluated based upon the examination of patient X-ray films by three experts consisting of one clinician and two radiologists. The decision of whether or not the treatment was effective was made by the majority based on independently reached conclusions by each expert, who were also blinded as to which treatment the patient had received. The streptomycin trial also included Sir Austin Bradford Hill who served as the trial statistician. Hill was recognized as the world's leading medical statistician and popularized the use of statistical methods in clinical trials, and who also attempted to improve the quality of their implementation and evaluation by publishing a series of 17 articles in *The Lancet* in 1937 (Hill et al. 2000).

With the success of the streptomycin trial, the MRC and Hill continued with further blinded, randomized, controlled comparative clinical trials (Ederer 2005; Days 2006): for example, chemotherapy of pulmonary tuberculosis in young adults (MRC 1952), an antihistaminic drug in the prevention and treatment of the common cold (MRC 1950), the use of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis (MRC 1954, 1955), and an anticoagulant to treat cerebrovascular disease (Hill et al. 1960). In United States, the first randomized controlled trail started in 1951 and was the US National Institute of Health study of the adrenocorticotropic hormone, cortisone and aspirin in the treatment of rheumatic heart disease in children (Rheumatics Fever Working Party 1960). Presently, a huge number of randomized controlled clinical trials are being conducted worldwide, with the number of clinical trials steadily increasing each year.

Although now commonplace, the fundamental elements of clinical trials, such as blinding, randomization, and control groups, did not just suddenly appear in the second quarter of the twentieth century. Evidence exists that a comparative concept for evaluating therapeutic efficacy with control groups has been known since ancient times (Ederer 2005; Day 2006). For example, Lilienfeld (1949) and Slotki (1951) cited the description of a nutritional experiment using a control group in the Book of Daniel from the Old Testament:

> **1.6**: Among these were some from Judah: Daniel, Hananiah, Mishael and Azariah. . . **1.8**: But Daniel resolved not to defile himself with the royal food and wine, and he asked the chief official for permission not to defile himself this way. . . . **1.11**: Daniel then said to the guard whom the chief official had appointed over Daniel, Hananiah, Mishael and Azariah. **1.12**: Please test your servants for ten days; Give us nothing but vegetables to eat and water to drink. **1.13**: Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see. **1.14**: So he agreed to this and tested them for ten days. **1.15**: At the end of the ten days they looked healthier and better nourished than any of the young men who ate the royal food. **1.16** So the guard took away their choice food and the wine they were to drink and gave them vegetables instead.

The above description is part of a story dating from approximately 800 BC when Daniel was taken captive by the ruler of Babylonia, Nebuchadnezzar. In order to refrain from eating royal meals containing meat (perhaps pork) and wine offered by Nebuchadnezzar, Daniel proposed a comparative evaluation and was rewarded when his test group fared better than the royal food group. Although it is difficult to confirm the accuracy of the account, it is clear that the comparative concept already existed when the Book of Daniel was written around 150 BC. In particular, it is

remarkable that the passage from the Book of Daniel mentioned not only the *choice of a control group* but the *use of a concurrent control group*. Unfortunately, this fundamental concept was not widely practiced until the latter half of the twentieth century (Ederer 2005; Day 2006).

Much later than the Book of Daniel, in the eighteenth and nineteenth centuries, there were some epoch-making clinical researches that formed the basis of the methodology used in current clinical trials. Before the modern clinical trial of the treatment of pulmonary tuberculosis with streptomycin mentioned above (Pocock 1984; Ederer 2005; Day 2006), the most famous historical example of a planned, controlled clinical trial involved six dietary treatments for scurvy on board a British ship. The trial was conducted by the ship's surgeon, James Lind, who was appalled by the ravages of scurvy which had claimed the lives of three quarters of the crew during the circumnavigation of the world by British admiral, George Anson (Lind 1753; Bull 1959; Pocock 1984; Mosteller 1981; Ederer 2005; Day 2006). In 1947, Lind conducted a comparative trial to establish the most promising "cure" for patients with scurvy using twelve individuals who had very similar symptoms on board the Salisbury. In addition to one common dietary supplement given to all of the patients, he assigned each of six pairs one of the following six dietary supplements:

1. Six spoonfuls of vinegar
2. A half-pint of sea water
3. A quart of cider
4. Seventy-five drops of vitriol elixir
5. Two oranges and one lemon
6. Nutmeg

Those patients who received the two oranges and one lemon were cured within approximately 6 days and were able to help nurse the other patients. Apart from the patients who improved somewhat after receiving the cider, Lind observed that the other remedies were ineffective. The reason for the success of Lind's trial was likely due to his knowledge of previous work by James Lancaster (Purchas 1625), who had served three teaspoons of lemon juice each day to sailors suffering from scurvy during the first expedition to India sent by the East India Company in 1601 (Mosteller 1981). Unfortunately, however, the British Navy did not supply lemon juice to its sailors until 1975, although conclusive results concerning the efficacy of such treatment had already been obtained much earlier (Bull 1959; Mosteller 1981).

The use of statistical concepts in clinical trials was also advocated earlier than the streptomycin trials. For example, Pierre Simon Laplace, a French mathematician and astronomer, mentioned the use of probability theory to determine the best treatment for the cure of a disease (Laplace 1814; Hill et al. 2000). Also, Pierre-Charles-Alexandre Louis, a French physician and pathologist, discussed the use of a "numerical method" for the assessment of treatments by constructing comparable groups of patients with similar degrees of a disease, i.e., to compare "like with like" (Louis 1837; Ederer 2005; Day 2006). Unfortunately, these suggestions were not earnestly acted upon until the streptomycin trial because in the eighteenth and nineteenth centuries, the investigators were more involved with the practice of medicine and less versed in the use of probability theory since saving patients' life was considered more important rather than collecting data from the aspect of ethics (Bull 1959; Hill et al. 2000).

Here, the history and development of clinical trials was very briefly traced. More detailed aspects of the history of clinical trials can be found in articles by Bull (1959), Armitage (1972, 1991), Lilienfeld (1982), Pocock (1984), Meinert (1986), Gail (1996), Ederer (2005) and Day (2006).

About the Author

Toshimitsu Hamasaki is Associate Professor of Department of Biomedical Statistics, Osaka University Graduate School of Medicine. He is the Elected member of International Statistical Institute. He has over 50 peer-reviewed publications roughly evenly split between applications and methods. He was awarded the Best Paper Prize (the Japanese Society of Computational Statistics, 1997) and Hida-Mizuno Prize (the Behaviormetric Society of Japan, 2003). He is the Editor-in-Chief of the *Journal of the Japanese Society of Computational Statistics* (2007–2010).

Cross References

►Biostatistics
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Design of Experiments: A Pattern of Progress
►Medical Research, Statistics in
►Medical Statistics

References and Further Reading

Armitage P (1972) History of randomised controlled trials. Lancet 299:1388

Armitage P (1991) Interim analysis in clinical trials. Stat Med 10: 925–937

Bayar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellen-berg JH et al (1976) Randomized clinical trials: perspective on some recent idea. New Engl J Med 295:74–80

Bull JP (1959) The historical development of the clinical trials. J Chron Dis 10:218–248

Byar DP (1991) Comment on "Ethics and statistics in randomized clinical trials" by R.M. Royall. Stat Sci 6:65–68
Cowan DH (1981) The ethics of trials of ineffective therapy. IRB: Rev Human Subjects Res 3:10–11
Day S (2006) The development of clinical trials. In: Machin D, Day S, Green S (eds) Textbook of clinical trials, 2nd edn. Wiley, Chichester, pp 5–11
Diehl HS, Baker AB, Cowan DW (1938) Cold vaccines: an evaluation based on a controlled study. J Am Med Assoc 11: 1168–1173
Ederer F (2005) Clinical trials, history of. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn. Wiley, Chichester, pp 864–874
Gail MH (1996) Statistics in action. J Am Stat Assoc 91:1–13
Gill DBEC (1996) Early controlled trials. Brit Med J 312:1298
Hart PD (1972) History of randomised controlled trials. Lancet 299:965
Hart PD (1996) Early controlled clinical trials. Brit Med J 312: 378–379
Hill AB, Marshall J, Shaw DA (1960) A controlled clinical trial of long-term anticoagulant therapy in cerebrovascular disease. Q J Med 29:597–608
Hill G, Forbes W, Kozak J, MacNeil I (2000) Likelihood and clinical trials. J Clin Epidemiol 53:223–227
Laplace PS (1814) Théori analytique des probabilités. Courcier, Paris
Lilienfeld AM (1949) Ceteris paribus: the evaluation of the clinical trial. Bull Hist Med 56:1–18
Lind J (1753) A treatise of the scurvy. Sands Murray & Cochran, Edinburgh
Louis PCA (1837) The applicability of statistics to the practice of medicine. London Medical Gazette 20:488–491
Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. Brit Med J 2:769–782
Medical Research Council (1950) Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold. Brit Med J 2:425–431
Medical Research Council (1952) Chemotherapy of pulmonary tuberculosis in young adults. Brit Med J 1:1162–1168
Medical Research Council (1954) A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis I. Brit Med J 1:1223–1227
Medical Research Council (1955) A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis II. Brit Med J 2:695–700
Meinert CL (1986) Clinical trials: design, conduct and analysis. Oxford University, New York
Mosteller F (1981) Innovation and evaluation. Science 211: 881–886
Pocock SJ (1984) Clinical trials: a practical approach. Wiley, Chichester
Purchas S (1625) Hakluytus posthumus or purchas his pilgrimes: contayning a histrory of the world in sea voyages and lande travells by englishmen and others. James MacLehose & Sons, Glasgow (reprinted, 1905)
Rheumatics Fever Working Party (1960) The evaluation of rheumatic heart disease in children: five years report f a co-operative clinical trials of ACTH, cortisone, and aspirin. Circulation 22:505–515
Royall RM (1991) Ethics and statics in randomized clinical trials. Stat Sci 6:52–88
Slotki JJ (1951) Daniel, Ezra, Nehemiah, Hebrew: text and english translation with introductions and commentary. Soncino Press, London
Smith R (1998) Fifty years of randomized controlled trials. Brit Med J 317:1166
Tukey JW (1977) Some thoughts on clinical trials, especially problems of multiplicity. Science 198:679–684
Zelen M (1979) A new design for randomized clinical trials. N Engl J Med 300:1242–1246

Clinical Trials: An Overview

Hiroyuki Uesaka
Osaka University, Osaka, Japan

A clinical trial is one type of clinical research where a procedure or drug is intentionally administered outside the realm of standard medical practice to human subjects with the aim of studying its effect on the human body. This includes medications, operations, psychotherapy, physiotherapy, rehabilitation, nursing, restricted diets, and the use of medical devices. The comparative study of two or more treatments, involving the random assignment of treatments to patients, is considered a clinical trial even if the study includes approved drugs or medical devices. This means that a clinical trial is an experiment which includes human subjects. It is necessary to distinguish clinical trials from observational studies which collect outcomes when executing a study treatment as an ordinary treatment.

Since clinical trials include human subjects, the ethical aspects, i.e., the rights, safety and well-being of individual research subjects, should take precedence over all other interests at all stages, from the planning of clinical trials to the reporting of results. Such ethical principles for clinical research are in accordance with the Declaration of Helsinki, "Ethical Principles for Medical Research Involving Human Subjects," issued by the World Medical Association (1964). In conducting a clinical trial, the study protocol should clearly describe the plan and content of the trial. The protocol must also be reviewed and approved by the ethics committee. Furthermore, the Declaration of Helsinki states: The protocol should contain a statement of the ethical considerations involved and indicate how the principles in the above declaration have been addressed. To protect the safety, well-being and rights of the human subjects participating in the trial, the Declaration indicates that potential subjects must be adequately informed of all

relevant aspects of the study which include aims, methods, the anticipated benefits and potential risks of the study and any discomfort that participation may entail. And it states: The potential subjects must be informed of their right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal. The voluntary agreement of a subject to participate after sufficient details have been provided is called informed consent. It is also recommended that the clinical trial be registered in a publicly accessible database, and the results from the trial should be made publicly available, regardless of whether the results are positive or negative.

Clinical trials for a new drug application, when they are conducted in the EU, Japan and/or the United States of America, must meet the requirements of the "Good Clinical Practice" (GCP) guideline (ICH Steering Committee 1996). The GCP guideline is a unified standard provided by the Europe, Japan and the United States in the framework of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use [http://www.ich.org/]. The GCP guideline provides protection for the safety, well-being and human rights of subjects in clinical trials in accordance with the Declaration of Helsinki. The GCP guideline also requires that people appointed by the sponsor, the so-called monitors, verify that the rights and well-being of all human subjects are being protected, that the reported trial data are accurate, complete, and verifiable from source documents, and that the conduct of the trial is in compliance with the currently approved protocol/amendment(s), with the GCP, and with the applicable regulatory requirement(s). This is referred to as trial monitoring in the GCP.

In planning a clinical trial, a protocol must be prepared, including descriptions of the trial justification, trial objectives, study treatments, the population to be studied as defined by the study inclusion and exclusion criteria, test treatments and treatment procedures, observed variables and observation procedures, specification of variables to assess treatment effect, collection of safety information, prohibited concomitant medications, discontinuation criteria for individual subjects, the number of subjects planned to be enrolled and the justification for such, statistical methods to be employed, data collection, quality control and quality assurance (ICH Steering Committee 1997). A case report form (CRF) should be prepared as well. The CRF is a document designed to record all of the required information to be reported according to the protocol. After a trial, a so-called clinical study report is prepared (ICH Steering Committee 1995). This is a document which contains clinical and statistical descriptions of the methods, rationale, results and analyzes of a specific clinical trial fully integrated into a single report. Clinical trials are conducted as collaborative activities involving many specialists, such as investigators, nurses, diagnostic testing specialists and other collaborators. Furthermore, regulators are involved in new drug applications. Therefore, the protocol, CRF and clinical study report should be clearly and accurately documented to be easily understood by those involved in the trial and by those who will make use of the trial results.

Clinical trials can be classified into several types depending on various features (ICH Steering Committee 1998; ICH Steering Committee 2000). First, a trail can be controlled or uncontrolled, this being determined by the presence of a control group. A controlled trial is a trial to compare the study treatment(s) with a control treatment that is either the current standard treatment, best supportive care, placebo, or some other treatment; an uncontrolled trial involves giving the same treatment to all of the subjects participating in the trial. The second feature involves the objective of a trial, either exploratory or confirmatory. A clinical trial that aims to generate or identify a research topic, or provide information to determine the specifics of a trial method is called an exploratory trial. A confirmatory trial is defined as an adequately controlled trial where hypotheses which were derived from earlier research or theoretical considerations are stated in advance and evaluated. Furthermore, a confirmatory trial generally includes three types of comparisons: a superiority trial, a non-inferiority trial, and an equivalence trial. A superiority trial is used to show the superiority of a test treatment over a control. A non-inferiority trial is designed to show that the efficacy or safety of the study treatment is no worse than that of the control. An equivalence trial serves to demonstrate that the test treatment is neither better nor worse than the control. The third aspect involves distinguishing between a pragmatic and an explanatory trial (Gent and Sackett 1979; Schwartz and Lellouch 1967). The objective of a pragmatic trial is to confirm effectiveness of the test treatment for those subjects who are assigned to the test treatment. An explanatory trial serves to establish a biological action for the treatment. Finally, the fourth characteristic focuses on the difference between a single- and a multi-center trial. The single-center trial is conducted by a single investigator, and the multi-center trial is co-conducted by multiple investigators at multiple study sites. Recently, many multi-center trials have been planned and conducted across not only a single country but also two or more countries. Such a multi-center trial is called a multinational trial.

The clinical development of a new drug advances in stages (ICH Steering Committee 1997). A safety trial is executed first to determine the maximum dose that can be safely administered to a subject. In most safety trials of the first use of a new drug in humans, the subjects are healthy volunteers. Administration of the study treatment begins from a dosage expected to be safe enough for normal healthy volunteers, and then the dosage is increased in stages. The pharmacokinetic profile is usually examined in the same trial. Pharmacokinetics investigates the process of drug disposition which usually consists of absorption, distribution, metabolism and excretion. This stage is called Phase I. The next stage is to determine the dosage range that can be safely administered to patients and at which sufficient effectiveness can be expected. The dosage that will be used in clinical treatment as well as the dose intervals are also clarified at this stage. This is called the Phase II. In the third stage, the efficacy and safety of the study treatment is confirmed in the target patient population. This stage is referred to as Phase III. The dose and dosage regimen which are confirmed to be efficacious and safe in phase III are then submitted to the regulatory authority to obtain marketing authorization of the new drug. After marketing authorization is obtained, the drug becomes widely used for clinical treatment. This stage is called Phase IV. During the phase III trials many restrictions are imposed to ensure the safety of the participating subjects. These include the necessity of physical examinations, collection of patient anamneses, regulation of concomitant medications, and clearly defined test treatment administration periods. However, in phase IV such restrictions are relaxed and the approved study treatment can be used by various patients under diverse conditions. Therefore, because the number of patients who are administered the newly approved drug increases rapidly, with patients often using the drug for very long times according to their disease condition, there is a real concern about harmful effects that have not been anticipated. Therefore, an investigation to clarify the safety and effectiveness of the treatment in daily life, an observational study, a large-scale trial, or a long-term trial, is conducted. Moreover, a clinical trial to compare the newly approved drug with other medicines that have been approved for the same indication may also be conducted.

The result of the trial should be scientifically valid. Clinical trial results are intended to be applied to a target population defined by inclusion and exclusion criteria for a given trial. The enrolled subjects should be a random sample from the target population so that the trial results can be applied to the target population. However a trial is conducted in a limited number of medical sites, and not all candidate subjects give informed consent. Therefore, whether or not the trial result can be generalized to the target population will depend on the study protocol and the actual execution procedure. Accordingly, it is preferable to execute the trial in a variety of medical institutions with a wide range of patients corresponding to the diversity of the target population to improve the possibility of generalizing to the target population. A controlled trial usually estimates the difference in response to treatments between treatment groups. As described above, the clinical trial participants are not a random sample of the target population. Therefore the true mean difference in the study population (all subjects who participate in the trial) will be estimated. This is accomplished by dividing the study population into two or more treatment groups which are assigned to different treatments, and then comparing the means of response to treatment between groups. The estimated mean difference is usually different from this true value. When random allocation of treatment to subjects is used, it is assumed that the departure from the true difference is probabilistic or random error. However, there is the possibility of systematic error due to the execution procedure of the trial. This systematic error is called bias (ICH Steering Committee 1998). The execution of treatment, evaluation of results, and/or subjects' reactions can be influenced if the people involved in a trial, such as investigators, relevant clinical staff or subjects, are aware of which treatment is assigned to subjects. Therefore, masking (blinding) and randomization are used to prevent participants from knowing which treatment is being allocated to which subjects. There are several levels of blinding: double-blind, single-blind, observer-blind and open-label. In a double-blind study neither the subjects, nor the investigator, nor any of the relevant clinical trial staff know who belongs to which treatment group. In a single-blind study only treatment assignments are unknown to the subjects or investigator and relevant clinical staff. In an observer-blind study treatment assignments are unknown to the observers who assess the subjects' conditions. In an open-label study treatment assignments are known to both investigators and subjects.

One of the typical methods of treatment assignment is to assign only one treatment to each subject, and then to compare the effects of the treatments between subject groups. This method is referred to as parallel group design. The other typical method is the cross-over design where one subject receives two or more treatments and an intra-subject comparison of treatments is done. It is necessary to select an appropriate design because bias can be caused by the design itself.

A clinical trial is an experiment with human beings as subjects. It is preferable that the number of subjects be as small as possible to protect the rights, health and welfare

of the subjects included in the trial. However, if the objective of the trial is not achieved, the reason for executing the trial is lost. Therefore, based on the estimated difference between the treatments, the trial should be designed to have sufficient precision to either detect such a difference if it truly exists, or to conclude that the difference is below a definite value based on concrete evidence. For this purpose, it is necessary to maintain high accuracy and precision in trials. To ensure the precision of a trial, it is important to consider the stratification of the study population, to make precise observations, and to secure a sufficient number of subjects.

The objective of these trials is to estimate beneficial and adverse effects, and to confirm a hypothesis about the effect of the study treatment. Even if the effect size of the test treatment is assumed to be of a given size, the true effect size may be less than assumed. When the gap between the actual and the assumed value is large, the planned number of subjects might be insufficient and, in some cases, many more subjects than originally planned will be needed. In such cases, a sequential design (Jennison and Turnbull 2000) and a more advanced adaptive design (Bretz et al. 2009) would be proposed.

About the Author

Dr. Hiroyuki Uesaka is a specially appointed Professor of The Center for Advanced Medical Engineering and Informatics, Osaka University, Suita, Japan. He has been working for pharmaceutical companies for about 40 years as a statistical expert of clinical trials. He authored more than 20 original statistical papers, and authored, coauthored and contributed to 6 books written in Japanese (including *Design and Analysis of Clinical Trials for Clinical Drug Development* (Asakura-Shoten, 2006) and *Handbook of Clinical Trials* (jointly edited with Dr. Toshiro Tango, Asakura-Shoten, 2006).

Cross References

►Biopharmaceutical Research, Statistics in
►Biostatistics
►Causation and Causal Inference
►Clinical Trials, History of
►Clinical Trials: Some Aspects of Public Interest
►Design of Experiments: A Pattern of Progress
►Equivalence Testing
►Hazard Regression Models
►Medical Research, Statistics in
►Medical Statistics
►Randomization
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines
►Statistics: Controversies in Practice
►Statistics: Nelder's view

References and Further Reading

Bretz F, Koenig F, Brannath W, Glimm E, Posch M (2009) Adaptive designs for confirmatory clinical trials. Stat Med 28:1181–1217

Gent M, Sackett DL (1979) The qualification and disqualification of patients and events in long-term cardiovascular clinical trials. Thromb Hemosta 41:123–134

ICH Steering Committee (1995) ICH harmonised tripartite guideline structure and content of clinical study reports. Recommended for adoption at Step 4 of the ICH Process on 30 November 1995

ICH Steering Committee (1996) ICH Harmonized tripartite guideline. Guideline for good clinical practice. Recommended for adoption at step 4 of the ICH process on 1 May 1996

ICH Steering Committee (1997) ICH Harmonized tripartite guideline. General considerations for clinical trials. Recommended for adoption at step 4 of the ICH process on 17 July 1997

ICH Steering Committee (1998) ICH Harmonized tripartite guideline. Statistical principles for clinical trials. Recommended for adoption at step 4 of the ICH process on 5 February 1998

ICH Steering Committee (2000) ICH Harmonized tripartite guideline. Recommended for adoption at step 4 of the ICH process on 20 July 2000

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman & Hall/CRC, Boca Raton

Schwartz D, Lellouch J (1967) Explanatory and pragmatic attitudes in therapeutical trials. J Chron Dis 20: 637–648

World Medical Association (1964) Declaration of Helsinki. Ethical principles for medical research involving human subjects 1964. Amended by the 59th WMA General Assembly, Seoul, Korea 2008, http://www.wma.net/

Clinical Trials: Some Aspects of Public Interest

Jagdish N. Srivastava
CNS Research Professor Emeritus
Colorado State University, Fort Collins, CO, USA

Medical experiments, often called "clinical trials," are obviously extremely important for the human race. Here, we shall briefly talk, in layman's language, about some important aspects of the same which are of great public interest.

Side Effect of Drugs

The side effects of allopathic drugs are notorious; death is often included in the same. For degenerative diseases (as opposed to infectious diseases, as in epidemics) it is not clear to the author whether any serious effort is being made by the pharmaceutical companies to develop drugs

which actually cure diseases; the trend seems to be at best to maintain people on drugs for a long time (even for the whole life). Mostly, people live under varying forms of painkiller-surgery regimes.

However, in many institutions (for example, departments doing research on nutrition) there are people who are genuinely interested in finding cures, though often they do not possess the resources they need. Many things, considered true by the public, are not quite so. Consider preservatives and other food additives that are legal. They are found in varying quantities in most foods, and many people do not pay attention to this at all, and consume unknown amounts each day. The thought that they have no side effects is based on relatively (time-wise) small experiments and extrapolations there from. It is probably true that if some food with a particular preservative or a (combination of the same with others) is consumed, it may not have any noticeable effect within a short period. But, the worry that many thinkers have is whether consuming food (all the time ignoring preservatives that it may contain) will have a disastrous effect (like producing cancer, heart attack, diabetes, stroke, etc.) 20, 30, 40, or 50 years earlier than it would have been expected for an additives-free diet. (The fact that, now, teenagers and young ones in their twenties are developing such diseases which, in an earlier age, were found mainly among seniors only, is alarming.) A full scale clinical trial (to study the long term effect of preservatives etc.) will take more than a century, and has not been done. Thus, extrapolations proclaiming that such additives are safe are based on guess work only, and are not necessarily scientifically sound. We live in an age when shelf life has become more important than human life.

It is not even clear whether the damage done from side effects and the painkiller-surgery policies is limited to the increase in the periods of sickness of people, the intensities of such sickness, and the reduction in the age at death. The bigger question is whether there is an effect on the progeny, and for how many generations. We recall that in the processes of natural selection in the theory of evolution, only the fittest may survive. Clearly, for the human race, only the policy that promotes the good of the general public corresponds to being fit for survival.

Contradictory Statements by Opposing Camps of Medical Researchers

Often, seemingly good scientists are found to be contradicting each other. For example, there may be a substance (say, an extract from some herbs) which may be claimed by some nature-cure scientists (based on their experiments) to positively affect some disease (relative to a placebo). However, some pharmaceutical scientists may claim that their experiments show that the drug is no better than the placebo. This is to say that, often in such cases, a close look may reveal that the two sets of experiments are not referring to the same situation. To illustrate, the subjects (people, on whom an experiment is done) in the first group may be people who just contracted the disease, these people being randomly assigned the drug or the placebo. In the second case, the subjects may be people who have had the disease for some time and have been taking painkillers. Now, the herbal drug may be quite effective on a body which is in a more primeval and natural state, and yet not work well in a body which has been corrupted by the chemicals in the painkiller. Clearly, that would explain the discrepancy and support the use of the herbal drug soon after the disease begins, simultaneously discouraging the use of painkillers etc. whose primary effect is to temporarily fool the mind into thinking that one is feeling better. Thus, it is necessary to examine a clinical trial closely rather than take its results on face value.

Large Clinical Trials: Meta-analysis

"Large" clinical trials are often touted as being very "informative." To illustrate, take the simple case of comparing two drugs *A* and *B* with respect to a placebo *C*. Now, how effective a drug is for a person may depend upon his or her constitution. On some people, *A* may be the best, on some *B*, and on others, all the three may be essentially useless. For me, even though I may not know the reality, suppose the reality is that *B* would be very effective with little negative side effect, *A* would be only somewhat effective but with a large negative side effect, and the effect of *C* would be small (being somewhat positive or somewhat negative depending on environmental factors). Suppose a trial is done in Arizona, involving 6,000 patients randomly divided into three equal groups, the result being that *A* is effective in 45% (cases in its group), *B* in 35%, and *C* in 5% cases. Clearly, here, the drug *A* wins. But, for me, what is the value of this information? I really need to know which drug would be best for me.

Now suppose a similar trial is done in Idaho and in California, the result for *A*, *B*, and *C* being 33%, 42%, 7%, and 54%, 52%, and 30% respectively in the two states. Does this help me in some way or does it simply add to the confusion? The drugs manufacturer, Mr. Gaines, would like "meta–analysis" (whose purpose is to combine the results in a legitimate and meaningful way), because his interest is in seeing the overall picture so that he can formulate an appropriate manufacturing policy for his company. However, the interest of the general public is different from that of Gaines, because each individual needs to know what is good for him or her personally. The individual's interest, in a sense, runs counter to ▶meta–analysis; he or she would be more interested in knowing what aspects of a person's

health make him or her more receptive to *A* or *B*. Instead of combining the data, more delineation needs to be done. In other words, one needs to connect the results with various features of the subjects and other related factors. Then we may gain knowledge not only on what proportion of subjects are positively affected by a drug, but what bodily features of people (or the food that they eat, or their lifestyle, or the environment around them, etc.) lead to this positive effect.

For example, in the above (artificial) data, it seems that *A* is better in a warm climate and *B* in cold. Where the climate is mild, all of them do well, and many people may recover without much of drugs. If I have been more used to a cold climate, *B* may be more effective on me. With this knowledge, even though *A* may turn out to be much better than *B* in the area where I live, *B* may be better for me individually.

(This leads us to the *philosophy* of statistical inference. Not only do we need to plan our experiments or investigations properly, we need to be careful in drawing inferences from the data obtained from them. According to the author, trying to find what a bunch of data "says" must involve in a relevant way the space of applications where such a finding will be made use of. Many scholars believe that, given a set of data, the "information" that the data contains is a fixed attribute of the data, and the purpose of inference is to bring out this attribute accurately. The author believes that the reason why the inference is sought (in particular, to what use or application the inference will be put) is also important, and should have a bearing on the inference drawn. This policy will give insight into the kind of information we need, what should receive more emphasis, etc. Clinical trials would really gain from this approach.)

Reducing Side Effects of Drugs

Studies are usually done using a "loss function" which tells how much "loss" shall we incur by adopting each of a set of policies. For example, we may have many drugs, several possible doses of a drug per day, many possible durations of time over which a drug is to be continued, etc. For each combination of these factors, the "loss" may be "the total time of absence from work," or "the total financial loss incurred because of sickness," or "the amount of fever," or "the blood pressure," etc. If the loss function involves only one variable (like "blood pressure"), it is "uni-dimensional." But if, many variables are involved simultaneously (like "blood pressure," "fever," "financial loss"), then it is called multi-dimensional. Usually, only one dimension is used or emphasized (like, "intensity of fever"). More theory needs to be developed on how to work with multi-dimensional loss functions.

Besides theory, we also need to develop good quantitative criteria for measuring "healthfulness." There can be various sectors. For example, we can have one criterion for the sector of upper digestive track, one for the middle, one for the colon, one for the respiratory system, one for bone diseases, one for the joints, one for nerves, one for cancerous growth, and so on. For each sector, the corresponding criterion will provide a measure of how healthy that sector is. Suppose we decided to have 25 such sectors. Then the loss function will be 25-dimensional. The drugs will be evaluated in each dimension, and the results will also be combined in various ways. The side effect of a drug in a particular sector will be caught more easily. When the drug is marketed, an assessment for each sector can be provided. Drugs with large effect in any sector can be rejected.

Experiments with Many Factors: Interactions

We make some technical remarks here. A large part of the field of statistical design of multi-factorial scientific experiments is concerned with the simplistic situation when there are either no interactions or the set of non-negligible interactions is essentially known (though the values of these interactions and the main effects are not known). However, in medical experiments, we can have interactions of even very high orders. Thus, for the field of multifactor clinical trials, we have to go beyond Plackett–Burman designs, and orthogonal arrays of small strength (such as 2). There is work available on search theory by the author and others, which would help. However, further work is needed in that field. Indeed, for vigorous full fledged research on how to cure diseases, the statistical theory of the design and analysis of multifactor multi-response experiments need to be developed much further beyond its current levels. However, the basics of the same are available in books such as Roy et al. (1970). For the reader who wishes to go deeper in the field of this article, some further references are provided below.

About the Author

Jagdish N. Srivastava is now CNS Research Professor Emeritus at Colorado State University, where he worked during 1966–2004 as Full Professor, and where he also once held a joint appointment in Philosophy. He was born in Lucknow, India on 20 June 1933. He did his Ph.D. in 1961 (at the University of North Carolina, Chapel Hill, NC, USA). He was President (Indian Society of Agricultural Statistics, 1977 Session; International Indian Statistical Association (IISA) (1993–1997); Forum for Interdisciplinary Mathematics (1994–1996). He visited institutions all over the world. His honors include the Fellowship of the American Statistical Association, Institute of Mathematical Statistics,

Institute of Combinatorial Mathematics and its Applications, IISA (Honorary), International Statistical Institute, and TWAS (The World Academy of Science for developing countries). He is the founder (1975) and Editor-in-Chief of the monthly *Journal of Statistical Planning and Inference*. Two volumes containing papers by experts from all over the world were published in honor of his 65th birthday. He discovered, among other results, the "Srivastava Codes," the "Srivastava Estimators," and the Bose-Srivastava Algebras. He has been interested in, and has contributed to, Science and Spirituality, and is a major researcher in Consciousness and the "Foundations of Reality" (including, in particular, "Quantum Reality").

Cross References

►Clinical Trials, History of
►Clinical Trials: An Overview
►Medical Research, Statistics in
►Medical Statistics
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines

References and Further Reading

Roy SN, Gnanadesikan R, Srivastava JN (1970) Analysis and design of quantitative multiresponse experiments. Pergamon, Oxford, England

Srivastava JN (1990) Modern factorial design theory for experimenters. In: Ghosh S (ed) Statistical design and analysis of industrial experiments. Marcel Dekker, New York, pp 311–406

Srivastava JN (1996) A critique of some aspects of experimental designs. In: Ghosh S, Rao CR (eds) Handbook of statistics, vol 13. Elsevier, Amsterdam, pp 309–341

Srivastava JN (2006) Foods, drugs and clinical trials: some outrageous issues. J Combinatorics Inf Syst Sci 31(1–4):365–378

Srivastava JN (2008) Some statistical issues concerning allopathic drugs for degenerative diseases. J Indian Soc Agric Stat 62: 120–125

Cluster Analysis: An Introduction

H. Charles Romesburg
Professor
Utah State University, Logan, UT, USA

Cluster analysis is the generic name for a variety of mathematical methods for appraising similarities among a set of objects, where each object is described by measurements made on its attributes. The input to a cluster analysis is a *data matrix* having t columns, one for each object, and n rows, one for each attribute. The (i, j)th element of the data matrix is the measurement of the ith attribute for the jth object. The output from a cluster analysis identifies groups of similar objects called *clusters*. A cluster may contain as few as one object, because an object is similar to itself.

Applications of cluster analysis are widespread because the need to assess similarities and dissimilarities among objects is basic to fields as diverse as agriculture, geology, market research, medicine, sociology, and zoology. For example, a hydrologist considers as the objects a set of streams, and for attributes describes each stream with a list of water quality measures. A cluster analysis of the data matrix identifies clusters of streams. The streams within a given cluster are similar, and any stream in one cluster is dissimilar to any stream in another cluster.

There are two types of cluster analysis. *Hierarchical cluster analysis* is the name of the collection of methods that produce a hierarchy of clusters in the form of a *tree*. The other type, *nonhierarchical cluster analysis*, is the name of the collection of methods that produce the number of clusters that the user specifies. For both types, computer software packages containing programs for the methods are available.

Let us illustrate the main features of hierarchical cluster analysis with an example where the calculations can be done by hand because the data matrix is small, five objects and two attributes, consisting of made-up data:

Data matrix						
		Object				
		1	2	3	4	5
Attribute	1	10	20	30	30	5
	2	5	20	10	15	10

To perform a hierarchical cluster analysis, we must specify: (1) a coefficient for assessing the similarity between any two objects, j and k; and (2) a clustering method for forming clusters.

For the first, let us choose the "Euclidean distance coefficient," e_{jk}. The smaller its value is, the more similar objects j and k are. If the value is zero, they are identical, i.e., maximally similar. For our example with $n = 2$ attributes, e_{jk} is the distance between object j and object k computed with the Pythagorean theorem. And for the clustering method, let us choose the "UPGMA method," standing for "unweighted pair-group method using arithmetic averages."

At the start of the cluster analysis, each object is considered to be a separate cluster. Thus with five objects, there are five clusters. For the five we compute the $5(5-1)/2 = 10$

values of e_{jk}. To demonstrate the calculation of one of these values, consider object 1 and object 5. The Euclidean distance e_{15} is

$$e_{15} = [(10-5)^2 + (5-10)^2]^{1/2} = 7.07.$$

In this manner, we compute the other values and put the ten in a list, from smallest, indicating the most similar pair of clusters (objects), to largest, indicating the least similar pair: $e_{34} = 5.0$, $e_{15} = 7.07$, $e_{24} = 11.2$, $e_{23} = 14.1$, $e_{12} = 18.0$, $e_{25} = 18.0$, $e_{13} = 20.6$, $e_{14} = 22.4$, $e_{35} = 25.0$, $e_{45} = 25.5$.

The two most similar clusters, 3 and 4, head the list, as the Euclidean distance between them is the smallest. Therefore,

Step 1 Merge clusters 3 and 4, giving 1, 2, (34), and 5 at the value of $e_{34} = 5.0$.

Next, for the four clusters – 1, 2, (34), and 5 – we obtain the $4(4-1)/2 = 6$ values of e_{jk}. Three of these values are unchanged by the clustering at step 1 and can be transcribed from the above list. The other three have to be computed according to the guiding principle of the UPGMA clustering method. It requires that we average the values of e_{jk} between clusters, like this:

$$e_{1(34)} = \tfrac{1}{2}(e_{13} + e_{14}) = \tfrac{1}{2}(20.6 + 22.4) = 21.5;$$
$$e_{2(34)} = \tfrac{1}{2}(e_{23} + e_{24}) = \tfrac{1}{2}(14.1 + 11.2) = 12.7;$$
$$e_{5(34)} = \tfrac{1}{2}(e_{35} + e_{45}) = \tfrac{1}{2}(25.0 + 25.5) = 25.3.$$

So, the six e_{jk} values listed in order of increasing distance are: $e_{15} = 7.07$, $e_{2(34)} = 12.7$, $e_{12} = 18.0$, $e_{25} = 18.0$, $e_{1(34)} = 21.5$, $e_{5(34)} = 25.3$. It follows that the two most similar clusters are 1 and 5, since the Euclidean distance between them is the smallest. Therefore,

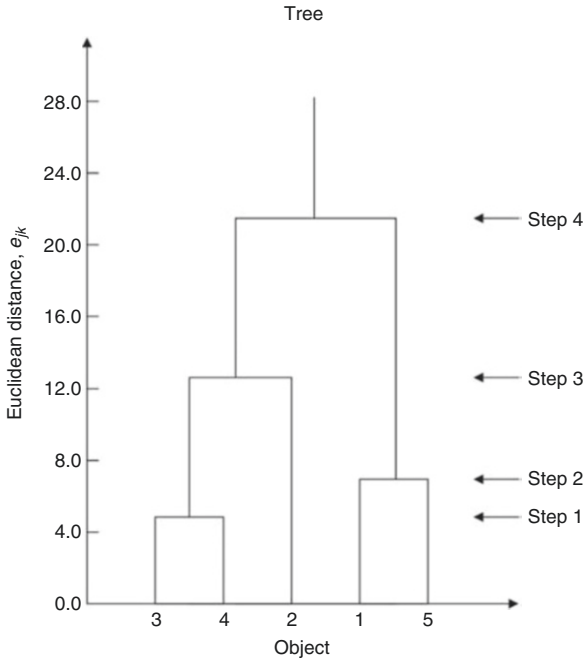
Step 2 Merge clusters 1 and 5, giving 2, (34), and (15) at the value of $e_{15} = 7.07$.

Before going to the next clustering step, we note that step 2 left the distance between clusters 2 and (34) unchanged at $e_{2(34)} = 12.7$. The two remaining distances are calculated according to the UPGMA clustering method by averaging the values of e_{jk} as follows:

$$e_{2(15)} = \tfrac{1}{2}(e_{12} + e_{25}) = \tfrac{1}{2}(18.0 + 18.0) = 18.0;$$
$$e_{(15)(34)} = \tfrac{1}{4}(e_{13} + e_{14} + e_{35} + e_{45}) = \tfrac{1}{4}(20.6 + 22.4 + 25.0 + 25.5) = 23.4.$$

The list of e_{jk} in increasing distance is now: $e_{2(34)} = 12.7$, $e_{2(15)} = 18.0$, $e_{(15)(34)} = 23.4$. The two most similar clusters, 2 and (34) head the list. Therefore,

Step 3 Merge clusters 2 and (34), giving (15) and (234) at the value of $e_{2(34)} = 12.7$.



Cluster Analysis: An Introduction. Fig. 1 Tree showing the hierarchy of similarities between the five objects specified by the data matrix in the text

At this point there are two clusters: (15) and (234). The average Euclidean distance between them is:

$$e_{(15)(234)} = \tfrac{1}{6}(e_{12} + e_{13} + e_{14} + e_{25} + e_{35} + e_{45}) = \tfrac{1}{6}(18.0 + 20.6 + 22.4 + 18.0 + 25.0 + 25.5) = 21.6.$$

The list of e_{jk} has only one value: $e_{(15)(245)} = 21.6$. Therefore,

Step 4 Merge clusters (15) and (234), giving (12345) at the value of $e_{(15)(234)} = 21.6$.

The calculations are finished. With each step, the tree (Fig. 1) has been growing. It summarizes the clustering steps, e.g., showing that the branches containing cluster (34) and cluster 2 join at an Euclidean distance value of 12.7.

The tree is a hierarchical ordering of similarities that begins at the tree's bottom where each object is separate, its own cluster. As we move to higher levels of e_{jk}, we become more tolerant and allow clusters to hold more than one object. When we reach the tree's top we are completely tolerant of the differences between objects, and all objects are considered as one cluster.

Suppose we took the five objects in the data matrix and plotted them on a graph with attribute 1 as one axis and

attribute 2 as the other. We would see that the distances between the objects suggest clusters that nearly match those in the tree. However, real applications typically have many attributes, often more than a hundred. In such cases, the researcher cannot grasp the inter-object similarities by plotting the objects in the high-dimension attribute space. Yet cluster analysis will produce a tree that approximates the inter-object similarities.

A tree is an old and intuitive way of showing a hierarchy. Witness the tree of life forms, the Linnaean classification system. At its bottom is the level of Species, at the next higher hierarchical level is the Genus, consecutively followed by levels of Order, Class, Phylum, and Kingdom.

A widely practiced way of creating a classification of objects is to perform a hierarchical cluster analysis of the objects. On the tree, draw a line perpendicular across the tree's axis, cutting it into branches, i.e., clusters. The objects in the clusters define the classes. Details may be found in Romesburg (2004) and in Sneath and Sokal (1973).

There are several general points to note about hierarchical cluster analysis:

1. There are various coefficients that can be used to assess the similarity between clusters. Of these, there are two types: dissimilarity coefficients and similarity coefficients. With a dissimilarity coefficient (as the Euclidean distance coefficient is), the smaller its value is, the more similar the two clusters are. Whereas with a similarity coefficient, the larger its value is, the more similar the two clusters are. An example of a similarity coefficient is the Pearson product moment correlation coefficient. But whether a dissimilarity coefficient or a similarity coefficient is used, a clustering method at each step merges the two clusters that are most similar.
2. Although the UPGMA clustering method (also called "average linkage clustering method") is perhaps most often used in practice, there are other clustering methods. UPGMA forms clusters based on the average value of similarity between the two clusters being merged. Another is the SLINK clustering method, short for "single linkage" clustering method, and sometimes called "nearest neighbor" clustering method. When two clusters are joined by it, their similarity is that of their most similar pair of objects, one in each cluster. Another is the CLINK clustering method, short for "complete linkage" clustering method, and sometimes called "furthest neighbor" clustering method. When two clusters are joined by it, their similarity is that of the most dissimilar pair of objects, one in each cluster. Another is Ward's clustering method, which assigns objects to clusters in such a way that a sum-of-squares index is minimized.
3. The data in the data matrix may be measured on a continuous scale (e.g., temperature), an ordinal scale (e.g., people's ranked preference for products), or on a nominal scale for unordered classes (e.g., people's sex coded as 1 = female, 0 = male).

   For an illustration of nominal scale measurement, suppose a military researcher takes a set of aircraft as the objects, and for their attributes records whether or not an aircraft can perform various functions. If the jth aircraft is able to perform the ith function, the (i, j)th element of the data matrix is coded with a "1"; if it is unable to perform the ith function, it is coded with a "0." In this way, the data matrix consists of zeroes and ones. A similarity coefficient, such as the one named "the simple matching coefficient," gives a numerical value for the similarity between any two aircraft. The cluster analysis produces a tree which shows which of the aircraft are functionally similar (belong to the same cluster) and which are functionally dissimilar (belong to different clusters).
4. Whenever the attributes of the data matrix are measured on a continuous scale, it is sometimes desired to standardize the data matrix. Standardizing recasts the units of measurement of the attributes as dimensionless units. Then the cluster analysis is performed on the standardized data matrix rather than on the data matrix. There are several alternative ways of standardizing (Romesburg 2004).
5. Commercial software packages for performing hierarchical cluster analysis include SPSS, SAS, CLUSTAN, and STATISTICA. Of these, SPSS is representative, allowing the user a choice of about 35 similarity/dissimilarity coefficients and seven clustering methods.
6. In the literature of cluster analysis, certain terms have synonyms. Among other names for the objects to be clustered are "cases," "individuals," "subjects," "entities," "observations," "data units," and "OTU's" (for "operational taxonomic units"). Among other names for the attributes are "variables," "features," "descriptors," "characters," "characteristics," and "properties." And among other names for the tree are the "dendrogram" and the "phenogram."

In contrast to hierarchical cluster analysis, nonhierarchical cluster analysis includes those clustering methods that do not produce a tree. The software packages mentioned above have programs for nonhierarchical cluster analysis. Perhaps the most-used nonhierarchical method is *K-means cluster analysis*. For it, the user specifies k, the number of clusters wanted, where k is an integer less than t, the number of objects. Software programs for K-means

cluster analysis usually differ a bit in their details, but they execute an iterative process to find clusters, which typically goes like this:

To begin the first iteration, the program selects k objects from the data matrix and uses them as k cluster seeds. The selection is made so that the Euclidean distances between the cluster seeds is large, which helps insure that the seeds cover all regions of the attribute space in which the objects reside.

Next, the program forms tentative clusters by sequentially assigning each remaining object to whichever cluster seed it is nearest to. As objects are assigned, the cluster seeds are recomputed and made to have the attribute values that are the average of those of the objects in the clusters. Hence, cluster seeds generally change as objects are tentatively assigned to clusters.

When the first iteration is finished, the resulting cluster seeds are taken as the k initial seeds to start the second iteration. Then the process is repeated, sequentially assigning the objects to their nearest cluster seed, and updating the seeds as the process moves along.

Finally, after a number of iterations, when the change in the cluster seeds is tolerably small from one iteration to the next, the program terminates. The k final clusters are composed of the objects associated with the k cluster seeds from the final iteration.

We now turn to the question, "Which is the better method for finding clusters – hierarchical cluster analysis or nonhierarchical cluster analysis?" The answer depends. Broadly speaking, researchers like having a choice of a large variety of similarity/dissimilarity coefficients, and like having the similarities among clusters displayed as a hierarchy in the form of a tree – two features that hierarchical methods offer but nonhierarchical methods do not. However, for hierarchical methods the amount of computation increases exponentially with the number of objects. Whereas for nonhierarchical methods the amount of computation increases less than exponentially because the methods do not require the calculation of similarities between all pairs of objects. In any event, all of the software packages mentioned above can handle very large data matrices for hierarchical methods and for nonhierarchical methods. For instance, according to the literature that accompanies CLUSTAN's hierarchical cluster analysis program, the limit to the size of a data matrix that at present can be processed in a reasonable time on a basic PC is in the neighborhood of 16,000 objects and 1,000 attributes. For more objects than that, CLUSTAN's nonhierarchical K-means program can handle as many as a million objects.

Books that provide detailed accounts of hierarchical cluster analysis and nonhierarchical cluster analysis include those by Aldenderfer and Blashfield (1984), Everitt (1993), and Romesburg (2004).

About the Author

Dr. H. Charles Romesburg is Professor of Environment and Society at Utah State University and holds degrees in Operations Research and Biostatistics, Nuclear Engineering, and Mechanical Engineering. He is the author of four books, including *Cluster Analysis for Researchers* (North Carolina: Lulu, 2004) and *Best Research Practices* (North Carolina: Lulu, 2009). He is an active and prolific researcher with numerous scientific articles to his credit. His publications in which he is the sole author have been cited more than 1,000 times. The Wildlife Society has awarded him its Wildlife Publication Award for his article "Wildlife Science: Gaining Reliable Knowledge."

Cross References

►Data Analysis
►Distance Measures
►Fuzzy Logic in Statistical Data Analysis
►Hierarchical Clustering
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Random Permutations and Partition Models

References and Further Reading

Aldenderfer MS, Blashfield RK (1984) Cluster analysis. Sage, Beverly Hills

Everitt B (1993) Cluster analysis. E. Arnold, London

Romesburg HC (2004) Cluster analysis for researchers. Lulu.com, North Carolina

Sneath PHA, Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman, San Francisco

Cluster Sampling

Janusz Wywial
Professor
Katowice University of Economics, Katowice, Poland

The quality of statistical inference is dependent not only on, for example, estimator construction but on the structure of a population and a sampling scheme too. For example, let the estimation of total wheat production in a population of farms be considered. The population of farms is divided into clusters corresponding to villages. This estimation can be based on the ordinary simple sample or on the cluster sample. Population units can be selected to the sample by means of

several sampling schemes. The units (i.e., farms) can be selected to the ordinary sample, or clusters of the units (i.e., villages) can be drawn to the cluster sample. The accuracy of the estimation depends on the sampling scheme and on the intraclass spread of a variable under study (wheat production). When should the ordinary simple sample be used and when should the cluster one?

Let us consider a fixed and finite population case. The fixed and finite population of the size N is denoted by $\Omega = \{\omega_1, \ldots, \omega_N\}$, where ω_k is an element (unit) of the population U. Let us assume that Ω is divided into G mutually disjoint clusters $\Omega_k\ (k = 1, \ldots, G)$ such that $\bigcup_{k=1}^{G} \Omega_k = \Omega$. The size of a cluster Ω_k is denoted by N_k. So, $0 \le N_k \le N$ and $\bigcup_{k=1}^{G} N_k = N$. Let $U = \{\Omega_1, \ldots, \Omega_G\}$ be the set of all clusters. The clusters are called units (elements) of the set U. The cluster sample is selected from the set U and it is denoted by $s = \{\Omega_{i_1}, \ldots, \Omega_{i_n}\}$. The size of s is denoted by n, where $0 \le n \le G$. Let $\mathbf{S}$ be the set (space) of samples. The cluster sample is a random one if it is selected from U according to some sampling design denoted by $P(s)$, where $P(s) \ge 0$ for $s \in \mathbf{S}$ and $\sum_{s \in S} P(s) = 1$. Let the inclusion probability of the first order be denoted by $\pi_k = \sum_{\{s: k \in s\}} P(s)$, $k = 1, \ldots, G$. A random sample is selected from a population by means of the so-called sampling scheme, which fulfills the appropriate sampling design. It is well known that a sample can be selected according to previously determined inclusion probabilities of the first order without any explicit definition of the sampling design. This inference simplifies our practical research. Frequently, the inclusion probabilities are determined as proportional to cluster sizes, so $\pi_k \propto N_k$ for $k = 1, \ldots, G$. In general, $\pi_k \propto x_k$, where $x_k > 0$ is the value of an auxiliary variable.

Let us note that it is possible to show that the well-known systematic sampling design is a particular case of the cluster sampling design. Moreover, the cluster sampling design is a particular case of two (or more) stage sampling designs.

In general, all known sampling designs and schemes can be applied to the cluster case. The examples of sampling designs and schemes are as follows: the simple cluster sample of fixed size n, drawn without replacement, is selected according to the following sampling design: $P(s) = 1/\binom{G}{g}$ for $s \in \mathbf{S}$. The inclusion probability of the first order is $\pi_k = \frac{g}{G}$. The sampling scheme fulfilling that sampling design is as follows: The first element (unit) of the set U is selected to the sample with the probability $1/G$, the next one with the probability $1/(G-1)$, the kth element with the probability $1/(G-k+1)$, and so on until the nth element of the sample.

The sampling scheme selecting with replacement units to the sample of fixed size n is as follows: Each element of U is selected with probabilities equal to p_k, where, for example, $p_k = x_k / \sum_{i \in U} x_i$. So, elements are independently drawn to the sample n times. In this case, the sampling design is defined in a straightforward manner. Particularly, if $p_k = 1/G$ for all $k = 1, \ldots, G$, the simple cluster sample drawn with replacement is selected according to the sampling design $P(s) = 1/G^n$. In this case, each element of U is selected with the probability $1/G$ to the sample of size n.

The so-called Poisson without replacement sampling scheme is as follows: The kth unit is selected with the probability p_k, $0 < p_k \le 1$, $k = 1, \ldots, G$. In this case, the sample size is not fixed because $0 \le n \le G$. There exists the so-called conditional without replacement sampling design of a fixed sample size, but unfortunately its sampling schemes are complicated, see, for example, Tillé (2006). Additionally, let us note that the cluster sample can be useful in the case of estimating the population mean.

It is well known that the precision of a population mean estimation, performed on the basis of the simple cluster sample, depends on the so-called intraclass (intracluster) correlation coefficient, see, for example, Hansen et al. (1953) or Cochran (1963).

Let us assume that sizes of clusters are the same and equal to M and $N = GM$. The ordinary variance of the variable is defined by $v = \frac{1}{N} \sum_{k=1}^{G} \sum_{j \in \Omega_k} (y_{kj} - \bar{y})^2$, where $\bar{y}_k = \frac{1}{N} \sum_{k=1}^{G} \sum_{j \in \Omega_k} y_{kj}$ is the cluster sample. The intraclass and the betweenclass variances are given by the expressions: $v_w = \frac{1}{G(M-1)} \sum_{k=1}^{G} \sum_{j \in \Omega_k} (y_{kj} - \bar{y}_k)^2$ and $v_b = \frac{1}{G-1} \sum_{k=1}^{G} (\bar{y}_k - \bar{y})^2$, respectively, where $\bar{y}_k = \frac{1}{M} \sum_{j \in \Omega_k} y_{kj}$. The intraclass correlation coefficient is defined by the following expression: $r_I = \frac{2}{N^2 v} \sum_{k=1}^{G} \sum_{i \ne j \in \Omega_k} \sum (y_{ki} - \bar{y})(y_{kj} - \bar{y})$. The coefficient ρ takes its value from the closed interval $[-1/(M-1), 1]$. The coefficient r_I can be rewritten in the following forms: $r_I = (v_b - v_w/M)/v$, $r_I = 1 - v_w/v$ or $r_I = (Mv_m/v - 1)/(M-1)$. The expressions lead to the conclusion that the intraclaas correlation coefficient is negative (positive) when the ordinary variance is smaller (larger) than the intraclass variance or equivalent if the ordinary variance is larger (smaller) than the betweenclass variance divided by M.

Let us note that it is well known that the simple *cluster sample* mean is a more accurate estimator of the population mean than the simple sample mean when the

intraclass correlation coefficient is negative. So, this leads to the conclusion that, if only possible, a population should be clustered in such a way that the intraclass correlation coefficient takes the smallest negative value. A more complicated case of unequal cluster sizes was considered, for example, by Konijn (1973). In this case, Särndal et al. (1992) considered the so-called homogeneity coefficient, which is the function of the intraclass variance. On the basis of the cluster sample, not only the estimation of population parameters is performed but also testing statistical hypothesis, see, for example, Rao and Scott (1981).

Cross References

►Adaptive Sampling
►Intraclass Correlation Coefficient
►Multistage Sampling
►Sample Survey Methods
►Sampling From Finite Populations

References and Further Reading

Cochran WG (1963) Sampling techniques. Wiley, New York
Hansen MH, Hurvitz WN, Madow WG (1953) Sample survey methods and theory, vols I and II. Wiley, New York
Konijn HS (1973) Statistical theory of sample survey and analysis. North-Holland, Amsterdam
Rao JNK, Scott A (1981) The analysis of categorical data from complex sample surveys: chi-square tests of goodnees of fit and independence in two-way tables. J Am Stat Assoc 76(374): 221–230
Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York/Berlin/Heidelberg/London/Paris/Tokyo/Hong Kong/Barcelona/Budapest
Tillé Y (2006) Sampling algorithms. Springer, New York

Coefficient of Variation

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

Coefficient of variation is a relative measure of dispersion and it may be considered in three different contexts: in probability, in a data set or in a sample.

In the first context it refers to distribution of a random variable X and is defined by the ratio

$$v = \frac{\sigma}{\mu} \tag{1}$$

where $\mu = EX$ and $\sigma = \sqrt{E(X - EX)^2}$. It is well defined if $EX > 0$. Moreover it is scale-invariant in the sense that cX has the same v for all positive c.

Data series $x = (x_1, \ldots, x_n)$ corresponds to distribution of a random variable X taking values x_i with probabilities $p_i = \frac{k_i}{n}$, for $i = 1, \ldots, n$, where k_i is the number of appearance of x_i in the series. In this case the formula (1) remains valid if we replace μ by $\overline{x} = \frac{1}{n}\sum_i x_i$ and σ by $\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$.

If $x = (x_1, \ldots, x_n)$ is a sample from a population, then the coefficient may be treated as a potential estimator of the coefficient of variation v in the whole population. Since $\frac{1}{n}\sum(x_i - \overline{x})^2$ is biased estimator of σ^2 in order to eliminate this bias we use the sample coefficient of variance in the form

$$\mathrm{v} = \frac{s}{\overline{x}}, \tag{2}$$

where $s = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$.

One can ask whether v is normalized, i.e., whether it takes values in the interval $[0, 1]$.

In spite of that v is well defined providing $\overline{x} > 0$ it seems reasonable to restrict oneself to the nonnegative samples x, i.e., satisfying the condition $x_i \geq 0$ for all i and $\sum_i x_i > 0$. Under this assumption the sample coefficient (2) of variance in the sample takes values in the interval $[0, \sqrt{n}]$ and the lower and upper bound is attained. Therefore it is not normalized.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Semi-Variance in Finance
►Standard Deviation
►Variance

References and Further Reading

Stępniak C (2007) An effective characterization of Schur-convex functions with applications. J Convex Anal 14:103–108

Collapsibility

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Collapsibility in Contingency Tables

Consider the I by J by K contingency table representing the joint distribution of three discrete variables X, Y, Z, the I by J marginal table representing the joint distribution of X and Y, and the set of conditional I by J subtables (strata) representing the joint distributions of X and Y within levels

Collapsibility. Table 1 Trivariate distribution with (a) strict collapsibility of $Y|X$ risk differences over Z, (b) collapsibility of $Y|X$ risk ratios when standardized over the Z margin, and (c) noncollapsibility of $Y|X$ odds ratios over Z. Table entries are cell probabilities

	$Z = 1$		$Z = 0$		Collapsed over Z	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$
$Y = 1$	0.20	0.15	0.10	0.05	0.30	0.20
$Y = 0$	0.05	0.10	0.15	0.20	0.20	0.30
Risks[a]	0.80	0.60	0.40	0.20	0.60	0.40
Differences	0.20		0.20		0.20	
Ratios	1.33		2.00		1.50	
Odds ratios	2.67		2.67		2.25	

[a]Probabilities of $Y = 1$ in column

of Z. A measure of association of X and Y is *strictly collapsible* across Z if it is constant across the strata (subtables) and this constant value equals the value obtained from the marginal table.

Noncollapsibility (violation of collapsibility) is sometimes referred to as ▶Simpson's paradox, after a celebrated article by Simpson (1951). This phenomenon had however been discussed by earlier authors, including Yule (1903); see also Cohen and Nagel (1934). Some statisticians reserve the term Simpson's paradox to refer to the special case of noncollapsibility in which the conditional and marginal associations are in opposite directions, as in Simpson's numerical examples. Simpson's algebra and discussion, however, dealt with the general case of inequality. The term "collapsibility" seems to have arisen in later work; see Bishop et al. (1975).

Table 1 provides some simple examples. The difference of probabilities that $Y = 1$ (the risk difference) is strictly collapsible. Nonetheless, the ratio of probabilities that $Y = 1$ (the risk ratio) is not strictly collapsible because the risk ratio varies across the Z strata, and the odds ratio is not at all collapsible because its marginal value does not equal the constant conditional (stratum-specific) value. Thus, collapsibility depends on the chosen measure of association.

Now suppose that a measure is not constant across the strata, but that a particular summary of the conditional measures does equal the marginal measure. This summary is then said to be *collapsible* across Z. As an example, in Table 1 the ratio of risks averaged over (standardized to) the marginal distribution of Z is

$$\Sigma_z P(Y = 1|X = 1, Z = z)P(Z = z)/\Sigma_z P(Y = 1|X = 0, Z = z)$$
$$P(Z = z) = \{-0.8(0.5) + 0.4(0.5)\}/\{-0.6(0.5) + 0.2(0.5)\} = 1.50,$$

which is equal to the marginal (crude) risk ratio. Thus, the risk ratio in Table 1 is collapsible under this particular weighting (standardization) scheme for the risks.

Various tests of collapsibility and strict collapsibility have been developed (e.g., Whittemore 1978; Asmussen and Edwards 1983; Ducharme and LePage 1986; Greenland and Mickey 1988; Geng 1989) as well as generalizations to partial collapsibility. The literature on graphical probability models distinguishes other types of collapsibility; see Frydenberg (1990), Whittaker (1990, Sect. 12.5) and Lauritzen (1996, Sect. 46.1) for examples. Both definitions given above are special cases of parametric collapsibility (Whittaker 1990).

Collapsibility in Regression Models

The above definition of strict collapsibility extends to regression contexts. Consider a generalized linear model (see ▶Generalized Linear Models) for the regression of Y on three vectors w, x, z:

$$g[E(Y|w, x, z)] = \alpha + w\beta + x\gamma + z\delta.$$

This regression is said to be collapsible for β over z if $\beta^* = \beta$ in the regression omitting z,

$$g[E(Y|w, x)] = \alpha^* + w\beta^* + x\gamma^*$$

and is noncollapsible otherwise. Thus, if the regression is collapsible for β over Z and β is the parameter of interest, Z need not be measured to estimate β. If Z is measured, however, tests of $\beta^* = \beta$ can be constructed (Hausman 1978; Clogg et al. 1995).

The preceding definition generalizes the original contingency-table definition to arbitrary variables. There is a technical problem with the above regression definition,

however: If the first (full) model is correct, it is unlikely that the second (reduced) regression will follow the given form; that is, most families of regression models are not closed under deletion of Z. For example, suppose Y is Bernoulli with mean p and g is the logit link function $\ln[p/(1-p)]$, so that the full regression is first-order logistic. Then the reduced regression will not follow a first-order logistic model except in special cases. One way around this dilemma (and the fact that neither of the models is likely to be exactly correct) is to define the model parameters as the asymptotic means of the maximum-likelihood estimators. These means are well-defined and interpretable even if the models are not correct (White 1994).

If the full model is correct, $\delta = 0$ implies collapsibility for β and γ over Z. Nonetheless, if neither β nor δ is zero, marginal independence of the regressors does not ensure collapsibility for β over Z except when g is the identity or log link (Gail et al. 1984; Gail 1986). Conversely, collapsibility can occur even if the regressors are associated (Whittemore 1978; Greenland et al. 1999). Thus, it is not generally correct to equate collapsibility over Z with simple independence conditions, although useful results are available for the important special cases of linear, log-linear, and logistic models (e.g., see Gail 1986; Wermuth 1987, 1989; Robinson and Jewell 1991; Geng 1992; Guo and Geng 1995).

Confounding Versus Noncollapsibility

Much of the statistics literature does not distinguish between the concept of confounding as a bias in effect estimation and the concept of noncollapsibility; for example, Becher (1992) defines confounding as $\beta^* \neq \beta$ in the regression models given above, in which case the elements of Z are called confounders. Similarly, Guo and Geng (1995) define Z to be a nonconfounder if $\beta^* = \beta$. Nonetheless, confounding as defined in the causal-modeling literature (*See* ►Confounding) may occur with or without noncollapsibility, and noncollapsibility may occur with or without confounding; see Greenland (1987, 1996) and Greenland et al. (1999) for examples. Mathematically identical conclusions have been reached by other authors, albeit with different terminology in which noncollapsibility is called "bias" and confounding corresponds to "covariate imbalance" (Gail 1986; Hauck et al. 1991).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Confounding and Confounder Control
►Simpson's Paradox

References and Further Reading

Asmussen S, Edwards D (1983) Collapsibility and response variables in contingency tables. Biometrika 70:567–578

Becher H (1992) The concept of residual confounding in regression models and some applications. Stat Med 11:1747–1758

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Clogg CC, Petkova E, Haritou A (1995) Statistical methods for comparing regression coefficients between models (with discussion). Am J Sociol 100:1261–1305

Cohen MR, Nagel E (1934) An introduction to logic and the scientific method. Harcourt Brace, New York

Ducharme GR, LePage Y (1986) Testing collapsibility in contingency tables. J R Stat Soc Ser B 48:197–205

Frydenberg M (1990) Marginalization and collapsibility in graphical statistical models. Ann Stat 18:790–805

Gail MH (1986) Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Moolgavkar SH, Prentice RL (eds) Modern statistical methods in chronic disease epidemiology. Wiley, New York, pp 3–18

Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71:431–444

Geng Z (1989) Algorithm AS 299. Decomposability and collapsibility for log-linear models. J R Stat Soc Ser C 38:189–197

Geng Z (1992) Collapsibility of relative risk in contingency tables with a response variable. J R Stat Soc Ser B 54:585–593

Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol 125:761–768

Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501

Greenland S, Mickey RM (1988) Closed-form and dually consistent methods for inference on collapsibility in $2 \times 2 \times K$ and $2 \times J \times K$ tables. J R Stat Soc Ser C 37:335–343

Greenland S, Robins J, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14:29–46

Guo J, Geng Z (1995) Collapsibility of logistic regression coefficients. J R Stat Soc Ser B 57:263–267

Hauck WW, Neuhas JM, Kalbfleisch JD, Anderson S (1991) A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol 44:77–81

Hausman J (1978) Specification tests in econometrics. Econometrica 46:1251–1271

Lauritzen SL (1996) Graphical models. Clarendon, Oxford

Neuhaus JM, Kalbfleisch JD, Hauck WW (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 59:25–35

Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression. Int Stat Rev 59:227–240

Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc Ser B 13:238–241

Wermuth N (1987) Parametric collapsibility and lack of moderating effects in contingency tables with a dichotomous response variable. J R Stat Soc Ser B 49:353–364

Wermuth N (1989) Moderating effects of subgroups in linear models. Biometrika 76:81–92
White HA (1994) Estimation, inference, and specification analysis. Cambridge University Press, New York
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York
Whittemore AS (1978) Collapsing multidimensional contingency tables. J R Stat Soc Ser B 40:328–340
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Comparability of Statistics

Peter Hackl
Professor
Vienna University of Business and Economics, Vienna, Austria

Comparison is the cognitive function that is basis for any measuring process and a frequent activity in everyday human life. Nowadays, comparisons of statistics over time and, even more demanding, cross-national and multilateral comparisons are a central element of economic and social analyzes. For politics and administration, for business and media, and for each citizen, comparisons are means to understand and assess the political, economic, and social processes of this world. This situation raises questions: Under which conditions are statistics comparable? Under which conditions is a comparison valid and leads to reliable results? What conclusions may be drawn and what are the risks implied by such conclusions?

Comparability: Definition and Assessment

Rathsmann-Sponsel and Sponsel (2001) describe "comparison" as a 7-digit relation $f(P, S, Z, K, V, A, B)$; P represents the comparing person, S and Z describe the comparing situation and the purpose of comparison, respectively; vector K stands for a number of criteria, V for a number of procedures, and A and B represent the characteristics of the objects to be compared, respectively. This rather formal view of psychologists indicates the complexity of the interaction between the individual and the objects to be compared. More visible becomes this complexity when the definition is applied to real situations, e.g., comparing the employment rates of two countries.

The employment rate is the number of persons aged 15–64 in employment as the share of the total population of the same age group. Obviously, the definition of "being in employment" and the exact meaning of "persons aged 15–64" are crucial for the value that is obtained for the employment rate. In addition, the statistical value is affected by the sampling design and other aspects of the data collection.

In general, statistics are based on concepts and definitions, and the value of a statistic is the result of a complex measurement process; *comparability is affected* by all these factors and, consequently, a wide range of facts must be considered. Moreover, the relevance of these facts depends on the purpose of comparison, the comparing situation, and other aspects of the comparison process. E.g., if the comparison of the employment rates of two countries is the basis for a decision on the allocation of subsidies for structural development, comparability is a more serious issue than in the case where the result of the comparisons does not have such consequences. These – and many other – characteristics must be taken into consideration when assessing differences between two employment rates.

Assessment of comparability has to take into account the multi-dimensionality of the conditions for comparability. Many aspects to be considered are qualitative, so that the corresponding dimensions cannot be measured on a metric scale. Moreover, important characteristics of the statistical products or the underlying measurement processes are often not available or uncertain.

Hence, in general, it is not feasible to give a comprehensive picture of comparability by means of a few metric measures. Alternatives are

- An indicator in form of a number between zero and one that indicates the degree of comparability, a one indicating perfect comparability.
- A rating of comparisons on an ordinal rating scale with a small number of points, a high value representing good comparability.

An example for a rating scale is the three point scale that is used for the "Overall assessment of accuracy and comparability" of indicators – such as the employment rate – within the Eurostat Quality Profiles; see Jouhette and Sproge (2005). This overall assessment is rated from "A" to "C". Grad "A" indicates that

- Data is collected from reliable sources applying high standards with regard to methodology/accuracy and is well documented in line with Eurostat metadata standard.
- The underlying data is collected on the basis of a common methodology for the European Union and,

where applicable, data for US and Japan can be considered comparable; major differences being assessed and documented.
- Data are comparable over time; impact of procedural or conceptual changes being documented.

This example illustrates:

- That the rating process reduces a high-dimensional information to a single digit.
- Where the characteristics of the statistics to be compared, the underlying measurement processes, and also conditions of the comparison process are crucial input elements for the rating of comparability.
- That the rating outcome has only the character of a label which the user might trust but which only reflects – perhaps vaguely – the result of a complex and subjective assessment process.
- That the rating outcome may miss to give appropriate weight to aspects that are important for a certain user.

A rating of the comparability on an ordinal rating scale has the advantage that it allows an easy communication about comparability.

The professional assessment of comparability requires:

- An adequate *competence of the scrutinizer*
- A careful *documentation of all characteristics* of the statistical products that are relevant for assessing the comparability

Generally, the scrutinizer will be different from the producer of a statistical product. This certainly will be the case in respect of cross-national comparisons. The producer has to provide a comprehensive set of metadata that documents all characteristics which are relevant for assessing the comparability. The outcome of this exercise might be an indicator of the types that are described above.

For the non-expert user, the assessment of the comparability of statistical products is hardly possible even when a well-designed set of all metadata is available that are relevant for assessing the comparability. Most users of the statistical products will have to rely on the professional assessment of the experts.

Comparability in Official Statistics

The integration of societies and the globalization of economies have the consequence that not only comparisons over time but especially cross-regional comparisons of statistical products are of increasing interest and importance. Political planning and decisions of supranational bodies need information that encompasses all involved nations. Multi-national enterprises and globally acting companies face the same problem.

Of even higher relevance for the need of comparability is the fact that statistical products are more and more used for *operational purposes*. Within the European Union, the process of integration of the member states into a community of states requires political measures in many areas. National statistical indicators are the basis for allocating a part of the common budget to member states, for administrating the regional and structural funds, for assessing the national performances with respect to the pact for stability and growth, and for various other purposes. It is in particular the European version of the System of National Accounts (SNA) ESA that plays such an operational role in various administrative processes of the European Union. The Millennium Development Goals and the Kyoto Protocol are other examples for the use of statistical indicators in defining political aims and assessing the corresponding progress. In all these cases, the comparability of the relevant statistical products is a core issue.

In the cross-national context, the responsibility for harmonizing cross-national concepts, definitions, and methodological aspects must be assigned to an authority with *supra-national competence*. Organizations like the UN, OECD, and Eurostat are engaged in the compilation of standards and the editing of recommendations, guidelines, handbooks, and training manuals, important means to harmonize statistical products and improve their comparability. Examples of standards are the Statistical Classifications of Economic Activities (ISIC) and the International Statistical Classification of Diseases (ICD). Principles and Recommendations for Population and Housing Censuses adopted by the Statistical Commission of the UN is an example for a standard methodology. Examples of standards on the European level are the NACE and CPA.

Within the European Union, standards and methods are laid down in regulations which are *legally binding* for the member states. E.g., the ESA 95 was approved as a Council Regulation in June 1996 and stipulates the member states to apply the SNA in a very concrete form. In working groups, experts from the member states are dealing with the preparation and implementation of such regulations; the harmonization of national statistical products is a central concern of these activities.

The important role that is attributed to statistical comparability within the ESS is stressed by the fact that the European Statistics Code of Practice (2005) contains Coherence and Comparability as one of its 15 principles. The corresponding indicators refer mainly to national aspects but also to the European dimension.

To assess the comparability of statistical products, national reports are essential that provide *metadata* for all related aspects of the statistical product. Standard formats for the documentation of metadata have been suggested by the International Monetary Fund in form of the General Data Dissemination Standard (GDDS) and the Special Data Dissemination Standard (SDDS).

It should be mentioned that standardizing concepts, definitions, and methods also has unfavorable effects; *comparability has a price*. An important means for improving harmonization are standards; however, they are never perfect and tend to get outdated over time. In particular the adaptation of standards to methodological progress might be a time-consuming task. Generally, standardization reduces flexibility and makes adaptations to new developments, especially of methodological alternatives, more difficult. This is especially true if standards are put into the form of regulations. It is even truer if such standards are implemented in order to ease the use for operational purposes, as it is the case in the ESS.

Conclusions

Lack of comparability may lead to erroneous results when statistical products are compared. The need for cross-national comparability is even more pronounced if statistical results are used for operational purposes as it is the case, e.g., in the European Union. Hence, comparability is an important quality aspect of statistical products. It is affected by the involved concepts and definitions, the measurement processes, and comparability may also depend on conditions of the comparison. The producer of a statistical product has to care that the conditions of comparability are fulfilled to the highest extent possible. In the cross-national context, international organizations like ►Eurostat are fostering the compilation of standards for concepts and definitions and of principles and standards for methods and processes in order to harmonize statistical products and improve their cross-national comparability.

For the assessment of comparability, a wide range of information is needed, as many aspects of the statistics to be compared but also of the purpose and conditions of the comparison have to be taken into account. No general rules are available that guarantee a valid assessment of comparability; only experts with profound knowledge can be expected to give a reliable assessment. For such an assessment, metadata which document all relevant characteristics are essential and have to be provided by the producer of the statistical product. For cross-national purposes, organizations like Eurostat have to care that the relevant metadata are provided by the respective producers. The user, e.g., the consumer of an economic or social analysis, has to trust that the analysts and experts made use of the involved statistics responsively.

About the Author

Dr. Peter Hackl was born 1942 in Linz, Austria. He is a Professor (since 1981) at the Department of Statistics, Vienna University of Business and Economics (Wirtschaftsuniversität); Head of the Division of Economic Statistics (since 1991). During the academic year 1988/1989 and during the summer term 1992, he was visiting professor at the University of Iowa, Iowa City. He was President of the Austrian Statistical Society (Österreichische Statistische Gesellschaft), (1995–2000). He was Vice-Dean of Studies (1995–2000) at the Wirtschaftsuniversität, member and deputy chairman (1999–2004) of the Statistikrat, the advisory board of Statistics Austria, the Austrian National Statistical Office; Chairman of the Committee for Quality Assurance. He was Director General of Statistics Austria (2005–2009). Professor Hackl is Elected member of the International Statistical Institute (since 1981). He has authored/coauthored about 100 articles in refereed journals in statistical theory and methods and applications of statistics. He was also editor of two books and has authored/coauthored 5 books. He was awarded an Honorary Doctorate, Stockholm University of Economics (1996).

Cross References

►Economic Statistics
►Eurostat
►National Account Statistics

References and Further Reading

European Commission (2005) Communication from the Commission to the European Parliament and to the Council on the independence, integrity and accountability of the national and Community statistical authorities, Recommendation on the independence, integrity and accountability of the national and Community statistical authorities, COM (2005) 217

European Statistics Code of Practice (2005) http://epp.eurostat.ec.europa.eu/portall/page/portal/quality/documents/code_practice.pdf

Eurostat (2009) ESS handbook for quality reports. European Communities, Luxembourg

Jouhette S, Sproge L (2005) Quality profiles for structural indicators EU LFS based indicators. In: 29th CEIES seminar, Expert meeting statistics "Structural indicators", Luxembourg, European Communities, pp 107–129

Rathsmann-Sponsel I, Sponsel R (2001) Allgemeine Theorie und Praxis des Vergleichens und der Vergleichbarkeit. Internet

Publikation für allgemeine und integrative Psychotherapie, www.sgipt.org/wisms/verglbk0.htm
Nederpelt V, Peter WM (2009) Checklist quality of statistical output. Statistics Netherlands, Den Haag/Heerlen

Complier-Average Causal Effect (CACE) Estimation

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

Imagine a simple randomized controlled trial evaluating a psychotherapeutic intervention for depression. Participants are randomized to one of two treatment groups. The first (the control condition) comprises treatment and management as usual (TAU). Participants in the second group are offered a course of individual cognitive behavior therapy (CBT) *in addition* to TAU. The outcome of the trial is evaluated by assessing the severity of depression in all of the participants 6 months after ▶randomization. For simplicity, we assume there are no missing outcomes. We find a difference between the average outcomes for the two groups to be about four points on the depression rating scale, a difference that is of only borderline clinical significance. This difference of four points is the well-known intention-to-treat (ITT) effect – it is the estimated effect treatment *allocation* (i.e., offering the treatment). This we will call ITT_{ALL}.

Participants in the control (TAU) condition did not get any access to CBT but we now discover that only about half of those offered psychotherapy actually took up the offer. Only 50% of the treatment group actually received CBT. So, it might be reasonable to now ask "What was the effect of receiving CBT?" or "What was the treatment effect in those who complied with the trial protocol (i.e., treatment allocation)?" Traditionally, trialists have been tempted to carry out what is called a "Per Protocol" analysis. This involves dropping the non-compliers from the treatment arm and comparing the average outcomes in the compliers with the average outcome in all of the controls. But this is not comparing like with like. There are likely to be selection effects (confounding) – those with a better (or worse) treatment-free prognosis might be more likely to turn up for their psychotherapy. The same criticism also applies to the abandonment of randomization altogether and comparing the average outcomes in those who received treatment with those who did not (a mixture of controls and non-compliers) in a so-called "As treated" analysis.

To obtain a valid estimate of the *receipt* of treatment we need to be able to compare the average of the outcomes in those who received CBT with the average of the outcomes of the control participants who *would have received* CBT had they been allocated to the treatment group. This is the Complier-Average Causal Effect (CACE) of treatment. How do we do this? The simplest approach is based on the realization that the ITT effect is attenuated estimate of the CACE, and that the amount of attenuation is simply the proportion of compliers (or would-be compliers) in the trial. The proportion of compliers (P_C) is simply estimated by the proportion of those allocated to the treatment group who actually receive CBT. We postulate that we have two (possibly hidden) classes of participant: Compliers and Non-compliers. Non-compliers receive no CBT irrespective of their treatment allocation. The intention-to-treat effects in the Compliers and Non-compliers are ITT_C (≡ CACE) and ITT_{NC}, respectively. It should be clear to the reader that $ITT_{ALL} = P_C ITT_C$ and $(1 - P_C)ITT_{NC}$.

To make use of this simple relationship, let's now assume that treatment allocation in the Non-compliers has no impact on their outcome (i.e., does not affect the severity of their depression). This assumption is often referred to as an exclusion restriction. It follows that $ITT_{ALL} = P_C ITT_C$ and that

$$CACE = ITT_C = ITT_{ALL}/P_C$$

So with 50% compliance, and estimated overall ITT effect of 4 units on the depression scale, the CACE estimate is 8 units – a result with much more promise to our clinicians. To get a standard error estimate we might apply a simple bootstrap (see ▶Bootstrap Methods). Note that CACE estimation is not a means of conjuring up a treatment effect from nowhere – if the overall ITT effect is zero so will the CACE be. If the overall ITT effect is not statistically-significant, the CACE will not be statistically-significant.

One last point: in a conventional treatment trial aiming to demonstrate efficacy, the ITT estimate will be conservative, but in a trial designed to demonstrate equivalence (or non-inferiority) it is the CACE estimate that will be the choice of the cautious analyst (we do not wish to confuse attenuation arising from non-compliance with differences in efficacy).

Here we have illustrated the ideas with the simplest of examples. And here we have also made the derivation of the CACE estimate as simple as possible without any detailed reference to required assumptions. An analogous procedure was first used by Bloom (1984) but its

formal properties were derived and compared with instrumental variable estimators of treatment effects by Angrist et al. (1996). Non-compliance usually has an implication for missing data – those that do not comply (or would-be Non-compliers) with their allocated treatment are also those who are less likely to turn up to have their outcome assessed. The links between CACE estimation and missing data models (assuming latent ignorability) are discussed by Frangakis and Rubin (1999). Generalizations of CACE methodology to estimation of treatment effects through the use of Principal Stratification have also been introduced by Frangakis and Rubin (2002).

About the Author

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

Cross References

►Instrumental Variables
►Principles Underlying Econometric Estimators for Identifying Causal Effects

References and Further Reading

Angrist JD, Imbens GW, Rubin DB (1996) Indentification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91:444–472

Bloom HS (1984) Accounting for no-shows in experimental evaluation designs. Evaluation Rev 8:225–246

Frangakis CE, Rubin DB (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika 86:365–379

Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58:21–29

Components of Statistics

Clive William John Granger†
Professor Emeritus
University of California-San Diego, San Diego, CA, USA

The two obvious subdivisions of statistics are: (a) Theoretical Statistics and (b) Practical Statistics.

1. The theoretical side is largely based on a mathematical development of probability theory (*see* ►Probability Theory: An Outline) applicable to data, particularly the asymptotic properties of estimates (*see* ►Properties of Estimators) which lead to powerful theorems such as the ►Central Limit Theorem. The aim is to put many practical approaches to data analysis (*see also* ►Categorical Data Analysis; ►Multivariate Data Analysis: An Overview; ►Exploratory Data Analysis; ►Functional Data Analysis) on a sound theoretical foundation and to develop theorems about the properties of these approaches. The theories are usually based on a number of assumptions that may or may not hold in practice.
2. Practical statistics considers the analysis of data, how the data can be summarized in useful fashions, and how relationships between sets of data from different variables can be described and interpreted. The amount and the quality of the data (*see* ►Data Quality) that is available are essential features in this area. On occasions data may be badly constructed or terms may be missing which makes analysis more complicated.

Descriptive statistics include means, variances, histograms, correlations, and estimates of quantiles, for example. There are now various types of statistics depending on the area of application. General statistics arose from considerations of gambling (*see* ►Statistics and Gambling), agriculture (*see* ►Agriculture, Statistics in; ►Analysis of Multivariate Agricultural Data), and health topics (*see* ►Medical research, Statistics in; ►Medical Statistics) but eventually a number of specialized areas arose when it was realized that these areas contained special types of data which required their own methods of analysis. Examples are:

1. Biometrics (*see* ►Biostatistics), from biological data which required different forms of measurement and associated tests.
2. ►Econometrics, for which ►Variables may or may not be related with a time gap; data can be in the form of ►Time Series (particularly in economies and finance) or in large panels (*see* ►Panel Data) in various parts of economics. The techniques developed over a wide range and the ideas have spread into other parts of statistics.
3. Psychometrics, where methods are required for the analysis of results from very specific types of experiments used in the area of psychology (*see* ►Psychology, Statistics in).

Other major areas of application such as engineering, marketing, and meteorology generally use techniques derived from methods in the areas mentioned above, but all have developed some area-specific methods.

About the Author

In 2003 Professor Granger was awarded the Nobel Memorial Prize in Economic Sciences (with Professor Robert

Engle) for methods of analyzing economic time series with common trends (cointegration). Granger was knighted in 2005.

Professor Granger had sent his contributed entry on June 2 2008, excusing himself for not writing a bit longer, "Lexicon" kind of paper: "I have never written anything for a 'Lexicon' before and so have failed in my attempt to be helpful, but I do attach a page or so. I wish you good luck with your effort." We are immensely thankful for his unselfish contribution to the prosperity of this project.

Cross References

►Statistics: An Overview
►Statistics: Nelder's view

Composite Indicators

Jože Rovan
Associate Professor, Faculty of Economics
University of Ljubljana, Ljubljana, Slovenia

Definition: A composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multidimensional concept that is being measured (OECD, Glossary of Statistical Terms).

Multidimensional concepts like welfare, well-being, human development, environmental sustainability, industrial competitiveness, etc., cannot be adequately represented by individual indicators. For that reason, composite indicators are becoming increasingly acknowledged as a tool for summarizing complex and multidimensional issues.

Composite indicators primarily arise in the following areas: economy, society, globalization, environment, innovation, and technology. A comprehensive list of indicators can be found at the following address: http:// composite-indicators.jrc.ec.europa.eu/FAQ.htm#List_of_Composite_Indicators_

The *Handbook on Constructing Composite Indicators: Methodology and User Guide* (OECD 2008) recommends a ten-step process of constructing composite indicators:

- **Theoretical framework** is the starting point of the process of constructing composite indicators, defining individual indicators (e.g., variables) and their appropriate weights, reflecting the structure of the investigated multidimensional concept. This step is crucial in construction process because it has the greatest impact on the relevance of the indicator of the investigated phenomena. For that reason, the constructors team should include, besides the statisticians, who play the major role, the experts and stakeholders from the topic of the composite indicator.

Due to the fact that many new multidimensional concepts do not have a generally agreed theoretical framework, transparency is essential in constructing credible indicators.

- **Data selection** should acquire analytically sound relevant indicators, having in mind their avaliability (country coverage, time, appropriate scale of measurement, etc.). Engagement of experts and stakeholders is recommended.
- **Imputation of missing data** (see ►Imputation) provides a complete dataset (single or. multiple). Inspection of presence of ►outliers in the dataset should not be omitted.
- **Multivariate analysis** reveals the structure of the considered dataset from two aspects: (a) units and (b) available individual indicators, using appropriate multivariate methods, e.g., ►principal component analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), Cronbach coefficient alpha, cluster analysis (see ►Cluster Analysis: An Introduction), ►correspondence analysis, etc. These methods are able to reveal internally homogenous groups of countries or groups of indicators and interpret the results.
- **Normalization procedures** are used to achieve comparability of variables of the considered dataset, taking into account theoretical framework and the properties of the data. The robustness of normalization methods against possible ►outliers must be considered.
- **Weighting and aggregation** should take into account the theoretical framework and the properties of the data. The most frequently used aggregation form is *a weighted linear aggregation rule* applied to a set of variables (OECD 2003). Weights shoud reflect the relative importance of individual indicators in a construction of the particular composite indicator.
- **Uncertainty and ►sensitivity analysis** are necessary to evaluate robustness of composite indicators and to improve transparency, having in mind selection of indicators, data quality, imputation of missing data, data normalization, weighting, aggregation methods, etc.
- **Back to the original data**, to (a) reconsider the relationships between composite indicator and the original

data set and to identify the most influential indicators and (b) compare profiled performance of the considered units to reveal what is driving the composite indicator results, and in particular whether the composite indicator is overly dominated by a small number of indicators.

- **Links to other indicators** identify the relationships between the composite indicator (or its dimensions) and other individual or composite indicators.
- **Visualization of results** should attract audience, presenting composite indicators in a clear and accurate way.

Following the above-mentioned guidelines, the constructors of composite indicators should never forget *that composite indicators should never be seen as a goal per se. They should be seen, instead, as a starting point for initiating discussion and attracting public interest and concern* (Nardo et al. 2005).

However, there is now general agreement about the usefulness of composite indicators: There is a strong belief among the constructors of composite indicators that such summary measures are meaningful and that they can capture the main characteristic of the investigated phenomena. On the other side, there is a scepticism among the critics of this approach, who believe that there is no need to go beyond an appropriate set of individual indicators. Their criticism is focused on the "arbitrary nature of the weighting process" (Sharpe 2004) in construction of the composite indicators.

Cross References

►Aggregation Schemes
►Imputation
►Multiple Imputation
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Scales of Measurement
►Sensitivity Analysis

References and Further Reading

An information server on composite indicators and ranking systems (methods, case studies, events) http://composite-indicators.jrc.ec.europa.eu/FAQ.htm#List_of_Composite_Indicators

Freudenberg M (2003) Composite indicators of country performance: a critical assessment, OECD science, technology and industry working papers, OECD Publishing, 2003/16

OECD, European Commission, Joint Research Centre (2008) Handbook on constructing composite indicators: methodology and user guide. OECD Publishing

OECD, Glossary of statistical terms (http://stats.oecd.org/glossary/index.htm)

OECD (2003) Composite indicators of country performance: a critical assessment, DST/IND(2003)5, Paris

Munda G, Nardo M (2005) Constructing consistent composite indicators: the issue of weights, EUR 21834 EN. Joint Research Centre, Ispra

Nardo M, Saisana M, Saltelli A, Tarantola S (2005) Tools for composite indicators building. european commission, EUR 21682 EN. Joint Research Centre, Ispra, Italy

Saltelli A (2007) Composite indicators between analysis and advocacy. Soc Indic Res 81:65–77

Sharpe A (2004) Literature review of frameworks for macro-indicators. Centre for the Study of Living Standards, Ottawa, Canada

Computational Statistics

Colin Rose
Director
Theoretical Research Institute, Sydney, NSW, Australia

What Is Computational Statistics?

We define *computational statistics* to be: ... 'statistical methods/results that are enabled by using computational methods'. Having set forth a definition, it should be stressed, first, that names such as *computational statistics* and *statistical computing* are essentially semantic constructs that do not have any absolute or rigourous structure within the profession; second, that there are any number of competing definitions on offer. Some are unsatisfactory because they focus purely on data or graphical methods and exclude symbolic/exact methods; others are unsatisfactory because they place undue emphasis on 'computationally-intensive methods' or brute force, almost as if to exclude well-written efficient and elegant algorithms that might be computationally quite simple. Sometimes, the difficulty is not in the execution of an algorithm, but in writing the algorithm itself.

Computational statistics can enable one:

- To work with arbitrary functional forms/distributions, rather than being restricted to traditional known textbook distributions.
- To simulate distributional properties of estimators and test statistics, even if closed-form solutions do not exist (*computational inference* rather than *asymptotic inference*).
- To compare statistical methods under different alternatives.
- To solve problems numerically, even if closed-form solutions are not possible or cannot be derived.
- To derive symbolic solutions to probability, moments, and distributional problems that may never have been solved before, and to do so essentially in real-time.

- To explore multiple different models, rather than just one model.
- To explore potentially good or bad ideas via simulation in just a few seconds.
- To choose methods that are theoretically appropriate, rather than because they are mathematically tractable.
- To check symbolic/exact solutions using numerical methods.
- To bring to life theoretical models that previously were too complicated to evaluate . . .

Journals and Societies

Important journals in the field include:

- Combinatorics, Probability & Computing
- Communications in Statistics – Simulation and Computation
- Computational Statistics
- Computational Statistics and Data Analysis
- Journal of Computational and Graphical Statistics
- Journal of the Japanese Society of Computational Statistics
- Journal of Statistical Computation and Simulation
- Journal of Statistical Software
- SIAM Journal on Scientific Computing
- Statistics and Computing

Societies include: the International Association for Statistical Computing (IASC – a subsection of the ISI), the American Statistical Association (Statistical Computing Section), the Royal Statistical Society (Statistical Computing Section), and the Japanese Society of Computational Statistics (JSCS) . . .

Computational statistics consists of three main areas, namely numerical, graphical and symbolic methods . . .

Numerical Methods

The numerical approach is discussed in texts such as Gentle (2009), Givens and Hoeting (2005), and Martinez and Martinez (2007); for Bayesian methods, see Bolstad (2009). Numerical methods include: Monte Carlo studies to explore asymptotic properties or finite sample properties, pseudo-random number generation and sampling, parametric density estimation, non-parametric density estimation, ▶bootstrap methods, statistical approaches to software errors, information retrieval, statistics of databases, high-dimensional data, temporal and spatial modeling, ▶data mining, model mining, statistical learning, computational learning theory and optimisation etc. . . . While optimisation itself is an absolutely essential tool in the field, it is very much a field in its own right.

Graphical Methods

Graphical methods are primarily concerned with either (a) viewing theoretical models and/or (b) viewing data/fitted models.

In the case of *theoretical* models, one typically seeks to provide understanding by viewing one, two or three variables, or indeed even four dimensions (using 3-dimensional plots animated over time, translucent graphics etc.).

Visualizing *data* is essential to data analysis and assessing fit; see, for instance, Chen et al. (2008). Special interest topics include smoothing techniques, kernel density estimation, multi-dimensional data visualization, clustering, exploratory data analysis, and a huge range of special statistical plot types. Modern computing power makes handling and interacting with large data sets with millions of values feasible . . . including live interactive manipulations.

Symbolic/Exact Methods

The 21st century has brought with it a conceptually entirely new methodology: symbolic/exact methods. Recent texts applying the symbolic framework include Andrews and Stafford (2000), Rose and Smith (2002), and Drew et al. (2008).

Traditional 20th century computer packages are based on numerical methods that tend to be designed much like a cookbook. That is, they consist of hundreds or even thousands of numerical recipes designed for specific cases. One function is written for one aspect of the Normal distribution, another for the LogNormal, etc. This works very well provided one stays within the constraints of the known common distributions, but unfortunately, it breaks down as soon as one moves outside the catered framework. It might work perfectly for random variable X, but not for X^2, nor $\exp(X)$, nor mixtures, nor truncations, nor reflections, nor folding, nor censoring, nor products, nor sums, nor . . .

By contrast, symbolic/exact methods are built on top of computer algebra systems . . . programs such as *Mathematica* and Maple that understand algebra and mathematics. Accordingly, symbolic algorithms can provide exact general solutions . . . not just for specific distributions/models. Symbolic computational statistical packages include mathStatica (2002–2010, based on top of *Mathematica*) and APPL (based on top of Maple).

Symbolic methods include: automated expectations for arbitrary distributions, probability, combinatorial probability, transformations of random variables, products of random variables, sums and differences of random variables, generating functions, inversion theorems, maxima/minima of random variables, symbolic and numerical maximum likelihood estimation (using exact methods),

curve fitting (using exact methods), non-parametric kernel density estimation (for arbitrary kernels), moment conversion formulae, component-mix and parameter-mix distributions, copulae, pseudo-random number generation for arbitrary distributions, decision theory, asymptotic expansions, ▶order statistics (for identical and non-identical parents), unbiased estimators (h-statistics, k-statistics, polykays), moments of moments, etc.

The Changing Notion of What is Computational Statistics

Just 10 or 20 years ago, it was quite common for people working in computational statistics to write up their own code for almost everything they did. For example, the *Handbook of Statistics 9: Computational Statistics* (see Rao 1993) starts out Chapter 1 by describing algorithms for sorting data. Today, of course, one would expect to find sorting functionality built into any software package one uses . . . indeed even into a word processor. And, of course, the 'bar' keeps on moving and evolving. Even in recent texts such as Gentle (2009), about half of the text (almost all of Part 1) is devoted to computing techniques such as fixed- and floating-point, numerical quadrature, numerical linear algebra, solving non-linear equations, optimisation etc., . . . techniques that Gentle et al. (2004, p. 5) call "*statistical computing*" but which are really just *computing*. Such methods lie firmly within the domain of computational science and/or computational mathematics . . . they are now built into any decent modern statistical/mathematical software package . . . they take years of work to develop into a decent modern product, and they require tens of thousands of lines of code to be done properly . . . all of which means that it is extremely unlikely that any individual would write their own in today's world. Today, one does not tend to build an airplane simply in order to take a flight. And yet many current texts are still firmly based in the older world of 'roll your own', devoting substantial space to routines that are (a) general mathematical tools such as numerical optimisation and (b) which are now standard functionality in modern packages used for computational statistics. While it is, of course, valuable to understand how such methods work (in particular so that one is aware of their limitations), and while such tools are absolutely imperative to carrying out the discipline of computational statistics (indeed, as a computer itself is) – these tools are now general mathematical tools and the days of building one's own are essentially long gone.

Future Directions

It is both interesting and tempting to suggest likely future directions.

(a) *Software packages*: At the present time, the computational statistics software market is catered for from two polar extremes. On the one hand, there are major general mathematical/computational languages such as *Mathematica* and Maple which provide best of breed general computational/numerical/graphical tools, and hundreds of high-level functional programming constructs to expand on same, but they are less than comprehensive in field-specific functionality. It seems likely such packages will further evolve by developing and growing tentacles into specific fields (such as statistics, combinatorics, finance, econometrics, biometrics etc.). At the other extreme, there exist narrow field-specific packages such as S-Plus, Gauss and R which provide considerable depth in field-specific functionality; in order to grow, these packages will likely need to broaden out to develop more general methods/general mathematical functions, up to the standard offered by the major packages. The software industry is nascent and evolving, and it will be interesting to see if the long-run equilibrium allows for both extremes to co-exist. Perhaps, all that is required is for a critical number of users to be reached in order for each eco-system to become self-sustaining.

(b) *Methodology*: It seems likely that the field will see a continuing shift or growth from *statistical inference* to *structural inference*, . . . from *data mining* to *model mining*, . . . from *asymptotic inference* to *computational inference*.

(c) *Parallel computing*: Multicore processors have already become mainstream, while, at the same time, the growth in CPU speeds appears to be stagnating. It seems likely then that parallel computing will become vastly more important in evolving computational statistics into the future. Future computational statistical software may also take advantage of GPUs (graphical processing units), though it should be cautioned that the latter are constrained in serious statistical work by the extremely poor numerical precision of current GPUs.

(d) *Symbolic methods*: Symbolic methods are still somewhat in their infancy and show great promise as knowledge engines i.e., algorithms that can derive exact theoretical results for arbitrary random variables.

(e) *On knowledge and proof*: Symbolic algorithms can derive solutions to problems that have never been posed before – they place enormous technological power into the hands of end-users. Of course, it is possible (though rare) that an error may occur (say in integration, or by mis-entering a model). In a sense,

this is no different to traditional reference texts and journal papers which are also not infallible, and which are often surprisingly peppered with typographical or other errors.

In this regard, the computational approach offers both greater exposure to danger, as well as the tools to avoid it. The "danger" is that it has become extremely easy to generate output in real-time. The sheer scale and volume of calculation has magnified, so that the average user is more likely to encounter an error, just as someone who drives a lot is more likely to encounter an accident. *Proving* that the computer's output is actually correct can be very tricky, or impractical, or indeed impossible for the average practitioner to do, just as the very same practitioner will tend to accept a journal result at face value, without properly checking it, even if they could do so. The philosopher, Karl Popper, argued that the aim of science should not be to prove things, but to seek to refute them. Indeed, the advantage of the computational statistical approach is that it is often possible to check one's work using two completely different methods: both numerical and symbolic. Here, numerical methods take on a new role of checking symbolic results. One can throw in some numbers in place of symbolic parameters, and one can then check if the solution obtained using symbolic methods (the exact theoretical solution) matches the solution obtained using numerical methods (typically, ►numerical integration or Monte Carlo methods, see ►Monte Carlo Methods in Statistics). If the numerical and symbolic solutions do *not* match, there is an obvious problem and we can generally immediately reject the theoretical solution (*a la* Popper). On the other hand, if the two approaches match up, we still do not have a proof of correctness . . . all we have is just one point of agreement in parameter space. We can repeat and repeat and repeat the exercise with different parameter values . . . and as we do so, we effectively build up, not an absolute proof in the traditional sense, but, appropriately for the statistics profession, an ever increasing degree of confidence . . . effectively a proof by probabilistic induction . . . that the theoretical solution is indeed correct. This is an extremely valuable (though time-consuming) skill to develop, not only when working with computers, but equally with textbooks and journal papers.

About the Author

For biography *see* the entry ►Bivariate distributions.

Cross References

►Bootstrap Asymptotics
►Bootstrap Methods
►Data Mining
►Monte Carlo Methods in Statistics
►Nonparametric Density Estimation
►Non-Uniform Random Variate Generations
►Numerical Integration
►Numerical Methods for Stochastic Differential Equations
►R Language
►Statistical Software: An Overview
►Uniform Random Number Generators

References and Further Reading

Andrews DF, Stafford JEH (2000) Symbolic computation for statistical inference. Oxford University Press, New York
Bolstad WM (2009) Understanding computational Bayesian statistics. Wiley, USA
Chen C, Härdle W, Unwin A (2008) Handbook of data visualization. Springer, Berlin
Drew JH, Evans DL, Glen AG, Leemis LM (2008) Computational probability. Springer, New York
Gentle JE (2009) Computational statistics. Springer, New York
Gentle JE, Härdle W, Mori Y (eds) (2004) Handbook of computational statistics: concepts and methods. Springer, Berlin
Givens GH, Hoeting JA (2005) Computational statistics. Wiley, New Jersey
Martinez WL, Martinez AR (2007) Computational statistics handbook with MATLAB, 2nd edn. Chapman & Hall, New York
mathStatica (2002–2010), www.mathStatica.com
Rao CR (1993) Handbook of statistics 9: computational statistics. Elsevier, Amsterdam
Rose C, Smith MD (2002) Mathematical statistics with Mathematica. Springer, New York

Conditional Expectation and Probability

Takis Konstantopoulos
Professor
Heriot-Watt University, Edinburgh, UK

In its most elementary form, the conditional probability $P(A|B)$ of an event A given an event B is defined by

$$P(A|B) = \frac{B(A \cap B)}{P(B)},$$

provided that $P(B) \neq 0$. This is a well-motivated definition, compatible both with the frequency interpretation of probability as well as with elementary probability on countable spaces. An immediate consequence of the definition is ▶*Bayes' theorem*: if $A_1, A_2, \ldots, A_n$ are mutually disjoint events whose union has probability one, then $P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$.

Suppose now that X, Y are random variables taking values in finite sets. We define the conditional distribution of X given Y by

$$P(X = x|Y = y) = \begin{cases} \frac{P(X=x,Y=y)}{P(Y=y)}, & \text{if } P(Y = y) \neq 0 \\ 0, & \text{if } P(Y = y) = 0. \end{cases}$$

The latter choice, i.e., 0/0 interpreted as 0, is both physically motivated and mathematically desirable. The object $P(X = x|Y = y)$ is a probability in x (i.e., it sums up to 1 over x) and a function of y. If X takes values in a set of real numbers then we can define the conditional expectation of X given $Y = y$ by

$$E(X|Y = y) = \sum_x xP(X = x|Y = y), \tag{1}$$

where the summation extends over all possible values x of X. This is a function of y, say $h(y) = E(X|Y = y)$. We can then talk about the conditional expectation of X given Y as the *random variable* $h(Y)$ obtained by substituting y by the random variable Y in the argument of $h(y)$. From this definition the following important property of $E(X|Y)$ is easily derived:

$$E\big[(X - E(X|Y)) \cdot g(Y)\big] = 0, \tag{2}$$

for any random variable $g(Y)$ which is a (deterministic) function of Y.

One can easily generalise the above to countably-valued random variables. However, defining conditional probability and expectation for general random variables cannot be done in the previous naive manner. One can mimic the definitions for random variables possessing density but this has two drawbacks: first, it is not easy to rigorously reconcile with the previous definitions; second, it is not easy to generalize. Instead, we resort to an axiomatic definition of conditional expectation, stemming directly from the fundamental property (2). It can be easily verified that, in the earlier setup, there is only one function $h(y)$ satisfying (2) for all functions $g(y)$, and this $h(y)$ is defined by (1).

The last observation leads us to the following definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a positive random variable (i.e., a measurable function $X : \Omega \to \mathbb{R}_+$). Let $\mathcal{G} \subset \mathcal{F}$ be another sigma-algebra. We say that $E(X|\mathcal{G})$ is the conditional expectation of X given $\mathcal{G}$ if (a) $E(X|\mathcal{G})$ is $\mathcal{G}$-measurable and (b) for all bounded $\mathcal{G}$-measurable random variables G, we have

$$E[XG] = E[E(X|\mathcal{G})G]. \tag{3}$$

Such an object exists and is *almost surely* unique. The latter means that if two random variables, H_1, H_2, say, satisfy $E[XG] = E[H_iG]$, $i = 1, 2$, for all G then $P(H_1 = H_2) = 1$. (Such H_i are called *versions* of the conditional expectation.) Existence is immediate by the ▶Radon–Nikodým theorem. Consider two measures on $(\Omega, \mathcal{G})$: the first one is P; the second one is $E[X1_G]$, $G \in \mathcal{G}$ (where 1_G is defined as 1 on G and 0 on $\Omega\backslash G$). When $P(G) = 0$ we have $E[X1_G] = 0$ and therefore the second measure is *absolutely continuous* with respect to the first. The Radon–Nikodým theorem ensures that the derivative (density) of the second measure with respect to the first exists and that it satisfies (3). This observation and string of arguments is due to Kolmogorov (1933), and it is through this that modern Probability Theory was established as a mathematical discipline having a natural connection with Measure Theory.

Having defined $E[X|\mathcal{G}]$ for positive X we can define it for negative X by reversing signs and for general X via the formula $E[X|\mathcal{G}] = E[\max(X, 0)|\mathcal{G}] + E[\min(X, 0)|\mathcal{G}]$, provided that wither $E[\max(X, 0)] < \infty$ or $E[\min(X, 0)] > -\infty$.

Given then two random variables X, Y (the first of which is real-valued, but the second may take values in fairly arbitrary spaces (such as a space of functions), we can define $E[X|Y]$ as $E[X|\sigma(Y)]$ where $\sigma(Y)$ is the σ-algebra generated by Y. It can be seen that this is entirely compatible with the initial definition (1).

Passing on to conditional probability, observe that if A is an event, the expectation of 1_A is precisely $P(A)$. By analogy, we define

$$P(A|\mathcal{G}) = E[1_A|\mathcal{G}].$$

For each event $A \in \mathcal{F}$, this is a random variable, i.e., a measurable function of $\omega \in \Omega$ which is defined almost surely uniquely (see explanation after formula (3)). For a real-valued random variable X we define the conditional distribution function $P(X \leq x|\mathcal{G})$ as $E[1_{X \leq x}|\mathcal{G}]$. We would like this to be a right-continuous non-decreasing function of x. Since, for each x, $P(X \leq x|\mathcal{G})$ is defined only almost surely, we need to show that we can, for each x, pick a version H_x of $P(X \leq x|\mathcal{G})$ in a way that the probability of the event $\{H_x \leq H_y \text{ if } x \leq y \text{ and } \lim_{\epsilon \downarrow 0} H_{x+\epsilon} = H_x\}$ is one. This can be done and $\{H_x\}_{x \in \mathbb{R}}$ is referred to as a *regular conditional distribution function* of X given $\mathcal{G}$. Informally (and in practice) it is denoted as $P(X \leq x|\mathcal{G})$. Regular conditional

probabilities exist not only for real random variables X but also for random elements X taking values in a Borel space Kallenberg (2002).

From this general viewpoint we can now go down again and verify everything we wish to have defined in an intuitive or informal manner. For instance, if (X, Y) is a pair of real-valued random variables having joint density $f(x,y)$ then, letting $f_Y(y) := \int_{\mathbb{R}} f(x,y)dx$, define

$$h(x|y) := \begin{cases} \frac{f(x,y)}{f_Y(y)}, & \text{if } f_Y(y) \neq 0 \\ 0, & \text{if } f_Y(y) = 0 \end{cases}.$$

Then the function $\int_{-\infty}^{x} h(s|Y)ds$ serves as a the regular conditional distribution of X given Y. We can also verify that $E(\max(X,0)|Y) = \int_0^\infty P(X > x|Y)dx$ (in the sense that the right-hand side is a version of the left) and several other elementary formulae.

It is important to mention an interpretation of $E(X|Y)$ as a projection. First recall the definition of projection in Euclidean space: Let x be a point (vector) in the space $\mathbb{R}^n$ and Π a hyperplane. We define the projection $\widehat{x}$ of x onto Π as the unique element of Π which has minimal distance from x. Equivalently, the angle between $x - \widehat{x}$ and any other vector g of Π must be 90°: this can be written as $\langle x - \widehat{x}, g\rangle = 0$, i.e., the standard inner product of $x - \widehat{x}$ and g is equal to 0. Next suppose that $E[X^2] < \infty$. Then it can be seen that

$$E[(X - E(X|\mathcal{G}))^2] = \min_G E[(X - G)^2],$$

where the minimum is taken over all $\mathcal{G}$-measurable random variables G with $E[G^2] < \infty$. The defining property (3) then says that the inner product between $X - E(X|\mathcal{G})$ and any G is zero, just as in Euclidean space. Keeping the geometric meaning in mind, we can devise (prove and interpret) several properties of the conditional expectation. We mention one below.

The *tower property*: If $\mathcal{G}_2 \subset \mathcal{G}_1$ are two sigma-algebras then

$$E[E(X|\mathcal{G}_1)|\mathcal{G}_2] = E[X|\mathcal{G}_2].$$

The geometric meaning is as follows: If Π_1 is a hyperplane (e.g., a plane in three dimensions) and Π_2 a hyperplane contained in Π_1 (e.g., a line on the plane) then we can find the projection onto Π_2 by first projecting onto Π_1 and then projecting the projection. The tower property holds for general random variables as long as conditional expectation can be defined, i.e., it does not require $E[X^2] < \infty$. Another interpretation of it is as follows: if $\mathcal{G}_1, \mathcal{G}_2$ represent states of knowledge (information, say) and $\mathcal{G}_1$ is wider than $\mathcal{G}_2$ (in the sense that $\mathcal{G}_2$ can be obtained from $\mathcal{G}_1$) then, in finding the conditional expectation of X given $\mathcal{G}_2$, the additional knowledge contained in $\mathcal{G}_1$ can be ignored. A particular form of this property is in the relation

$$E[E(X|\mathcal{G})] = E[X].$$

Another important property is that $E[GX|\mathcal{G}] = GE[X|\mathcal{G}]$ if G is $\mathcal{G}$-measurable. On the other hand, if Z is independent of (X, Y) then $E[X|Y, Z] = E[X|Y]$. For further properties, see Williams (1989). In particular, if X and Y are independent then $E[X|Y] = E[X]$, i.e., it is a constant.

For *normal* random variables, the geometric picture completely characterizes what we can do with them. Recall that a random variable X is centred normal if it has finite variance σ^2 and if for all constants a, b there is a constant c such that cX has the same distribution as $aX' + bX''$ where X', X'' are independent copies of X. It follows that $a^2 + b^2 = c^2$ and that X has density proportional to $e^{-x^2/2\sigma^2}$. We say that X is normal if $X - E[X]$ is centred normal. We say that a collection of random variables $\{X_t\}_{t\in T}$, with T being an arbitrary set, is (jointly) normal if for any $t_1, \ldots, t_n$, and any constants $c_1, \ldots, c_n$, the random variable $c_1X_{t_1} + \cdots + c_nX_{t_n}$ is normal. It follows that if $\{X, Y_1, \ldots, Y_k\}$ are jointly normal then $E[X|Y_1, \ldots, Y_k] = E[X|\sigma(Y_1, \ldots, Y_k)] = a_1Y_1 + \cdots + a_kY_k + b$ where the constants can be easily computed by (3). The *Kalman filter property* says that if $\{X, Y_1, Y_2\}$ are centred jointly normal such that Y_1 and Y_2 are independent then $E[X|Y_1, Y_2] = E[X|Y_1] + E[X|Y_2]$. The geometric interpretation of this is: to project a vector onto a plane defined by two orthogonal lines, we project to each line and then add the projections. The Kalman filter is one of the important applications of Probability to the fields of Signal Processing, Control, Estimation, and Inference (Catlin 1989).

By the term *conditioning* in Probability we often mean an effective application of the tower property in order to define a probability measure or to compute the expectation of a functional. For example, if $X_1, X_2, \ldots$ are i.i.d. positive random variables and an N is a geometric random variable, say $P(N = n) = \alpha^{n-1}(1 - \alpha)$, $n = 1, 2, \ldots$, then $E[\theta^{X_1+\cdots+X_N}] = E[E(\theta^{X_1+\cdots+X_N}|N)]$. But $E[\theta^{X_1+\cdots+X_N}|N = n] = E[\theta^{X_1+\cdots+X_n}] = (E[\theta^{X_1}])^n$, by independence. Hence $E[\theta^{X_1+\cdots+X_N}|N] = (E[\theta^{X_1}])^N$ and so $E[\theta^{X_1+\cdots+X_N}] = E[(E[\theta^{X_1}])^N] = (1 - \alpha)/(1 - \alpha E[\theta^{X_1}])$. Conditional expectation and probability are used in defining various classes of ►stochastic processes such as ►*martingales* and ►*Markov chains* (Williams 1989). Conditional probability is a fundamental object in ►*Bayesian statistics* (Williams 2001). Other applications are in the field of Financial Mathematics where the operation of taking conditional expectation of a future random variable with respect to the sigma-algebra of all events prior to the

current time t plays a fundamental role. In fact, it can be said that the notion of conditioning, along with that of independence and coupling, are the cornerstones of modern probability theory and its widespread applications.

About the Author

For biography *see* the entry ►Radon–Nikodým Theorem.

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bivariate Distributions
►Expected Value
►Radon–Nikodým Theorem

References and Further Reading

Catlin D (1989) Estimation, control, and the discrete Kalman filter. Springer, New York
Kallenberg O (2002) Foundations of modern probability, 2nd edn. Springer, New York
Kolmogorov A (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Julius Springer, Berlin (English translation by Chelsea, New York, 1956)
Williams D (1989) Probability with Martingales. Cambridge University Press, Cambridge
Williams D (2001) Weighing the odds: a course in probability and statistics. Cambridge University Press, Cambridge

Confidence Distributions

William E. Strawderman
Professor
Rutgers University, Newark, NJ, USA

A Confidence Distribution (CD), $H(X,\theta)$, for a parameter is a function of the data, X, and the parameter in question, θ, such that: (a) for each data value X, $H(X,.)$ is a (continuous) cumulative distribution function for the parameter, and (b) for the true parameter value, θ_0, $H(.,\theta_0)$ has a uniform distribution (see ►Uniform Distribution in Statistics) on the interval (0,1). The concept of CD has its historic roots in Fisher's fiducial distribution (Fisher 1930), although, in its modern version, it is a strictly frequentist construct (Schweder and Hjort 2002, 2003; Singh et al. 2005, 2007, and see also Efron 1993). The CD carries a great deal of information pertinent to a variety of frequentist inferences and may be used for the construction of confidence intervals, tests of hypotheses and point estimation.

For instance, the αth quantile of $H(X,.)$, is the upper end of a $100(1-\alpha)$ percent one sided confidence interval for θ, and also the interval formed by the sth and tth quantiles $(s<t)$ is a $100(t-s)$ percent confidence interval. These properties indicate that a confidence distribution is, in a sense, a direct frequentist version of Fisher's fiducial distribution.

Similarly, a level α test of the one-sided hypothesis $K_0: \theta \le \theta_0$ versus $K_1: \theta > \theta_0$ is given by rejecting K_0 when $H(X,\theta_0) \le \alpha$, and an analogous result holds for testing $K_0: \theta \ge \theta_0$ versus $K_1: \theta < \theta_0$. Additionally, for testing the two sided hypothesis $K_0: \theta = \theta_0$ versus $K_1: \theta \ne \theta_0$, the rejection region $2\{\min(H(X,\theta_0), 1-H(X,\theta_0))\} \ge \alpha$ gives an α level test.

The CD may also be used in a natural ways to construct a point estimate of θ. Perhaps the most straightforward estimator is the median of $H(X,.)$, which is median unbiased, and under mild conditions, consistent. Another obvious estimator, $\tilde{\theta} = \int \theta(\partial \mathrm{H}(\mathrm{X},\theta)/\partial\theta)\mathrm{d}\theta$ is also consistent under weak conditions.

One particularly simple way to construct a CD is via a pivotal quantity, $\psi(X,\theta)$, a function of X and θ whose cumulative distribution function, $G(.)$ under the true θ does not depend on θ. Then $G(\Psi(X,\theta))$ is a CD provided $\Psi(X,\theta)$ is increasing in θ. Such quantities are easy to construct in invariant models such as location or scale models. Here is a prototypical example in a normal location model. Suppose $X_i \sim N(\theta,1)$, for $i = 1,\cdots,n$, are iid. Then $\Psi(X_1,\cdots,X_n,\theta) = (\overline{X}-\theta) \sim N(0,1/n)$ so that $H(X_1,\cdots,X_n,\theta) = \Phi^{-1}\left(\sqrt{n}(\theta-\overline{X})\right)$ is a CD.

Another common construction is based on a series of one sided α-level tests of $K_0: \theta \le \theta_0$ versus $K_1: \theta > \theta_0$. If the function $P[\theta_0, X]$ is a p-value for each value of θ_0, then typically $P[\theta_0,.]$ has a uniform distribution for each value of θ_0, and hence $H(X,\theta) = P[\theta, X]$ is a CD.

The above discussion can be extended naturally to include the notion of an asymptotic CD by replacing (b) above, with the requirement that $H(.,\theta_0)$ approaches a uniform distribution on $(0,1)$ weakly as the sample size approaches infinity, and dropping the continuity requirement in (a). Profile likelihoods (see, e.g., Efron 1993; Schweder and Hjort 2002; Singh et al. 2007), and Bootstrap Distributions (see Efron 1998; Singh et al. 2005, 2007) are asymptotic CD's under weak conditions.

It can also be extended to include nuisance parameters. For example, in the case of a sample from a normal population with unknown mean and variance, the usual t-pivot can be used to construct a CD for the mean, while the usual chi-square pivot can be used to construct a CD for the variance.

See Schweder and Hjort (2002, 2003) or Singh et al. (2005, 2007), for more detailed discussion on construction,

properties and uses of *CD*'s. In particular Singh et al. (2005) discusses the combination of information from independent sources via *CD*'s.

About the Author

William E. Strawderman is Professor of Statistics and Biostatistics at Rutgers University, and past chair of the Department of Statistics. He is a Fellow of IMS and ASA and an elected member if ISI and has served on the councils of IMS and ISBA, as Chair of the ASA Section on Bayesian Statistics. He had been on the editorial boards of the Annals of Statistics, JASA, and the IMS Lecture Notes series. He has won the Distinguished Alumni Award in Science of the Graduate School at Rutgers University, and the ASA's Youden Award in Interlaboratory Studies.

Cross References

►Bootstrap Methods
►Confidence Interval
►Data Depth
►Fiducial Inference

References and Further Reading

Efron B (1993) Bayes and likelihood calculations from confidence intervals. Biometrika 80:3–26

Efron B (1993) Empirical Bayes methods for combining likelihoods (with discussion), (1998). J Am Stat Assoc 91:538–565

Fisher RA (1930) Inverse Probability. Proc Camb Philos Soc 26: 528–535

Schweder T, Hjort NL (2002) Confidence and likelihood. Scand J Stat 29:309–332

Schweder T, Hjort NL (2003) Frequentist analogues of priors and posteriors. In: Econometrics and the philosophy of economics. Princeton University Press, Princeton, NJ, pp 285–317

Singh K, Xie M, Strawderman WE (2005) Combining information from independent sources through confidence distributions. Ann Stat 33:159–183

Singh K, Xie M, Strawderman WE (2007) Confidence distribution (CD)-distribution estimator of a parameter. In: Regina Liu, William Strawderman, and Cun-Hui Zhang (eds) Complex datasets and inverse problems. IMS Lecture Notes-Monograph Series, 54, pp 132–150

Confidence Interval

Michael Smithson
Professor
The Australian National University, Canberra, ACT, Australia

A $100(1-\alpha)\%$ confidence interval is an interval estimate around a population parameter θ that, under repeated random samples of size N, is expected to include θ's true value $100(1-\alpha)\%$ of the time. The confidence interval thereby indicates the precision with which a population parameter is estimated by a sample statistic, given N and α. For many statistics there are also methods of constructing *confidence regions,* which are multivariate versions of simultaneous confidence intervals.

The *confidence level,* $100(1-\alpha)\%$, is chosen a priori. A *two-sided* confidence interval uses a lower limit L and upper limit U that each contain θ's true value $100(1-\alpha/2)\%$ of the time, so that together they contain θ's true value $100(1-\alpha)\%$ of the time. This interval often is written as $[L, U]$, and sometimes writers combine a confidence level and interval by writing $\Pr(L \leq \theta \leq U) = 1-\alpha$. In some applications, a *one-sided* confidence interval is used, primarily when only one limit has a sensible meaning or when interest is limited to bounding a parameter estimate from one side only.

The confidence interval is said to be an inversion of its corresponding significance test because the $100(1-\alpha)\%$ confidence interval includes all hypothetical values of the population parameter that cannot be rejected by its associated significance test using a Type I error-rate criterion of α. In this respect, it provides more information than a significance test does. Confidence intervals become narrower with larger sample size and/or lower confidence levels. Narrower confidence intervals imply greater statistical power for the corresponding significance test, but the converse does not always hold.

The limits L and U are derived from a sample statistic (often the sample estimate of θ) and a sampling distribution specifying a probability for each value that the sample statistic can take. Thus L and U also are sample statistics and will vary from one sample to another. This fact underscores a crucial point of interpretation regarding a confidence interval, namely that we cannot claim that a particular interval has a $1-\alpha$ probability of containing the population parameter value.

A widespread practice regarding two-sided confidence intervals is to place L and U so that α is evenly split between the lower and upper tails. This is often a matter of convention, but can be dictated by criteria that statisticians have used for determining the "best" possible confidence interval. One such criterion is simply narrowness. It is readily apparent, for instance, that if a sampling distribution is symmetric and unimodal then for high confidence levels the shortest $100(1-\alpha)\%$ confidence interval constructed from that distribution is one that allocates $\alpha/2$ to the tails outside of the lower and upper limits.

Other criteria for evaluating confidence intervals are as follows. A $100(1-\alpha)\%$ confidence interval is *exact* if it can be expected to contain the relevant parameter's true value $100(1-\alpha)\%$ of the time. When approximate intervals

are used instead, if the rate of coverage is greater than $100(1-\alpha)\%$ then the interval is *conservative*; if the rate is less then the interval is *liberal*. The $100(1-\alpha)\%$ interval that has the smallest probability of containing values other than the true parameter value is said to be *uniformly most accurate*. A confidence interval whose probability of including any value other than the parameter's true value is less than or equal to $100(1-\alpha)\%$ is *unbiased*.

Example 1 Suppose that a standard IQ test has been administered to a random sample of $N = 25$ adults from a large population with a sample mean of 103 and standard deviation $s = 10$. We will construct a two-sided 95% confidence interval for the mean, μ. The limits U and L must have the property that, given a significance criterion of α, sample size of 25, mean of 103 and standard deviation of 10, we could reject the hypotheses that $\mu > 103 + U$ or $\mu < 103 - L$ but not $L \le \mu \le U$.

The sampling distribution of the t-statistic defined by $t = \dfrac{\overline{X} - \mu}{s_{err}}$ is a t-distribution with $df = N - 1 = 24$. When $df = 24$ the value $t_{\alpha/2} = 2.064$ standard-error units above the mean cuts $\alpha/2 = .025$ from the upper tail of this t-distribution, and likewise $-t_{\alpha/2} = -2.064$ standard-error units below the mean cuts $\alpha/2 = .025$ from the lower tail. The sample standard error is $s_{err} = s/\sqrt{N} = 4.128$. So a t-distribution around $U = 103 + (2.064)(4.128) = 107.13$ has .025 of its tail below 103, while a t-distribution around $L = 103 - (2.064)(4.128) = 98.87$ has 0.025 of its tail above 103. These limits fulfill the above required property, so the 95% confidence interval for μ is $[98.87, 107.13]$. Thus, we cannot reject hypothetical values of μ that lie between 98.87 and 107.13, using $\alpha = .05$.

Example 2 (transforming one interval to obtain another) Cohen's d for two independent samples is defined by $\delta = (\mu_1 - \mu_2)/\sigma_p$, where μ_1 and μ_2 are the means of two populations from which the samples have been drawn and σ_p is the population pooled standard deviation. This quantity has a noncentral t distribution with a noncentrality parameter $\Delta = \delta[N_1N_2/(N_1 + N_2)]^{1/2}$, where N_1 and N_2 are the sizes of the two samples. The sample t-statistic is the sample estimate of Δ. Suppose a two-condition between-subjects experiment with $N_1 = N_2 = 40$ yields $t(78) = 3.45$. Using an appropriate algorithm (Smithson 2003) we can find the 95% confidence interval for Δ, which is $[1.407, 5.473]$. Because δ and Δ are monotonically related by $\delta = \Delta/[N_1N_2/(N_1 + N_2)]^{1/2}$, we can obtain a 95% confidence interval for δ by applying this formula to the lower and upper limits of the interval for Δ. The sample estimate of δ is $d = t/[N_1N_2/(N_1 + N_2)]^{1/2} = 3.45/4.472 = 0.771$, and applying the same transformation to the limits of the interval for Δ gives an interval of $[0.315, 1.224]$ for δ.

About the Author

Michael Smithson is a Professor in the Department of Psychology at The Australian National University in Canberra, and received his Ph.D. from the University of Oregon. He is the author of *Confidence Intervals* (2003), *Statistics with Confidence* (2000), *Ignorance and Uncertainty* (1989), and *Fuzzy Set Analysis for the Behavioral and Social Sciences* (1987), coauthor of *Fuzzy Set Theory: Applications in the Social Sciences* (2006), and coeditor of *Uncertainty and Risk: Multidisciplinary Perspectives* (2008) and *Resolving Social Dilemmas: Dynamic, Structural, and Intergroup Aspects* (1999). His other publications include more than 120 refereed journal articles and book chapters. His primary research interests are in judgment and decision making under uncertainty and quantitative methods for the social sciences.

Cross References

- ▶Confidence Distributions
- ▶Decision Trees for the Teaching of Statistical Estimation
- ▶Effect Size
- ▶Fuzzy Logic in Statistical Data Analysis
- ▶Margin of Error
- ▶Sample Size Determination
- ▶Statistical Fallacies: Misconceptions, and Myths
- ▶Statistical Inference
- ▶Statistical Inference: An Overview

References and Further Reading

Altman DG, Machin D, Bryant TN, Gardner MJ (2000) Statistics with confidence: confidence intervals and statistical guidelines, 2nd edn. British Medical Journal Books, London

Smithson M (2003) Confidence intervals. Sage University Papers on Quantitative Applications in the Social Sciences, 139. Sage, Thousand Oaks

Confounding and Confounder Control

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Introduction

The word *confounding* has been used to refer to at least three distinct concepts. In the oldest and most widespread usage, confounding is a source of bias in estimating causal

effects. This bias is sometimes informally described as a mixing of effects of extraneous factors (called confounders) with the effect of interest, and important in causal inference (*see* ▶Causation and Causal Inference). This usage predominates in nonexperimental research, especially in epidemiology and sociology. In a second and more recent usage originating in statistics, confounding is a synonym for a change in an effect measure upon stratification or adjustment for extraneous factors (a phenomenon called *noncollapsibility* or *Simpson's paradox; see* ▶Simpson's Paradox; Collapsibility). In a third usage, originating in the experimental-design literature, confounding refers to inseparability of main effects and interactions under a particular design (*see* ▶Interaction).

The three concepts are closely related and are not always distinguished from one another. In particular, the concepts of confounding as a bias in effect estimation and as noncollapsibility are often treated as equivalent, even though they are not. Only the first usage, confounding as a bias, will be described here; for more detailed coverage and comparisons of concepts see, Greenland et al. (1999a), Pearl (2009), and Greenland et al. (2008).

Confounding as a Bias in Effect Estimation

In the first half of the nineteenth century, John Stuart Mill described the problem of confounding in causal inference; he acknowledged the seventeenth century scientist Francis Bacon as a forerunner in dealing with these issues (Mill 1843, Chap. III). Mill listed a key requirement for an experiment intended to determine causal relations:

> "...none of the circumstances [of the experiment] that we do know shall have effects susceptible of being *confounded* with those of the agents whose properties we wish to study" (emphasis added) (Mill 1843, Chap. X).

In Mill's time the word "experiment" referred to an observation in which some circumstances were under the control of the observer, as it still is used in ordinary English, rather than to the notion of a comparative trial. Nonetheless, Mill's requirement suggests that a comparison is to be made between the outcome of our "experiment" (which is, essentially, an uncontrolled trial) and what we would expect the outcome to be if the agents we wish to study had been absent. If the outcome is not as one would expect in the absence of the study agents, then Mill's requirement ensures that the unexpected outcome was not brought about by extraneous "circumstances" (factors). If, however, those circumstances do bring about the unexpected outcome, and that outcome is mistakenly attributed to effects of the study agents, then the mistake is one of confounding (or confusion) of the extraneous effects with the agent effects.

Much of the modern literature follows the same informal conceptualization given by Mill. Terminology is now more specific, with "treatment" used to refer to an agent administered by the investigator and "exposure" often used to denote an unmanipulated agent. The chief development beyond Mill is that the expectation for the outcome in the absence of the study exposure is now almost always explicitly derived from observation of a control group that is untreated or unexposed.

Confounding typically occurs when natural or social forces or personal preferences affect whether a person ends up in the treated or control group, and these forces or preferences also affect the outcome variable. While such confounding is common in observational studies, it can also occur in randomized experiments when there are systematic improprieties in treatment allocation, administration, and compliance. A further and somewhat controversial point is that confounding (as per Mill's original definition) can also occur in perfect randomized trials due to *random* differences between comparison groups (Fisher 1935; Rothman 1977); this problem will be discussed further below.

The Potential-Outcome Model of Confounding

Various models of confounding have been proposed for use in statistical analyses. Perhaps the one closest to Mill's concept is based on the *potential-outcome* or counterfactual model for causal effects (*see* ▶Causation and Causal Inference). Suppose we wish to consider how a health-status (outcome) measure of a population would change in response to an intervention (population treatment). More precisely, suppose our objective is to determine the effect that applying a treatment x_1 had or would have on an outcome measure μ relative to applying treatment x_0 to a specific target population A. For example, this population could be a cohort of breast-cancer patients, treatment x_1 could be a new hormone therapy, x_0 could be a placebo therapy, and the measure μ could be the 5-year survival probability. The treatment x_1 is sometimes called the *index* treatment; and x_0 is sometimes called the *control* or *reference* treatment (which is often a standard or placebo treatment).

The potential-outcome model posits that, in population A, μ will equal μ_{A1} if x_1 is applied, μ_{A0} if x_0 is applied; the causal effect of x_1 relative to x_0 is defined as the change from μ_{A0} to μ_{A1}, which might be measured as $\mu_{A1} - \mu_{A0}$, or if μ is strictly positive, μ_{A1}/μ_{A0}. If A is given treatment x_1,

then μ will equal μ_{A1} and μ_{A1} will be observable, but μ_{A0} will be unobserved.

Suppose now that μ_{B0} is the value of the outcome μ observed or estimated for a population B that was administered treatment x_0. If this population is used as a substitute for the unobserved experience of population A under treatment x_0, it is called the control or reference population. *Confounding* is said to be present if $\mu_{A0} \neq \mu_{B0}$, for then there must be some difference between populations A and B *other than treatment difference* that is affecting μ.

If confounding is present, a naive (crude) association measure obtained by substituting μ_{B0} for μ_{A0} in an effect measure will not equal the effect measure, and the association measure is said to be *confounded.* Consider $\mu_{A1} - \mu_{B0}$, which measures the association of treatments with outcomes *across* the populations. If $\mu_{A0} \neq \mu_{B0}$, then $\mu_{A1} - \mu_{B0}$ is said to be *confounded* for $\mu_{A1} - \mu_{A0}$, which measures the effect of treatment x_1 on population A. Thus, to say an association measure $\mu_{A1} - \mu_{B0}$ is confounded for an effect measure $\mu_{A1} - \mu_{A0}$ is to say these two measures are not equal.

Dependence of Confounding on the Outcome Measure and the Population

A noteworthy aspect of the potential-outcome model is that confounding depends on the outcome measure. For example, suppose populations A and B have a different 5-year survival probability μ under placebo treatment x_0; that is, suppose $\mu_{B0} \neq \mu_{A0}$ so that $\mu_{A1} - \mu_{B0}$ is confounded for the actual effect $\mu_{A1} - \mu_{A0}$ of treatment on 5-year survival. It is then still possible that 10-year survival, v, under the placebo would be identical in both populations; that is v_{A0} could still equal v_{B0}, so that $v_{A1} - v_{B0}$ is not confounded for the actual effect of treatment on 10-year survival. Lest one think this situation unlikely, note that we should generally expect no confounding for 200-year survival, since no known treatment is likely to raise the 200-year survival probability of human patients above zero.

Even though the presence of confounding is dependent on the chosen outcome measure, as defined above its presence does not depend on how the outcome is contrasted between treatment levels. For example, if Y is binary so that $\mu = E(Y)$ is the Bernoulli parameter or risk $\Pr(Y = 1)$, then the risk difference $\mu_{A1} - \mu_{B0}$, risk ratio μ_{A1}/μ_{B0}, and odds ratio $\{\mu_{A1}/(1-\mu_{A1})\}/\{\mu_{B0}/(1-\mu_{B0})\}$ are all confounded under exactly the same circumstances. In particular, and somewhat paradoxically, confounding may be absent even if the odds ratio changes upon covariate adjustment, i.e., even if the odds ratio is noncollapsible (Greenland and Robins 1986; Greenland et al. 1999a, 2008; *see* ►Collapsibility).

A second noteworthy point is that confounding depends on the target population. The preceding example, with A as the target, had different 5-year survivals μ_{A0} and μ_{B0} for A and B under placebo therapy, and hence $\mu_{A1} - \mu_{B0}$ was confounded for the effect $\mu_{A1} - \mu_{A0}$ of treatment on population A. A lawyer or ethicist may also be interested in what effect the hormone treatment would have had on population B. Writing μ_{B1} for the (unobserved) outcome under treatment, this effect on B may be measured by $\mu_{B1} - \mu_{B0}$. Substituting μ_{A1} for the unobserved μ_{B1} yields $\mu_{A1} - \mu_{B0}$. This measure of association is confounded for $\mu_{B1} - \mu_{B0}$ (the effect of treatment x_1 on 5-year survival in population B) if and only if $\mu_{A1} \neq \mu_{B1}$. Thus, the same measure of association, $\mu_{A1} - \mu_{B0}$, may be confounded for the effect of treatment on neither, one, or both of populations A and B, and may or may not be confounded for the effect of treatment on other targets such as the combined population $A \cup B$.

Confounders (Confounding Factors) and Covariate Imbalance

The potential-outcome model is that it invokes no explicit differences (imbalances) between populations A and B with respect to circumstances or covariates that might influence μ. (Greenland and Robins 1986, 2009). Clearly, if μ_{A0} and μ_{B0} differ, then A and B must differ with respect to factors that influence μ. This observation has led some authors to define confounding as the presence of such covariate differences between the compared populations (Stone 1993). This is incorrect, however, because confounding is only a consequence of these covariate differences. In fact, A and B may differ profoundly with respect to covariates that influence μ, and yet confounding may be absent. In other words, a covariate difference between A and B is a necessary but not sufficient condition for confounding, as can be seen when the impact of covariate differences may balance each other out, leaving no confounding.

Suppose now that populations A and B differ with respect to certain covariates, and that these differences have led to confounding of an association measure for the effect measure of interest. The responsible covariates are then termed *confounders* of the association measure. In the above example, with $\mu_{A1} - \mu_{B0}$ confounded for the effect $\mu_{A1} - \mu_{A0}$, the factors responsible for the confounding (i.e., the factors that led to $\mu_{A0} \neq \mu_{B0}$) are the confounders.

It can be deduced that a variable cannot be a confounder unless it can affect the outcome parameter μ within treatment groups and it is distributed differently among the compared populations (e.g., see Yule 1903, who uses terms such as "fictitious association" rather than confounding). These two necessary conditions are sometimes

offered together as a definition of a confounder. Nonetheless, counterexamples show that the two conditions are not sufficient for a variable with more than two levels to be a confounder (Greenland et al. 1999a). Note that the condition of affecting the outcome parameter is a causal assertion and thus relies on background knowledge for justification (Greenland and Robins 1986; Robins 2001; Pearl 2009).

Control of Confounding

Prevention of Confounding

An obvious way to avoid confounding is estimating $\mu_{A1} - \mu_{A0}$ is to obtain a reference population B for which μ_{B0} is known to equal μ_{A0}. Such a population is sometimes said to be *comparable* to or *exchangeable* with A with respect to the outcome under the reference treatment. In practice, such a population may be difficult or impossible to find. Thus, an investigator may attempt to construct such a population, or to construct exchangeable index and reference populations. These constructions may be viewed as design-based methods for the control of confounding.

Perhaps no approach is more effective for preventing confounding by a known factor than *restriction*. For example, gender imbalances cannot confound a study restricted to women. However, there are several drawbacks: restriction on enough factors can reduce the number of available subjects to unacceptably low levels, and may greatly reduce the generalizability of results as well. *Matching* the treatment populations on confounders overcomes these drawbacks, and, if successful, can be as effective as restriction. For example, gender imbalances cannot confound a study in which the compared groups have identical proportions of women. Unfortunately, differential losses to observation may undo the initial covariate balances produced by matching.

Neither restriction nor matching prevents (although it may diminish) imbalances on unrestricted, unmatched, or unmeasured covariates. In contrast, ►*randomization* offers a means of dealing with confounding by covariates not accounted for by the design. It must be emphasized, however, that this solution is only probabilistic and subject to severe constraints in practice. Randomization is not always feasible or ethical, and many practical problems (such as differential loss and noncompliance) can lead to confounding in comparisons of the groups actually receiving treatments x_1 and x_0.

One somewhat controversial solution to noncompliance problems is *intent-to-treat analysis,* which defines the comparison groups A and B by treatment assigned rather than treatment received. Confounding may, however, affect even intent-to-treat analyses, and (contrary to widespread misperceptions) the bias in those analyses can exaggerate the apparent treatment effect (Robins 1998). For example, the assignments may not always be random, as when blinding is insufficient to prevent the treatment providers from protocol violations. And, purely by bad luck, randomization may itself produce allocations with severe covariate imbalances between the groups (and consequent confounding), especially if the study size is small (Fisher 1935; Rothman 1977). *Blocked* (matched) randomization can help ensure that random imbalances on the blocking factors will not occur, but it does not guarantee balance of unblocked factors.

Adjustment for Confounding

Design-based methods are often infeasible or insufficient to prevent confounding. Thus, there has been an enormous amount of work devoted to analytic adjustments for confounding. With a few exceptions, these methods are based on observed covariate distributions in the compared populations. Such methods can successfully control confounding only to the extent that enough confounders are adequately measured. Then, too, many methods employ parametric models at some stage, and their success may thus depend on the faithfulness of the model to reality. These issues cannot be covered in depth here, but a few basic points are worth noting. The simplest and most widely trusted methods of adjustment begin with *stratification* on confounders. A covariate cannot be responsible for confounding within internally homogeneous strata of the covariate. For example, gender imbalances cannot confound observations within a stratum composed solely of women. More generally, comparisons within strata cannot be confounded by a covariate that is unassociated with treatment within strata. This is so, whether the covariate was used to define the strata or not. Thus, one need not stratify on all confounders in order to control confounding; it suffices to stratify on a balancing score (such as a propensity score) that yields strata in which the confounders are unassociated with treatment.

If one has accurate background information on relations among the confounders, one may use this information to identify sets of covariates statistically sufficient for adjustment, for example by using causal diagrams or conditional independence conditions (Pearl 1995, 2009; Greenland et al. 1999ab; Glymour and Greenland 2008). Nonetheless, if the stratification on the confounders is too coarse (e.g., because categories are too broadly defined), stratification may fail to adjust for much of the confounding by the adjustment variables.

One of the most common adjustment approaches today is to enter suspected confounders into a model for the outcome parameter μ. For example, let μ be the mean (expectation) of an outcome variable of interest Y,

let X be the treatment variable of interest, and let Z be a suspected confounder of the $X - Y$ relation. Adjustment for Z is often made by fitting a generalized-linear model (see ▶Generalized Linear Models) $g(\mu) = g(\alpha + \beta x + \gamma z)$ or some variant, where $g(\mu)$ is a strictly increasing function such as the natural log $\ln(\mu)$, as in log-linear modeling, or the logit function $\ln\{\mu/(1-\mu)\}$, as in ▶logistic regression; the estimate of β that results is then taken as the Z-adjusted estimate of the X effect on $g(\mu)$.

An oft-cited advantage of model-based adjustment is that it allows adjustment for more variables and in finer detail than stratification. If however the form of the fitted model cannot adapt well to the true dependence of Y on X and Z, such model-based adjustments may fail to adjust for confounding by Z. For example, suppose Z is symmetrically distributed around zero within X levels, and the true dependence is $g(\mu) = g(\alpha + \beta x + \gamma z^2)$; then using the model $g(\mu) = g(\alpha + \beta x + \gamma z)$ will produce little or no adjustment for Z. Similar failures can arise in adjustments based on models for treatment probability (propensity scores). Such failures can be minimized or avoided by using reasonably flexible models, by carefully checking each fitted model against the data, and by combining treatment-probability and outcome models to produce *doubly robust* effect estimators (Hirano et al. 2003; Bang and Robins 2005).

Finally, if (as is often done) a variable used for adjustment is not a confounder, bias may be introduced by the adjustment (Greenland and Neutra 1980; Greenland et al. 1999b; Hernán et al. 2002; Pearl 2009). The form of this bias often parallels *selection bias* familiar to epidemiologists, and tends to be especially severe if the variable is affected by both the treatment and the outcome under study, as in classic *Berksonian bias* (Greenland 2003). In some but not all cases the resulting bias is a form of confounding within strata of the covariate (Greenland et al. 1999b); adjustment for covariates affected by treatment can produce such confounding, even in randomized trials (Cox 1958, Chap. 2; Greenland 2003).

Confounded Mechanisms Versus Confounded Assignments

If the mechanism by which the observational units come to have a particular treatment is independent of the potential outcomes of the units, the mechanism is sometimes described as *unconfounded* or *unbiased* for μ (Rubin 1991; Stone 1993); otherwise the mechanism is confounded or biased. Randomization is the main practical example of such a mechanism. Graphical models (*see* ▶Causal Diagrams) provide an elegant algorithm for checking whether the graphed mechanism is unconfounded within strata of covariates (Pearl 1995, 2009; Greenland et al. 1999b; Glymour and Greenland 2008). Note however that in typical epidemiologic usage the term "confounded" refers to the result of a single assignment (the study group actually observed), not the behavior of the mechanism. Thus an unconfounded mechanism can by chance produce confounded assignments.

The latter fact resolves a controversy about adjustment for baseline (pre-treatment) covariates in randomized trials. Although Fisher asserted that randomized comparisons were "unbiased," he also pointed out that particular assignments could be confounded in the single-trial sense used in epidemiology; see Fisher (1935, p. 49). Resolution comes from noting that Fisher's use of the word "unbiased" referred to the design and corresponds to an unconfounded assignment mechanism; it was not meant to guide analysis of a given trial (which has a particular assignment). Once the trial is underway and the actual treatment allocation is completed, the unadjusted treatment-effect estimate will be biased conditional on the observed allocation if the baseline covariate is associated with treatment in the allocation and the covariate affects the outcome; this bias can be removed by adjustment for the covariate (Rothman 1977; Greenland and Robins 1986, 2009; Greenland et al. 1999a).

Confounder Selection

An essential first step in the control of confounding is to identify which variables among those measured satisfied the minimal necessary conditions to be a confounder. This implies among other things that the variables cannot be affected by exposure or outcome; it thus excludes intermediate variables and effects of exposure and disease, whose control could introduce Berksonian bias. This initial screening is primarily a subject-matter decision that requires consideration of the causal ordering of the variables. Relatively safe candidate confounders will be "pre-treatment" covariates (those occurring before treatment or exposure), which have the advantage that they cannot be intermediates or effects of exposure and outcome. Exceptions occur in which control of certain pre-treatment variables introduce bias (Pearl 1995, 2009; Greenland et al. 1999b), although the bias so introduced may be much less than the confounding removed (Greenland 2003).

Variables that pass the initial causal screening are sometimes called "potential confounders." Once these are identified, the question arises as to which must be used for adjustment. A common but unjustified strategy is to select confounders to control based on a test (usually a significance test) of each confounder's association with the treatment X (a test of imbalance) or with the outcome Y, e.g.,

using stepwise regression. Suppose Z is a pre-treatment covariate (potential confounder). The strategy of testing the Z association with X arises from a confusion of two distinct inferential problems:

1. Do the treated ($X = 1$) evince larger differences from the untreated ($X = 0$) with respect to Z than one should expect from a random (or unconfounded) assignment mechanism?
2. Should we control for Z to estimate the treatment effect?

A test of the $X - Z$ association addresses question (a), but not (b). For (b), the "large-sample" answer is that control is advisable, regardless of whether the $X - Z$ association is random. This is because an imbalance produces bias conditional on the observed imbalance, even if the imbalance derived from random variation.

The mistake of significance testing for confounding lies in thinking that one can ignore an imbalance if it is from random variation. Random assignment only guarantees valid performance of statistics over all possible treatment allocations. It does not however guarantee validity conditional on the observed Z imbalance, even though any such imbalance must be random in a randomized trial. Thus the $X - Z$ test addresses a real question (one relevant to a field methodologist studying determinants of response/treatment), but is irrelevant to the second question (b) (Greenland and Neutra 1980; Robins and Morgenstern 1987; Greenland et al. 1999a).

The case of testing the Z association with Y devolves in part to whether one trusts prior (subject-matter) knowledge that Z affects Y (or is a proxy for a cause of Y) more than the results of a significance test in one's own limited data. There are many examples in which a well-known risk factor exhibits the expected association with Y in the data, but for no more than chance reasons or sample-size limitations, that association fails to reach conventional levels of "significance" (e.g., Greenland and Neutra 1980). In such cases there is a demonstrable statistical advantage to controlling Z, thus allowing subject-matter knowledge to over-ride nonsignificance (Robins and Morgenstern 1987).

Another problematic strategy is to select a potential confounder Z for control based on how much the effect estimate changes when Z is controlled. Like the testing methods described above, it also lacks formal justification and can exhibit poor performance in practice (Maldonado and Greenland 1993). The strategy can also mislead if the treatment affects a high proportion of subjects and one uses a "noncollapsible" effect measure (one that changes upon stratification even if no confounding is present), such as an odds ratio or rate ratio (Greenland and Robins 1986; Greenland 1996; Greenland et al. 1999a).

In practice, there may be too many variables to control using conventional methods, so the issue of confounder selection may seem pressing. Nonetheless, hierarchical-Bayesian or other shrinkage methods may be applied instead. These methods adjust for all the measured confounders by estimating the confounder effects using a prior distribution for those effects. See Greenland (2000, 2008) for details. Some of these methods (e.g., the Lasso; Tibshirani 1996) may drop certain variables entirely, and thus in effect result in confounder selection; unlike significance-testing based selection, however, this selection has a justification in statistical theory.

About the Author

Dr. Greenland is Professor of Epidemiology, UCLA School of Public Health, and Professor of Statistics, UCLA College of Letters and Science. He has published over 300 scientific papers, two of which have been cited over 500 times. Professor Greenland is a coeditor (with Kenneth J.h Rothman) of the highly influential text in the field of epidemiology, *Modern Epidemiology*. This book has received the highest number of citations among all texts and papers in the field, over 8,000 (M. Porta et al. 2006). *Book citations: influence of epidemiologic thought in the academich community*, Rev Saúde Púbica, 40, p. 50; Lippincott-Raven 1998. Currently, he is an Associate editor for *Statistics in Medicine*. He was Chair, Section in Epidemiology, American Statistical Association (2005–2007). He is Chartered Statistician and Fellow, Royal Statistical Society (1993), and a Fellow American Statistical Association (1998).

Cross References

►Bias Analysis
►Causal Diagrams
►Causation and Causal Inference
►Collapsibility
►Exchangeability
►Interaction
►Simpson's Paradox
►Statistical Methods in Epidemiology

References and Further Reading

Bang H, Robins J (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61:962–972
Cox DR (1958) Planning of experiments. Wiley, New York
Fisher RA (1935) The Design of experiments. Oliver & Boyd, Edinburgh

Glymour MM, Greenland S (2008) Causal diagrams. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia, pp 183–209
Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501
Greenland S (2000) When should epidemiologic regressions use random coefficients? Biometrics 56:915–921
Greenland S (2003) Quantifying biases in causal models. Epidemiology 14:300–307
Greenland S (2008) Variable selection and shrinkage in the control of multiple confounders. Am J Epidemiol 167:523–529. Erratum 1142
Greenland S, Neutra RR (1980) Control of confounding in the assessment of medical technology. Int J Epidemiol 9: 361–367
Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419
Greenland S, Robins JM (2009) Identifiability, exchangeability, and confounding revisited. Epidemiol Perspect Innov (online journal) 6:4
Greenland S, Robins JM, Pearl J (1999a) Confounding and collapsibility in causal inference. Stat Sci 14:29–46
Greenland S, Pearl J, Robins JM (1999b) Causal diagrams for epidemiologic research. Epidemiology 10:37–48
Greenland S, Rothman KJ, Lash TL (2008) Measures of effect and measures of association. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia, pp 51–70
Hernán M, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation. Am J Epidemiol 155:176–184
Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189
Maldonado G, Greenland S (1993) A simulation study of confounder-selection strategies. Am J Epidemiol 138: 923–936
Mill JS (1843) A system of logic, ratiocinative and inductive. Reprinted by Longmans. Green & Company, London, 1956
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710
Pearl J (2009) Causality, 2nd edn. Cambridge University Press, New York
Robins JM (1998) Correction for non-compliance in equivalence trials. Stat Med 17:269–302
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Morgenstern H (1987) The foundations of confounding in epidemiology. Comput Math Appl 14:869–916
Rothman KJ (1977) Epidemiologic methods in clinical trials. Cancer 39:1771–1775
Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47:1213–1234
Stone R (1993) The assumptions on which causal inference rest. J R Stat Soc B 55:455–466
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58:267–288
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Contagious Distributions

Seng Huat Ong[1], Choung Min Ng
[1]Professor
University of Malaya, Kuala Lumpur, Malaysia

The term contagious distribution was apparently first used by Neyman (1939) for a discrete distribution that exhibits clustering or contagious effect. The classical Neyman Type *A* distribution is one well-known example. However, contagious distributions are used nowadays to describe a plethora of distributions, many of which possess complicated probability distribution expressed in terms of special functions (see, for instance, Johnson et al. 2005).

It is instructive to give an account of the derivation of the Neyman Type *A* distribution as developed by Neyman (1939) in his paper "On a new class of contagious distributions applicable in entomology and bacteriology." Neyman wanted to model the distribution of larvae on plots in a field. He assumed that the number of clusters of eggs per unit area, N, followed a Poisson distribution with mean θ denoted by $Poi(\theta)$, while the number of larvae developing from the egg clusters $X_i, i = 1, 2, \ldots, N$ is distributed as another Poisson distribution $Poi\ (\phi)$. Mathematically, this may be expressed as follows:

$$S_N = X_1 + X_2 + \ldots + X_N$$

where S_N is the total number of larvae per unit area. The distribution of S_N is then a Neyman Type *A*.

The above model is known as a *true contagion* model, where the occurrence of "a favourable event depends on the occurrence of the previous favorable events" (Gurland 1959). Among other terms used for the distribution arising from this model are *generalized, clustered, stopped*, or *stopped-sum* distribution (see Douglas 1980; Johnson et al. 2005). It is convenient to represent the distribution of S_N concisely by

$$S_N \sim U_1 \vee U_2,$$

which reads S_N distribution is a U_1 distribution generalized by a U_2 distribution. As an example, the Neyman Type *A* distribution for S_N, above, is

$$\text{Neyman Type } A \sim Poi\,(\theta) \vee Poi\,(\phi).$$

In addition, the probability generating function (pgf) of S_N distribution is

$$E\left[z^{S_N}\right] = g_1(g_2(z)),$$

where $g_i(z)$ is the pgf for the corresponding $U_i, i = 1, 2$ distribution.

Next, using the same example as above but instead, the number of larvae per unit area is now considered to be distributed as a Poisson distribution, $Poi\,(k\theta)$, where due to heterogeneity, the mean number of eggs that hatched into larvae is assumed to vary with k following a Poisson distribution $Poi\,(\phi)$. The distribution for the number of larvae per unit area is again a Neyman Type A.

The model that gives rise to a distribution as in the preceding formulation is known as an *apparent contagion* model. Generally, this distribution arises when a parameter of a U_1 distribution is a random variable Ω that follows a U_2 distribution of its own. This type of distribution is also known as a *mixed, mixture*, or *compound* distribution (see Ord 1971; Johnson et al. 2005). A compound (mixture) distribution can be represented by

$$U_1 \underset{\Omega}{\wedge} U_2,$$

which means that the U_1 distribution is compounded by U_2 distribution (the distribution of Ω). U_2 is known as the compounding (mixing) distribution. Thus, the Neyman Type A distribution formulated through compounding is represented by

$$\text{Neyman Type } A \sim Poi\,(k\theta) \underset{k}{\wedge} Poi\,(\phi)$$

The pgf of a compound (mixture) distribution is

$$\int_{\omega} g_1\,(z|\,\omega)\; dF_2\,(\omega)$$

where $g_1(.)$ is the pgf for the U_1 and $F_2(.)$ is the cumulative distribution function for U_2. The class of mixed Poisson distributions is a well-known class of compound distributions that has found applications in many areas of study including biology, sociology, and medicine.

Note that both given examples of contagion models lead to the Neyman Type A distribution. The relationship between the generalized and compound distributions is given by the following theorem:

Theorem 1 (Gurland 1957) Let U_1 be a random variable with pgf $[h\,(z)]^{\theta}$, where θ is a given parameter. Suppose now θ is regarded as a random variable. Then, whatever be U_2

$$U_1 \wedge U_2 \sim U_2 \vee U_1. \qquad (1)$$

This relation shows that it may not be possible to distinguish between the two types of contagion directly from the data (Gurland 1959).

Contagious distributions have been studied by many researchers including Feller (1943), Skellam (1952), Beall and Rescia (1953), Gurland (1958), Hinz and Gurland (1970), Khatri (1971), and Hill (1993), creating a rich literature in this field. The readers are referred to Ord (1972, p. 126) for a list of generalized and compound Poisson distributions such as Polya-Aeppli, negative binomial, and Hermite distributions. Other references for generalized and compound distributions can be found in Douglas (1980, Chaps. 4 and 5) and Johnson et al. (2005, Chaps. 8 and 9). These references also describe statistical inference for the contagious distributions. Recent review articles on this subject are Gupta and Ong (2005) and Karlis and Xekalaki (2005).

Cross References

►Mixture Models
►Multivariate Statistical Distributions

References and Further Reading

Beall G, Rescia RR (1953) A generalization of Neyman's contagious distributions. Biometrics 9:354–386
Douglas JB (1980) Analysis with standard contagious distributions. International Co-operative Publishing House, Burtonsville
Feller W (1943) On a general class of "contagious" distributions. Ann Math Stat 14:389–400
Gupta RC, Ong SH (2005) Analysis of long-tailed count data by Poisson mixtures. Commun Stat 34(3):557–574
Gurland J (1957) Some interrelations among compound and generalized distributions. Biometrika 44:265–268
Gurland J (1958) A generalized class of contagious distributions. Biometrics 14:229–249
Gurland J (1959) Some applications of the negative binomial and other contagious distributions. Am J Public Health 49:1388–1399
Hill DH (1991) Response and sequencing errors in surveys: a discrete contagious regression analysis. J Am Stat Assoc 88:775–781
Hinz P, Gurland J (1970) A test of fit for the negative binomial and other contagious distributions. J Am Stat Assoc 65: 887–903
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken
Karlis D, Xekalaki E (2005) Mixed Poisson distributions. Int Stat Rev 73:35–58
Khatri CG (1971) On multivariate contagious distributions. Sankhya 33:197–216
Neyman J (1939) On a new class of "contagious" distributions, applicable in entomology and bacteriology. Ann Math Stat 10: 35–57
Ord JK (1972) Families of frequency distributions. Griffin, London
Skellam JG (1952) Studies in statistical ecology: I. spatial pattern. Biometrika 39:346–362

Continuity Correction

Rabi Bhattacharya
Professor of Mathematics
The University of Arizona, Tucson, AZ, USA

According to the central limit theorem (*CLT*) (see ►Central Limit Theorems), the distribution function F_n of a normalized sum $n^{-1/2}(X_1 + \ldots + X_n)$ of n independent random variables $X_1, \ldots, X_n$, having a common distribution with mean zero and variance $\sigma^2 > 0$, converges to the distribution function Φ_σ of the normal distribution with mean zero and variance σ^2, as $n \to \infty$. We will write Φ for Φ_1 for the case $\sigma = 1$. The densities of Φ_σ and Φ are denoted by ϕ_σ and ϕ, respectively. In the case $X_j's$ are discrete, F_n has jumps and the normal approximation is not very good when n is not sufficiently large. This is a problem which most commonly occurs in statistical tests and estimation involving the normal approximation to the binomial and, in its multi-dimensional version, in Pearson's frequency ►chi-square tests, or in tests for association in categorical data. Applying the *CLT* to a binomial random variable T with distribution $B(n,p)$, with mean np and variance $npq(q = 1 - p)$, the normal approximation is given, for integers $0 \le a \le b \le n$, by

$$P(a \le T \le b) \approx \Phi\left((b - np)/\sqrt{npq}\right) - \Phi\left((a - np)/\sqrt{npq}\right). \quad (1)$$

Here $\approx$ indicates that the difference between its two sides goes to zero as $n \to \infty$. In particular, when $a = b$, the binomial probability $P(T = b) = C_b^n p^b q^{n-b}$ is approximated by zero. This error is substantial if n is not very large. One way to improve the approximation is to think graphically of each integer value b of T being uniformly spread over the interval $\left[b - \frac{1}{2}, b + \frac{1}{2}\right]$. This is the so called *histogram approximation*, and leads to the *continuity correction* given by replacing $\{a \le T \le b\}$ by $\left\{a - \frac{1}{2} \le T \le b + \frac{1}{2}\right\}$

$$P\left(a - \frac{1}{2} \le T \le b + \frac{1}{2}\right) \approx \Phi\left(\left(b + \frac{1}{2} - np\right)/\sqrt{npq}\right) - \Phi\left(\left(a - \frac{1}{2} - np\right)/\sqrt{npq}\right). \quad (2)$$

To give an idea of the improvement due to this correction, let $n = 20, p = 0.4$. Then $P(T \le 7) = 0.4159$, whereas the approximation (1) gives a probability $\Phi(-0.4564) = 0.3240$, and the continuity correction (2) yields $\Phi(-0.2282) = 0.4177$. Analogous continuity corrections apply to the Poisson distribution with a large mean.

For a precise mathematical justification of the continuity correction consider, in general, i.i.d. integer-valued random variables $X_1, \ldots, X_n$, with lattice span 1, mean μ, variance σ^2, and finite moments of order at least four. The distribution function $F_n(x)$ of $n^{-1/2}(X_1 + \ldots + X_n)$ may then be approximated by the ►*Edgeworth expansion* (See Bhattacharya and Ranga 1976, p. 239, or Gnedenko and Kolmogorov 1954, p. 213)

$$F_n(x) = \Phi_\sigma(x) - n^{-\frac{1}{2}} S_1\left(n\mu + n^{\frac{1}{2}}x\right)\phi_\sigma(x) + n^{-\frac{1}{2}}\mu_3/(6\sigma^3)(1 - x^2/\sigma^2)\phi_\sigma(x) + O(n^{-1}), \quad (3)$$

where $S_1(y)$ is the right continuous periodic function $y - \frac{1}{2}$ (mod 1) which vanishes when $y = \frac{1}{2}$. Thus, when a is an integer and $x = (a - n\mu)/\sqrt{n}$, replacing a by $a + \frac{1}{2}$ (or $a - \frac{1}{2}$) on the right side of (3) gets rid of the discontinuous term involving S_1.

Consider next the continuity correction for the *(Mann-Whitney-)Wilcoxon two sample test* (see ►Wilcoxon–Mann–Whitney Test). Here one wants to test nonparametrically if one distribution G is stochastically larger than another distribution F, with distribution functions $G(.), F(.)$. Then the null hypothesis is $H_0 : F(x) = G(x)$ for all x, and the alternative is $H_1 : G(x) \le F(x)$ for all x, with strict inequality for some x. The test is based on independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from the two unknown continuous distributions F and G, respectively. The test statistic is W_s = the sum of the ranks of the $Y_j's$ in the combined sample of $m + n$ $X_i's$ and $Y_j's$. The test rejects H_0 if $W_s \ge c$, where c is chosen such that the probability of rejection under H_0 is a given level α. It is known (see Lehmann 1975, pp. 5–18) that W_s is asymptotically normal and $E(W_s) = \frac{1}{2}n(m + n + 1)$, $Var(W_s) = mn(m + n + 1)/12$. Since W_s is integer-valued, the continuity correction yields

$$P(W_s \ge c|H_0) = P\left(W_s \ge c - \frac{1}{2}|H_0\right) \approx 1 - \Phi(z), \quad (4)$$

where $z = \left(c - \frac{1}{2} - \frac{1}{2}n(m + n + 1)\right)/\sqrt{mn(m + n + 1)/12}$.

As an example, let $m = 5$, $n = 7$, $c = 54$. Then $P(W_s \ge 54 \mid H_0) = 0.101$, and its normal approximation is $1 - \Phi(1.380) = 0.0838$. The continuity correction yields the better approximation $P(W_s \ge 54 \mid H_0) = P(W_s \ge 53.5 \mid H_0) \approx 1 - \Phi(1.299) = 0.0097$.

The continuity correction is also often used in 2×2 *contingency tables* for testing for association between two categories. It is simplest to think of this as a two-sample problem for comparing two proportions p_1, p_2 of individuals with a certain characteristic (e.g., smokers) in two populations (e.g., men and women), based on two independent random samples of sizes n_1, n_2 from the two populations, with $n = n_1 + n_2$. Let r_1, r_2 be the numbers in the samples possessing the characteristic. Suppose first that we

wish to test $H_0 : p_1 = p_2$, against $H_1 : p_1 < p_2$. Consider the test which rejects H_0, in favor of H_1, if $r_2 \geq c(r)$, where $r = r_1 + r_2$, and $c(r)$ is chosen so that the conditional probability (under H_0) of $r_2 \geq c(r)$, given $r_1 + r_2 = r$, is α. This is the uniformly most powerful unbiased (*UMPU*) test of its size (See Lehmann 1959, pp. 140–146, or Kendall and Stuart 1973, pp. 570–576). The conditional distribution of r_2, given $r_1 + r_2 = r$, is multinomial, and the test using it is called *Fisher's exact test*. On the other hand, if $n_i p_i \geq 5$ and $n_i(1 - p_i) \geq 5$ $(i = 1, 2)$, the normal approximation is generally used to reject H_0. Note that the (conditional) expectation and variance of r_2 are $n_2 r/n$ and $n_1 n_2 r(n-r)/[n^2(n-1)]$, respectively (See Lehmann 1975, p. 216). The normalized statistic t is then

$$t = [r_2 - n_2 r/n]/\sqrt{n_1 n_2 r(n-r)/[n^2(n-1)]}, \quad (5)$$

and H_0 is rejected when t exceeds $z_{1-\alpha}$, the $(1 - \alpha)th$ quantile of Φ. For the continuity correction, one subtracts $\frac{1}{2}$ from the numerator in (5), and rejects H_0 if this adjusted t exceeds $z_{1-\alpha}$. Against the two-sided alternative $H_1 : p_1 \neq p_2$, Fisher's *UMPU* test rejects H_0 if r_2 is either too large or too small. The corresponding continuity corrected t rejects H_0 if either the adjusted t, obtained by subtracting $\frac{1}{2}$ from the numerator in (5), exceeds $z_{1-\alpha/2}$, or if the t adjusted by adding $\frac{1}{2}$ to the numerator in (5) is smaller than $-z_{1-\alpha/2}$. This may be compactly expressed as

$$\text{Reject } H_0 \text{ if } V \equiv (n-1)\left[\,|\, r_1 n_2 - r_2 n_1 \,|\, - \frac{1}{2}n\right]^2 \Big/ (n_1 n_2 r(n-r)) > \chi^2_{1-\alpha}(1), \quad (6)$$

where $\chi^2_{1-\alpha}(1)$ is the $(1 - \alpha)$th quantile of the ▶chi-square distribution with one degree of freedom. This two-sided continuity correction was originally proposed by F. Yates in 1934, and it is known as *Yates' correction*. For numerical improvements due to the continuity corrections above, we refer to Kendall and Stuart (1973, pp. 575–576) and Lehmann (1975, pp. 215–217). For a critique, see Connover (1974). If the sampling of n units is done at random from a population with two categories (men and women), then the *UMPU* test is still the same as Fisher's test above, conditioned on fixed marginals n_1,(and, therefore, n_2) and r.

Finally, extensive numerical computations in Bhattacharya and Chan (1996) show that the chisquare approximation to the distribution of *Pearson's frequency chi-square* statistic is reasonably good for degrees of freedom 2 and 3, even in cases of small sample sizes, extreme asymmetry, and values of expected cell frequencies much smaller than 5. One theoretical justification for this may be found in the classic work of Esseen (1945), which shows that the error of chisquare approximation is $O(n^{-d/(d+1)})$ for degrees of freedom d.

Acknowledgments

The author acknowledges support from the NSF grant DMS 0806011.

About the Author

For biography *see* the entry ▶Random Walk.

Cross References

▶Binomial Distribution
▶Chi-Square Test: Analysis of Contingency Tables
▶Wilcoxon–Mann–Whitney Test

References and Further Reading

Bhattacharya RN, Chan NH (1996) Comparisons of chisquare, Edgeworth expansions and bootstrap approximations to the distribution of the frequency chisquare. Sankhya Ser A 58:57–68
Bhattacharya RN, Ranga Rao R (1976) Normal approximation and asymptotic expansions. Wiley, New York
Connover WJ (1974) Some reasons for not using Yates' continuity correction on 2 × 2 contingency tables. J Am Stat Assoc 69:374–376
Esseen CG (1945) Fourier analysis of distribution functions: a mathematical study of the Laplace–Gaussian law. Acta Math 77:1–125
Gnedenko BV, Kolmogorov AN (1954) Limit distributions of sums of independent random variables. English translation by K.L. Chung, Reading
Kendall MG, Stuart A (1973) The advanced theory of statistics, vol 2, 3rd edn. Griffin, London
Lehmann EL (1959) Testing statistical hypotheses. Wiley, New York
Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. (With the special assistance of D'Abrera, H.J.M.). Holden-Day, Oakland

Control Charts

Alberto Luceño
Professor
University of Cantabria, Santander, Spain

Introduction

A control chart is a graphical statistical device used to monitor the performance of a repetitive process. Control charts were introduced by Shewhart in the 1920s while working for Western Electric and Bell Labs and, since then, they have been routinely used in Statistical Process Control (SPC). According to Shewhart, control charts are useful to define the standard to be attained for a process, to help

attaining that standard, and to judge whether that standard has been reached.

Variability and Its Causes

Any manufacturing or business process shows some degree of variability. This is obviously true when little effort has been made to try to keep the process stable around a target, but it continues to be true even when a lot of effort has already been dedicated to stabilize the process. In other words, the amount of variability can be reduced (as measured, for example, by the output standard deviation), but cannot be eliminated completely. Therefore, some knowledge about the types of variability that can be encountered in practice and the causes of this variability is necessary.

Concerning the types of variability, one must recognize at least the difference between stationary and non-stationary behavior, the former being desirable, the latter undesirable. A stationary process has fixed mean, variance and probability distribution, so that it is difficult (if not impossible) to perfectly attain this desirable state in practice. A non-stationary process does not have fixed mean, variance or probability distribution, so that its future behavior is unpredictable. Moreover, any natural process, when left to itself, tends to be non-stationary, sometimes in the long run, but most often in the short run. Consequently, some control effort is almost always necessary to, at least, induce stationarity in the process. Control charts are useful for this purpose.

Concerning the causes of variability, the most obvious facts are that there are a lot of causes, that many of them are unknown, and, consequently, that they are difficult to classify. Nevertheless, Shewhart suggested that it is conceptually useful to classify the causes of variability in two groups: common causes and special causes. Common causes are those that are still present when the process has been brought to a satisfactory stationary state of control; they can be described as chance variation, because the observed variation is the sum of many small effects having different causes. Special causes are those that have larger effects and, hence, have the potential to send the process out of control; hopefully, they can eventually be discovered (assigned) and permanently removed from the system.

Control charts are useful tools to detect the presence of special causes of variation worthy of removal. They do so by modelling the likely performance of a process under the influence of the common causes of variation, so that the unexpected behavior (and possible non-stationarity) of the process caused by the emergence of a special cause at any time can be detected efficiently.

Shewhart Charts

When Shewhart presented his control charts, he did not claim any mathematical or statistical optimality for such charts, but he did demonstrate that the cost of controlling a process could often be reduced by using control charts. Consequently, Shewhart control charts are much more justifiable for their practical benefits than for their theoretical properties.

A Basic Chart

Bearing this in mind, a Shewhart control chart for a measurable quality characteristic is constructed in the following way. (1) Select the frequency of sampling and the sample size; e.g., take $n = 4$ observations every 2 h. (2) Calculate the sample average $\bar{X}_t$ for every time interval t (e.g., every 2 h) and plot $\bar{X}_t$ versus t for all the values of t at hand. By doing so, one obtains a run chart. (3) Add a center line (CL) to the run chart. The ordinate of this horizontal line can be a target value for the quality characteristic, a historical mean of past observations, or simply the mean of the observations at hand. (4) Add an upper control limit (UCL) and a lower control limit (LCL). These horizontal lines are usually situated symmetrically around the CL and at a distance of three times the standard deviation of the statistics plotted in the run chart (e.g., three times the standard deviation of $\bar{X}_t$).

This chart is used at every time interval t to take the decision of whether the process should be considered to be in a state of economic control or not. The usual decision rule is: (1) Decide that the process stays in control at time t if the plotted statistics ($\bar{X}_t$) lies between the UCL and the LCL, and continue plotting. (2) Declare an out of control situation otherwise; in this case, a search for an assignable cause should be started, which hopefully will eventually lead to the identification of this cause and its permanent removal from the system. This type of control procedure is sometimes called process monitoring, or process surveillance, and is a part of SPC. Figure 1 shows a Shewhart chart for a random sample of values of $\bar{X}_t$ having mean $\mu = 50$ and standard deviation $\sigma = 2$. The chart does not show any alarm.

Some Modifications of the Basic Chart

Under certain theoretical assumptions, the basic chart can claim some type of optimality. However, it may not be completely satisfactory in practice. Consequently, the form of the basic chart and how it is used can be modified in many different ways. For example, the control limits could not be symmetrically placed around the CL or could not necessarily lie at three standard deviations from the CL.

Warning limits situated at two standard deviations from the CL could also be plotted. Lines at one standard deviation from the CL could be added. The decision rule could correspondingly be modified using, for example, the so called Western Electric rules, etc.

The usefulness of these modifications of the basic chart should be judged, in each particular application, on the bases of the economical or practical advantages they provide. In doing so, the costs of declaring that the process is in control when in fact is not, and vice versa, usually play a role. The elapsed time since the process starts to be out of control until this state is detected can also play a role (true alarm), as well as the time between consecutive declarations of out of control situations when the process stays in control (false alarm rate). These elapsed random times are usually called run lengths (RLs) and their means are called average run lengths (ARLs). Clearly, the frequency distribution (or probability distribution) of the RL will depend on whether the process is in control (RL for false alarms) or out of control (RL for true alarms), and the ARL for false alarms should be much larger than the ARL for true alarms.

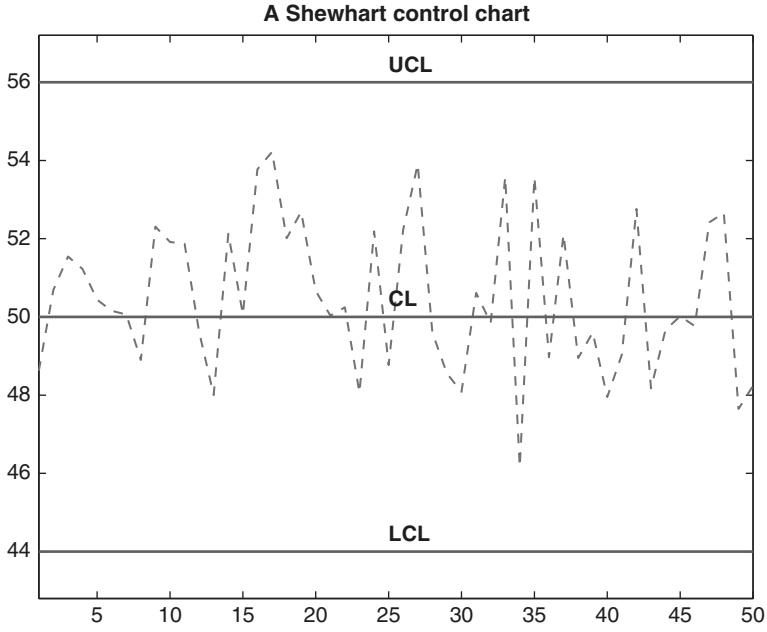
Some More Basic Charts

Control of the mean of a measurable quality characteristic is important, but a process can also be out of control because of excessive variation around its mean. Therefore, in addition to the basic $\bar{X}$ chart, previously described, it is customary to simultaneously run a chart to control the range (R-chart) or the standard deviation (S-chart) of the observations taken every time interval t.

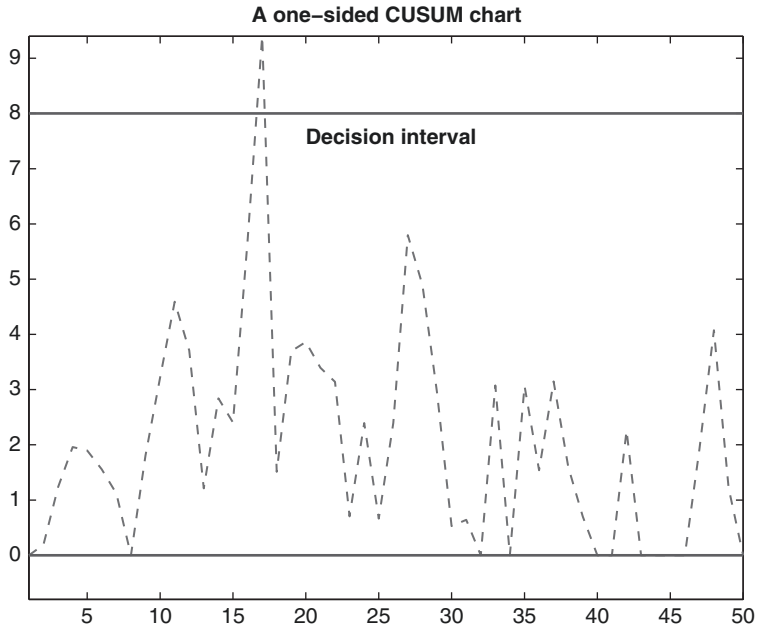
Similarly, when the quality characteristic is not measurable, one can use a p-chart or an np-chart to control the fraction nonconforming for each time interval t, or a c-chart or a u-chart to control the total numbers (counts) of nonconforming items for each period t.

Some Other Types of Control Charts

Basic Shewhart charts are useful to detect relatively large and sporadic deviations from the state of control. However, the control of a process may be jeopardized also by small but persistent deviations from the state of control. The Western Electric rules may be considered as one of many attempts to tackle this problem. However, a more formal approach was suggested by Page (1954, 1957) by introducing the cumulative sum (CUSUM) charts. Moreover, the introduction of the exponentially weighted moving average (EWMA) charts provided an alternative procedure. More recently, cumulative score (CUSCORE) charts, specialized in detecting particular types of deviation from the state of control, have also been suggested (e.g., by Box and Ramírez 1992; Box and Luceño 1997; Box et al. 2009).



Control Charts. Fig. 1 An example of a Shewhart chart



Control Charts. Fig. 2 An example of a one-sided CUSUM chart for the same data as in Fig. 1

CUSUM Charts

To be able to efficiently detect small persistent deviations from target occurring before and at period t, some use of recent observations is necessary. CUSUM charts do so by using the following statistics:

$$S_t^+ = \max\left[S_{t-1}^+ + (\bar{X}_t - k^+); 0\right];$$
$$S_t^- = \max\left[S_{t-1}^- + (-\bar{X}_t - k^-); 0\right]; \quad (1)$$

where k^+ and k^- are called reference values. The process is considered to be in control until the period t at which one of the inequalities $S_t^+ > h^+$ or $S_t^- > h^-$ becomes true, where h^+ and h^- are called decision intervals. At this time, an alarm is declared, and the search for a special cause (or assignable cause, in Deming's words) should begin.

The reference values and decision intervals of the chart are often chosen in the light of the theoretical ARLs that they produce when the process is on target and when the process is out of target by an amount of D times the standard deviation of X_t (or, equivalently, $D\sqrt{n}$ times the standard deviation of $\bar{X}_t$).

If only one of the statistics in (1) is used, the CUSUM chart is called one-sided; if both are used, the CUSUM is called two-sided. The theoretical evaluation of the run length distributions for two-sided CUSUM charts is considerably more difficult than for their one-sided counterparts. Figure 2 shows a one-sided CUSUM chart based on S_t^+, with reference value $\mu + 0.25\sigma$ and decision interval at 4σ, for the sample used in Fig. 1. This chart produces an alarm at $t = 17$.

EWMA Charts

EWMA charts use recent data in a different way than CUSUM charts. The EWMA statistic is

$$\tilde{X}_t = (1-\lambda)\tilde{X}_{t-1} + \lambda\bar{X}_t, \quad (2)$$

where $0 < \lambda < 1$, but most often $0.1 \le \lambda \le 0.4$. The EWMA statistic at time t is an average of all observations taken at time t and before, in which each observation receives a weight that decreases exponentially with its age. In other words, Eq. (2) can be written as

$$\tilde{X}_t = \lambda[\bar{X}_t + (1-\lambda)\bar{X}_{t-1} + (1-\lambda)^2\bar{X}_{t-2} + \cdots]. \quad (3)$$

The smaller the value of λ, the smoother the chart.

The process is usually considered to be in control until the period t at which $|\tilde{X}_t|$ reaches three times the standard deviation of the EWMA statistic $\tilde{X}_t$. It can be shown that the variance of $\tilde{X}_t$ is the product of the variance of $\bar{X}_t$ by a factor $\lambda[1-(1-\lambda)^{2t}]/(2-\lambda)$, where $t = 0$ is the origin of the chart. When an alarm is triggered, the search for a special cause should start.

Information about the above mentioned charts and many possible variants can be found in the bibliography that follows.

About the Author

Professor Luceño was awarded 1998 Brumbaugh Award of the American Society for Quality jointly with Professor George E.P. Box. He is a co-author (with G.E.P. Box) of the well known text *Statistical Control By Monitoring and Feedback Adjustment* (John Wiley & Sons, 1997), and (with G.E.P. Box and M.A. Paniagua-Quiñones) ñ (John Wiley & Sons, 2009). He is currently Associate Editor of *Quality Technology and Quantitative Management*, and *Quality Engineering*.

Cross References

►Acceptance Sampling
►Industrial Statistics
►Multivariate Statistical Process Control
►Six Sigma
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Box GEP, Luceño A (1997) Statistical control by monitoring and feedback adjustment. Wiley, New York
Box GEP, Ramírez JG (1992) Cumulative score charts. Qual reliab Eng Int 8:17–27
Box GEP, Luceño A, Paniagua-Quiñones MA (2009) Statistical control by monitoring and adjustment, 2nd edn. Wiley, New York
Deming WE (1986) Out of the crisis. Massachusetts Institute of Technology, Center for Advanced Engineering Studies, Cambridge
Khattree R, Rao CR (eds) (2003) Handbook of statistics 22: statistics in industry. Elsevier, Amsterdam
Luceño A, Cofiño AS (2006) The random intrinsic fast initial response of two-sided CUSUM charts. Test 15:505–524
Luceño A, Puig-Pey J (2000) Evaluation of the run-length probability distribution for CUSUM charts: assessing chart performance. Technometrics 42:411–416
Montgomery DC (2005) Introduction to statistical quality control, 5th edn. Wiley, New York
NIST/SEMATECH (2009) e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/
Page ES (1954) Continuous inspection schemes. Biometrika 41: 100–114
Page ES (1957) On problems in which a change in a parameter occurs at an unknown point. Biometrika 44:248–252
Ryan TP (1989) Statistical methods for quality improvement. Wiley, New York
Ruggery F, Kenetts RS, Faltin FW (eds) (2007) Encyclopedia of statistics in quality and reliability. Wiley, New York
Shewhart WA (1931) Economic control of quality of manufacturing product. Van Nostrand Reinhold, Princeton, NJ. Republished by Quality Press, Milwaukee, 1980
Western Electronic Company (1956) Statistical quality control handbook. Western Electric Corporation, Indianapolis

Convergence of Random Variables

Pedro J. Rodríguez Esquerdo
Professor, Head
University of Puerto Rico, San Juan, Puerto Rico

C

Introduction

The convergence of a sequence of random variables (RVs) is of central importance in probability theory and in statistics. In probability, it is often desired to understand the long term behavior of, for example, the relative frequency of an event, does it converge to a number? In what sense does it converge? In statistics, a given estimator often has the property that for large samples the values it takes are distributed around and are close to the value of the desired parameter. In many situations the distribution of this estimator can be approximated by a well known distribution, which can simplify the analysis. Thus it is necessary to understand the types of convergence of such sequences, and conditions under which they occur.

Four modes of convergence are presented here.

1. *Weak convergence*, also called *convergence in distribution* or *convergence in law*, refers to the conditions under which a sequence of distribution functions converges to a cumulative distribution function (cdf).
2. A second mode is *convergence in probability*, which studies the limiting behavior of the sequence of probabilities that for each n, a RV deviates by more than a given quantity from a limiting RV.
3. *Convergence with probability one*, or almost sure convergence, studies the conditions under which the probability of a set of points in the sample space for which a sequence of RVs converges to another RV is equal to one.
4. *Convergence in the rth mean* refers to the convergence of a sequence of expected values. As it is to be expected, there are some relations between the different modes of converge.

The results here are explained, for their formal proof, the reader is referred to the included references. In general, the RVs $\{X_n\}$ cannot be assumed to be independent or identically distributed. For each value of the subscript n, the distribution of X_n may change (Casella and Berger 2002). In many cases, however, the sequence of *dfs* converge to another df.

A large amount of literature exists on the convergence of random variables. An excellent reference for understanding the definitions and relations is Rohatgi (1976). For

a discussion of some of these modes of convergence and as they apply to statistics, see Casella and Berger (2002). Chow and Teicher (1997), Loeve (1976) and Dudley (1989) present a more formal and general approach to the concept of convergence of random variables. In this paper, the notation $\{X_n\}$ is used to represent the sequence $X_1, X_2, X_3, \ldots$

Convergence in Distribution

Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P), and let $\{F_n\}$ be the corresponding sequence of cdfs. Let X be a RV with cdf F. The sequence $\{X_n\}$ is said to *converge in distribution* to X if $\lim_{n\to\infty} F_n(x) = F(x)$ at every point where $F(x)$ is continuous. This type of convergence is sometimes also called *convergence in law* and denoted $X_n \xrightarrow{\mathcal{L}} X$. A sequence of distribution functions does not have to converge, and when it does:

1. *The limiting function does not have to be a cdf itself.* Consider the sequence given by $F_n(x) = 0$ if $x < n$ and $F_n(x) = 1$ if $x \geq n$; $n = 1, 2, \ldots$ Then, at each real value x, $F_n(x) \to 0$, which is not a cdf.
2. *Convergence in distribution does not imply that the sequence of moments converges.* For $n = 1, 2, \ldots$, consider a sequence of cdfs $\{F_n\}$ defined by $F_n(x) = 0$, if $x < 0$; $F_n(x) = 1 - 1/n$, for $0 \leq x < n$; and $F_n(x) = 1$ for $x \geq n$. The sequence of cdfs converges to the cdf $F(x) = 1$ for $x \geq 0$, and $F(x) = 0$ otherwise. For each n, the cdf F_n corresponds to a discrete RV X_n that has probability function (pf) given by $P\{X_n = 0\} = 1 - 1/n$ and $P\{X_n = n\} = 1/n$. The limiting cdf F, corresponds to a RV X with pf $P(X = 0) = 1$. For $k \geq 1$, the kth moment of X_n is $E\left(X_n^k\right) = 0(1 - 1/n) + n^k(1/n) = n^{k-1}$. Finally, $E(X^k) = 0$, so that $E\left(X_n^k\right)$ does not converge to $E(X^k)$.
3. *Convergence in distribution does not imply convergence of their pfs or probability density functions* (pdfs). Let a and b be fixed real numbers, and $\{X_n\}$ a sequence of RVs with pfs given by $P\{X_n = x\} = 1$ for $x = b + a/n$ and $P\{X_n = x\} = 0$ otherwise. None of the pfs assigns any probability to the point $x = b$. Then $P\{X_n = x\} \to 0$, which is not a pf, but the sequence of cdfs $\{F_n\}$ of the RVs X_n converges to a cdf, $F(x) = 1$ for $x \geq b$ and $F(x) = 0$ otherwise.
4. *For integer valued RVs, its sequence of pfs converges to another pf if and only if the corresponding sequence of RVs converges in distribution.*
5. *If a sequence of RVs $\{X_n\}$ converges in distribution to X and c is a real constant, then $\{X_n + c\}$, and $\{cX_n\}$ converge in distribution to $\{X + c\}$, and $\{cX\}$, respectively.*

Convergence in Probability

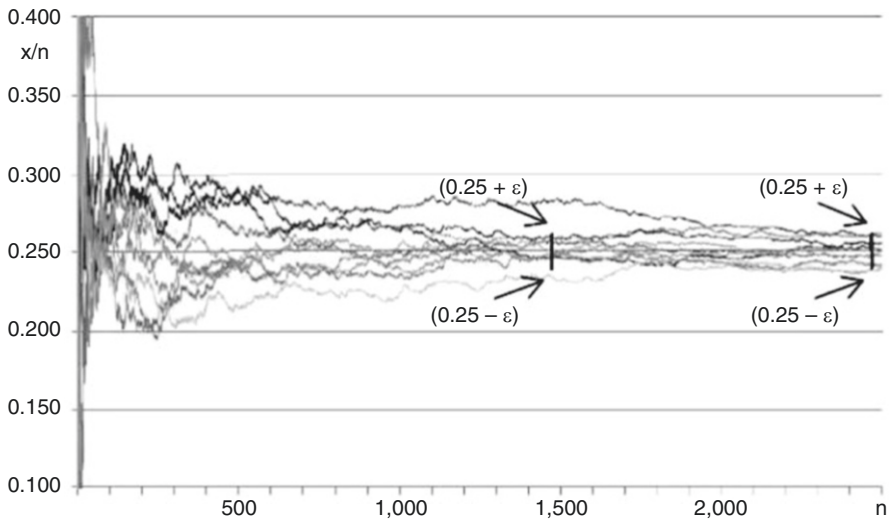
Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P). The sequence $\{X_n\}$ is said to *converge in probability* to a RV X, denoted by $X_n \xrightarrow{p} X$, if for every real number $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0.$$

Convergence in probability of $\{X_n\}$ to the RV X refers to the convergence of a sequence of probabilities, real numbers to 0. It means that the probability that the distance between X_n and X is larger than $\varepsilon > 0$ tends to 0 as the n increases to infinity. It does not mean that given $\varepsilon > 0$, we can find N such that $|X_n - X| < \varepsilon$ for all $n \geq N$.

Convergence in probability, behaves in many respects as one would expect with respect to common arithmetic operations and under continuous transformations. The following results hold (Rohatgi 1976):

1. $X_n \xrightarrow{p} X$ if and only if $X_n - X \xrightarrow{p} 0$.
2. If $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, then $P\{X = Y\} = 1$.
3. If $X_n \xrightarrow{p} X$, then $X_n - X_m \xrightarrow{p} 0$, as $n, m \to \infty$.
4. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$, and $X_n - Y_n \xrightarrow{p} X - Y$.
5. If $X_n \xrightarrow{p} X$ and k is a real constant then $kX_n \xrightarrow{p} kX$.
6. If $X_n \xrightarrow{p} k$ then $X_n^2 \xrightarrow{p} k^2$.
7. If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$; a, b real constants, then $X_n Y_n \xrightarrow{p} ab$.
8. If $X_n \xrightarrow{p} 1$ then $1/X_n \xrightarrow{p} 1$.
9. If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$; a, b real constants, $b \neq 0$, then $X_n/Y_n \xrightarrow{p} a/b$.
10. If $X_n \xrightarrow{p} X$ and Y is a RV then $X_n Y \xrightarrow{p} XY$.
11. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n Y_n \xrightarrow{p} XY$.
12. Convergence in probability is stronger than convergence in distribution; that is, if $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{\mathcal{L}} X$.
13. Let k be a real number, then convergence in distribution to k implies convergence in probability to k, that is, if $X_n \xrightarrow{\mathcal{L}} k$ then $X_n \xrightarrow{p} k$.
14. In general, convergence in distribution does not imply convergence in probability. For an example, consider the identically distributed RVs $X, X_1, X_2, \ldots$ with sample space $\{0, 2\}$, such that for every n, $P(X_n = 0, X = 0) = P(X_n = 2, X = 2) = 0$ and $P(X_n = 2, X = 0) = P(X_n = 0, X = 2) = {}^1/_2$. Because X, X_n, are identically distributed RVs, $X_n \xrightarrow{\mathcal{L}} X$, but $P\{|X_n - X| > {}^1/_2\} \geq P\{|X_n - X| = 2\} = 1 \neq 0$. (Rohatgi 1976).



Ten series of 2,500 trials each, of a Binomial (4, 0.25) RV X were simulated. The ratio of the running total of successes x, to the number of trials n is plotted for each series. For X/n to converge in probability to 0.25 implies for this experiment, that as n increases, for fixed ε, the probability of observing a series outside the interval $(0.25 - \varepsilon, 0.25 + \varepsilon)$, will decrease to zero. It does not mean there is a value N such that all the series that we can possibly simulate n will be found inside the interval for all $n > N$.

Convergence of Random Variables. Fig. 1 Illustration of convergence in probability

15. Convergence in probability does not imply that the kth moments converge, that is, $X_n \xrightarrow{p} X$ does not imply that $E\left(X_n^k\right) \to E(X^k)$ for any integer $k > 0$. This is illustrated by the example in (14) above.

Figure 1 illustrates the concept of convergence in probability for series of sample means of RVs from a Binomial(4, .025) distribution. The following results further relate convergence in distribution and convergence in probability. Let $\{X_n, Y_n\}, n = 1, 2, \ldots$ be a sequence of pairs of random variables, and let c be a real number.

16. If $|X_n - Y_n| \xrightarrow{p} 0$ and $Y_n \xrightarrow{\mathcal{L}} Y$, then $X_n \xrightarrow{\mathcal{L}} Y$.
17. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$. This is also true for the difference $X_n - Y_n$.
18. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$ then $X_n Y_n \xrightarrow{\mathcal{L}} cX$ (for $c \neq 0$) and $X_n Y_n \xrightarrow{p} 0$ (for $c = 0$).
19. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$ then $X_n/Y_n \xrightarrow{\mathcal{L}} X/c$ (for $c \neq 0$).

Almost Sure Convergence

Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P). The sequence $\{X_n\}$ is said to *converge to X with probability one* or almost surely, denoted $X_n \xrightarrow{as} X$ if

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

Almost sure convergence of a sequence of RVs $\{X_n\}$ to an RV X, means that the probability of the event $\left\{\omega;\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}$ is one (see also ▶Almost Sure Convergence of Random Variables). That is, the set of all points ω in the sample space Ω, where $X_n(\omega)$ converges to $X(\omega)$, has probability one. It is not required that the sequence of functions $\{X_x(\omega)\}$ converge to the function $X(\omega)$ pointwise, for all ω in the sample space, only that the set of such ω has probability one.

1. Convergence almost surely implies convergence in probability. If the sequence of random variables $\{X_n\}$ converges almost surely to X then it converges in probability to X.

2. Skorokhod's representation theorem shows that if a sequence of RVs $\{X_n\}$ converges in distribution to an RV X, then there exists a sequence of random variables $\{Y_n\}$, identically distributed as $\{X_n\}$ such that $\{Y_n\}$ converges almost surely to a RV Y, which itself is identically distributed as X (Dudley 1989).
3. Continuity preserves convergence in distribution, in probability, and almost sure convergence. If X_n converges in any of these modes to X, and f is a continuous function defined on the real numbers, then $f(X_n)$ converges in the same mode to $f(X)$.
4. If $\{X_n\}$ is a strictly decreasing sequence of positive random variables, such that X_n converges in probability to 0, then X_n converges almost surely to 0.
5. Convergence in probability does not imply convergence almost surely. Consider (Casella and Berger 2002) the sample space given by the interval $[0, 1]$, and the uniform probability distribution. Consider the RV $X(\omega) = \omega$ and let $\{X_n\}$ be defined by

$$X_1(\omega) = \omega + I_{[0,1]}(\omega), \quad X_2(\omega) = \omega + I_{[0,1/2]}(\omega),$$
$$X_3(\omega) = \omega + I_{[1/2,1]}(\omega), \quad X_4(\omega) = \omega + I_{[0,1/3]}(\omega),$$
$$X_5(\omega) = \omega + I_{[1/3,2/3]}(\omega), \quad X_6(\omega) = \omega + I_{[2/3,1]}(\omega),$$

and so on. Here $I_A(\omega)$ is the indicator function of the set A. Then $\{X_n\}$ converges in probability to X, but does not converge almost surely since the value $X_n(\omega)$ alternates between ω and $\omega + 1$ infinitely often.

Convergence in the *r*th Mean

Definition Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P) The sequence $\{X_n\}$ is said to *converge to X in the rth mean*, $r \geq 1$, if $E(|Xn|^r) < \infty$, $E(|X|^r) < \infty$ and $\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0$.

1. When $r = 1$ we say that $\{X_n\}$ converges in the mean, while for $r = 2$, we say that $\{X_n\}$ converges in the mean square.
2. If a sequence $\{X_n\}$ converges in the rth mean, and $s < r$, then $\{X_n\}$ converges in the sth mean. For example, convergence in the mean square implies convergence in the mean. This means that if the variances of a sequence converge, so do the means.
3. Convergence in the rth mean implies convergence in probability, if $\{X_n\}$ converges in the rth mean to X, then $\{X_n\}$ converges in probability to X. However, the converse is not true. For an example, consider the sequence $\{X_n\}$ with probability function defined by $P(X_n = 0) = 1 - \frac{1}{n^3}$ and $P(X_n = n) = \frac{1}{n^3}$ for $r > 0$. (Rohatgi 1976).

About the Author

Dr. Pedro J. Rodríguez Esquerdo is Professor at the Department of Mathematics, College of Natural Sciences and is currently Professor and Head, Institute of Statistics and Computer Information Science, College of Business Administration, both at the University of Puerto Rico, Rio Piedras. He is past Associate Dean for Academic Affairs at the University of Puerto Rico (1988–1993). Professor Rodríguez Esquerdo has served on several advisory boards, including the Advisory Board of Gauss Research Laboratory (2000–) in San Juan, Puerto Rico. He has been a consultant on education, technology, statistics, mathematics, and intellectual property law. He designed and maintains his statistics education web site www.eduesta.org since 1997. Prof. Rodríguez Esquerdo coauthored a book with professors Ana Helvia Quintero and Gloria Vega, *Estadística Descriptiva* (Publicaciones Puertorriqueñas, 1997), and is currently participating in a distance learning project in applied mathematics.

Cross References

►Almost Sure Convergence of Random Variables
►Binomial Distribution
►Central Limit Theorems
►Ergodic Theorem
►Glivenko-Cantelli Theorems
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Probability Theory: An Outline
►Random Variable
►Strong Approximations in Probability and Statistics
►Uniform Distribution in Statistics
►Weak Convergence of Probability Measures

References and Further Reading

Casella G, Berger RL (2002) Statistical inference. Duxbury, Pacific Grove

Chow Y, Teicher H (1997) Probability theory: independence, interchangeability, martingales, 3rd edn. Springer, New York

Dudley RM (1989) Real analysis and probability. Chapman & Hall, New York

Loeve M (1977) Probability theory, vol I, 4th edn. Springer, New York

Rohatgi VK (1976) An introduction to probability theory and mathematical statistics. Wiley, New York

Cook's Distance

R. Dennis Cook
Professor
University of Minnesota, Minneapolis, MN, USA

Introduction

Prior to 1975 there was little awareness within statistics or the applied sciences generally that a single observation can influence a statistical analysis to a point where inferences drawn with the observation included can be diametrically opposed to those drawn without the observation. The recognition that such *influential observations* do occur with notable frequency began with the 1977 publication of *Cook's Distance*, which is a means to assess the influence of individual observations on the estimated coefficients in a linear regression analysis (Cook 1977). Today the detection of influential observations is widely acknowledged as an important part of any statistical analysis and Cook's distance is a mainstay in linear regression analysis. Generalizations of Cook's distance and of the underlying ideas have been developed for application in diverse statistical contexts. Extensions of Cook's distance for linear regression along with a discussion of surrounding methodology were presented by Cook and Weisberg (1982).

Cook's distance and its direct extensions are based on the idea of contrasting the results of an analysis with and without an observation. Implementation of this idea beyond linear and ▶generalized linear models can be problematic. For these applications the related concept of *local influence* (Cook 1986) is used to study the touchiness of an analysis to local perturbations in the model or the data. Local influence analysis continues to be an area of active investigation (see, for example, Zhu et al. 2007).

Cook's Distance

Consider the linear regression of a response variable Y on p predictors $X_1, \ldots, X_p$ represented by the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \boldsymbol{\varepsilon}_i,$$

where $i = 1, \ldots, n$ indexes observations, the β's are the regression coefficients and ε is an error that is independent of the predictors and has mean 0 and constant variance σ^2. This classic model can be represented conveniently in matrix terms as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\mathbf{Y} = (Y_i)$ is the $n \times 1$ vector of responses, $\mathbf{X} = (X_{ij})$ is the $n \times (p+1)$ matrix of predictor values X_{ij}, including a constant column to account for the intercept β_0, and $\boldsymbol{\varepsilon} = (\varepsilon_i)$ is the $n \times 1$ vector of errors. For clarity, the ith response Y_i in combination with its associated values of the predictors $X_{i1}, \ldots, X_{ip}$ is called the ith *case*. Let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimator of the coefficient vector $\boldsymbol{\beta}$ based on the full data and let $\widehat{\boldsymbol{\beta}}_{(i)}$ denote the OLS estimator based on the data after removing the ith case. Let s^2 denote estimator of σ^2 based on the OLS fit of the full dataset – s^2 the residual sum of squares divided by $(n-p-1)$.

Cook (1977) proposed to assess the influence of the ith case on $\widehat{\boldsymbol{\beta}}$ by using a statistic D_i, which subsequently became known as Cook's distance, that can be expressed in three equivalent ways:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T \mathbf{X}^T \mathbf{X} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(p+1)s^2} \tag{1}$$

$$= \frac{(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})^T (\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})}{(p+1)s^2} \tag{2}$$

$$= \frac{r_i^2}{p+1} \times \frac{h_i}{1-h_i}. \tag{3}$$

The first form (1) shows that Cook's distance measures the difference between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(i)}$ using the inverse of the contours of the estimated covariance matrix $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ of $\widehat{\boldsymbol{\beta}}$ and scaling by the number of terms $(p+1)$ in the model. The second form shows that Cook's distance can be viewed also as the squared length of the difference between the $n \times 1$ vector of fitted values $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ based on the full data and the $n \times 1$ vector of fitted values $\widehat{\mathbf{Y}}_{(i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(i)}$ when $\boldsymbol{\beta}$ is estimated without the ith case.

The final form (3) shows the general characteristics of cases with relatively large values of D_i. The ith *leverage* h_i, $0 \le h_i \le 1$, is the ith diagonal of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ that puts the "hat" on $\mathbf{Y}$, $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. It measures how far the predictor values $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ for the ith case are from the average predictor value $\overline{\mathbf{X}}$. If $\mathbf{X}_i$ is far from $\overline{\mathbf{X}}$ then the ith case will have substantial pull on the fit, h_i will be near its upper bound of 1, and the second factor of (3) will be very large. Consequently, D_i will be large unless the first factor in (3) is small enough to compensate. The second factor tells us about the leverage or pull that $\mathbf{X}_i$ has on the fitted model, but it does not depend on the response and thus says nothing about the actual fit of the ith case. That goodness of fit information is provided by r_i^2 in first factor of (3): r_i is the *Studentized residual* for the ith case – the ordinary residual for the ith case divided by $s\sqrt{1-h_i}$. The squared Studentized residual r_i^2 will be large when Y_i does not fit the model and thus can be regarded as an *outlier*, but it says nothing about leverage. In short, the

first factor gives information on the goodness of the fit of Y_i, but it says nothing about leverage, while the second factor gives the leverage information but says nothing about goodness of fit. When multiplied, these factors combine to give a measure of the influence of the ith case.

The Studentized residual r_i is a common statistic for testing the hypothesis that Y_i is not an outlier. That test is most powerful when h_i is small, so $\mathbf{X}_i$ is near $\overline{\mathbf{X}}$, and least powerful when h_i is relatively large. However, leverage or pull is weakest when h_i is small and strongest when h_i is large. In other words, the ability to detect ►outliers is strongest where the outliers tend to be the least influential and weakest where the outliers tend to be the most influential. This gives another reason why influence assessment can be crucial in an analysis.

Cook's distance is not a test statistic and should not by itself be used to accept cases or reject cases. It may indicate an anomalous case that is extramural to the experimental protocol or it may indicate the most important case in the analysis, one that points to a relevant phenomenon not reflected by the other data. Cook's distance does not distinguish these possibilities.

Illustration

The data that provided the original motivation for the development of Cook's distance came from an experiment on the absorption of a drug by rat livers. Nineteen rats were given various doses of the drug and, after a fixed waiting time, the rats were sacrificed and the percentage Y of the dose absorbed by the liver was measured. The predictors were dose, body weight and liver weight. The largest absolute Studentized residual is $\max |r_i| = 2.1$, which is unremarkable when adjusting for multiple testing. The case with the largest leverage 0.85 has a modest Studentized residual of 0.80, but a relatively large Cook's distance of 0.93 – the second largest Cook's distance is 0.27. Body weight and dose have significant effects in the analysis of the full data, but there are no significant effects after the influential case is removed. It is always prudent to study the impact of cases with relatively large values of D_i and all case for which $D_i > 0.5$. The most influential case in this analysis fits both of these criteria. The rat data are discussed in Cook and Weisberg (1999) and available from the accompanying software.

Acknowledgments

Research for this article was supported in part by National Science Foundation Grant DMS-0704098.

About the Author

Dennis Cook is Full Professor, School of Statistics, University of Minnesota. He served a ten-year term as Chair of the Department of Applied Statistics, and a three-year term as Director of the Statistical Center, both at the University of Minnesota. He has served as Associate Editor of the *Journal of the American Statistical Association* (1976–1982; 1988–1991; 2002–2005), *The Journal of Quality Technology*, *Biometrika* (1991–1993), *Journal of the Royal Statistical Society, Seies B* (1992–1997) and *Statistica Sinica* (1999–2005). He is a three-time recipient of the Jack Youden Prize for Best Expository Paper in *Technometrics* as well as the Frank Wilcoxon Award for Best Technical Paper. He received the 2005 COPSS Fisher Lecture and Award. He is a Fellow of ASA and IMS, and an elected member of the ISI.

Cross References

►Influential Observations
►Regression Diagnostics
►Robust Regression Estimation in Generalized Linear Models
►Simple Linear Regression

References and Further Reading

Cook RD (1977) Detection of influential observations in linear regression. Technometrics 19:15–18. Reprinted in 2000 under the same title for the Technometrics Special 40th Anniversary Issue 42, 65–68

Cook RD (1986) Assessment of local influence (with discussion). J R Stat Soc Ser B 48:133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London/New York. This book is available online without charge from the University of Minnesota Digital Conservancy: http://purl.umn.edu/37076

Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York

Zhu H, Ibrahim JG, Lee S, Zhang H (2007) Perturbation selection and influence measures in local influence analysis. Ann Stat 35:2565–2588

Copulas

Carlo Sempi
Professor, Dean of the Faculty of Mathematical, Physical and Natural Sciences
Università del Salento, Lecce, Italy

Copulas were introduced by Sklar in 1959 (Sklar 1959). In a statistical model they capture the dependence structure of the random variables involved, whatever the distribution

functions of the single random variables. They also allow the construction of families of bivariate or multivariate distributions.

The definition of the notion of copula relies on those of d-box (Definition 1) and of H-volume (Definition 2). Here, and in the following, we put $\mathbb{I} := [0,1]$.

Definition 1 *Let* $\boldsymbol{a} = (a_1, a_2, \ldots, a_d)$ *and* $\mathrm{b} = (b_1, b_2, \ldots, b_d)$ *be two points in* $\overline{\mathbb{R}}^d$, *with* $0 \le a_j \le b_j \le 1$ $(j \in \{1, 2, \ldots, d\})$; *the* d-box $[\mathbf{a}, \mathbf{b}]$ *is the cartesian product*

$$[\mathbf{a}, \mathbf{b}] = \prod_{j=1}^{d} [a_j, b_j],$$

Definition 2 *For a function* $H : \overline{\mathbb{R}}^d \to \overline{\mathbb{R}}$, *the* H-volume V_H *of the* d-box $[\mathbf{a}, \mathbf{b}]$ *is defined by*

$$V_H([\mathbf{a}, \mathbf{b}]) := \sum_{\mathbf{v}} sign(\mathbf{v})\, H(\mathbf{v}),$$

where the sum is taken over the 2^d *vertices* $\mathbf{v}$ *of the box* $[\mathbf{a}, \mathbf{b}]$; *here*

$$sign(\mathbf{v}) = \begin{cases} 1, & \text{if } v_j = a_j \text{ for an even number of indices,} \\ -1, & \text{if } v_j = a_j \text{ for an odd number of indices.} \end{cases}$$

Definition 3 *A function* $C_d : \mathbb{I}^d \to \mathbb{I}$ *is a* d-copula *if*

(a) $C_d(x_1, x_2, \ldots, x_d) = 0$, *if* $x_j = 0$ *for at least one index* $j \in \{1, 2, \ldots, d\}$;

(b) *when all the arguments of* C_d *are equal to* 1, *but for the* j*-th one, then*

$$C_d(1, \ldots, 1, x_j, 1, \ldots, 1) = x_j;$$

(c) *the* V_{C_d}*-volume of every* d-box $[\mathbf{a}, \mathbf{b}]$ *is positive,* $V_{C_d}([\mathbf{a}, \mathbf{b}]) \ge 0$.

The set of d*-copulas* $(d \ge 2)$ *is denoted by* C_d; *in particular, the set of (bivariate) copulas is denoted by* C_2.

Property (c) is usually referred to as the "d-increasing property of a d-copula". Thus every copula is the restriction to the unit cube $\mathbb{I}^d$ of a distribution function that concentrates all the probability mass on $\mathbb{I}^d$ and which has uniform margins (and this may also serve as an equivalent definition).

It is possible to show that C_d is a compact set in the set of all continuous functions from $\mathbb{I}^d$ into $\mathbb{I}$ equipped with the product topology, which corresponds to the topology of pointwise convergence. Moreover, in C_d pointwise and uniform convergence are equivalent.

Every d-copula satisfies the Fréchet–Hoeffding bounds: for all $x_1, \ldots, x_d$ in $\mathbb{I}$, one has

$$W_d(x_1, \ldots, x_d) \le C(x_1, \ldots, x_d) \le M_d(x_1, \ldots, x_d), \quad (1)$$

where

$$W_d(x_1, \ldots, x_d) := \max\{0, x_1 + \cdots + x_d - d + 1\}$$
$$M_d(x_1, \ldots, x_d) := \min\{x_1, \ldots, x_d\}.$$

Also relevant is the "independence copula"

$$\Pi_d(x_1, \ldots, x_d) := \prod_{j=1}^{d} x_j.$$

While Π_d and M_d are copulas for every $d \ge 2$, W_d is a copula only for $d = 2$, although the lower bound provided by (1) is the best possible.

- Π_d is the distribution function of the random vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ whose components are independent and uniformly distributed on $\mathbb{I}$.
- M_d is the distribution function of the vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ whose components are uniformly distributed on $\mathbb{I}$ and such that $U_1 = U_2 = \cdots = U_d$ almost surely.
- W_2 is the distribution function of the vector $\mathbf{U} = (U_1, U_2)$ whose components are uniformly distributed on $\mathbb{I}$ and such that $U_1 = 1 - U_2$ almost surely.

The importance of copulas for the applications in statistics stems from Sklar's theorem.

Theorem 1 (Sklar 1959) *Let* H *be a* d*-dimensional distribution function with margins* $F_1, F_2, \ldots, F_d$, *and let* A_j *denote the range of* F_j, $A_j := F_j(\overline{\mathbb{R}})$ $(j = 1, 2, \ldots, d)$. *Then there exists a* d*-copula* C, *uniquely defined on* $A_1 \times A_2 \times \cdots \times A_d$, *such that, for all* $(x_1, x_2, \ldots, x_d) \in \overline{\mathbb{R}}^d$,

$$H(x_1, x_2, \ldots, x_d) = C(F_1(t_1), F_2(t_2), \ldots, F_d(t_d)). \quad (2)$$

Conversely, if $F_1, F_2, \ldots, F_d$ *are distribution functions, and if* C *is any* d*-copula, then the function* $H : \overline{\mathbb{R}}^d \to \mathbb{I}$ *defined by* (2) *is a* d*-dimensional distribution function with margins* $F_1, F_2, \ldots, F_d$.

For a compact and elegant proof of this result see (Rüschendorf 2009).

The second ("converse") part of Sklar's theorem is especially important in the construction of statistical models, since it allows to proceed in two separate steps:

- Choose the one-dimensional distribution functions F_1, $F_2, \ldots, F_d$ that describe the behavior of the individual statistical quantities (random variables) $X_1, X_2, \ldots, X_d$ that appear in the model.
- Fit these in (2) by means of a copula C that captures the dependence relations among $X_1, X_2, \ldots, X_d$.

These two steps are independent in the sense that, once a copula C has been chosen, any choice of the distribution functions $F_1, F_2, \ldots, F_d$ is possible.

It should be stressed that the copula whose existence is established in Sklar's theorem is uniquely defined only when the distribution functions have no discrete component; otherwise, there are, in general, several copulas that coincide on $A_1 \times A_2 \times \cdots \times A_d$ and which satisfy (2). This lack of uniqueness may have important consequences when dealing with the copula of random variables (see, e.g., (Marshall 1996)).

The introduction of copulas in the statistical literature has allowed an easier way to construct models by proceeding in two separate steps: (i) the specification of the marginal laws of the random variables involved and (ii) the introduction of a copula that describes the dependence structure among these variables. In many applications (mainly in Engineering) this has allowed to avoid the mathematically elegant and easy-to-deal, but usually unjustified, assumption of independence.

In view of possible applications, it is important to have at one's disposal a stock of copulas. Many families of bivariate copulas can be found in the books by Nelsen (2006), by Balakrishnan and Lai (2009) and Jaworski et al. (2010). Here we quote only the gaussian, the meta-elliptical (Fang et al. 2002) and the extreme-value copulas (Ghoudi et al. 1998). A popular family of copulas is provided by the *Archimedean* copulas, which, in the two-dimensional case, are represented in the form

$$C_\varphi(s,t) = \varphi^{[-1]}\left(\varphi(s) + \varphi(t)\right),$$

where the *generator* $\varphi : [0,1] \to [0,+\infty]$ is continuous, strictly decreasing, convex and $\varphi(1) = 0$, and $\varphi^{[-1]}$ is the *pseudo-inverse* of φ, defined by $\varphi^{[-1]}(t) := \varphi^{-1}(t)$, for $t \in [0, \varphi(0)]$, and by 0, for $t \in [\varphi(0), +\infty]$. These copulas depend on a function of a single variable, the generator φ; as a consequence, the statistical properties of a pair of random variables having C_φ as their copula are easily computed in terms of φ (Genest and MacKay 1986; Nelsen 2006). For the multivariate case the reader is referred to the paper by McNeil and Nešlehová (2009), where the generators of a such a copula are completely characterized.

Notice, however, that the choice of a symmetric copula, in particular of an Archimedean one, means that the random variables involved are exchangeable, if they have the same distribution. The effort to avoid this limitation motivates the recent great interest in the construction of nonsymmetric copulas (see, e.g., Liebscher (2008)).

It must also be mentioned that many methods of construction for copulas have been introduced; here we mention

- Ordinal sums (Mesiar and Sempi 2010);
- Shuffles of Min (Mikusiński et al. 1992) and its generalization to an arbitrary copula (Durante et al. 2009);
- The $*$-product (Darsow et al. 1992) and its generalization (Durante et al. 2007a);
- Transformations of copulas, $C_h(u,v) := h^{[-1]}\left(C(h(u),h(v))\right)$, where the function $h : \mathbb{I} \to \mathbb{I}$ is concave (Durante and Sempi 2005);
- Splicing of symmetric copulas (Durante et al. 2007b; Nelsen et al. 2008);
- Patchwork copulas (De Boets and De Meyer 2007; Durante et al. 2009);
- Gluing of copulas (Siburg and Stoimenov 2008).

A strong motivation for the development of much of copula theory in recent years has come from their applications in Mathematical Finance (see, e.g., (Embrechts et al. 2003), in Actuarial Science (Free and Valdez 1998), and in Hydrology (see, e.g., (Genest and Favre 2007; Salvadori et al. 2007)).

About the Author

Carlo Sempi received his Ph.D. in Applied Mathematics in 1974, University of Waterloo, Canada (his advisor was Professor Bruno Forte). He was Chairman of the Department of Mathematics, Università del Salento (2002–2008). Currently, he is Faculty of Mathematical, Physical and Natural Sciences, University of Salento. Professor Sempi has (co-)authored about 75 refereed papers; many of these papers are on Copulas (some written with the leading authorities on this subject (including Abe Sklar and Roger Nelsen). He was the organizer of the conference "Meeting Copulæ: the 50th anniversary," Lecce, June 2009.

Cross References

►Bivariate Distributions
►Copulas in Finance
►Copulas: Distribution Functions and Simulation
►Measures of Dependence
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation
►Non-Uniform Random Variate Generations
►Quantitative Risk Management
►Statistical Modeling of Financial Markets

References and Further Reading

Balakrishnan N, Lai C-D (2009) Continuous bivariate distributions, 2nd edn. Springer, New York

Darsow W, Nguyen B, Olsen ET (1992) Copulas and Markov processes. Illinois J Math 36:600–642

De Baets B, De Meyer H (2007) Orthogonal grid constructions of copulas. IEEE Trans Fuzzy Syst 15:1053–1062

Durante F, Sempi C (2005) Copula and semicopula transforms. Int J Math Math Sci 4:645–655
Durante F, Klement EP, Quesada-Molina JJ (2007a) Remarks on two product-like constructions for copulas. Kybernetika (Prague) 43:235–244
Durante F, Kolesárová A, Mesiar R, Sempi C (2007b) Copulas with given diagonal sections: novel constructions and applications. Int J Uncertain Fuzziness Knowledge-Based Syst 15: 397–410
Durante F, Rodríguez Lallena JA, Úbeda Flores M (2009) New constructions of diagonal patchwork copulas. Inf Sci 179:3383–3391
Durante F, Sarkoci P, Sempi C (2009) Shuffles of copulas. J Math Anal Appl 352:914–921
Embrechts P, Lindskog F, McNeil A (2003) Modelling dependence with copulas and applications to risk management. In: Rachev S (ed) Handbook of heavy tailed distributions in finance, Chapter 8. Elsevier, Amsterdam, pp 329–384
Fang HB, Fang KT, Kotz S (2002) The meta-elliptical distributions with given marginals. J Multivariate Anal 82:1–16
Free EW, Valdez EA (1998) Understanding relationships using copulas. N Am J Actuar 2:1–25
Genest C, Favre AC (2007) Everything you always wanted to know about copula modeling but were afraid to ask. J Hydrol Eng 12(4):347–368
Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283
Ghoudi K, Khoudraji A, Rivest L-P (1998) Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. Canad J Stat 26:187–197
Jaworski P, Durante F, Härdle W, Rychlik T (eds) (2010) Copula theory and its applications. Springer, Berlin
Liebscher E (2008) Construction of asymmetric multivariate copulas. J Multivariate Anal 99:2234–2250
Marshall AW (1996) Copulas, marginals, and joint distributions, In: Rüschendorf L, Schweizer B, Taylor MD (eds) Distributions with fixed marginals and related topics, Institute of Mathematical Statistics, Lecture Notes – Monograph Series vol 28, Hayward, pp 213–222
McNeil AJ Nešlehová J (2009) Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. Ann Stat 37:3059–3097
Mesiar R, Sempi C (2010) Ordinal sums and idempotents of copulas. Mediterr J Math 79:39–52
Mikusiński P, Sherwood H, Taylor MD (1992) Shuffles of min. Stochastica 13:61–74
Nelsen RB (2006) An introduction to copulas, Lecture Notes in Statistics 139, 2nd edn. Springer, New York
Nelsen RB, Quesada Molina JJ, Rodríguez Lallena JA, Úbeda Flores M (2008) On the construction of copula and quasi-copulas with given diagonal sections. Insurance Math Econ 42:473–483
Rüschendorf L (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. J Statist Plan Inference 139:3921–3927
Salvadori G, De Michele C, Kottegoda NT, Rosso R (2007) Extremes in nature. An approach using copulas, Water Science and Technology Library, vol 56. Springer, Dordrecht
Siburg KF, Stoimenov PA (2008) Gluing copulas. Commun Stat Theory and Methods 37:3124–3134
Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Copulas in Finance

Cherubini Umberto
Associate Professor of Mathematical Finance, MatematES
University of Bologna, Bologna, Italy

C

Introduction

Correlation trading denotes the trading activity aimed at exploiting changes in correlation or more generally in the dependence structure of assets or risk factors. Likewise, correlation risk is defined as the exposure to losses triggered by changes in correlation. The copula function technique, which enables analyzing the dependence structure of a joint distribution independently from the marginal distributions, is the ideal tool to assess the impact of changes in market comovements on the prices of assets and the amount of risk in a financial position. As far as the prices of assets are concerned, copula functions enable pricing multivariate products consistently with the prices of univariate products. As for risk management, copula functions enable assessing the degree of diversification in a financial portfolio as well as the sensitivity of risk measures to changes in the dependence structure of risk factors. The concept of consistency between univariate and multivariate prices and risk factors is very similar, and actually parallel, to the problem of compatibility between multivariate probability distributions and distribution of lower dimensions. In finance, this concept is endowed with a very practical content, since it enables designing strategies involving univariate and multivariate products with the aim of exploiting changes in correlation.

Copulas and Spatial Dependence in Finance

Most of the applications of copula functions in finance are limited to multivariate problems in a cross-sectional sense (as econometricians are used to saying), or in a spatial sense (as statisticians prefer). In other words, almost all the applications have to do with the dependence structure of different variables (prices or losses in the case of finance) at the same date. The literature on applications like these is too large to be quoted here in detail, and we refer the reader to the bibliography below and to those in Bouyé et al. (2000) and Cherubini et al. (2004) for more details.

Pricing Applications

Standard asset pricing theory is based on the requirement that the prices of financial products must be such that no arbitrage opportunities can be exploited, meaning that no financial strategy can be built yielding positive return

with no risk. The price consistent with absence of arbitrage opportunities is also known as the "*fair value*" of the product. The fundamental theorem of finance states that this amounts to assert that there must exist a probability measure, called *risk-neutral* measure, under which the expected future returns on each and every asset must be zero, or, which is the same, that the prices of financial assets must be endowed with the martingale property, when measured with that probability measure. Then, the price of each asset promising some payoff in the future must be the expected value with respect to the same probability measure. This implies that if the payoff is a function of one risk factor only, the price is the expected value with respect to a univariate probability measure. If the payoff is instead a function of more than one variable, then it must be computed by taking expectations with respect to the joint distribution of the risk factors. Notice that this implies that there must be a relationship of price consistency between the prices of univariate and multivariate products, and more generally there must be *compatibility* relationships (in the proper meaning of the term in statistics) among prices. This is particularly true for derivative products promising some payments contingent on a multivariate function of several assets. These are the so-called basket derivative products, which are mainly designed on common equity stock (*equity derivatives*), or insurance policies against default of a set of counterparties (*credit derivatives*). The same structure may be used for products linked to commodities or actuarial risks. There are also products called "*hybrids*" that include different risk factors (such as market risk, i.e., the risk of market movements and default of some obligors) in the same product. For the sake of illustration, we provide here two standard examples of basket equity and credit derivatives:

Example 1 (*Altiplano Note*) These are so-called *digital* products, that is, paying a fixed sum if some event takes place at a given future date T. Typically, the event is defined as a set of stocks or market indexes, and the product pays the fixed sum if all of them are above some given level, typically specified as a percentage of the initial level. The price of this product is of course the joint *risk-neutral* probability that all the assets be above a specified level at time $T : Q(S_1(T) > K_1, S_2(T) > K_2, \ldots, S_m(T) > K_m)$, where K_i are the levels (so-called *strike* prices). Consider now that we can actually estimate the marginal distributions from the option markets, so that we can price each $Q_i(S_i(T) > K_i)$. As a result, the only reason why one wants to invest in the multivariate digital product above instead of on a set of univariate ones is to exploit changes in correlation among the assets. To put it in other terms, the value of a multivariate product can increase even if the prices of all the univariate products remain unchanged, and this may occur if the correlation increases. Copula functions are ideal tools to single out this effect.

Example 2 (Collateralized Debt Obligation (*CDO*)) Today it is possible to invest in portfolios of credit derivatives. In nontechnical terms, we can buy and sell insurance ("*protection*" is the term in market jargon) on the first x% losses on defaults of a set of obligors (called "names"). This product is called $0 - x\%$ *equity tranche* of a portfolio of credit losses. For the sake of simplicity assume 100 names and a 0–1% equity tranche, and assume that in case of default, each loss is equal to 1. So, this tranche pays insurance the first time a default occurs (it is also called a *first-to-default* protection). Again, we can easily recover the univariate probabilities of default from other products, namely the so-called credit default swap (*CDS*) market. So, we can price the protection for every single name in the basket. The price of the *first-to-default* must then be compatible with such prices. In fact, with respect to such prices, the multivariate product is different only because it allows to invest in correlation. Again, the equity tranche can increase in value even though the values of single-insurance *CDS* for all the names remain constant, provided that the correlation of defaults increase. Even in this case, copula functions provide the ideal tool to evaluate and trade the degree of dependence of the events of default.

Risk Management

The risk manager faces the problem of measuring the exposure of the position to different risk factors. In the standard practice, he transforms the financial positions in the different assets and markets into a set of *exposures* (*buckets*, in jargon) to a set of risk factors (*mapping* process). The problem is then to estimate the joint distribution of losses $L_1, L_2, L_3, \ldots, L_k$, on these exposures and define a risk measure on this distribution. Typical measures are *Value-at-Risk* (*VaR*) and *Expected Shortfall* (*ES*) defined as

$$VaR(L_i) \equiv \inf(x : H_i(L_i) > 1-\alpha) \qquad ES \equiv E(L_i | L_i \geq VaR)$$

where $H_i(.)$ is the marginal probability distribution of loss L_i. The risk measure of the overall portfolio will analogously be

$$VaR\left(\sum_{i=1}^{k} L_i\right) \equiv \inf\left(x : H\left(\sum_{i=1}^{k} L_i\right) > 1-\alpha\right)$$
$$ES \equiv E\left(\sum_{i=1}^{k} L_i | L_i \geq VaR\right)$$

where $H(.)$ is now the probability distribution of the sum of losses. It is clear that the relationship between

univariate and multivariate risk measures is determined by the dependence structure linking the losses themselves. Again, copula functions are used to merge these risk measures together. Actually, if the $\max(.,.)$ instead of the sum were used as the aggregation operator, the risk measure would use the copula function itself as the aggregated distribution of losses.

Copula Pricing and Arbitrage Relationships

Using copula functions is very easy to recover arbitrage relationships (i.e., consistency, or compatibility relationships) among prices of multivariate assets. These relationships directly stem from links between copula functions representing the joint distribution of a set of events and those representing the joint distribution of the complement sets. A *survival copula* is defined as

$$Q(S_1 > K_1, S_2 > K_2, \ldots, S_m > K_m) = \overline{C}(1-u_1, 1-u_2, \ldots, 1-u_m)$$

The relevance of this relationship in finance is clear because it enforces the so-called put-call parity relationships. These relationships establish a consistency link between the price of products paying out if all the events in a set take place and products paying out if none of them take place. Going back to the Altiplano note above, we may provide a straightforward check of this principle.

Example 3 (Put-Call Parity of Altiplano Notes) Assume an Altiplano Note like that in Example 1, with the only difference that the fixed sum is paid if all the assets S_i are below (instead of above) the same predefined thresholds K_i. Clearly, the value of the product will be $Q(S_1(T) \leq K_1, S_2(T) \leq K_2, \ldots, S_m(T) \leq K_m)$. Given the marginal distributions, the dependence structure of this product, which could be called *put*, or *bearish*, Altiplano should be represented by a copula, while the price of the *call* or *bullish* Altiplano in Example 1 should be computed using the survival copula. It can be proved that if this is not the case, one could exploit arbitrage profits (see Cherubini and Luciano 2002; Cherubini and Romagnoli 2009).

Copulas and Temporal Dependence in Finance

So far, we have described correlation in a spatial setting. The flaw of this approach, and of copula applications to finance in general, is that no consistency link is specified, among prices with the same underlying risk factors, but payoffs at different times. We provide three examples here, two of which extend the equity and credit products cases presented above, while the third one refers to a problem arising in risk management applications. Research on this topic, as far as applications to finance are concerned, is at an early stage, and is somewhat covered in the reference bibliography below.

Example 4 (Barrier Altiplano Note) Assume an Altiplano Note with a single asset, but paying a fixed sum at the final date if the price of that asset S remains above a given threshold K on a set of different dates $\{t_1, t_2, t_3, \ldots, t_n\}$. This product can be considered multivariate just like that in Example 1, by simply substituting the concept of *temporal dependence* for that of *spatial dependence*. Again, copula functions can be used to single out the impact of changes in temporal dependence on the price of the product. For some of these products, it is not uncommon to find the so-called memory feature, according to which the payoff is paid for all the dates in the set at the first time that the asset is above the threshold.

Example 5 (Standard Collateralized Debt Obligations) (CDX, iTraxx) In the market there exist *CDO* contracts, like those described in Example 2 above, whose terms are standardized, so that they may be of interest for a large set of investors. These products include 125 "names" representative of a whole market (*CDX* for the US and *iTraxx* for Europe), and on these markets people may trade tranches buying and selling protection on 0–3%, 3–6%, and so on, according to a schedule, which is also standardized. So, for example, you may buy protection against default of the first 3% of the same 125 names, but for a time horizon of 5 or 10 years (the standard maturities are 5, 7, and 10 years). For sure you will pay more for the 10 years insurance than for the 5 years insurance on the same risk. How much more will depend on the relationship between the losses which you may incur in the first 5 years and those that you may face in the remaining 5 years. Clearly, temporal dependence cannot be avoided in this case and it is crucial in order to determine a consistency relationship between the price of insurance against losses on a term of 5 years and those on a term of 10 years. This consistency relationship is known as the *term structure* of *CDX* (or *iTraxx*) premia.

Example 6 (Temporal aggregation of risk measures) We may also think of a very straightforward problem of temporal dependence in risk management, which arises whenever we want to compute the distribution of losses over different time horizons. An instance in which this problem emerges is when one wants to apply risk measures to compare the performance of managed funds over different investment horizons. The same problem arises whenever we have to establish a dependence structure between risk factors that are measured with different time frequencies.

To take the typical example, assume you want to study the dependence structure between market risk and credit risk in a portfolio. The risk measures of market risk are typically computed on losses referred to a period of 1 or 10 days, while credit risk losses are measured in periods of months. Before linking the measures, one has to modify the time horizon of one of the two in order to match that of the other one. The typical "*square root rule*" used in the market obviously rests on the assumption of independent losses with Gaussian distribution, but this is clearly a coarse approximation of reality.

Financial Prices Dynamics and Copulas

The need to extend copulas to provide a representation of both spatial and temporal dynamics of financial prices and risk factors has led to the rediscovery of the relationship between ►copulas and the Markov process (see ►Markov Processes) that was first investigated by Darsow et al. (1992). Actually, even though the Markov assumption may seem restrictive for general applications, it turns out to be consistent with the standard *Efficient Market Hypothesis* paradigm. This hypothesis postulates that all available information must be embedded in the prices of assets, so that price innovations must be unpredictable. This leads to models of asset prices driven by independent increments, which are Markovian by construction. For these reasons, this approach was rediscovered both for pricing and financial econometrics applications (Cherubini et al. 2008, 2009; Cherubini and Romagnoli 2010; Ibragimov 2009; Chen 2009).

We illustrate here the basic result going back to Darsow et al. (1992) with application to asset prices. We assume a set of $\{S_1, S_2, \ldots, S_m\}$ assets and a set of $\{t_0, t_1, t_2, \ldots, t_n\}$ dates, and a filtered probability space generated by the prices and satisfying the usual conditions. Denote S_i^j the price of asset i at time j. First, define the product of two copulas as

$$A * B(u,v) \equiv \int_0^1 \frac{\partial A(u,t)}{\partial t}\frac{\partial B(t,v)}{\partial t}dt$$

and the extended concept of "star-product" as

$$A * B(u_1, u_2, \ldots, u_{m+n-1}) \equiv \int_0^{u_m} \frac{\partial A(u_1, u_2, \ldots, u_{m-1}, t)}{\partial t} \times \frac{\partial B(t, u_{m+1}, u_{m+2}, \ldots, u_{m+n-1})}{\partial t} dt$$

Now, Darsow et al. proved that a stochastic process S_i is a first order ►Markov chain if and only if there exists a set of bivariate copula functions $T_i^{j,j+1}, j = 1, 2, \ldots, n$, such that the dependence among $\{S_i^1, S_i^2, \ldots, S_i^n\}$ can be written as

$$G_i^j\left(u_i^1, u_i^2, \ldots, u_i^j\right) = T_i^{1,2}\left(u_i^1, u_i^2\right) * T_i^{2,3}\left(u_i^1, u_i^2\right) \ldots * T_i^{j-1,j}\left(u_i^1, u_i^2\right)$$

The result was extended to general Markov processes of order k by Ibragimov (2009). Within this framework, Cherubini et al. (2009) provided a characterization of processes with independent increments. The idea is to represent the price S^j (or its logarithm) as $S^{j-1} + Y^j$. Assume that the dependence structure between S^{j-1} and Y^j is represented by copula $C(u,v)$. Then, the dependence between S^{j-1} and S^j may be written as

$$T^{j-1,j}(u,v) = \int_0^u D_1 C\left(w, F_Y\left(F_{S,j}^{-1}(v) - F_{S,j-1}^{-1}(w)\right)\right)\mathrm{d}w$$

where D_1 represents partial derivative with respect to the first variable, $F_Y(.)$ denotes the probability distribution of the increment, and the distribution $F_{S,k}(.)$ the probability distribution of S^k. The probability distribution of S^k is obtained by taking the marginal

$$F(S^j \le s) = T^{j-1,j}(1,v) = \int_0^1 D_1 C\left(w, F_Y\left(s - F_{S,j-1}^{-1}(w)\right)\right)\mathrm{d}w$$

This is a sort of extension of the concept of convolution to the case in which the variables in the sum are not independent. Of course, the case of independent increments is readily obtained by setting $C(u,v) = uv$. The copula linking S^{j-1} and S^j becomes in this case

$$T^{j-1,j}(u,v) = \int_0^u F_Y\left(F_{S,j}^{-1}(v) - F_{S,j-1}^{-1}(w)\right)\mathrm{d}w$$

A well-known special case is

$$T^{j-1,j}(u,v) = \int_0^u \Phi\left(\Phi^{-1}(v) - \Phi^{-1}(w)\right)\mathrm{d}w$$

with $\Phi(x)$ the standard normal distribution, which yields the dependence structure of a Brownian motion (see ►Brownian Motion and Diffusions) upon appropriate standardization. As for pricing applications, Cherubini et al. (2008) applied the framework to temporal dependence of losses and the term structure of *CDX* premia, and Cherubini and Romagnoli (2010) exploited the model to price barrier Altiplanos. This stream of literature, which applies copulas to modeling stochastic processes in discrete time, casts a bridge to a parallel approach, that directly applies copulas to model dependence among

stochastic processes in continuous time: this is the so-called Lévy copula approach (Kallsen and Tankov 2006). Both these approaches aim at overcoming the major flaw of copula functions as a static tool and unification of them represents the paramount frontier issue in this important and promising field of research.

About the Author

Umberto Cherubini is Associate Professor of Mathematical Finance at the University of Bologna. He is the author or coauthor of about 50 papers and is the coauthor of 5 books, of which two are in Italian and *Copula Methods in Finance* (with E. Luciano and W. Vecchiato, 2004), *Structured Finance: The Object Oriented Approach* (with G. Della Lunga, 2007), and *Fourier Methods in Finance* (with G. Della Lunga, S. Mulinacci and P. Rossi, 2010), all with John Wiley, Finance Series. Chichester, UK.

Cross References

►Copulas
►Copulas: Distribution Functions and Simulation
►Quantitative Risk Management
►Statistical Modeling of Financial Markets

References and Further Reading

Bouyé E, Durrleman V, Nikeghbali A, Riboulet G, Roncalli T (2000) Copulas for finance: a reading guide and some applications. Groupe de Recherche, Opérationnelle, Crédit Lyonnais. Working paper

Chen X, Wu SB, Yi Y (2009) Efficient estimation of copula-based semiparametric Markov models. Ann Stat 37(6B):4214–4253

Cherubini U, Luciano E (2002) Bivariate option pricing with copulas. Appl Math Finance 9:69–86

Cherubini U, Romagnoli S (2009) Computing the volume of n-dimensional copulas. Appl Math Finance 16(4):307–314

Cherubini U, Romagnoli S (2010) The dependence structure of running maxima and minima: results and option pricing applications. Mathematical Finance 20(1):35–58

Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. Wiley Finance Series, Chichester

Cherubini U, Mulinacci S, Romagnoli S (2008) A copula-based model of the term structure of CDO tranches. In: Hardle WK, Hautsch N, Overbeck L (eds) Applied quantitative finance. Springer, Berlin, pp 69–81

Cherubini U, Mulinacci S, Romagnoli S (2009) A copula-based model of speculative price dynamics. University of Bologna, Berlin

Darsow WF, Nguyen B, Olsen ET (1992) Copulas and Markov processes. Illinois J Math 36:600–642

Embrechts P, Lindskog F, McNeil AJ (2003) Modelling dependence with copulas with applications to risk management. In: Rachev S (ed) A handbook of heavy tailed distributions in finance. Elsevier, Amsterdam, pp 329–384

Gregory J, Laurent JP (2005) Basket defaults, CDOs and factor copulas. J Risk 7(4):103–122

Ibragimov R (2005) Copula based characterization and modeling for time series. Econom Theory 25:819–846

Kallsen J, Tankov P (2006) Characterization of dependence of multidimensional Lévy processes using Lévy copulas. J Multivariate Anal 97:1151–1172

Li D (2000) On default correlation: a copula function approach. J Fixed Income 9:43–54

Patton A (2002) Modelling asymmetric exchange rate dependence. Int Econ Rev 47(2):527–556

Rosemberg J (2003) Non parametric pricing of multivariate contingent claims federal reserve bank of New York staff reports, n. 162

Van Der Goorberg R, Genest C, Werker B (2005) Bivariate option pricing using dynamic copula models. Insur Math Econ 37:101–114

Copulas: Distribution Functions and Simulation

Pranesh Kumar
Professor
University of Northern British Columbia, Prince George, BC, Canada

Introduction

In multivariate data modelling for an understanding of stochastic dependence the notion of correlation has been central. Although correlation is one of the omnipresent concepts in statistical theory, it is also one of the most misunderstood concepts. The confusion may arise from the literary meaning of the word to cover any notion of dependence. From mathematics point of view, correlation is only one particular measure of stochastic dependence. It is the canonical measure in the world of ►multivariate normal distributions and in general for spherical and elliptical distributions. However empirical research in many applications indicates that the distributions of the real world seldom belong to this class. We collect and present ideas of copula functions with applications in statistical probability distributions and simulation.

Dependence

We denote by (X, Y) a pair of real-valued nondegenerate random variables with finite variances σ_x^2 and σ_y^2 respectively. The correlation coefficient between X and Y is the standardized covariance σ_{xy}, i.e., $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, $\rho \in [-1, 1]$. The correlation coefficient is a measure of linear depen-hh dence only. In case of independent random variables, correlation is zero. Embrechts, McNeil and Straumann (1999) have discussed that in case of imperfect linear dependence, i.e., $-1 < \rho < 1$, misinterpretations of correlation are possible. Correlation is not ideal for a dependence measure and

causes problems when there are heavy-tailed distributions. Independence of two random variables implies they are uncorrelated but zero correlation does not in general imply independence. Correlation is not invariant under strictly increasing linear transformations. Invariance property is desirable for the statistical estimation and significance testing purposes. Further correlation is sensitive to ►outliers in the data set. The popularity of linear correlation and correlation based models is primarily because it is often straightforward to calculate and manipulate them under algebraic operations. For many ►bivariate distributions it is simple to calculate variances and covariances and hence the correlation coefficient. Another reason for the popularity of correlation is that it is a natural measure of dependence in multivariate normal distributions and more generally in multivariate spherical and elliptical distributions. Some examples of densities in the spherical class are those of the multivariate t-distribution and the ►logistic distribution.

Another class of dependence measures is rank correlations. They are defined to study relationships between different rankings on the same set of items. Rank correlation measures the correspondence between two rankings and assess their significance. Two commonly used measures of concordance are Spearman's rank correlation (ρ_s) and Kendall's rank correlation (τ). Assuming random variables X and Y have distribution functions F_1 and F_2 and joint distribution function F, Spearman's rank correlation $\rho_s = \rho(F_1(X), F_2(Y))$ where ρ is the linear correlation coefficient. If (X_1, Y_1) and (X_2, Y_2) are two independent pairs of random variables from the distribution function F, then the Kendall's rank correlation is $\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$. The main advantage of rank correlations over ordinary linear correlation is that they are invariant under monotonic transformations. However rank correlations do not lend themselves to the same elegant variance–covariance manipulations as linear correlation does since they are not moment-based.

A measure of dependence like linear correlation summarizes the dependence structure of two random variables in a single number. Scarsini (1984) has detailed properties of copula based concordance measures. Another excellent discussion of dependence measures is by Embrecht et al. (1999). Let $D(X, Y)$ be a measure of dependence which assigns a real number to any real-valued pair of random variables (X, Y). Then dependence measure $D(X, Y)$ is desired to have properties: (i) Symmetry: $D(X, Y) = D(Y, X)$; (ii) Normalization: $-1 \leq D(X, Y) \leq +1$; (iii) Comonotonic or Countermonotonic: The notion of comonotonicity in probability theory is that a random vector is comonotonic if and only if all marginals are non-decreasing functions (or non-increasing functions) of the same random variable. A measure $D(X, Y)$ is comonotonic if $D(X, Y) = 1 \iff X, Y$ or countermonotonic if $D(X, Y) = -1 \iff X, Y$; (iv) For a transformation T strictly monotonic on the range of X, $D(T(X), Y) = D(X, Y)$, $T(X)$ increasing or $D(T(X), Y) = -D(X, Y)$, $T(X)$ decreasing.

Linear correlation ρ satisfies properties (i) and (ii) only. Rank correlations fulfill properties (i)–(iv) for continuous random variables X and Y. Another desirable property is: (v) $D(X, Y) = 0 \iff X, Y$ (Independent). However it contradicts property (iv). There is no dependence measure satisfying properties (iv) and (v). If we desire property (v), we should consider dependence measure $0 \leq D^*(X, Y) \leq +1$. The disadvantage of all such dependence measures $D^*(X, Y)$ is that they can not differentiate between positive and negative dependence (Kimeldorf and Sampson 1978; Tjøstheim 1996).

Copulas

►Copulas have recently emerged as a means of describing joint distributions with uniform margins and a tool for simulating data. They express joint structure among random variables with any marginal distributions. With a copula we can separate the joint distribution into marginal distributions of each variable. Another advantage is that the conditional distributions can be readily expressed using the copula. An excellent introduction of copulas is presented in Joe (1997) and Nelsen (2006). Sklar's theorem (1959) states that any multivariate distribution can be expressed as the k-copula function $C(u_1, \ldots, u_i, \ldots, u_k)$ evaluated at each of the marginal distributions. Copula is not unique unless the marginal distributions are continuous. Using probability integral transform, each continuous marginal $U_i = F_i(x_i)$ has a uniform distribution (see ►Uniform Distribution in Statistics) on $I \in [0, 1]$ where $F_i(x_i)$ is the cumulative integral of $f_i(x_i)$ for the random variable $X_i \in (-\infty, \infty)$. The k-dimensional probability distribution function F has a unique copula representation $F(x_1, x_2, \ldots, x_k) = C(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)) = C(u_1, u_2, \ldots, u_k)$. The joint probability density function is written as $f(x_1, x_2, \ldots, x_k) = \Pi_{i=1}^{k} f_i(x_i) \times c(F_1(x_1), F_2(x_2), \ldots, F_k(x_k))$ where $f_i(x_i)$ is each marginal density and coupling is provided by copula density $c(u_1, u_2, \ldots, u_k) = \partial^k C(u_1, u_2, \ldots, u_k)/\partial u_1 \partial u_2 \ldots \partial u_k$ if it exists. In case of independent random variables, copula density $c(u_1, u_2, \ldots, u_k)$ is identically equal to one. The importance of the above equation $f(x_1, x_2, \ldots, x_k)$

is that the independent portion expressed as the product of the marginals can be separated from the function $c(u_1, u_2, \ldots, u_k)$ describing the dependence structure or shape. The dependence structure summarized by a copula is invariant under increasing and continuous transformations of the marginals. This means that suppose we have a probability model for dependent insurance losses of various kinds. If our interest now lies in modelling the logarithm of these losses, the copula will not change, only the marginal distributions will change.

The simplest copula is independent copula $\Pi := C(u_1, u_2, \ldots, u_k) = u_1 u_2 \ldots u_k$ with uniform density functions for independent random variables. Another copula example is the Farlie–Gumbel–Morgenstern (FGM) bivariate copula. The general system of FGM bivariate distributions is given by $F(x_1, x_2) = F_1(x_1) \times F_2(x_2) [1 + \rho (1 - F_1(x_1)) (1 - F_2(x_2))]$ and copula associated with this distribution is a FGM bivariate copula $C(u, v) = uv [1 + \rho (1 - u) (1 - v)]$. A widely used class of copulas is Archimedean copulas which has a simple form and models a variety of dependence structures. Most of the Archimedean copulas have closed-form solutions. To define an Archimedean copula, let ϕ be a continuous strictly decreasing convex function from $[0,1]$ to $[0, \infty]$ such that $\phi(1) = 0$ and $\phi(0) = \infty$. Let ϕ^{-1} be the pseudo inverse of ϕ. Then a k-dimensional Archimedean copula is $C(u_1, u_2, \ldots, u_k) = \phi^{-1} [\phi(u_1) + \ldots + \phi(u_k)]$. The function ϕ is known as a generator function. Thus any generator function satisfying $\phi(1) = 0$; $\lim_{x \to 0} \phi(x) = \infty$; $\phi'(x) < 0$; $\phi''(x) > 0$ will result in an Archimedean copula. For an example, generator function $\phi(t) = \left(t^{-\theta} - 1\right)/\theta$, $\theta \in [-1, \infty) \backslash \{0\}$ results in the bivariate Clayton copula $C(u_1, u_2) = \max\left[\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, 0\right]$. The copula parameter θ controls the amount of dependence between X_1 and X_2.

The Frécht–Hoeffding bounds for copulas: The lower bound for k-variate copula is $W(u_1, u_2, \ldots, u_k) := \max\{1 - n + \sum_{i=1}^{k} u_i, 0\} \leq C(u_1, u_2, \ldots, u_k)$. The upper bound for k-variate copula is $C(u_1, u_2, \ldots, u_k) \leq \min_{i \in \{1,2,\ldots,k\}} u_i =: M(u_1, u_2, \ldots, u_k)$. For all copulas, the inequality $W(u_1, \ldots, u_k) \leq C(u_1, \ldots, u_k) \leq M(u_1, \ldots, u_k)$ is satisfied. This inequality is well known as the Frécht-Hoeffding bounds for copulas. Further, W and M are copulas themselves. It may be noted that the Frécht-Hoeffding lower bound is not a copula in dimension $k > 2$. Copulas M, W and Π have important statistical interpretations (Nelson, 2006). Given a pair of continuous random variables (X_1, X_2), (i) copula of (X_1, X_2) is $M(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely increasing function of the other; (ii) copula of (X_1, X_2) is $W(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely decreasing function of the other and (iii) copula of (X_1, X_2) is $\Pi(u_1, u_2) = u_1 u_2$ if and only if X_1 and X_2 are independent.

Three famous measures of concordance Kendall's τ, Spearman's ρ_s and Gini's index γ could be expressed in terms of copulas (Schweizer and Wolff 1981) $\tau = 4 \int \int_{I^2} C(u_1, u_2)\, dC(u_1, u_2) - 1$, $\rho_s = 12 \int \int_{I^2} u_1 u_2\, dC(u_1, u_2) - 3$ and $\gamma = 2 \int \int_{I^2} (|u_1 + u_2 - 1| - |u_1 - u_2|)\, dC(u_1, u_2)$. It may however be noted that the linear correlation coefficient ρ cannot be expressed in terms of copula.

The tail dependence indexes of a multivariate distribution describe the amount of dependence in the upper right tail or lower left tail of the distribution and can be used to analyze the dependence among extremal random events. Tail dependence describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. Joe (1997) defines tail dependence: If a bivariate copula $C(u_1, u_2)$ is such that $\lambda_U := \lim_{u \to 1} \frac{[1 - 2u + C(u,u)]}{(1-u)}$ exists, then $C(u_1, u_2)$ has upper tail dependence for $\lambda_U \in (0,1]$ and no upper tail dependence for $\lambda_U = 0$. Similarly lower tail dependence in terms of copula is defined $\lambda_L := \lim_{u \to 0} \frac{C(u,u)}{u}$. Copula has lower tail dependence for $\lambda_L \in (0,1]$ and no lower tail dependence for $\lambda_L = 0$. This measure is extensively used in extreme value theory. It is the probability that one variable is extreme given that other is extreme. Tail measures are copula-based and copula is related to the full distribution via quantile transformations, i.e., $C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$ for all $u_1, u_2 \in (0,1)$ in the bivariate case.

Simulation

Simulation in statistics has a pivotal role in replicating and analysing data. Copulas can be applied in simulation and Monte Carlo studies. Johnson (1987) discusses methods to generate a sample from a given joint distribution. One such method is a recursive simulation using the univariate conditional distributions. The conditional distribution of U_i given first $i - 1$ components is $C_i(u_i | u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1} C_i(u_1, \ldots, u_i)}{\partial u_1 \ldots \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C_{i-1}(u_1, \ldots, u_{i-1})}{\partial u_1 \ldots \partial u_{i-1}}$. For $k \geq 2$, procedure is as follows: (i) Simulate a random number $u_1$from Uniform (0,1); (ii) Simulate value u_2 from the conditional distribution $C_2(u_2|u_1)$;(iii) Continue in this way; (iv) Simulate a value u_k from $C_k(u_k|u_1, \ldots, u_{k-1})$.

We list some important contributions in the area of copulas under the reference section.

Acknowledgments

Author wish to thank the referee for the critical review and useful suggestions on the earlier draft. This work was

supported by the author's discovery grant from the *Natural Sciences and Engineering Research Council of Canada (NSERC)* which is duly acknowledged.

About the Author

Dr. Pranesh Kumar is Professor of Statistics in the University of Northern British Columbia, Prince George, BC, Canada. He has held several international positions in the past: Professor and Head, Department of Statistics, University of Transkei, South Africa; Associate Professor, Bilkent University, Ankara, Turkey; Associate Professor, University of Dar-es-Salaam, Tanzania; Visiting Professor, University of Rome, Italy; Visiting Senior Researcher at the Memorial University of Newfoundland, Canada; Associate Professor, Indian Agricultural Statistics Research Institute, New Delhi. Dr. Kumar has published his research in many prestigious professional journals. He holds membership in several professional international societies including the International Statistical Institute. Dr. Kumar have membership of the editorial boards of the *Journal of Applied Mathematics and Analysis*, *Journal of Mathematics Research*, *Journal of the Indian Society of Agricultural Statistics and JNANABHA* which has reciprocity agreement with the American Mathematical Society. He has (co-)authored about 75 refereed papers.

Cross References

►Copulas
►Copulas in Finance
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation

References and Further Reading

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

Cuadras CM, Fortiana J, Rodríguez Lallena JA (2002) Distributions with given marginals and statistical modelling. Kluwer, Dodrecht

Embrechts P, McNeil A, Straumann D (1997) Correlation and dependence in risk management: properties and pitfalls. Risk 12(5):69–71

Fang K-T, Kotz S, Ng K-W (1987) Symmetric multivariate and related distributions. Chapman & Hall, London

Frank MJ (1979) On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19:194–226

Frécht M (1951) Sue les tableaux de corrélation dont les marges son données. Ann Univ Lyon Sect A 9:53–77

Genest C (1987) Frank's family of bivariate distributions. Biometrika 74:549–555

Genest C, Mackay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

Genest C, Rivest L (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043

Genest C, Ghoudi K, Rivest L (1995) A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82:543–552

Gumbel EJ (1960) Bivariate exponential distributions. J Am Stat Assoc 55:698–707

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

Hutchinson TP, Lai CD (1990) Continuous bivariate distributions emphasizing applications. Rumsby Scientific, Adelaide, South Australia

Joe H (1997) Multivariate models and dependent concepts. Chapman & Hall, New York

Johnson ME (1987) Multivariate statistical simulation. Wiley, New York

Kimeldorf G, Sampson AR (1978) Monotone dependence. Ann Stat 6:895–903

Marshall AW, Olkin I (1988) Families of multivariate distributions. J Am Stat Assoc 83:834–841

Nelsen R (2006) An introduction to copulas. Springer, New York

Nelsen RB, Quesada Molina JJ, Rodríguez Lallena JA, Úbeda Flores M (2001) Bounds on bivariate distribution functions with given margins and measures of association. Commun Stat Theory Meth 30:1155–1162

Scarsini M (1984) On measures of concordance. Stochastica 8:201–219

Schweizer B, Sklar A (1983) Probabilistic metric spaces. North Holland, New York

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

Sklar A (1959) Fonctions de répartition án dimensional et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Tjøstheim D (1996) Measures of dependence and tests of independence. Statistics 28:249–284

Cornish–Fisher Expansions

Vladimir V. Ulyanov
Professor
Lomonosov Moscow State University, Moscow, Russia

Introduction

In statistical inference it is of fundamental importance to obtain the sampling distribution of statistics. However, we often encounter situations where the exact distribution cannot be obtained in closed form, or even if it is obtained, it might be of little use because of its complexity. One practical way of getting around the problem is to provide reasonable approximations of the distribution function and its quantiles, along with extra information on their possible errors. This can be accomplished with the help of Edgeworth and Cornish–Fisher expansions. Recently, interest in Cornish–Fisher expansions has increased

because of intensive study of VaR (Value at Risk) models in financial mathematics and financial risk management (see Jaschke (2002)).

Expansion Formulas

Let X be a univariate random variable with a continuous distribution function F. For $\alpha : 0 < \alpha < 1$, there exists x such that $F(x) = \alpha$, which is called the (lower) $100\alpha\%$ *point* of F. If F is strictly increasing, the inverse function $F^{-1}(\cdot)$ is well defined and the $100\alpha\%$ point is uniquely determined. We also speak of "quantiles" without reference to particular values of α meaning the values given by $F^{-1}(\cdot)$.

Even in the general case, when $F(x)$ is not necessarily continuous nor is it strictly increasing, we can define its inverse function by the formula

$$F^{-1}(u) = \inf\{x; F(x) > u\}.$$

This is a right-continuous nondecreasing function defined on the interval (0,1) and $F(x_0) \geq u_0$ if $x_0 = F^{-1}(u_0)$.

Let $F_n(x)$ be a sequence of distribution functions and let each F_n admit the ▶Edgeworth expansion (EE) in the powers of $\epsilon = n^{-1/2}$ or n^{-1}:

$$\begin{aligned} F_n(x) &= G_{k,n}(x) + O(\epsilon^k) \quad \text{with} \\ G_{k,n}(x) &= G(x) + \{\epsilon a_1(x) + \cdot + \epsilon^{k-1} a_{k-1}(x)\} g(x), \end{aligned} \tag{1}$$

where $g(x)$ is the density function of the limiting distribution function $G(x)$. An important approach to the problem of approximating the quantiles of F_n is to use their asymptotic relation to those of G's. Let x and u be the corresponding quantiles of F_n and G, respectively. Then we have

$$F_n(x) = G(u). \tag{2}$$

Write $x(u)$ and $u(x)$ to denote the solutions of (2) for x in terms of u and u in terms of x, respectively [i.e. $u(x) = G^{-1}(F_n(x))$ and $x(u) = F_n^{-1}(G(u))$]. Then we can use the EE (1) to obtain formal solutions $x(u)$ and $u(x)$ in the form

$$x(u) = u + \epsilon b_1(u) + \epsilon^2 b_2(u) + \cdots \tag{3}$$

and

$$u(x) = x + \epsilon c_1(x) + \epsilon^2 c_2(x) + \cdots. \tag{4}$$

Cornish and Fisher (1937) and Fisher and Cornish (1946) obtained the first few terms of these expansions when G is the standard normal distribution function (i.e., $G = \Phi$). We call both (3) and (4) the *Cornish–Fisher expansions*, (CFE). Concerning CFE for random variables obeying limit laws from the family of Pearson distributions see Bol'shev (1963). Hill and Davis (1968) gave a general algorithm for obtaining each term of CFE when G is an analytic function:

Theorem 1 *Assume that the distribution function G is analytic. Then the following relation for x and u satisfying $F_n(x) = G(u)$ holds:*

$$x = u - \sum_{r=1}^{\infty} \frac{1}{r!} \left\{-[g(u)]^{-1} d_u\right\}^{r-1} \left[\{z_n(u)\}^r / g(u)\right], \tag{5}$$

where $d_u = d/du$ and $z_n(u) = F_n(u) - G(u)$.

A similar relation can be written for u as a function of x.

In many statistical applications, $F_n(x)$ is known to have an asymptotic expansion of the form

$$F_n(x) = G(x) + g(x)\left[n^{-a} p_1(x) + n^{-2a} p_2(x) + \cdots\right],$$

where $p_r(x)$ may be polynomials in x and $a = 1/2$ or 1. Then the formulas (5) can be written as

$$x = u - \sum_{r=1}^{\infty} \frac{1}{r!} d_{(r)} \{g_n(u)\}^r, \tag{6}$$

where $q_n(u) = n^{-a} p_1(u) + n^{-2a} p_2(u) + \cdots$,
$m(x) = -g'(x)/g(x)$,

$d_{(1)}$ = the identity operator,
$d_{(r)} = \{m(u) - d_u\}\{2m(u) - d_u\}\cdots\{(r-1)m(u) - d_u\}$,
$r = 2, 3, \ldots$

The rth term in (6) is of order $O(n^{-ra})$.

It is a tedious process to rewrite (6) in the form of (3) and to express the adjustment terms $b_k(u)$ directly in terms of the cumulants (see Hill and Davis (1968)). Lee and Lin developed a recurrence formula for $b_k(u)$, which is implemented in the algorithm AS269 (see Lee and Lin (1992, 1993)).

Usually the CFE are applied in the following form with $k = 1, 2$, or 3:

$$x_k(u) = u + \sum_{j=1}^{k-1} \epsilon^j b_j(u) + O(\epsilon^k), \tag{7}$$

In order to find the explicit expressions for $b_1(u)$ and $b_2(u)$ we substitute (7) with $k = 2$ to (1) and using (2) we have

$$\begin{aligned} F_n(x) &= F_n(u + \epsilon b_1 + \epsilon^2 b_2 + \cdots) \\ &= G(u + \epsilon b_1 + \epsilon^2 b_2) + g(u + \epsilon b_1 + \epsilon^2 b_2) \\ &\quad \times \{\epsilon a_1(u + \epsilon b_1) + \epsilon^2 a_2(u)\} + O(\epsilon^2). \end{aligned}$$

By Taylor's expansions for G, g, and a_1, we obtain

$$\begin{aligned} F_n(x) =& G(u) + \epsilon g(u)\{b_1 + a_1(u)\} \\ &+ \epsilon^2 \left[g(u) b_2 + \frac{1}{2} g'(u) b_1^2 + g(u) a_1'(u) b_1 \right. \\ &\left. + g(u) a_2(u) + g'(u) b_1 a_1(u)\right] + O(\epsilon^3), \end{aligned}$$

which should be $G(u)$. Therefore,

$$b_1 = -a_1(u),$$
$$b_2 = \frac{1}{2}\{g'(u)/g(u)\}a_1^2(u) - a_2(u) + a_1'(u)a_1(u).$$

An application of general formulas (6) in the case of normal limit distribution see the entry ►Edgeworth Expansion.

Suppose that

$$F_n(x) = G_f(x) + \frac{f\gamma}{4n}[G_f(x) + 2G_{f+2}(x) + G_{f+4}(x)] + \frac{f}{960n^2}\sum_{j=0}^{4}(-1)^j c_j G_{f+2j}(x) + o(n^{-2}),$$

where $G_f(x) = \Pr\{\chi_f^2 \le x\}$; that is, the distribution function of the ►chi-square distribution with f degrees of freedom, γ is a constant, and c_j are constants such that $\sum_{j=0}^{4}(-1)^j c_j = 0$. Then $G(u) = G_f(u)$,

$$g(u) = g_f(u) = \left[\Gamma(f/2)2^{f/2}\right]^{-1} u^{f/2-1}\exp(-u/2),$$
$$m(u) = -g_f'(u)/g_f(u) = \frac{1}{2} - \frac{1}{u}\left(\frac{f}{2} - 1\right).$$

Thus, we can write

$$q_n(u) = -\frac{u(u-f-2)}{2n(f+2)} - \frac{u}{48n^2(f+2)(f+4)(f+6)}[c_4u^3 + (c_4 - c_3)(f+6)u^2 + (c_1 - c_0)(f+4)(f+6)u - c_0(f+2)(f+4)(f+6)] + o(n^{-2}).$$

Therefore, we obtain

$$x = u - q_n(u) - [\gamma^2 u(u-f-2)/\{16n^2(f+2)^2\}] \times \{u^2 - 2(f+4)u + (f+2)^2\} + o(n^{-2}).$$

The upper and lower bounds for the quantiles $x = x(u)$ and $u = u(x)$, satisfying the equation (2), i.e.

$$\underline{x}_n(u) \ge x(u) \ge \bar{x}_n(u), \quad \underline{u}_n(x) \ge u(x) \ge \bar{u}_n(x)$$

were obtained for some special distributions by Wallace (1959).

Validity of Cornish–Fisher Expansions

In applications, the CFE are usually used in the form (7). It is necessary to remember that the approximations for α-quantiles provided by the CFE

(i) become less and less reliable for $\alpha \to 0$ and $\alpha \to 1$;
(ii) do not necessarily improve (converge) for a fixed F_n and increasing order of approximation k.

Let x_α and x_α^* be the upper $100\alpha\%$ points of F_n and $G_{k,n}$ from (1), respectively; that is, they satisfy

$$F_n(x_\alpha) = G_{k,n}\left(x_\alpha^*\right) = 1 - \alpha.$$

The approximate quantile x_α^* based on the Edgeworth expansion is available in numerical form but cannot be expressed in explicit form. Suppose that the remainder term, $R_{k,n}(x) = F_n(x) - G_{k,n}(x)$, is such that

$$|R_{k,n}| \le \epsilon^n C_k.$$

Then

$$\left|F_n\left(x_\alpha^*\right) - (1-\alpha)\right| = \left|F_n\left(x_\alpha^*\right) - G_{k,n}\left(x_\alpha^*\right)\right| \le \epsilon^n C_k.$$

This gives an error bound for the absolute differences between the probabilities based on the true quantiles and their approximations.

The other validity of the CFE was obtained by considering the distribution function $\tilde{F}_{k,n}$ of

$$\tilde{X} = U + \sum_{j=1}^{k-1}\epsilon^j b_j(U),$$

where U is the standard normal variable. Takeuchi and Takemura (1988) showed that if $|F_n(x) - G_{k,n}(x)| = o(\epsilon^{k-1})$, then $|F_n(x) - \tilde{F}_{k,n}| = o(\epsilon^{k-1})$.

Function of Sample Mean

Usually the conditions that are sufficient for validity of EE are sufficient as well for validity of CFE. Under the conditions of section "►Function of Sample Means" in the entry ►Edgeworth Expansion and in its notation we have (see Hall (1992)):

$$\sup_{\epsilon<\alpha<1-\epsilon}\left|x_\alpha - u_\alpha - \sum_{j=1}^{k-2}\frac{b_j(u_\alpha)}{n^{j/2}}\right| = o\left(\frac{1}{n^{(k-2)/2}}\right),$$

where $x_\alpha = \inf\{x; \Pr(\sqrt{n}H(\bar{\mathbf{Y}})/\sigma \le x) > \alpha\}$, $u_\alpha = \Phi^{-1}(\alpha)$, ϵ is any constant in (0,1/2) and b_j's are polynomials depending on Q_j's.

Error Bounds

It is possible to get error bounds for approximation given by the CFE provided we have error bounds for EE. For simplicity, we give error bounds for the first-order CFE (see Chap. 5 in Fujikoshi et al. (2010)):

Theorem 2 *Suppose that for the distribution function of U we have*

$$F(x) \equiv Pr\{U \le x\} = G(x) + R_1(x),$$

where for remainder term $R_1(x)$ there exists a constant c_1 such that

$$|R_1(x)| \le d_1 \equiv c_1\epsilon.$$

Let x_α and u_α be the upper $100\alpha\%$ points of F and G, respectively; that is,

$$Pf\{U \le x_\alpha\} = G(u_\alpha) = 1 - \alpha.$$

Then, for any α such that $1 > \alpha > d_1$ and $1 > \alpha + d_1$:

1. $u_{\alpha+d_1} \le x_\alpha \le u_{\alpha-d_1}$.
2. $|x_\alpha - u_\alpha| \le d_1 / g(u_{(1)})$, *where*

$$g(u_1) = \min_{u \in [u_{\alpha+d_1}, u_{\alpha-d_1}]} g(u).$$

About the Author

DSc Vladimir V. Ulyanov is Professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics at Lomonosov Moscow State University. He has received the State Prize of the USSR for Young Scientists (1987), Alexander von Humboldt Research Fellowship, Germany (1991–1993), JSPS Research Fellowship, Japan (1999, 2004). He was visiting Professor/Researcher at Bielefeld University, Germany, University of Leiden, Université de Paris V, University of Hong Kong, Institute of Statistical Mathematics in Tokyo, National University of Singapore, the University of Melbourne etc. He is a member of the Bernoulli Society. Professor Ulyanov is the author of more than 50 journal articles and a book *Multivariate Statistics: High-Dimensional and Large-Sample Approximations* (with Y. Fujikoshi and R. Shimizu, John Wiley and Sons, 2010).

Cross References

►Edgeworth Expansion
►Multivariate Statistical Distributions

References and Further Reading

Bol'shev LN (1963) Asymptotically Pearson transformations. Theor Probab Appl 8:121–146
Cornish EA Fisher RA (1937) Moments and cumulants in the specification of distributions. Rev Inst Int Stat 4:307–320
Fisher RA, Cornish EA (1946) The percentile points of distributions having known cumulants. J Am Stat Assoc 80:915–922
Fujikoshi Y, Ulyanov VV, Shimizu R (2010) Multivariate statistics : high-dimensional and large-sample approximations. Wiley Series in Probability and Statistics. Wiley, Hoboken
Hall P (1992) The bootstrap and Edgeworth expansion. Springer, New York
Hill GW, Davis AW (1968) Generalized asymptotic expansions of Cornish–Fisher type. Ann Math Stat 39:1264–1273
Jaschke S (2002) The Cornish–Fisher-expansion in the context of delta-gamma-normal approximations. J Risk 4(4):33–52
Lee YS, Lin TK (1992) Higher-order Cornish–Fisher expansion. Appl Stat 41:233–240
Lee YS, Lin TK (1993) Correction to algorithm AS269: higher-order Cornish–Fisher expansion. Appl Stat 42:268–269
Takeuchi K, Takemura A (1988) Some results on univariate and multivariate Cornish–Fisher expansion: algebraic properties and validity, Sankhyā A 50:111–136
Wallace DL (1959) Bounds on normal approximations to Student's and the chisquare distributions. Ann Math Stat 30:1121–1130

Correlation Coefficient

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

Introduction

A covariance term loosely aims at capturing some essence of *joint dependence* between two random variables. A correlation coefficient is nothing more than an appropriately scaled version of the covariance.

Section "Population Correlation Coefficient" introduces the concepts of a covariance and the population correlation coefficient. Section "Correlation Coefficient and Independence" highlights some connections between the correlation coefficient, independence, and dependence.

Section "A Sample Correlation Coefficient" summarizes the notion of a sample correlation coefficient and its distribution, both exact and large-sample approximation, due to Fisher (1915; 1925). Section "Partial Correlations" gives a brief summary of the concept of partial correlation coefficients.

Population Correlation Coefficient

A covariance term tries to capture a sense of *joint dependence* between two real valued random variables. A correlation coefficient, however, is an appropriately scaled version of a covariance.

Definition 1 *The covariance between two random variables X_1 and X_2, denoted by $Cov(X_1, X_2)$, is defined as*

$$Cov(X_1, X_2) = E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right]$$

$$\text{or equivalently } E\left[X_1 X_2\right] - \mu_1 \mu_2,$$

where $\mu_i = E(X_i)$, $i = 1, 2$ and $E\left[X_1 X_2\right]$, μ_1, μ_2 are assumed finite.

Definition 2 *The correlation coefficient between two random variables X_1 and X_2, denoted by ρ_{X_1,X_2}, is defined as*

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2},$$

whenever one has $0 < \sigma_1^2 = V(X_1) < \infty$ *and* $0 < \sigma_2^2 = V(X_2) < \infty$.

One may note that we do not explicitly assume $-\infty < Cov(X_1, X_2) < \infty$. In view of the assumption $0 < \sigma_1^2, \sigma_2^2 < \infty$, one can indeed claim the finiteness of $Cov(X_1, X_2)$ by appealing to Cauchy–Schwartz inequality. It should also be clear that

$$Cov(X_1, X_2) = Cov(X_2, X_1) \text{ and } Cov(X_1, X_1) = V(X_1),$$

as long as those terms are finite.

Two random variables X_1, X_2 are respectively called negatively correlated, uncorrelated, or positively correlated if and only if ρ_{X_1,X_2} *is neg*ative, zero or positive.

Theorem 1 *Consider random variables* X_1 *and* X_2 *and assume that* $0 < V(X_1), V(X_2) < \infty$. *Then, we have the following results:*

1. *Let* $Y_i = c_i + d_i X_i$ *w.p.*1 *where* $-\infty < c_i < \infty$ *and* $0 < d_i < \infty$ *are fixed numbers,* $i = 1, 2$. *Then,* $\rho_{Y_1,Y_2} = \rho_{X_1,X_2}$.
2. $|\rho_{X_1,X_2}| \le 1$, *the equality holds if and only if* $X_1 = a + bX_2$ *w.p.*1 *for some real numbers a and b.*

More details can be found from Mukhopadhyay (2000, Sect. 3.4).

Correlation Coefficient and Independence

If ρ_{X_1,X_2} is finite and X_1, X_2 are independent, then $\rho_{X_1,X_2} = 0$. Its converse is not necessarily true. In general, $\rho_{X_1,X_2} = 0$ may not imply independence between X_1, X_2. An example follows.

Example 1 Let X_1 be $N(0,1)$ and $X_2 = X_1^2$. Then, $Cov(X_1, X_2) = 0$, and surely $0 < V(X_1), V(X_2) < \infty$, so that $\rho_{X_1,X_2} = 0$. But, X_1 and X_2 are dependent variables. More details can be found from Mukhopadhyay (2000, Sect. 3.7). The recent article of Mukhopadhyay (2010) is relevant here.

Now, we state an important result which clarifies the role of zero correlation in a bivariate normal distribution.

Theorem 2 *Suppose that* (X_1, X_2) *has the* $N_2\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$ *distribution where* $-\infty < \mu_1, \mu_2 < \infty$, $0 < \sigma_1, \sigma_2 < \infty$ *and* $-1 < \rho(= \rho_{X_1,X_2}) < 1$. *Then,* X_1 *and* X_2 *are independent if and only if* $\rho = 0$.

Example 2 A zero correlation coefficient implies *independence* not merely in the case of a bivariate normal distribution. Consider random variables X_1 and X_2 whose joint probability distribution puts mass only at four points $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. Now, if $Cov(X_1, X_2) = 0$, then X_1 and X_2 must be independent.

A Sample Correlation Coefficient

We focus on a bivariate normal distribution. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid $N_2\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$ where $-\infty < \mu_1, \mu_2 < \infty, 0 < \sigma_1^2, \sigma_2^2 < \infty$ and $-1 < \rho < 1, n \ge 2$. Let us denote

$$\overline{X} = n^{-1}\Sigma_{i=1}^n X_i \quad \overline{Y} = n^{-1}\Sigma_{i=1}^n Y_i$$

$$S_1^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2 \quad S_2^2 = (n-1)^{-1}\Sigma_{i=1}^n (Y_i - \overline{Y})^2$$

$$S_{12} = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \quad r = S_{12}/(S_1 S_2).$$

Here, r is called the *Pearson* (or *sample*) *correlation coefficient*. This r is customarily used to estimate ρ.

The probability distribution of r is complicated, particularly when $\rho \ne 0$. But, even without explicitly writing the pdf of r, it is simple enough to see that the distribution of r can not involve μ_1, μ_2, σ_1^2 and σ_2^2.

Francis Galton introduced a numerical measure, r, which he termed "reversion" in a lecture at the Royal Statistical Society on February 9, 1877 and later called "regression." The term "cor-relation" or "correlation" probably appeared first in Galton's paper to the Royal Statistical Society on December 5, 1888. At that time, "correlation" was defined in terms of deviations from the median instead of the mean. Karl Pearson gave the definition and calculation of correlation r in 1897. In 1898, Pearson and his collaborators discovered that the standard deviation of r happened to be $(1-\rho^2)/\sqrt{n}$ when n was large. "Student" derived the "probable error of a correlation coefficient" in 1908. Soper (1913) gave large-sample approximations for the mean and variance of r which performed better than those proposed earlier by Pearson. Refer to DasGupta (1980) for more historical details.

The unsolved problem of finding the exact pdf of r for normal variates came to R. A. Fisher's attention via Soper's 1913 paper. The pdf of r was published in the year 1915 by Fisher for all values of $\rho \in (-1,1)$. Fisher, at the age of 25, brilliantly exploited the n-dimensional geometry to come up with the solution, reputedly within one week. Fisher's genius immediately came into limelight. Following the publication of Fisher's results, however, Karl Pearson set up a major cooperative study of the correlation. One will notice that in the team formed for this cooperative project (Soper et al. 1917) studying the distribution of the sample correlation coefficient, the young Fisher was not included. This happened in spite of the fact that Fisher was right there and he already earned quite some fame. Fisher felt hurt as he was left out of this project. One thing led to another. RA. Fisher and Karl Pearson continued criticizing each other even more as each held on to his own philosophical stand.

We will merely state the pdf of r when $\rho = 0$. This pdf is given by

$$f(r) = c\,(1-r^2)^{\frac{1}{2}(n-4)} \text{ for } -1 < r < 1,$$

where $c = \Gamma\left(\frac{1}{2}(n-1)\right)\left\{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}(n-2)\right)\right\}^{-1}$ for $n \geq 3$. Using a simple transformation technique, one can easily derive the following result:

$r(n-2)^{1/2}(1-r^2)^{-1/2}$ has the Student's t distribution with $(n-2)$ degrees of freedom when $\rho = 0$.

Fisher's geometric approach (1915) also included the exact pdf of r in the form of an infinite power series for all values of $\rho \neq 0$. One may also look at Rao (1973, pp. 206–209) for a non-geometric approach.

Large-Sample Distribution

But, now suppose that one wishes to construct an *approximate* $100(1-\alpha)\%$ confidence interval for ρ, $0 < \alpha < 1$. In this case, one needs to work with the *non-null* distribution of r. We mentioned earlier that the *exact* distribution of r, when $\rho \neq 0$, was found with an ingenious geometric technique by Fisher (1915). That exact distribution being very complicated, Fisher (1915) proceeded to derive the following asymptotic distribution when $\rho \neq 0$:

$$\sqrt{n}(r-\rho) \xrightarrow{\mathcal{L}} N\left(0, (1-\rho^2)^2\right) \text{ as } n \to \infty.$$

For a proof, one may look at Sen and Singer (1993, pp. 134–136) among other sources.

One should realize that a variance stabilizing transformation may be useful here. We may invoke Mann-Wald Theorem (see Mukhopadhyay 2000, pp. 261–262) by requiring a suitable function $g(.)$ such that the asymptotic variance of $\sqrt{n}\,[g(r) - g(\rho)]$ becomes free from ρ. That is, we want to have:

$$g'(\rho)(1-\rho^2) = k, \text{ a constant.}$$

So, $g(\rho) = k\int \frac{1}{(1-\rho^2)}d\rho$. Hence, we rewrite

$$g(\rho) = \tfrac{1}{2}k\int\left\{\tfrac{1}{1-\rho} + \tfrac{1}{1+\rho}\right\}d\rho = \tfrac{1}{2}k\log\left\{\tfrac{1+\rho}{1-\rho}\right\} + \text{constant.}$$

It is clear that we should look at the transformations:

$$U = \frac{1}{2}\log\left\{\frac{1+r}{1-r}\right\} \text{ and } \xi = \frac{1}{2}\log\left\{\frac{1+\rho}{1-\rho}\right\},$$

and consider the asymptotic distribution of $\sqrt{n}[U-\xi]$. Now, we can claim that

$$\sqrt{n}[U-\xi] \xrightarrow{\mathcal{L}} N(0,1) \text{ as } n \to \infty,$$

since with $g(\rho) = \frac{1}{2}\log\left\{\frac{1+\rho}{1-\rho}\right\}$, one has $g'(\rho) = \frac{1}{1-\rho^2}$. That is, for large n, we should consider the following pivot:

$\sqrt{n}[U-\xi]$, which is approximately $N(0,1)$ for large n.

These transformations can be *equivalently* stated as

$$U = \tanh^{-1}(r) \text{ and } \xi = \tanh^{-1}(\rho),$$

which are referred to as Fisher's Z transformations introduced in 1925.

Fisher obtained the first four moments of $\tanh^{-1}(r)$ which were later updated by Gayen (1951). It turns out that the variance of $\tanh^{-1}(r)$ is approximated better by $\frac{1}{n-3}$ rather than $\frac{1}{n}$ when n is moderately large. Hence, in many applications, one uses an alternate pivot (for $n > 3$):

$$\sqrt{n-3}\left[\tanh^{-1}(r) - \tanh^{-1}(\rho)\right], \text{ which is approximately } N(0,1),$$

for large n whatever be ρ, $-1 < \rho < 1$.

For large n, one customarily uses Fisher's Z transformations to come up with an *approximate* $100(1-\alpha)\%$ confidence interval for ρ. Also, to test a null hypothesis $H_0 : \rho = \rho_0$, for large n, one uses the test statistic

$$Z_{calc} = \sqrt{n-3}\left[\tanh^{-1}(r) - \tanh^{-1}(\rho_0)\right]$$

and comes up with an *approximate* level α test against an appropriate alternative hypothesis. These are customarily used in all areas of statistical science whether the parent population is bivariate normal or not.

Partial Correlations

Suppose that in general $\mathbf{X} = (X_1, \ldots, X_p)$ has a p-dimensional probability distribution with all pairwise correlations finite. Now, ρ_{X_i,X_j} will simply denote the correlation coefficient between X_i, X_j based on their joint bivariate distribution derived from the distribution of $\mathbf{X}$, for any $i \neq j = 1, \ldots, p$.

Next, $\rho_{X_i,X_j.X_k}$ is simply the correlation coefficient between X_i, X_j based on their joint bivariate conditional distribution given X_k that is derived from the distribution of $\mathbf{X}$, for any $i \neq j \neq k = 1, \ldots, p$.

Similarly, $\rho_{X_i,X_j.X_k,X_l}$ is simply the correlation coefficient between the pair of random variables X_i, X_j based on their joint bivariate conditional distribution given X_k, X_l derived from the distribution of $\mathbf{X}$, for any $i \neq j \neq k \neq l = 1, \ldots, p$. Clearly, one may continue further like this.

Such correlation coefficients $\rho_{X_i,X_j.X_k}$, $\rho_{X_i,X_j.X_k,X_l}$ are referred to as partial correlation coefficients. Partial correlation coefficients have important implications in multiple linear regression analysis. One may refer to Ravishanker and Dey (2002, pp. 160–164) among other sources.

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Autocorrelation in Regression
►Intraclass Correlation Coefficient
►Kendall's Tau
►Measures of Dependence
►Rank Transformations
►Spurious Correlation
►Tests of Independence
►Weighted Correlation

References and Further Reading

DasGupta S (1980) Distributions of the correlation coefficient. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 9–16
Fisher RA (1915) Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. Biometrika 10:507–521
Fisher RA (1925) Theory of statistical estimation. Proc Camb Phil Soc 22:700–725
Gayen AK (1951) The frequency distribution of the product moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika 38:219–247
Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York
Mukhopadhyay N (2010) When finiteness matters: Counterexamples to notions of covariance, correlation, and independence. The Amer. Statistician, in press
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Ravishanker N, Dey DK (2002) A first course in linear model theory. Chapman & Hall/CRC Press, Boca Raton
Sen PK, Singer JO (1993) Large sample methods in statistics. Chapman & Hall, New York
Soper HE (1913) On the probable error of the correlation coefficient to a second approximation. Biometrika 9:91–115
Soper HE, Young AW, Cave BM, Lee A, Pearson K (1917) On the distribution of the correlation coefficient in small samples. A cooperative study. Biometrika 11:328–413
"Student" (Gosset WS) (1908) The probable error of a mean. Biometrika 6:1–25

Correspondence Analysis

Jörg Blasius
Professor
University of Bonn, Bonn, Germany

Correspondence analysis (CA) has been developed in the 1960s in France by Jean-Paul Benzécri and his collaborators; it is the central part of the French "Analyse des Données," or in English, geometric data analysis (cf. Benzécri et al. 1973; Greenacre 1984, 2007; Lebart et al. 1984; Le Roux and Rouanet 2004). The method can be applied to any data table with nonnegative entries. The main objective of CA is to display rows and columns of data tables in two-dimensional spaces, called "maps." This kind of data description via visualization reflects a way of thinking that is typical for the social sciences in France, especially associated with the name of Pierre Bourdieu, and of many statisticians in the 1970s and 1980s in France, who at that time published almost only in French. The philosophy behind their work can be expressed by the famous quotation of Jean-Paul Benzécri who pointed out that "The model must follow the data, and not the other way around." Instead of limiting the data to restrictive and subjectively formulated statistical models, they show the importance of the data and of the features in the data themselves. The discussion outside of France started with the textbooks by Greenacre (1984) and Lebart et al. (1984).

CA translates deviations from the independence model in a contingency table into distances as the following brief introduction shows. In the simple case, there is a two-way table $\mathbf{N}$ with I rows and J columns. In cases where the data are from survey research, the cells n_{ij} of $\mathbf{N}$ contain the frequencies of a bivariate cross-tabulation of two variables, with $\sum_{ij} n_{ij} = n$. Dividing n_{ij} by the sample size n provides the percentages of the total p_{ij}, or, for the entire table, with the $(I \times J)$ correspondence matrix $\mathbf{P}$. Thereby, $\mathbf{r} = \mathbf{P1}$ is the vector of the "row masses," or the "average column profile" with elements $r_i = n_{i+}/n$, and $\mathbf{c} = \mathbf{P}^T\mathbf{1}$ is the vector of "column masses" or the "average row profile" with elements $c_j = n_{+j}/n$; $\mathbf{D}_r$ and $\mathbf{D}_c$ are the diagonal matrices of the row and column masses, respectively.

The matrix of row profiles can be defined as the rows of the correspondence matrix $\mathbf{P}$ divided by their respective row masses, $D_r^{-1}P$; for the matrix of columns profiles yields PD_c^{-1}. As a measure of similarity between two row profiles (or between two column profiles, respectively), a weighted Euclidian or chi-square distance in the metric D_r^{-1} (or, D_c^{-1}, respectively) is used. For chi-square calculations, the weighted deviations from independence over all cells of the contingency table are used. For each cell, the unweighted deviation of the observed from the expected value can be calculated by $(n_{ij} - \hat{n}_{ij})$, with $\hat{n}_{ij} = (n_{i+} \times n_{+j})/n$. Dividing $(n_{ij} - \hat{n}_{ij})$ by n provides with $(p_{ij} - r_i c_j)$, or, in matrix notation, $(\mathbf{P} - \mathbf{rc}^T)$, with the unweighted deviations from the independence model for the entire table.

To fulfill the chi-square statistic, this matrix is weighted by the product of the square root of the row and column masses to give the standardized residuals $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$, or in matrix notation, the $(I \times J)$ matrix of

standardized residuals $S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$. The similarity to chi-square analysis and total inertia as a measure for the variation in the data table, which is defined as $\sum_{ij} s_{ij}^2 = \frac{\chi^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$, becomes apparent. Applying singular value decomposition to $\mathbf{S}$ results in $SVD(\mathbf{S}) = U\mathbf{\Gamma}V^T$, where $\mathbf{\Gamma}$ is a diagonal matrix with singular values in descending order $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_S > 0$, with S = rank of $\mathbf{S}$. The columns from $\mathbf{U}$ are the left singular vectors, the columns from $\mathbf{V}$ are the right singular vectors, with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

The connection between SVD as used in CA and the well-known canonical decomposition is shown by $S^TS = V\mathbf{\Gamma}\mathbf{U}^TU\mathbf{\Gamma}\mathbf{V}^T = V\mathbf{\Gamma}^2\mathbf{V}^T = V\mathbf{\Lambda}V^T$, with $SS^T = U\mathbf{\Gamma}\mathbf{V}^TV\mathbf{\Gamma}\mathbf{U}^T = U\mathbf{\Gamma}^2\mathbf{U}^T = U\mathbf{\Lambda}U^T$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_S > 0$; $\chi^2/n = \sum_s \lambda_s$ = total inertia, since trace $(\mathbf{SS}^T)$ = trace $(\mathbf{S}^T\mathbf{S})$ = trace $(\mathbf{\Gamma}^2)$ = trace $(\mathbf{\Lambda})$.

As in ▶principal component analysis (PCA), the first axis is chosen to explain the maximum variation in the data; the second axis captures the maximum of the remaining variation, and so on. Again, analogous to PCA, it is possible to interpret the variable categories in relation to the axes, which can be considered the latent variables. And furthermore, as in PCA and other data reduction methods, only the s major components are used for interpretation. The number of interpretable dimensions depends on criteria such as the eigenvalue criteria, theory (how many latent variables can be substantively interpreted), or the scree test (for more details, see Blasius 1994).

For the graphical representation, we use $F = D_r^{-1/2}U\mathbf{\Gamma}$ providing the principal coordinates of the rows, and $G = D_c^{-1/2}V\mathbf{\Gamma}$ providing the principal coordinates of the columns (for further details see Greenacre 1984, 2007). The maps drawn on the basis of principal coordinates are called "symmetric maps." In the full space, the distances between the rows and the distances between the columns can be interpreted as Euclidian distances, whereas the distances between the rows and the columns are not defined.

As in PCA, the input data can be factorized. Understanding correspondence analysis as a model (see, e.g., van der Heijden et al. 1989, 1994), the row and column coordinates can be used for recomputing the input data. Adding the latent variables successively models the deviations from independency. This is similar to the loglinear model and other modeling approaches such as the latent class model or the log-multiplicative model (see, e.g., van der Heijden et al. 1989, 1994; Goodman 1991). In loglinear analysis, for example, these deviations are modeled by using higher-order interaction effects; in correspondence analysis latent variables are used. For any cell yields $n_{ij} = nr_ic_j\left(1 + \sum_{s=1}^{S} f_{is}g_{js}/\gamma_s\right)$, or $p_{ij} = r_ic_j\left(1 + \sum_{s=1}^{S} f_{is}g_{js}/\gamma_s\right)$, and in matrix notation $P = rc^T + D_rF\mathbf{\Gamma}^{-1}G^TD_c$. The left part of the equation reflects the independence model and the right part, the modeling from independency by including the S factors in successive order. Including all factors in the model fully reconstructs the original data table $\mathbf{N}$.

The interpretation of CA is similar to the one of PCA, both methods provide eigenvalues and their explained variances, factor loadings, and factor values. While PCA is restricted to metric data, CA can be applied to any kind of data table with nonnegative entries, among others, to indicator and Burt matrices – in these two cases the method is called multiple correspondence analysis (MCA).

Whereas simple correspondence analysis is applied to a single contingency table or to a stacked table, MCA uses the same algorithm to an indicator or a Burt matrix. In the case of survey research, input data to simple CA is usually a matrix of raw frequencies of one or more contingency tables. In this context, there is usually one variable to be described, for example, preference for a political party, and one or more describing variables, for example, educational level and other sociodemographic indicators such as age groups, gender, and income groups. The number of variables can be quite high, apart from theoretical considerations there is no real limitation by the method. In the given case, each of the describing variables is cross-tabulated with the variable to be described in order to investigate the importance of this association. Concatenating, or stacking the tables before applying CA allows to visualize and interpret several relationships of "preferred political party" with the sociodemographic indicators in the same map.

Applying CA to the indicator matrix $\mathbf{Z}$ (=MCA), the table of input data has as many rows as there are respondents, and as many columns as there are response alternatives in all variables included in the analysis. A "1" in a given row indicates the respondent who chose that specific response category; otherwise there is a "0" for "specific response category not chosen." Considering all categories of all variables provides row sums that are constant and equal to the number of variables, the column sums reflect the marginals. An alternative to the indicator matrix as input to MCA is the Burt matrix $\mathbf{B}$. This matrix can either be generated by cross-tabulating all variables by all variables, including the cross-tabulations of the variables by themselves, and stacking them row- and column-wise. Further, $\mathbf{B}$ can be computed by multiplying the transposed indicator matrix by itself, that is $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$. The solutions from $\mathbf{Z}$ can be directly converted to those of $\mathbf{B}$ by rescaling the solution; for example, the squared eigenvalues of $\mathbf{Z}$

are equal to those of **B**. As it is true for PCA, MCA contains all first-order interaction effects, the method can be understood as a generalization of PCA to categorical data.

Taking a two-way contingency table with $I = 7$ rows, $J = 5$ columns, and $n = 500$ cases as an example, input data of the simple CA would be the frequencies of the (7×5) cross-table. Turning to MCA, input data is an indicator matrix with 500 rows (the number of cases) and 12 columns (the number of variable categories). MCA is also known under the names "homogeneity analysis" (see Gifi 1990; Heiser and Meulman 1994), "dual scaling" (Nishisato 1980, 2007), and "quantification of qualitative data III" (Hayashi 1954); CA procedures are available in all major statistic packages as well as in *R* (Greenacre and Nenadić 2006). For details regarding the history of CA and related methods, we refer to Nishisato (2007, Chap. 3).

CA employs the concept of inertia: the farther the categories are from the centroid along a given axis (squared distances) and the higher their masses (their marginals), the more the categories determine the geometric orientation of that axis. In the graphical solution, the locations of all variable categories can be compared to each other (except in simple CA and using symmetric maps, in this case the distances between rows and columns are not defined), short distances imply high similarities and long distances imply high dissimilarities. For all dimensions, CA supplies principal inertias that can be interpreted as canonical correlation coefficients (they are the singular values of the solution, i.e., the square roots of the eigenvalues), correlation coefficients between the item categories and the latent variables as well as scores for all item categories and all respondents.

There are several extensions of simple CA and MCA. With respect to the Burt matrix **B**, it is apparent that most of the variation in this super matrix is caused by the main diagonal blocks. These sub-matrices contain the cross-tabulations of the variables by themselves; the main diagonal elements of them contain the marginals of the variables while their off-diagonal elements are equal to zero. Excluding this variation in an iterative procedure and visualizing the variation of the off-diagonal blocks of **B** only is the objective of joint correspondence analysis. The aim of subset correspondence analysis is to concentrate on some response categories only, while excluding others from the solution. For example, applying subset MCA to a set of variables, the structure of non-substantive responses ("don't know," "no answer") can be analyzed separately, or these responses can be excluded from the solution while concentrating on the substantive responses. Variables can also be included in the model as supplementary or passive ones; in this case they do not have any impact on the geometric orientation of the axes but they can be interpreted together with the active variables. CA can not only be applied to single and stacked contingency tables or to indicator matrix, it can also be used to analyze rank and metric data, multiple responses, or squared tables. The statistical background and examples of these kinds of data can be found in the textbook of Greenacre (2007) as well as in the readers of Greenacre and Blasius (1994, 2006), and Blasius and Greenacre (1998).

About the Author

Jörg Blasius is the President of RC33 (Research Committee of Logic and Methodology in Sociology) at ISA (International Sociological Association) (2006–2010). Together with Michael Greenacre (Barcelona) he founded CARME (Correspondence Analysis and Related Methods Network) and edited three books on Correspondence Analysis, in June 2009 they had a special issue on this topic in *Computational Statistics and Data Analysis* (further coeditors: Patrick Groenen and Michel van de Velden).

Cross References

►Data Analysis
►Distance Measures
►Multivariate Data Analysis: An Overview

References and Further Reading

Benzécri JP et al (1973) L'analyse des Données. L'analyse des Correspondances. Dunod, Paris

Blasius J (1994) Correspondence analysis in social science research. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 23–52

Blasius J, Greenacre M (eds) (1998) Visualization of categorical data. Academic, London

Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester

Goodman LA (1991) Measures, models, and graphical display in the analysis of cross-classified data (with discussion). J Am Stat Assoc 86:1085–1138

Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic, London

Greenacre MJ (2007) Correspondence analysis in practice. Chapman & Hall, Boca Raton

Greenacre MJ, Blasius J (eds) (1994) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London

Greenacre MJ, Blasius J (eds) (2006) Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton

Greenacre MJ, Oleg Nenadić (2006) Computation of multiple correspondence analysis, with code in R. In: Greenacre M, Blasius J (eds) Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton, pp 523–551

Hayashi C (1954) Multidimensional quantification – with the applications to the analysis of social phenomena. Ann Inst Stat Math 5:231–245

Heiser WJ, Meulman JJ (1994) Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 179–209
Lebart L, Morineau A, Warwick KM (1984) Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley, New York
Le Roux B, Rouanet H (2004) Geometric data analysis. North Holland, Amsterdam
Nishisato S (1980) Analysis of categorical data: dual scaling and its applications. University of Toronto Press, Toronto
Nishisato S (2007) Multidimensional nonlinear descriptive analysis. Chapman & Hall, Boca Raton
Van der Heijden PGM, de Falguerolles A, de Leeuw J (1989) A combined approach to contingency table analysis using correspondence analysis and loglinear analysis. Appl Stat 38:249–292
Van der Heijden PGM, Mooijaart A, Takane Y (1994) Correspondence analysis and contingency table models. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 79–111

C_p Statistic

Colin Mallows
Basking Ridge, Avayalabs, NJ, USA

The C_p statistic was invented by C. Mallows in 1963. It facilitates the comparison of many subset-regression models, by giving for each model an unbiased estimate of the (scaled) total mean-square-error for that model. There is an associated graphical technique called the "C_p plot" in which values of C_p (one for each subset of regressors) are plotted against p.

The problem in choosing a subset-regression model for predicting a response is that including too many unnecessary terms will add to the variance of the predictions, while including too few will result in biased predictions.

In more detail, if we have n observations, and k regressors are available (possibly including a constant term), let P denote some subset of these. (Usually if a constant term is to be considered, this will appear in each subset). Let p be the number of regressors in the subset P. Then C_p (for the P-subset model) is defined to be

$$C_p = \frac{RSS_P}{s^2} - n + 2p$$

where RSS_P is the residual sum of squares for this P-model, and s^2 is an estimate of the residual variance when all relevant terms are included in the model. Usually this is taken to be RSS_K where K is the set of all available regressors.

Under the usual assumptions, that the vector of observations y equals $v + z$ where v is the vector of true means, and the z's are independent with mean zero and constant variance σ^2, $s^2 C_p$ is an unbiased estimate of $\sigma^2 E(J_P)$ where J_P is $|\hat{v}_P - v|^2$, and where $\hat{v}_P$ is the estimate of v that is obtained by fitting the P model. Thus J_P is a measure of the adequacy for prediction of the P model. This result holds even when the true model v is not expressible in terms of the available regressors. However in this case we cannot use the residual sum of squares from the full (K) model as an estimate of σ^2.

The C_p statistic is often used to guide selection of a subset-model, but this cannot be recommended; while for each P separately, C_p gives an unbiased estimate of the scaled mean-square error for that subset, this is not true if the subset is chosen to minimise C_p. In fact this approach can lead to worse results than are obtained by simply fitting all available regressors. In a 1995 paper, Mallows has attempted to quantify this effect.

The C_p statistic is similar to ►Akaike's Information criterion.

About the Author

Colin L. Mallows spent 40 years at AT&T Bell Labs and one of its descendants, AT&T Labs. Since retiring he has been a consultant at another descendant, Avaya Labs. He is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the Royal Statistical Society. He has served on several committees of the IMS and ASA. He was an Associate editor of JASA (1966–1972). He has been COPSS Fisher Lecturer and ASA Deming Lecturer, and has received the ASA Wilks Medal. He has written over 150 papers, and has edited two books.

Cross References

►Akaike's Information Criterion
►General Linear Models

References and Further Reading

Daniel C, Wood FS (1980) Fitting equations to data, rev edn. Wiley, New York
Gorman JW, Toman RJ (1966) Selection of variables for fitting equations to data. Technometrics 8:27–51
Mallows CL (1973) Some comments on Cp. Technometrics 15:661–675
Mallows CL (1995) More comments on Cp. Technometrics 37:362–372

Cramér–Rao Inequality

Maarten Jansen[1], Gerda Claeskens[2]
[1]Université libre de Bruxelles, Brussels, Belgium
[2]K. U. Leuven, Leuven, Belgium

The Cramér–Rao Lower Bound

The Cramér–Rao inequality gives a lower bound for the variance of an unbiased estimator of a parameter. It is named after work by Cramér (1946) and Rao (1945). The inequality and the corresponding lower bound in the inequality are stated for various situations. We will start with the case of a scalar parameter and independent and identically distributed random variables $X_1, \ldots, X_n$, with the same distribution as X.

Denote $\boldsymbol{X} = (X_1, \ldots, X_n)$ and denote the common probability mass function or probability density function of X at a value x by $f(x;\theta)$ where $\theta \in \Theta$, which is a subset of the real line $\mathbb{R}$ and $x \in \mathbb{R}$. Denote the support of X by R, that is, $R = \{x : f(x;\theta) > 0\}$.

Assumptions

1. The partial derivative $\frac{\partial}{\partial\theta} \log f(x;\theta)$ exists for all $\theta \in \Theta$ and all $x \in R$ and it is finite. This is equivalent to stating that the Fisher information value $I_X(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(X;\theta)\right)^2\right]$ is well defined, for all $\theta \in \Theta$.
2. The order of integration and differentiation is interchangeable in $\int \frac{\partial}{\partial\theta} \log f(x;\theta) dx$. If the support of X, that is, the set R, is finite, then the interchangeability is equivalent with the condition that the support does not depend on θ. A counter-example on uniformly distributed random variables is elaborated below.

The Cramér–Rao inequality

Under assumptions (i) and (ii), if $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, this means that $E[\hat{\theta}] = \theta$, then

$$\mathrm{var}(\hat{\theta}) \geq 1/\left[n \cdot I_X(\theta)\right].$$

The lower bound in this inequality is called the Cramér–Rao lower bound.

The proof starts by realizing that the correlation of the score $V = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f_X(X_i;\theta)$ and the unbiased estimator $\hat{\theta}$ is bounded above by 1. This implies that $\left(\mathrm{var}(V) \cdot \mathrm{var}(\hat{\theta})\right)^{1/2} \geq \mathrm{cov}(V, \hat{\theta})$. The assumptions are needed to prove that the expected score $E(V)$ is zero. This implies that the covariance $\mathrm{cov}(V, \hat{\theta}) = 1$, from which the stated inequality readily follows.

A second version of the Cramér–Rao inequality holds if we estimate a functional $\kappa = H(\theta)$. Under assumptions (i) and (ii), if $\boldsymbol{X}$ is a sample vector of independent observations from random variable X with density function $f(x;\theta)$ and $\hat{\kappa} = h(\boldsymbol{X})$ is an unbiased estimator of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ, then

$$\mathrm{var}(\hat{\kappa}) \geq \left[\frac{dH(\theta)}{d\theta}\right]^2 / \left[n \cdot I_X(\theta)\right].$$

Similar versions of the inequality can be phrased for observations that are independent but not identically distributed.

In the case of a vector parameter $\boldsymbol{\theta}$, the variance of the single parameter estimator $\mathrm{var}(\hat{\theta})$ is replaced by the covariance matrix of the estimator vector $\Sigma_{\hat{\boldsymbol{\theta}}}$. This matrix is bounded by a matrix expression containing the inverse of the Fisher information matrix, where bounded means that the difference between the covariance matrix and its "upper bound" is a negative semidefinite matrix.

The Cramér–Rao inequality is important because it states what the best attainable variance is for unbiased estimators. Estimators that actually attain this lower bound are called efficient. It can be shown that maximum likelihood estimators asymptotically reach this lower bound, hence are asymptotically efficient.

Cramér–Rao and UMVUE

If $\boldsymbol{X}$ is a sample vector of independent observations from the random variable X with density function $f_X(x;\theta)$ and $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, then $\mathrm{var}(\hat{\theta}) = 1/\left[n \cdot I_X(\theta)\right] \Leftrightarrow \hat{\theta} = aV + b$ with probability one, where V is the score and a and b are some constants. This follows from the proof of the Cramér–Rao inequality: the lower bounded is reached if the correlation between the score and the estimator is one. This implies that $\mathrm{var}\left(\frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}}\right) = 0 \Rightarrow \frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}} = c$ almost surely for some constant c. We here used the notation σ_X to denote the standard deviation of a random variable X.

The coefficients a and b may depend on θ, but $\hat{\theta}$ should be observable without knowing θ.

If a and b exist such that $\hat{\theta}$ is unbiased and observable, then $\hat{\theta}$ has the smallest possible variance among all unbiased estimators: it is then certainly the uniformly minimum variance unbiased estimator (UMVUE).

It may, however, be well possible that no a and b can be found. In that case, the UMVUE, if it exists, does not reach the Cramér-Rao lower bound. In that case, the notion of *sufficiency* can be used to find such UMVUE.

Counter example: estimators for the upperbound of uniform data

Let $X \sim \text{unif}[0,a]$, so $f_X(x) = \frac{1}{a} I(0 \le x \le a)$, where $I(c \le x \le d)$ is the *indicator* function of the interval $[c,d]$. We want to estimate a. The maximum likelihood estimator (MLE) is $\hat{a}_{\text{MLE}} = \max_{i=1,\ldots,n} X_i$, which is biased. Define $\hat{a}_u = \frac{n}{n-1}\hat{a}_{\text{MLE}}$, which is unbiased. The method of moments leads to an estimator $\hat{a}_{\text{MME}} = 2\overline{X}$, which is also unbiased. The score is $V_i = \frac{\partial}{\partial a} \log f_X(X_i;a) = -\frac{1}{a}$. This is a constant (so, not a random variable), whose expected value is of course *not zero*. This is because the partial derivative and expectation cannot be interchanged, as the boundary of the support of X depends on a. As a consequence, the Cramér–Rao lower bound is *not* valid here. We can verify that $\text{var}(\hat{a}_{\text{MLE}}) = \frac{n}{(n+2)(n+1)^2} a^2$ and $\text{var}(\hat{a}_u) = \frac{1}{n(n+2)} a^2$. This is (for $n \to \infty$) one order of magnitude smaller than $\text{var}(\hat{a}_{\text{MME}}) = \frac{1}{3n} a^2$ and also one order of magnitude smaller than what you would expect for an unbiased estimator if the Cramér–Rao inequality would hold.

A Bayesian Cramér–Rao Bound

It should be noted that biased estimators can have variances below the Cramér–Rao lower bound. Even the MSE (mean squared error), which equals the sum of the variance and the squared bias can be lower than the Cramér–Rao lower bound (and hence lower than any unbiased estimator could attain). A notable example in this respect is Stein's phenomenon on shrinkage rules (Efron and Morris 1977).

In practice, large classes of estimators, for example most nonparametric estimators, are biased. An inequality that is valid for biased or unbiased estimators is due to van Trees (1968, p. 72), see also Gill and Levit (1995) who developed multivariate versions of the inequality.

We assume that the parameter space Θ is a closed interval on the real line and denote by g some probability distribution on Θ with density $\lambda(\theta)$ with respect to the Lebesgue measure. This is where the Bayesian flavor enters. The θ is now treated as a random variable with density λ. We assume that λ and $f(x;\cdot)$ are absolutely continuous and that λ converges to zero at the endpoints of the interval Θ. Moreover we assume that $E[\frac{\partial}{\partial\theta} \log f(X;\theta)] = 0$. We denote $I(\lambda) = E[\{\log \lambda(\theta)\}^2]$ and have that $E[I_X(\theta)] = \int I_X(\theta) g(\theta) d\theta$. Then, for an estimator $\hat{\theta} = \hat{\theta}(X)$, it holds that

$$E[\{\hat{\theta} - \theta\}^2] \ge \frac{1}{E[I_X(\theta)] + I(\lambda)}.$$

A second form of this inequality is obtained for functionals $\kappa = H(\theta)$. Under the above assumptions, for an estimator $\hat{\kappa} = h(\boldsymbol{X})$ of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ,

$$E[\{\hat{\kappa} - H(\theta)\}^2] \ge \frac{\{E[\frac{d}{d\theta} H(\theta)]\}^2}{E[I_X(\theta)] + I(\lambda)}.$$

About the Authors

For the biography of Maarten Jansen *see* the entry ▶Nonparametric Estimation.

For the biography of Gerda Claeskens *see* the entry ▶Model Selection.

Cross References

▶Estimation
▶Minimum Variance Unbiased
▶Sufficient Statistical Information
▶Unbiased Estimators and Their Applications
▶Uniform Distribution in Statistics

References and Further Reading

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton

Efron B, Morris C (1977) Stein's paradox in statistics. Scient Am 236:119–127

Gill RD, Levit BY (1995) Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. Bernoulli 1(1–2): 59–79

Rao C (1945) Information and the accuracy attainable in the estimation of statistical parameters. Bull Calcutta Math Soc 37:81–89

van Trees HL (1968) Detection, estimation and modulation theory: part I. Wiley, New York

Cramér–Von Mises Statistics for Discrete Distributions

Michael A. Stephens
Professor Emeritus
Simon Fraser University, Burnaby, B.C., Canada

Introduction

Cramér–von Mises statistics are well established for testing fit to continuous distributions; see Anderson (2010) and Stephens (2010), both articles in this encyclopedia. In this paper, the corresponding statistics for testing discrete distributions will be described.

Consider a discrete distribution with k cells labeled $1, 2, \ldots, k$, and with probability p_i of falling into cell i. Suppose n independent observations are given; let o_i be the observed number of observations and $e_i = np_i$ be the expected number in cell i. Let $S_j = \sum_{i=1}^{j} o_i$ and $T_j = \sum_{i=1}^{j} e_i$.

Then S_j/n and $H_j = T_j/n$ are the cumulated histograms of observed and expected values and correspond to the empirical distribution function $F_n(z)$ and the cumulative distribution function $F(.)$ for continuous distributions. Suppose $Z_j = S_j - T_j, j = 1, 2, \ldots, k$; the weighted mean of the Z_i is $\bar{Z} = \sum_{j=1}^{k} Z_j t_j$, where $t_j = (p_j + p_{j+1})/2$, with $p_{k+1} = p_1$. The modified Cramér–von Mises statistics are then defined as follows:

$$W_d^2 = n^{-1} \sum_{j=1}^{k} Z_j^2 t_j; \tag{1}$$

$$U_d^2 = n^{-1} \sum_{j=1}^{k} (Z_j - \bar{Z})^2 t_j; \tag{2}$$

$$A_d^2 = n^{-1} \sum_{j=1}^{k} Z_j^2 t_j / \{H_j(1 - H_j)\}. \tag{3}$$

note that $Z_k = 0$ in these summations, so that the last term in W_d^2 is zero. The last term in A_d^2 is of the form $0/0$, and is set equal to zero.

The well-known Pearson χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{k} (o_i - e_i)^2 / e_i.$$

Statistics corresponding to the Kolmogorov–Smirnov statistics (see ▶Kolmogorov-Smirnov Test) for continuous observations are

$$D_d^+ = \max_j (Z_j)/\sqrt{n}, D_d^- = \max_j (-Z_j)/\sqrt{n},$$

$$D_d = \max_j |Z_j|/\sqrt{n}.$$

Comments on the Definitions

1. Several authors have examined distributions of the Kolmogorov–Smirnov family, see Pettitt and Stephens (1977) and Stephens (1986) for tables and references. In general, for continuous data, the Kolmogorov–Smirnov statistic is less powerful as an omnibus test statistic than the Cramér–von Mises family; limited Monte Carlo studies suggest that this holds also for D_d.
2. The Cramér–von Mises and Kolmogorov–Smirnov statistics take into account the order of the cells, in contrast to the Pearson χ^2 statistic.
3. Use of t_j in these definitions ensures that the value of the statistic does not change if the cells are labelled in reverse order.

   For instance, in testing the ▶binomial distribution, one statistician might record the histogram of successes, and another the histogram of failures; or in a test involving categorical data such as the tones of a photograph, the histogram of cells with light to dark observations might be recorded, or *vice versa*.
4. The statistic U_d^2 is intended for use with a discrete distribution around a circle, since its value does not change with different choices of origin; this is why p_{k+1} is set equal to p_1.

Matrix Formulation

To obtain asymptotic distributions it is convenient to put the above definitions into matrix notation. Let a prime, e.g., Z', denote the transpose of a vector or matrix. Let $\mathbf{I}$ be the $k \times k$ identity matrix, and let p' be the $1 \times k$ vector $(p_1, p_2, \ldots, p_k)$. Suppose $\mathbf{D}$ is the $k \times k$ diagonal matrix whose j-th diagonal entry is $p_j, j = 1, \ldots, k$ and let $\mathbf{E}$ be the diagonal matrix with diagonal entries t_j, and $\mathbf{K}$ be the diagonal matrix whose (j,j)-th element is $K_{jj} = 1/\{H_j(1 - H_j)\}, j = 1, \ldots, k-1$ and $K_{kk} = 0$. Let o_i and e_i be arranged into column vectors $\mathbf{o}, \mathbf{e}$ (so that, for example, the j-th component of $\mathbf{o}$ is $o_j, j = 1, \ldots, k$). Then $Z = Ad$, where $d = o - e$ and A is the $k \times k$ partial- sum matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}.$$

The definitions become

$$W_d^2 = Z'EZ/n, \tag{4}$$

$$U_d^2 = Z'(I - E11')E(I - 11'E)Z/n, \tag{5}$$

$$A_d^2 = Z'EKZ/n, \tag{6}$$

$$X^2 = (d'D^{-1}d)/n = Z'A^{-1'}D^{-1}A^{-1}Z/n. \tag{7}$$

Asymptotic Theory

All Parameters Known

All four statistics above are of the general form $S = Y'MY$, where $Y = Z/\sqrt{n}$ and $\mathbf{M}$ is symmetric. For W_d^2, $M = E$, for U_d^2, $M = (I - E11')E(I - 11'E)$, and for A_d^2, $M = EK$. Also Y has mean 0. Suppose its covariance matrix is $\mathbf{\Sigma}_y$, to be found below; then S may be written

$$S = Y'MY = \sum_{i=1}^{k-1} \lambda_i (w_i' Y)^2, \tag{8}$$

where λ_i are the $k-1$ non-zero eigenvalues of $M\boldsymbol{\Sigma}_y$ and w_i are the corresponding eigenvectors, normalized so that $w_i'\boldsymbol{\Sigma}_y w_j = \delta_{ij}$ where δ_{ij} is 1 if $i=j$ and 0 otherwise.

As $n \to \infty$, the s_i tend to standard normal, and they are independent; the limiting distribution of S is that of S_∞ where

$$sinf S_\infty = \sum_{i=1}^{k-1} \lambda_i s_i^2 \tag{9}$$

which is a sum of independent weighted χ_1^2 variables.

Recall that $Y = Z/\sqrt{n} = Ad/\sqrt{n}$; its covariance $\boldsymbol{\Sigma}_y$ is found as follows. Calculate the $k \times k$ matrix

$$\boldsymbol{\Sigma}_0 = D - pp'; \tag{10}$$

this is the covariance matrix of $(o-e)/\sqrt{n}$. Then $\boldsymbol{\Sigma}_y = A\boldsymbol{\Sigma}_0 A'$, with entries $\boldsymbol{\Sigma}_{y,ij} = \min(H_i, H_j) - H_i H_j$.

For the appropriate M for the statistic required, the eigenvalues λ_i, $i = 1, \cdots, k$ of $M\boldsymbol{\Sigma}_y$ are used in (9) to obtain the limiting distribution of the statistic. The limiting distributions have been examined in detail in Choulakian et al. (1994).

Parameters Unknown

Cramér–von Mises statistics when the tested distribution contains unknown parameters θ_i have been investigated by Lockhart et al. (2007). The θ_i must be estimated efficiently, for example by maximum likelihood (ML). Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$ is the vector of m parameters.

The log-likelihood is (omitting irrelevant constants)

$$L^* = \sum_{i=1}^{k} o_i \log p_i,$$

and p_i contains the unknown parameters. The ML estimation consists of solving the m equations

$$\frac{\partial L^*}{\partial \theta_j} = \sum_{i=1}^{k} \frac{o_i}{p_i} \frac{\partial p_i}{\partial \theta_j} = 0,$$

for $j = 1, \ldots, m$.

Let $\hat{\boldsymbol{\theta}}$ be the ML estimate of $\boldsymbol{\theta}$, let $\hat{p}$ be the estimate of p, evaluated using $\hat{\boldsymbol{\theta}}$, and let $\hat{e}$ be the estimated vector of expected values in the cells, with components $\hat{e}_j = n\hat{p}_j$. Then let $\hat{d} = (o - \hat{e})$ and $\hat{Z} = A\hat{d}$.

Define a k by m matrix B with entries

$$B_{i,j} = \partial p_i / \partial \theta_j$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, m$. The matrix $B'D^{-1}B$ is the Fisher Information matrix for the parameter estimates. Define $V = (B'D^{-1}B)^{-1}$. The asymptotic covariance of $\hat{\boldsymbol{\theta}}$ is then V/n, the covariance of $\hat{d}/\sqrt{n}$ is $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_0 - BVB'$, where $\boldsymbol{\Sigma}_0$ is defined in (10), and the covariance of $\hat{Z}/\sqrt{n} = A\hat{d}/\sqrt{n} = \hat{Y}$ is

$$\boldsymbol{\Sigma}_u = A\boldsymbol{\Sigma}_d A'.$$

Then, as in the previous section, where parameters were known, the weights λ_i in the asymptotic distribution (9) are the k eigenvalues of $M\boldsymbol{\Sigma}_u$ for the appropriate M for the statistic required.

In practice, in order to calculate the statistics, using (4–7), the various vectors and matrices must be replaced by their estimates where necessary. For example, let matrix $\hat{D}$ be D with p replaced by $\hat{p}$ and similarly obtain $\hat{B}, \hat{E}, \hat{V}, \hat{K}$ and $\hat{\boldsymbol{\Sigma}}_0$ using estimates in an obvious way. The eigenvalues will also be found using the estimated matrices $\hat{\boldsymbol{\Sigma}}_u$ and $\hat{M}$. Consistent estimates of the λ_i will be obtained and (9) used to find the estimated asymptotic distribution.

Thus the steps are :

1. Calculate $\hat{V} = (\hat{B}'\hat{D}^{-1}\hat{B})^{-1}$.
2. Calculate $\hat{\boldsymbol{\Sigma}}_d = \hat{\boldsymbol{\Sigma}}_0 - \hat{B}\hat{V}\hat{B}'$ and $\hat{\boldsymbol{\Sigma}}_u = A\hat{\boldsymbol{\Sigma}}_d A'$.
3. For the statistic required, let $\hat{M}$ be the estimate of the appropriate M. Find the $k-1$ eigenvalues of $\hat{M}\hat{\boldsymbol{\Sigma}}_u$, or (equivalently) those of the symmetric matrix $\hat{M}^{1/2}\hat{\boldsymbol{\Sigma}}_u\hat{M}^{1/2}$ and use them in (9) to obtain the asymptotic distribution.

For practical purposes, percentage points of S_∞ using exact or estimated λs, can be used for the distributions of the statistics for finite n; this has been verified by many Monte Carlo studies. One therefore needs good approximate points in the upper tail of S_∞; these can be found from the percentage points of $S1$, where $S1$ has the distribution $a + b\chi_p^2$, and the a, b, p are chosen so that the first three cumulants of $S1$ match those of S_∞ in (9). These cumulants are $\kappa_j = 2^{j-1}(j-1)! \sum_{i=1}^{k-1} \lambda_i^j$. In particular, the mean κ_1 is $\sum_{i=1}^{k-1} \lambda_i$, the variance κ_2 is $\sum_{i=1}^{k-1} 2\lambda_i^2$ and κ_3 is $8\sum_{i=1}^{k-1} \lambda_i^3$. Then for the $S1$ approximation, $b = \kappa_3/(4\kappa_2)$, $p = 8\kappa_2^3/\kappa_3^2$, and $a = \kappa_1 - bp$. This approximation is generally accurate in the upper tail, at levels $\alpha < 0.15$. More accurate points can be obtained by the method of Imhof (1961).

About the Author

Michael A. Stephens is Professor Emeritus of Mathematics and Statistics at Simon Fraser University in Burnaby, British Columbia, Canada. Prior to that he taught at several universities including McGill, Nottingham, McMaster, and Toronto, and was a visiting professor at Stanford, Wisconsin-Madison, and Grenoble. He has (co-)authored over 100 papers on the analysis of directional data, continuous proportions, curve-fitting, and tests of fit. Professor Stephens was President of the Statistical Society of Canada in 1983. He is a Fellow of the Royal Statistical Society, and

his honors include membership in the International Statistical Institute, and fellowships of the American Statistical Association and the Institute of Mathematical Statistics. Dr. Stephens received the B.Sc. degree (1948) from Bristol University and A.M. degree (1949) in physics from Harvard University, where he was the first Frank Knox Fellow, and Ph.D. degree (1962) from the University of Toronto. In 1989 he was awarded the Gold Medal, Statistical Society of Canada for two main areas of research: analysis of directional data, and statistical theory and methods associated with goodness of fit.

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Exact Goodness-of-Fit Tests Based on Sufficiency
►Kolmogorov-Smirnov Test
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Anderson TW (2010) Anderson–Darling tests of goodness-of-fit. Article in this encyclopedia
Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann Math Stat 23:193–212
Choulakian V, Lockhart RA, Stephens MA (1994) Cramer–von Mises Tests for discrete distributions. Can J Stat 22:125–137
Darling DA (1955) The Cramér–Smirov test in the parametric case. Ann Math Stat 26:1–20
Imhof JP (1961) Computing the distribution of quadratic forms in normal variables. Biometrika 48:419–426
Lockhart RA, Spinelli JJ, Stephens MA (2007) Cramér–von Mises statistics for discrete distributions with unknown parameters. Can J Stat 35:125–133(9)
Pettitt AN, Stephens MA (1977) The Kolmogorov–Smirov test for discrete and grouped data. Technometrics 19(2):205–210
Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369
Stephens MA (1986) Tests based on EDF statistics. In: D'Agostino R, Stephens MA (eds) Chap. 4 in Goodness-of-fit techniques. Marcel Dekker, New York
Stephens MA (2010) EDF tests of fit. Article in this encyclopedia

Cross Classified and Multiple Membership Multilevel Models

Harvey Goldstein
Professor of Social Statistics
University of Bristol, Bristol, UK

Hierarchically Structured Data

Interesting real life data rarely conform to classical text book assumptions about data structures. Traditionally these assumptions are about observations that can be modelled with independently, and typically identically, distributed "error" terms. More often than not, however, the populations that generate data samples have complex structures where measurements on data units are not mutually independent, but depend on each other through complex structural relationships. For example, a household survey of voting preferences will typically show variation among households and voting constituencies (constituencies and households differ on average in their political preferences). This implies that the replies from individual respondents within a household or constituency will be more alike than replies from individuals in the population at large. Another example of such "hierarchically structured data" would be measurements on students in different schools, where, for example, schools differ in terms of the average attainments of their students. In epidemiology we would expect to find differences in such things as fertility and disease rates across geographical and administrative areas.

Techniques for modelling such data have come to be known as "multilevel" or "hierarchical data" models and basic descriptions of these are dealt with in other articles (*see* ►Multilevel Analysis). In the present article we shall consider two particular extensions to the basic multilevel model that allow us to fit structures that have considerable complexity and are quite commonly found, especially in the social and medical sciences.

The Basic Multilevel Model

A simple multilevel model for hierarchical data structures with normally distributed responses can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij},\ u_j \overset{iid}{\sim} N(0, \sigma_u^2),\ e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2). \quad (1)$$

This might be applied to a sample, say, of school students where i indexes students (level 1), who are grouped within schools (level 2). The response y might be an attainment measure and x a predictor such as a prior test score. Often referred to as a "variance components" model this may be extended in a number of ways to better fit a data set. For example, we may introduce further covariates and we may allow the coefficients of such covariates to vary at level 2, so that, say, β_1, may vary from school to school. Another possibility is to allow the level 1 variance to depend on a set of explanatory variables, so that, for example, we can allow the variance between male students to be different from that for female students. We can have several responses that are correlated leading to a multivariate model, and we can consider non-normal responses, such as binary ones, in order to fit generalised linear multilevel models. We can also have several further levels; for example schools may be

grouped within school boards or authorities, so yielding a three level structure. Goldstein (2010) provides further details and discusses estimation methods.

Cross Classified Structures

The above only describes purely hierarchical models. In practice, however, data structures are often more complicated. Consider an educational example where students move through both their primary and secondary education with the response being attainment at the end of secondary school. For any given primary school, students will generally move to different secondary schools, and any given secondary school will draw students from a number of primary schools. We therefore have a *cross classification* of primary by secondary schools where each cell of the classification will be populated by students (some may be empty). When we model such a structure we have a contribution to the response that is the sum of an effect from the primary and an effect from the secondary school attended by a student. A basic, variance components, cross classified model may be written as

$$\begin{aligned} y_i^{(1)} &= \beta_0 + \beta_1 x_{1i} + u^{(2)}_{\text{primary school}(i)} + u^{(3)}_{\text{secondary school}(i)} \\ &\quad + u^{(1)}_{\text{student}(i)} \\ u^{(2)}_{\text{primary school}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(2)}), \\ u^{(3)}_{\text{secondary school}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(3)}) \\ u^{(1)}_{\text{student}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(1)}),\ i = 1, \ldots, N. \end{aligned} \tag{2}$$

We have changed the notation to make it more general and flexible. The superscript refers to the set of units, or classification; 1 being students, 2 primary school and 3 secondary school. Model (2) thus assumes that there are separate, additive, contributions from the primary and the secondary school attended. As with the simple hierarchical model we can extend (2) in several ways by introducing random coefficients, complex variance structures and further cross classifications and levels. There are many examples where cross classified structures are important. Thus, for example, students will generally be grouped by the neighborhood where they live and this will constitute a further classification. In a repeated measures study where there is a sample of subjects and a set of raters or measurers, if the subjects are rated by different people at each occasion we would have a cross classification of subjects by raters.

Multiple Membership Structures

In many circumstances units can be members of more than one higher level unit at the same time. An example is friendship patterns where at any time individuals can be members of more than one friendship group. In an educational system students may attend more than one school over time. In all such cases we shall assume that for each higher level unit to which a lower level unit belongs there is a known weight (summing to 1.0 for each lower level unit), which represents, for example, the amount of time spent in the higher level unit. The choice of weights may be important but is beyond the scope of this article. For more details about choosing weights see Goldstein et al. (2007).

Using the general notation we used for cross classifications we can write a basic variance components multiple membership model as

$$\begin{aligned} y_i^{(1)} &= \beta_0 + \beta_1 x_i + \sum_{j \in school(i)} w^{(2)}_{i,j} u^{(2)}_{(j)} + u^{(1)}_i \\ u^{(2)}_{(j)} &\sim N(0, \sigma^2_{u(2)}),\ u^{(1)}_{(i)} \sim N(0, \sigma^2_{u(1)}) \\ \sum_{j \in school(i)} w^{(2)}_{i,j} &= 1. \end{aligned} \tag{3}$$

This assumes that the total contribution from the level 2 units (schools) is a weighted sum over all the units of which the level 1 unit has been a member. Thus, for example, if every student spends half their time in one school and half their time in another (randomly selected) then the variance at level 2 will be

$$\text{var}(0.5u^{(2)}_{j_1} + 0.5u^{(2)}_{j_2}) = \sigma^2_{u(2)}/2. \tag{2}$$

Thus, a failure to account for the multiple membership of higher level units in this case will lead us to treat the estimate of the level 2 variance, $\sigma^2_{u(2)}/2$ as if it were a consistent estimate of the true level 2 variance $\sigma^2_{u(2)}$. More generally, ignoring a multiple membership structure will lead to an underestimation of the higher level variance.

Finally, we can combine cross classified and multiple membership structures within a single model and this allows us to handle very complex structures. An example where the response is a binary variable is given in Goldstein (2010, Chap. 13). It is possible to use maximum likelihood estimation for these models, but apart from small scale datasets, MCMC estimation is more efficient and flexible. The MLwiN software package (Rasbash et al. 2009; Browne 2009. http://www.cmm.bristol.ac.uk) is able to fit these models.

About the Author

Professor Goldstein is a chartered statistician, has been editor of the *Royal Statistical Society's Journal, Series A*, a member of the Society's Council and was awarded the Society's Guy medal in silver in 1998. He was elected a member of the International Statistical Institute in 1987, and a Fellow of the British Academy in 1996. He was awarded an honorary doctorate by the Open University in 2002.

His most important research interest is in the methodology of multilevel modelling. He has had research funding for this since 1986 and has been involved in the production (with Jon Rasbash and William Browne) of a widely used software package (MLwiN) and made a number of theoretical developments. These include multiple membership models, multilevel structural equation models and more recently the use of multilevel multivariate latent normal models and especially their application to missing data problems. The major text on multilevel modelling is his book *Multilevel Statistical Models* (New York, Wiley, Fourth Edition, 2010). He has also written extensively on the use of statistical modelling techniques in the construction and analysis of educational tests. The implications of adopting such models have been explored in a series of papers since 1977.

Cross References

►Multilevel Analysis
►Psychology, Statistics in

References and Further Reading

Browne WJ (2009) MCMC estimation in MLwiN. Version 2.10. Bristol, Centre for Multilevel Modelling, University of Bristol

Goldstein H (2010) Multilevel statistical models, 4th edn. New York, Wiley

Goldstein H, Burgess S, McConell B (2007) Modelling the effect of pupil mobility on school differences in educational achievement. J R Stat Soc Ser A 170(4):941–954

Rasbash J, Steele F, Browne W, Goldstein H (2009) A user's guide to MLwiN version 2.10. Bristol, Centre for Multilevel Modelling, University of Bristol

Cross-Covariance Operators

Charles R. Baker
Professor Emeritus
University of North Carolina, Chapel Hill, NC, USA

This article will initially treat joint probability measures and their associated cross-covariance operators. Subsequently, attention will be shifted to three examples of problems on capacity of information channels.

Cross-covariance operators were introduced in Baker (1970) as a tool in solving a basic problem in information theory, and treated more extensively in Baker (1973). Related results are in Gualtierotti (1979) and Fortet (1995). The empahsis in Baker (1973) was in two directions: showing the added power of analysis obtained by introducing the cross-covariance operator of a joint measure, and providing new results for actually computing likelihood ratios for joint measures. Applications to date have included results on absolute continuity of probability measures, mutual information for pairs of ►stochastic processes, and analysis of information capacity for communication channels. More recently, there has been interest in this topic by researchers in machine learning, who have applied theory from Baker (1973) in a number of interesting publications (e.g., Fukumizu et al. 2004, 2009; Gretton et al. 2005).

The joint measures to be discussed are probability measures on the product of two real separable Hilbert spaces, H_1 and H_2, with Borel sigma fields θ_1. and θ_2. Denote the inner products by $<,>_1$ on H_1 and $<,>_2$ on H_2. H_1 x H_2 is then a real separable Hilbert space under the inner product defined by $<(x,u),(v,y)>_{12} = <x,v>_1 + <u,y>_2$. Next, introduce a joint measure π_{12} on the measurable space (H_1 x H_2, θ_1 x θ_2). Only strong second-order probability measures will be considered: those joint measures ∂_{12} such that $E_{\partial_{1,2}}|||(x,y)|||^2_{12} = \int_{H_1xH_2}(||x||^2_1 + ||y||^2_2)d\partial_{12}(x,y)$ is finite. All Gaussian measures on H_1xH_2 are strong second order, as are their projections on H_1 and H_2. From the measure π_{12} one has projections π_i on (H_i, θ_i), $i = 1, 2$. Let m_1 and m_2 denote the mean elements and R_1 and R_2 the covariance operators of π_1 and π_2.

The first result of note is the definition and properties of the cross-covariance operator for the joint measure π_{12}. Denoting that operator by C_{12}, it is defined for all (u,v) in H_1xH_2 by

$$<C_{12}v, u>_1 = \int_{H_1xH_2} <x - m_1, u>_1 <y - m_2, v>_2 \, d\pi_{12}(x, y).$$

Theorem 1 *C_{12} has representation $C_{12} = R_1^{1/2}VR_2^{1/2}$, $V: H_2 \to H_1$ a unique linear operator having $||V|| \le 1$ and $P_1VP_2 = V$, P_i the projection of H_i onto the closure of range(R_i).* ⊡

Next, we turn to the definition and properties of the covariance operator $\mathbf{R}_{12}$ of π_{12}. By direct computation (Baker 1973), one can show that this operator is defined on every element (u,v) in H_1xH_2 by

$$\begin{aligned}\mathbf{R}_{12}(u, v) &= (R_1u + C_{12}v, R_2v + C_{12}^*u) \\ &= (R_1 \otimes R_2)(u, v) + (C_{12}^* \otimes C_{12})(u, v).\end{aligned}$$

We now give a result that illustrates both similarity and difference between a joint measure and the usual measure.as defined on one of the spaces H_1 or H_2. We define a self-adjoint operator $\mathbf{V}$ in H_1xH_2 by $\mathbf{V}(u,v) = (Vv, V^*u) = (V^* \otimes V)(u,v)$ and denote by $\mathbf{I}$ the identity operator in H_1xH_2; it is shown in Baker (1973) that $||\mathbf{V}|| \le 1$ and that the non-zero eigenvalues of VV^* are squares of the non-zero eigenvalues of $\mathbf{V}$.

Theorem 2 *The covariance operator* $\boldsymbol{R}_{12}$ *of the measure* π_{12} *on* $H_1 x H_2$ *has representation* $\boldsymbol{R}_{12} = \boldsymbol{R}_{1\otimes 2}^{1/2}(\boldsymbol{I}+\boldsymbol{V})\boldsymbol{R}_{1\otimes 2}^{1/2}$, *where* $\boldsymbol{R}_{1\otimes 2}$ *is the covariance operator of the product measure* $\pi_1 \otimes \pi_2$. *If* π_{12} *is Gaussian, then* π_{12} *and* $\pi_1 \otimes \pi_2$ *are mutually absolutely continuous if and only if* $\boldsymbol{V}$ *is Hilbert-Schmidt with* $||\boldsymbol{V}|| < 1$, *and otherwise orthogonal.*⊡

The preceding results give some of the basic properties of the covariance operator of a joint measure, and it is seen that the cross-covariance operator is an essential component in the definition and properties of the covariance operator. In Baker (1973), considerable attention is given to Gaussian joint measures. However, it should be noted that the definition of the cross-covariance operator and its relation to the covariance operator hold for any strong-second order joint probability measure When the joint measure at hand is not Gaussian, one still has the cross-covariance operator available as well as the mean and the covariance operator. These functions can frequently be estimated from data and used to develop suboptimum but effective operations using (for example) second moment criteria.

We now turn to a brief introduction to three problems on the capacity of a Gaussian channel without feedback (Baker 1978, 1987; Baker and Chao 1996a, b). The cited papers provide examples of the use of results from Baker (1973) in applications to information theory. The definition of the channel capacity is as follows. We have a joint measure $\pi_{S,AS+N}$ where S is the actual signal, AS is the transmitted coded signal (from a measurable space (Ω, Θ)) and AS+N is the received waveform of signal+noise from a measurable space (Ψ, Γ). The (average) mutual information will be finite if $\pi_{S,AS+N}$ is absolutely continuous with respect to its product measure $\pi_{S\otimes AS+N}$, and its value is then given by

$$\int_{\Omega x \Psi} \log[(d\pi_{S,AS+N}/d\pi_{S\otimes AS+N})(x,y)]d\pi_{S,AS+N}(x,y).$$

The transmitted signal AS and the received AS+N can vary with choices by the coder (and the jammer in the third example below), and the channel capacity is the supremum of the mutual information over all admissible S and AS+N.pairs.

In each case, the transmitted signal has a constraint given in terms of the ambient noise process. When the constraint on the transmitted signal is given in terms of the channel noise covariance, one says that the channel is "matched" (coder constraint is matched to the channel noise covariance) (Baker 1978). The second type of channel is "mismatched" (the signal constraint is not given in terms of the channel noise covariance) (Baker 1987). The third class is the jamming channel without feedback, wherein the noise in the channel consists of a known Gaussian ambient noise (nature's contribution) plus an independent noise that is under the control of a hostile jammer (Baker and Chao 1996a, b). In this channel, there is a constraint on the jammer's noise as well as one on the coder's.transmitted signal.

In the jamming channel, the jammer has no constraints on the choice of the probability distributions of the noise at his command. However, it is known (Ihara 1978) that if the channel noise due to nature is Gaussian, then the information capacity is minimized by the jammer choosing (among all processes satisfying the constraints) a Gaussian process. Thus, the original problem becomes a problem involving an ambient Gaussian noise (which is used to calculate the coder's constraint) and an independent Gaussian process (jamming) giving the covariance constraint that the jammer uses.

Cross References

►Canonical Analysis and Measures of Association
►Measure Theory in Probability
►Statistical Signal Processing
►Statistical View of Information Theory

References and Further Reading

Baker CR (1970) Mutual information for Gaussian processes. SIAM J Appl Math 19(2):451–458
Baker CR (1973) Joint measures and cross-covariance operators. Trans Am Math Soc 186:273–289
Baker CR (1978) Capacity of the Gaussian channel without feedback. Inf Cont 37:70–89
Baker CR (1987) Capacity of the mismatched Gaussian channel. IEEE Trans Inf Theory 33:802–812
Baker CR, Chao IF (1996a) Information capacity of channels with partially unknown noise. I. Finite-dimensional channels. SIAM J Appl Math 56:946–963
Baker CR, Chao IF (1996b) Information capacity of channels with partially unknown noise. II. Infinite-dimensional channels. SIAM J Cont Optimization 34:1461–1472
Fortet RM (1995) Vecteurs, fonctions et distributions aleatoires dans les espaces de Hilbert. Hermes, Paris (see pp. 331 ff.)
Fukumizu K, Bach FR, Jordan MJ (2004) Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. J Mach Learning Res 5:73–99
Fukumizu K, Bach FR, Jordan MJ (2009) Kernel dimensionality reduction in regression. Ann Stat 37:1871–1905
Gretton A, Bousquet O, Smola AJ, Schölkopf B (2005) Measuring statistical dependence with Hilbert–Schmidt norms. MPI Technical Report, Max Planck Institute for Biological Cybernetics. Tübingen, Germany' Report 140
Gualtierotti AF (1979) On cross-covariance operators. SIAM J Appl Math 37(2):325–329
Ihara S (1978) On the capacity of channels with additive non-Gaussian noise. Inf Cont 37:34–39