# R

## R Language

Robert Gentleman[1], Wolfgang Huber[2],
Vincent J. Carey[3]
[1]Genentech, South San Francisco, CA, USA
[2]Heidelberg, Germany
[3]Brigham and Women's Hospital, Boston, MA, USA

R is a widely used open source language for scientific computing and visualization. It is based on the S language (S: An Interactive Environment for Data Analysis and Graphics, R. A. Becker and J. M. Chambers, Wadsworth, 1984), but with a few paradigms adopted from the Lisp family of languages.

R began its life in 1992, when Ross Ihaka and Robert Gentleman started a project that ultimately evolved into what it is now. In the early days, their main goal was to develop something that was like S, but which had clearer underlying semantics. Around the same time, other major changes were taking place: the world wide web was quickly gaining steam, and a new open source operating system named Linux (with major components from the GNU project) was becoming a popular tool for academic researchers. With these advances, it made sense to make the software more widely available, and hence it needed a name. In part to reflect its heritage, and in part to reflect their contributions, Ross and Robert chose to call it R. At this time, R was still primitive and had restricted capabilities, but a number of other scientists realized its potential, and shortly after its release the R-core group was formed. The activities and hard work of this group of contributors was what really made the breakthrough, and due to their efforts, R quickly become more stable, reliable and forward looking.

R is now under constant development by a team of approximately 20 individuals (essentially members of R-core) and has a fairly consistent 6 month release cycle. The core language is extended through add-on packages which can be obtained and installed in a local version of R, thereby customizing it to a user's interests. These packages are perhaps one of the main advantages of R, since a wide variety of statistical, computational and visualization methods are available. These add-on packages are often written by experts in the methodology, and that has served well to ensure the high quality of the outputs. However, users should realize that the availability of packages on sites such as cran.r-project.org or www.bioconductor.org does not imply any endorsement of their scientific quality except perhaps by their authors. Textbooks, the mailing list archives, and the scientific publications that sometimes accompany a package are good places for users to derive a judgment on packages' suitability for their needs.

R has served to bring scientific computing into many peoples hands. Many statisticians world-wide use R for their teaching and research. R has become widely used in many other fields as well, physics, chemistry, sociology, and notably biology. It is used in a wide variety of industries, Google, Microsoft, various pharmaceutical companies, banks and many investment houses. Its flexibility and the ability to relatively easily code new algorithms is likely to be one of the reasons that R has seen such wide-spread adoption. While reliable estimates of the number of users are hard to obtain increases in download traffic, in frequency of posting to the email help list suggest that the user base is continuing to grow quite rapidly.

R has made scientific computing easier for many sophisticated users and it has also brought in people who are new to the field. Balancing the diverse needs of the community is a particular problem. Discussions on the mailing lists can range from the very philosophical (often surrounding variants of object oriented programming) to the somewhat simpler (but often repeated) bug report that R's numerical capabilities are questionable. For all classes of new users it may be helpful to realize that R has a long history, the actual numerical code used for most of the applications is more than 10 years old and much of it is even older. Large parts of that code have been widely tested for years and it is somewhat unlikely that it fails to perform as intended in any really obvious way (as one person put it, you may be new to R, but R is not new). The fact that R has a long history (somewhat longer than that of Java, for example) means that changes to the way that functions work (even when we know that the original version

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

was not optimal) are not likely to happen – there is simply too much code and too many users invested in the way it currently works and one must have very good reasons to modify code.

New users of R, or any other programming language, that want to do scientific programming should be conversant with the basics of computer arithmetic. To quote from "The Elements of Programming Style" by Kernighan and Plauger: 10.0 times 0.1 is hardly ever 1.0. The issue is one of representing a real number in the allocated memory of a computer, this can only be done exactly for a small subset of all numbers and for all others some rounding is needed. Interested readers should consult a good book on numerical computing and David Goldberg (1991), "What Every Computer Scientist Should Know About Floating-Point Arithmetic," ACM Computing Surveys, 23/1, 5–48, which is available online at: http://docs.sun.com/source/806-3568/ncg_goldberg.html.

As noted above the R language is largely based on the S language that was developed at Bell Laboratories by John Chambers and colleagues during the 1970s and 1980s. While a major goal was providing an interactive environment for performing statistical, and more generally scientific, computing there were other motivations for their work. One of the guiding philosophies of John's work was the notion that scientists needed to use computers to solve problems, and that if the computing environment was suitably conducive they would gradually evolve into being programmers. The main reason is that while there are many commonalities between problems, there is always some need for additional programming and the tweaking of inputs or outputs. Thus, one hopes that the language will actually help to develop the next generation of computational experts by converting some set of its users into programmers. It is worth emphasizing again, that the benefits of having a scientist conversant with, and invested in a method typically means that the method will provide the correct outputs. There is potential for the implementation to be sub-optimal, but that can generally be overcome if the method gets adopted for wide-spread use.

In his book, Algorithms + Data Structures = Programs, N. Wirth describes the fundamental notion that computer programs rely on both a set of algorithms and a set of data structures. R contains a very rich set of algorithms, but of equal importance is the ability of the user to create and use data structures that are appropriate to the problem at hand. R has a very rich and extensible collection of data structures and well designed data structures can greatly simplify many programming problems. Common examples are specific data structures do hold dates and data structures to hold time series objects. Both of these specialized containers are widely used and their use greatly simplifies many programming problems. Specialized methods can be written to deal with the specific implementations and users are then free to worry about other problems (and not converting month–day–year representations into something numeric).

In our work in computational biology we found that the complexity of most experiments was very high and the users were typically spending a great deal of their time doing very basic data management. A very typical example comes from the analysis of microarrays on some set of samples. The arrays provide us with a very large (generally 10s of thousands) of measurements one genes for each sample and at the same time we would have a separate set of data describing the characteristics of the sample. Most users would then spend some time arranging that the order of the columns in the microarray data was the same as that of the rows in the sample characteristic data (somewhat peculiarly microarray data are stored with samples as columns, while most other statistical data is stored with the samples as rows). That is fine until subsets are needed or one decides to do a permutation test (see ▶Permutation Tests) for some hypothesis. At that point, depending on whether samples or genes are being permuted different operations are needed. This task is both tedious and has the potential to be done incorrectly in ways that are hard to detect. The rather simple expedient of defining a new data structure that contains both arrays, and where subsetting is defined and implemented to work appropriately greatly simplifies the analysts job and has the additional effect of making it much more likely that the right answer is obtained.

This observation leads us to another arena in which R is taking an important role: that of reproducibility in scientific computing. This issue arises often due to the fact that the analysis of any reasonably large and complex data set is error prone. The chance that mistakes are made, steps omitted increases as the number of people involved in the analysis grows and as the number of software tools increases. A dynamic document is a document that consists of both text and computer code. In greatly simplified terms the document is processed and the computer code is evaluated. An output document is created where each block of computer code is replaced with its output. Typically the computer code is used to produce the figures and tables that are needed for the final document. The final document can then be submitted as a paper to a journal or as an internal report within a group or company. The advantage of the approach is that anyone with access to the raw document and the data can reproduce the document and more importantly they can understand how every figure and table was produced.

## Cross References

►Computational Statistics
►Statistical Software: An Overview

# Radon–Nikodým Theorem

Takis Konstantopoulos[1], Zurab Zerakidze[2], Grigol Sokhadze[2]
[1]Professor
Heriot-Watt University, Edinburgh, UK
[2]Professor
Javakhishvili Tbilisi State University, Tbilisi, Georgia

The theorem is concerned with the existence of density (derivative) of one measure with respect to another. Let $(\Omega, \mathscr{F})$ be a measurable space, i.e., a set $\Omega$ together with a $\sigma$-algebra $\mathscr{F}$ of subsets of $\Omega$. Suppose that $\nu$, $\mu$ are two $\sigma$-finite positive measures on $(\Omega, \mathscr{F})$ such that $\nu$ is absolutely continuous (denoted by $\nu \ll \mu$) with respect to $\mu$, i.e., if $\mu(A) = 0$ for some $A \in \mathscr{F}$ then $\nu(A) = 0$. The Radon–Nikodým theorem states that these exists a $\mu$-integrable function $f : \Omega \to \mathbb{R}_+$ such that

$$\nu(A) = \int_A f(\omega)\mu(d\omega), \quad A \in \mathscr{F}.$$

Moreover, $f$ is $\mu$-a.e. unique, in the sense that if $f'$ also satisfies the above then the $\mu$-measure of the points $\omega$ such that $f(\omega) \neq f'(\omega)$ equals zero. The function $f$ is called Radon–Nikodým derivative of $\mu$ with respect to $\nu$ and this is often denoted by

$$f(\omega) = \frac{d\nu}{d\mu}(\omega).$$

The standard proof is as follows. First, assume that $\mu(\Omega) < \infty$. Denote by $G$ the class of all non-negative $\mu$-integrable functions $g$ such that

$$\int_A g(\omega)\mu(d\omega) \le \nu(A), \quad A \in \mathscr{F}.$$

Let $c$ be the supremum of the set numbers $\{\int_\Omega g d\mu : g \in G\}$, and choose a sequence $g_n$ of elements of $G$ such that $\lim_{n\to\infty} \int_\Omega g_n d\mu = \int_\Omega g d\mu$. Observe that if $g', g''$ are elements of $G$ then so is their maximum $\max(g', g'')$. This observation, together with the monotone convergence theorem, allows us to conclude that $f = \sup_n g_n$ is also a member of $G$ and $\int_\Omega f d\mu = c$. This shows that $\int_A f d\mu \le \nu(A)$ for all $A \in \mathscr{F}$. To show that the difference is actually zero we need to use the Hahn decomposition of a signed measure. Details can be found in Kallenberg (2002, pp. 28–30). The general case for a $\sigma$-finite $\mu$ follows easily by taking an increasing sequence $\Omega_n$ with $\mu(\Omega_n) < \infty$ and $\cup_n \Omega_n = \Omega$, and by applying the previous construction to each $\Omega_n$.

The theorem was proved by Johann Radon (1913) in 1913 for the case $\Omega = \mathbb{R}^n$ and generalized by Otton Nikodým (1930) in 1930 in its present form. The Radon–Nikodým derivative possesses the following properties:

1. Linearity: $\frac{d(c_1\nu_1 + c_2\nu_2)}{2\mu} = c_1\frac{d\nu_1}{d\mu} + c_2\frac{d\nu_2}{d\mu}$, $c_1, c_2 \in \mathbb{R}$.
2. Change of measure: If $\nu \ll \mu$ and $g$ is a $\nu$-integrable function then $\int_\Omega g d\nu = \int_\Omega g\frac{d\nu}{d\mu}d\mu$.
3. Chain rule: If $\lambda \ll \nu \ll \nu$ then $\frac{d\lambda}{d\mu} = \frac{d\lambda}{d\nu}\frac{d\nu}{d\mu}$.
4. Inverse rule: If $\nu \ll \mu$ and $\mu \ll \nu$ then $\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu}\right)^{-1}$.

It is worth noting that a more general statement holds, known as *Lebesgue decomposition*: Let $\nu$, $\mu$ be $\sigma$-finite measures on $(\Omega, \mathscr{F})$. Then there exists a unique measure $\nu_a \ll \mu$ and a unique measure $\nu_s \perp \mu$ (singular with respect to $\mu$) such that $\nu = \nu_a + \nu_s$.

Also note that the $\sigma$-finiteness condition cannot be dropped. For example, if $\Omega = \mathbb{R}$, $\mathscr{F}$ the $\sigma$-algebra of Borel sets, $\mu$ the counting measure and $\nu$ the Lebesgue measure, we certainly have $\nu \ll \mu$ but a density of $\nu$ with respect to $\mu$ does not exist.

The Radon–Nikodým theorem has numerous applications in many areas of modern mathematics. We mention a few below.

**1. Conditional expectation.** Let $(\Omega, \mathscr{F}, P)$ be a probability space, $X$ a non-negative random variable with $EX = \int_\Omega X dP < \infty$, and $\mathscr{G}$ a sub-$\sigma$-algebra of $\mathscr{F}$. The notion of conditional expectation $E(X|\mathscr{G})$ of $X$ given $\mathscr{G}$ was introduced by A.N. Kolmogorov (1933) in 1933 by means of the Radon–Nikodým derivative as follows. Consider the measure $\nu(A) = \int_A X dP$, $A \in \mathscr{G}$. Clearly, $\nu \ll P$ on $\mathscr{G}$. According to the Radon–Nikodým theorem there is a $\mathscr{G}$-measurable function $E(X|\mathscr{G})$ which satisfies the relation

$$\int_A E(X|\mathscr{G})dP = \int_A X dP, \quad A \in \mathscr{G}.$$

More generally, we define $E(X|\mathscr{G}) = E(X^+|\mathscr{G}) - E(X^-|\mathscr{G})$. For further information see Konstantopoulos (2009).

**2. ►Martingales.** If $P, Q$ are two probability measures on the same measurable space $(\Omega, \mathscr{F})$ such that $Q \ll P$ then, for any sub-$\sigma$-algebra $\mathscr{G} \subset \mathscr{F}$ we have $Q \ll P$ on $(\Omega, \mathscr{G})$. If we denote by $\left(\frac{dQ}{dP}\right)_{\mathscr{F}}$ and $\left(\frac{dQ}{dP}\right)_{\mathscr{G}}$ the two Radon–Nikodým derivatives we have the consistency property $E\left[\left(\frac{dQ}{dP}\right)_{\mathscr{F}}|\mathscr{G}\right] = \left(\frac{dQ}{dP}\right)_{\mathscr{G}}$.

In fact, if $\mathscr{F}_n$ is an increasing sequence of sub-$\sigma$-algebras generating $\mathscr{F}$, then $E\left[\left(\frac{dQ}{dP}\right)_{\mathscr{F}}|\mathscr{F}_n\right]$ is a uniformly integrable martingale (Williams 1989) whose limit (a.s. and in $L_1$) equals $\left(\frac{dQ}{dP}\right)_{\mathscr{F}}$.

**3. Kullback–Leibler divergence and Hellinger distance.** In Theoretical Statistics, the notion of ▶*Kullback–Leibler divergence* was introduced by Solomon Kullback and Richard Leibler (1951) in 1951. This is a generalization of the notion of ▶entropy for two distributions. If $\mu$ and $\nu$ are two probability measures on the same space with $\nu \ll \mu$, the Kullback–Leibler divergence or distance

$$Q(\nu \parallel \mu) = -\int_\Omega \log\left(\frac{d\nu}{d\mu}\right) d\mu$$

measures the relative variability of $\nu$ with respect to $\mu$. The quantity is always non-negative (owing to the convexity of $-\log$ and Jensen's inequality) but not symmetric. The Kulback–Leibler divergence is used in Information Theory (Cover and Thomas 1991) to define the mutual information between two random variables.

The *Hellinger distance* $H(\mu,\nu)$ between two probability measures $\mu$ and $\nu$ which are absolutely continuous to a third probability measure $\lambda$ is defined by

$$H^2(\mu,\nu) = \frac{1}{2}\int_\Omega \left(\sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}}\right)^2 d\lambda$$

and we have $H^2(\mu,\nu) \le Q(\nu \parallel \mu)$.

**4. Densities on $\mathbb{R}^n$.** Let $f$, $g$ be densities of two probability measures $P, Q$, respectively, on $\mathbb{R}^n$. Then $Q \ll P$ if and only if there are versions of $f$ and $g$ such that $\{x : f(x) > 0\} \subset \{x : g(x) > 0\}$. In this case, $\frac{dQ}{dP}(x) = \frac{g(x)}{f(x)}$. For example, if $P$ is the law of $n$ i.i.d. standard normal random variables $(\xi_1,\dots,\xi_n)$, and if $Q$ is the law of $(\xi_1+\mu_1,\dots,\xi_n+\mu_n)$, for some constants $\mu_1,\dots,\mu_n$, then $\frac{dQ}{dP}(x_1,\dots,x_n) = \exp\sum_{j=1}^n\left(\mu_j x_j - \frac{1}{2}\mu_j^2\right)$.

**5. The Radon–Nikodým derivative between two Brownian motions.** This is an infinite-dimensional generalisation of the previous example. Define the probability measure $P_{T,\mu}$, on the space $\Omega$ of continuous functions $\omega : [0,T] \to \mathbb{R}$, to be the law of a Brownian motion (see ▶Brownian Motion and Diffusions) with drift $\mu$ and unit variance. We have $P_{T,\mu} \ll P_{T,0}$ and $\frac{dP_{T,\mu}}{dP_{T,0}}(\omega) = e^{\mu\omega(T)-\frac{1}{2}\mu^2 T}$. Moreover, the consistency (martingale) property $E_{T,0}\left[\frac{dP_{T,\mu}}{dP_{T,0}}|\mathscr{F}_t\right] = \frac{dP_{t,\mu}}{dP_{t,0}}$, $t \le T$, holds. Here $\mathscr{F}_t$ is the $\sigma$-algebra generated by $(\omega(s), s \le t)$. A further generalisation of this is the *Cameron–Martin–Girsanov theorem* (Cameron and Martin 1944). Let $(X_t)$ be a measurable $(\mathscr{F}_t)$-adapted process such that $Z_t := \exp\left\{\int_0^t X_s dW_s - \frac{1}{2}\int_0^t X_s^2 ds\right\}$ is defined and is a martingale. Define $Q$ on $(\Omega,\mathscr{F}_T)$ by $\frac{dQ}{dP_{T,0}} = Z_T$. Then the law of the process $\left(W_t - \int_0^t X_s ds, 0 \le t \le T\right)$ on $(\Omega,\mathscr{F}_T,Q)$ is again $P_{T,0}$. More general results on the absolute convergence of Gaussian measures and the calculation of a density function are studied, e.g., by Feldman (1958), Ibragimov and Rozanov (1970), Zerakidze (1969) and Yadrenko (1980). Results on diffusions and general processes appear, e.g., in Liptser and Shiryaev (1974), Gikhman and Skorokhod (1971–1975). Smooth measures were studied by Bell (1991), Daletskii and Sokhadze (1988), Bogachev (2008), Kulik and Pilipenko (2000), among others.

**6. The Radon–Nikodým derivative between two ▶Poisson processes.** Let $P_\lambda$ be the law of a rate-$\lambda$ homogeneous Poisson process on a bounded measurable set $S \subset \mathbb{R}^n$ with Lebesgue measure $|S|$. The $P_\lambda$ is a probability measure on the space $\Omega$ of integer-valued random measures with no multiple points. For any $0 < \lambda, \mu < \infty$ we have that $P_\lambda \ll P_\mu$ with Radon–Nikodým derivative

$$\frac{dP_\lambda}{dP_\mu}(\omega) = \left(\frac{\lambda}{\mu}\right)^{\omega(S)} \cdot e^{-(\lambda-\mu)|S|}.$$

To see this, it is sufficient to show that for any bounded measurable $f : \Omega \to \mathbb{R}$ we have $E_\lambda[\exp\int_S f(x)\omega(dx)] = E_\mu\left[\left(\frac{\lambda}{\mu}\right)^{\omega(S)} \cdot e^{-(\lambda-\mu)|S|} \cdot \exp\int_S f(x)\omega(dx)\right]$, something that is easily verifiable by means of the Poisson characteristic functional $E_\lambda[\exp\int_S f(x)\omega(dx)] = \exp\lambda\int_S(e^{f(x)}-1)dx$. Note that if $\widehat{P}_\lambda$ is the image of $P_\lambda$ on $\mathbb{Z}_+$ under the mapping $\omega \mapsto \omega(S)$ then the formula above says that $\frac{dP_\lambda}{dP_\mu}(\omega) = \frac{d\widehat{P}_\lambda}{d\widehat{P}_\mu}(\omega(S))$. We also note that if $S$ is not bounded, e.g., if $S = \mathbb{R}^n$, the above fails to hold because $P_\lambda \perp P_\mu$ if $\lambda \ne \mu$.

**7. The Radon–Nikodým derivative between two Markov jump processes.** Consider a Markov jump process in a countable state space $S$ with transition rates $q_{x,y}$ such that $q(x) := -q(x,x) = \sum_y q_{x,y} < \infty$ for all $x \in S$, and initial distribution $\mu$. Let $Q$ be the matrix with entries $q_{x,y}$. In other words, $Q$ and $\mu$ define a probability measure $P_{\mu,Q}$ on the space $\Omega$ of right-continuous piecewise-constant functions $\omega : [0,T] \to S$. We only consider finite time horizon $T$. We change $\mu, Q$ to $\widetilde{\mu}, \widetilde{Q}$ in a way that $\mu \ll \widetilde{\mu}$ and $q_{x,y} = 0$ whenever $\widetilde{q}_{x,y} = 0$. Then $P_{\mu,Q} \ll P_{\widetilde{\mu},\widetilde{Q}}$ and

$$\frac{dP_{\mu,Q}}{dP_{\widetilde{\mu},\widetilde{Q}}}(\omega) = \frac{\mu(\omega(0))}{\widetilde{\mu}(\omega(0))}\exp\left\{-\int_0^T (q(\omega(s)) - \widetilde{q}(\omega(s)))ds\right\} \cdot \prod_{x\ne y}\left(\frac{q_{x,y}}{\widetilde{q}_{x,y}}\right)^{N_T(\omega,x,y)},$$

where $N_T(\omega, x, y)$ is the total number of points $s \le T$ such that $\omega(s-) = x, \omega(s) = y$.

**8. The Esscher transform.** Let $(X_t, t \ge 0)$ be a Lévy process (see ►Lévy Processes), i.e. a stochastic process with values in $\mathbb{R}$ which is continuous in probability and has stationary-independent increments. Assume that the Laplace exponent $\psi(\beta) = \log E \exp(\beta X_1)$ is defined for $\beta$ belonging to a non-trivial interval. Let $Z_t^\beta := \exp\{\beta X_t - \psi(\beta)t\}$ and define a new measure $P^\beta$ via the Radon–Nikodým derivative $\frac{dP^\beta}{dP}\big|_{\mathscr{F}_t} = Z_t^\beta$, where $\mathscr{F}_t = \sigma(X_s, s \le t)$. This derivative is known as the *Esscher transform* and leads to a natural generalisation of the Cameron–Martin–Girsanov theorem: The process $(X_t)$ is still a Lévy process under $P^\beta$. See Kyprianou (2006) for its use in Fluctuation Theory.

**9. Palm probability.** Let $(\Omega, \mathscr{F}, P)$ be a probability space and $M : (\Omega \times \mathbb{R}^d) \to \mathbb{R}_+$ be measurable in the first argument and a locally finite probability measure in the second. We call such an $M$ a random measure on $\mathbb{R}^d$. Assume that $\lambda(B) = EM(B)$ is a locally finite measure. Define the Campbell measure $C(A, B) = E[\mathbf{1}_A M(B)]$, $A \in \mathscr{F}, B \in \mathscr{B}$, where $\mathscr{B}$ is the class of Borel sets on $\mathbb{R}^d$, and observe that $C(A, \cdot) \ll \lambda$ for each $A \in \mathscr{F}$. The Radon–Nikodým derivative $P^x(A) = \frac{dC(A,\cdot)}{d\lambda}(x)$ has a version which is a probability measure on $(\Omega, \mathscr{F})$ and is called *Palm probability*. If $M$ is a simple point process (see ►Point Processes), i.e. $M(\omega, \cdot)$ takes values in $\mathbb{Z}_+$ such that $M(\omega, \{x\}) \in \{0, 1\}$, for all $x$ and $\omega$, then $P^x(A)$ gives the probability of $A$ given that $M$ places a unit mass at the point $x$. The concept is most useful for stationary random measures (Kallenberg 2002).

## About the Authors

Takis Konstantopoulos is Professor of Probability in the School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh. He is also a member of the Maxwell Institute for Mathematical Sciences. He received a BSc from the National Technical University of Athens, Greece and a M.Sc. and Ph.D. (1989) from the University of California, Berkeley, USA. He has held a charge de recherche position in INRIA, Sophia Antipolis, France. He has served as a faculty member in the University of Texas at Austin in Electrical Engineering and in Mathematics. He has been Professor of Mathematics (2002–2005) at the University of Patras Greece where he also served as Director of the Probability and Statistics division. His research interests include stochastic processes, limit theorems, stochastic networks, applied probability and pure mathematics. He is advisor of the Centre for Education in Sciences (Patras, Greece). He is an organizer of an international program on "stochastic processes in communication sciences" at the Newton Institute for Mathematical Sciences, Cambridge, UK, January–July 2010.

Professor Zerakidze was Head of the Higher Mathematics department of Gori State University. He was also Head of a Mathematical Society of Gori region (2005–2007). He has been awarded with the Order of Honour by the President of Georgia (March 17th 2000). His work "The divisible family of measures" was included in the Big Russian Encyclopedia in the section "Probability and the mathematical statistics" page 533, Moscow, 1999.

Professor Grigol Sokhadze received his Ph.D. in Mathematics in 1992. He has (co-)authored about 100 papers in probability and statistics.

## Cross References

►Brownian Motion and Diffusions
►Conditional Expectation and Probability
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Kullback-Leibler Divergence
►Martingales
►Measure Theory in Probability
►Poisson Processes
►Stochastic Processes: Applications in Finance and Insurance

## References and Further Reading

Bell D (1991) Transformations of measures on an infinite-dimensional vector space. Seminar on stochastic processes (Vancouver 1990). Prog Probab 24:15–25
Bogachev V (2008) Differentiable measures and the malliavin calculus (in Russian). R & C Dynamics, Moscow
Cameron RH, Martin WT (1944) Transformation of Wiener integrals under translations. Ann Math 45:386–396
Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York
Daletskii J, Sokhadze G (1988) Absolute continuity of smooth measures (in Russian). Funct Anal Appl 22(2):77–88
Feldman I (1958) Equivalence and perpendicularity of Gaussian processes. Pac J Math 8:699–708
Gikhman I, Skorokhod A (1971–1975) Theory of stochastic processes, vol 1–3. Nauka, Moscow (in Russian)
Ibragimov I, Rozanov J (1970) Gaussian random processes (in Russian). Nauka, Moscow
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)
Konstantopoulos T (2009) Conditional expectation and probability. This encyclopedia

Kulik A, Pilipenko A (2000) Nonlinear transformations of smooth measures in infinite-dimensional spaces. Ukran Math J 52(9):1226–1250
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Kyprianou AE (2006) Introductory lectures on fluctuations of Lévy processes with applications. Springer, Heidelberg
Liptser R, Shiryaev A (1974) Statistics of random processes (in Russian). Nauka, Moscow
Nikodým O (1930) Sur une généralisation des intégrales de M. J Radon Fundamenta Mathematicae 15:131–179
Radon J (1913) Theorie und Anwendungen der absolut additiven Mengenfunktionen. Sitzber, der Math.Naturwiss. Klasse der Kais. Akademie der Wiss. Wien, 112 Bd. Abt II a/2
Williams D (1989) Probability with martingales. Cambridge University Press, Cambridge
Yadrenko M (1980) Spectral theory of random fields (in Russian). Visha Shkola, Kiev
Zerakidze Z (1969) On the equivalence of distributions of Gaussian fields (in Russian). In: Proceedings of the Tbilisi institute of applied mathematics. Tbilisi, vol 2, pp 215–220

# Random Coefficient Models

Nicholas T. Longford
Universitat Pompeu Fabra, Barcelona, Spain

Independence of the observations is a key assumption of many standard statistical methods, such as ▶analysis of variance (ANOVA) and ordinary regression, and some of its extensions. Common examples of data structures that do not fit into such a framework arise in longitudinal analysis, in which observations are made on subjects at subject-specific sequences of time points, and in studies that involve subjects (units) ocurring naturally in clusters, such as individuals within families, schoolchildren within classrooms, employees within companies, and the like. The assumption of independence of the observations is not tenable, because observations within a cluster are likely to be more similar than observations in general. Such similarity can be conveniently represented by a positive correlation (dependence).

This section describes an adaptation of the ordinary regression for clustered observations. Such observations require two indices, one for elements within clusters, $i = 1, \ldots, n_j$, and another for clusters, $j = 1, \ldots, m$. Thus, we have $n = n_1 + \cdots + n_m$ elementary units and $m$ clusters. The ordinary regression model

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, \qquad (1)$$

with the usual assumptions of normality, independence and equal variance (homoscedasticity) of the deviations $\varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, i.i.d., implies that the regressions within the clusters $j$ have a common vector of coefficients $\boldsymbol{\beta}$. This restriction can be relaxed by allowing the regressions to differ in their intercepts. A practical way of defining such a model is by the equation

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j + \varepsilon_{ij}, \qquad (2)$$

where $\delta_j$ is a random sample from a centred normal distribution, $\delta_j \sim \mathcal{N}\left(0, \sigma_B^2\right)$, i.i.d., independent from the $\varepsilon$'s. With this model, the within-cluster regressions are parallel; their intercepts are $\beta_0 + \delta_j$, but the coefficients on all the other variables in $\mathbf{x}$ are common to the clusters. A more appealing interpretation of the model is that observations in a cluster are correlated,

$$\mathrm{cor}\left(y_{i_1,j}, y_{i_2,j}\right) = \frac{\sigma_B^2}{\sigma^2 + \sigma_B^2},$$

because they share the same deviation $\delta_j$. Further relaxation of how the within-cluster regressions differ is attained by allowing some (or all) the regression slopes to be specific to the clusters. We select a set of variables in $\mathbf{x}$, denoted by $\mathbf{z}$, and assume that the regressions with respect to these variables differ across the clusters, but are constant with respect to the remaining variables;

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\boldsymbol{\delta}_j + \varepsilon_{ij}, \qquad (3)$$

where $\boldsymbol{\delta}_j$ is a random sample from a multivariate normal distribution (see ▶Multivariate Normal Distributions) $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_B)$, independent from the $\varepsilon$'s. We say that the variables in $\mathbf{z}$ are associated with (cluster-level) variation. The variance of an observation $y_{ij}$, without conditioning on the cluster $j$, is

$$\mathrm{var}\left(y_{ij}\right) = \sigma^2 + \mathbf{x}_{ij}\boldsymbol{\Sigma}_B\mathbf{x}_{ij}^{\top}.$$

We refer to $\sigma^2$ and $\mathbf{z}_{ij}\boldsymbol{\Sigma}_B\mathbf{z}_{ij}^{\top}$ as the *variance components* (at the elementary and cluster levels, respectively). The principle of invariance with respect to linear transformations of $\mathbf{z}$ implies that the intercept should always be included in $\mathbf{z}$, unless $\mathbf{z}$ is empty, as in the model in (1). The function $V(\mathbf{z}) = \mathbf{z}\boldsymbol{\Sigma}_B\mathbf{z}^{\top}$, over the feasible values of $\mathbf{z}$, defines the *pattern* of variation, and it can be described by its behaviour (local minima, points of inflection, and the like). By way of an example, suppose $\mathbf{z}$ contains the intercept and a single variable $z$. Denote the variances in $\boldsymbol{\Sigma}_B$ by $\sigma_0^2$ and $\sigma_z^2$, and the covariance by $\sigma_{0z}$. Then

$$V(\mathbf{z}) = \sigma_0^2 + 2z\sigma_{0z} + z^2\sigma_z^2, \qquad (4)$$

and this quadratic function has a unique minimum at $z^* = -\sigma_{0z}/\sigma_z^2$, unless $\sigma_z^2 = 0$, in which case we revert to the model in (2) in which $V(\mathbf{z})$ is constant.

The model in (3) is fitted by maximum likelihood (ML) which maximizes the log-likelihood function

$$l\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_B\right) = -\frac{1}{2}\sum_{j=1}^{m}\Big[\log\left\{\det\left(\mathbf{V}_j\right)\right\} + \left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)^{\top}\mathbf{V}_j^{-1}\left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)\Big],$$

in which $\mathbf{V}_j$ is the variance matrix of the observations in cluster $j$, $\mathbf{y}_j$ the vector of the outcomes for the observations in cluster $j$, and $\mathbf{X}_j$ the corresponding regression design matrix formed by vertical stacking of the rows $\mathbf{x}_{ij}$, $i = 1, \ldots, n_j$. The variation design matrices $\mathbf{Z}_j, j = 1, \ldots, m$, are defined similarly; with them, $\mathbf{V}_j = \sigma^2\mathbf{I}_{n_j} + \mathbf{Z}_j\boldsymbol{\Sigma}_B\mathbf{Z}_j^{\top}$, where $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix. For ML solutions, see Longford (1993) and Goldstein (2000). These and other algorithms are implemented in most standard statistical packages.

►Model selection entails two tasks, selecting a set of variables to form $\mathbf{x}$ and selecting its subset to form $\mathbf{z}$. The variables in $\mathbf{x}$ can be defined for elements or clusters; the latter can be defined as being constant within clusters. Inclusion of cluster-level variables in $\mathbf{z}$ does not have an interpretation in terms of varying regression coefficients, so associating them with variation is in most contexts not meaningful. However, the identity in (4) and its generalisations for $\boldsymbol{\Sigma}_B$ with more than two rows and columns indicate that $\mathbf{z}$ can be used for modelling variance heterogeneity. The likelihood ratio test statistic and various information criteria can be used for selecting among alternative models, so long as one is a submodel of the other; that is, the variables in both $\mathbf{x}$ and $\mathbf{z}$ of one model are subsets of (or coincide with) their counterparts in the other model.

Random coefficients can be applied to a range of models much wider than ordinary regression. In principle, we can conceive any *basis model*, characterized by a vector of parameters, which applies to every cluster. A subset of these parameters is constant across the clusters and the remainder varies according to a model for cluster-level variation. The latter model need not be a multivariate normal distribution, although suitable alternatives to it are difficult to identify. The basis model itself can be complex, such as a random coefficient model itself. This gives rise to three- or, generally, *multilevel models*, in which elements are clustered within two-level units, these units in three-level units, and so on. Generalized linear mixed models have ►generalized linear models (McCullagh and Nelder 1989) as their basis.

Random coefficient models are well suited for analysing surveys in which clusters arise naturally as a consequence of the organisation (design) of the survey and the way the studied population is structured. They can be applied also in settings in which multiple observations are made on subjects, as in longitudinal studies (Molenberghs and Verbeke 2000). In some settings it is contentious as to whether the clusters should be regarded as fixed or random. When they are assumed to be random the (random coefficient) models are often more parsimonious than their fixed-effects (ANCOVA) models, because the number of parameters involved does not depend on the number of clusters.

## About the Author

Dr. Longford has been a visiting lecturer and visiting Associate Professor in Spain, Sudan, Germany, Denmark, USA, Brazil and New Zealand. He was a President of the Princeton-Trenton (NJ) Chapter of ASA.

## Cross References

►Cross Classified and Multiple Membership Multilevel Models
►Linear Mixed Models
►Multilevel Analysis
►Sensometrics
►Statistical Analysis of Longitudinal and Correlated Data
►Testing Variance Components in Mixed Linear Models

## References and Further Reading

Goldstein H (2000) Multilevel statistical models, 2nd edn. Edward Arnold, London
Longford NT (1993) Random coefficient models. Oxford University Press, Oxford
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

# Random Field

Mikhail P. Moklyachuk
Professor
Kyiv National Taras Shevchenko University, Kyiv, Ukraine

*Random field* $X(t)$ on $D \subset \mathbb{R}^n$ (i.e., $t \in D \subset \mathbb{R}^n$) is a function whose values are random variables for any $t \in D$. The dimension of the coordinate is usually in the range from one to four, but any $n > 0$ is possible. A one-dimensional

random field is usually called a stochastic process. The term "random field" is used to stress that the dimension of the coordinate is higher than one. Random fields in two and three dimensions are encountered in a wide range of sciences and especially in the earth sciences, such as hydrology, agriculture, and geology. Random fields where $t$ is a position in space-time are studied in turbulence theory and in meteorology.

Random field $X(t)$ is described by its finite-dimensional (cumulative) distributions

$$F_{t_1,\ldots,t_k}(x_1,\ldots,x_k) = P\{X(t_1) < x_1,\ldots,X(t_k) < x_k\}, k = 1,2\ldots$$

The cumulative distribution functions are by definition left-continuous and nondecreasing. Two requirements on the finite-dimensional distributions must be satisfied. The symmetry condition

$$F_{t_1,\ldots,t_k}(x_1,\ldots,x_k) = F_{t_{\pi 1},\ldots,t_{\pi k}}(x_{\pi 1},\ldots,x_{\pi k}),$$

where $\pi$ is a permutation of the index set $\{1,\ldots,k\}$. The compatibility condition

$$F_{t_1,\ldots,t_{k-1}}(x_1,\ldots,x_{k-1}) = F_{t_1,\ldots,t_k}(x_1,\ldots,x_{k-1},\infty).$$

*Kolmogorov Existence Theorem* states: If a system of finite-dimensional distributions $F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)$, $k = 1,2\ldots$, satisfies the symmetry and compatibility conditions, then there exists on some probability space a random field $X(t)$, $t \in D$, having $F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)$, $k = 1,2\ldots$, as its finite-dimensional distributions.

The expectation (mean value) of a random field is by definition the Stieltjes integral

$$m(t) = EX(t) = \int_{\mathbb{R}^1} x dF_t(x).$$

The (auto-)covariance function is also expressed as the Stieltjes integral

$$B(t,s) = E(X(t)X(s)) - m(t)m(s) = \iint_{\mathbb{R}^2} xy dF_{t,s}(x,y) - m(t)m(s),$$

whereas the variance is $\sigma^2(t) = B(t,t)$.

*Gaussian random fields* play an important role due to several reasons: the specification of their finite-dimensional distributions is simple, they are reasonable models for many natural phenomena, and their estimation and inference are simple.

A Gaussian random field is a random field where all the finite-dimensional distributions are ▶multivariate normal distributions. Since multivariate normal distributions are completely specified by expectations and covariances, it suffices to specify $m(t)$ and $B(t,s)$ in such a way that the symmetry condition and the compatibility condition hold true. The expectation can be arbitrarily chosen, but the covariance function must be positive-definite to ensure the existence of all finite-dimensional distributions.

*Wiener sheet* (*Brownian sheet*) is a Gaussian random field $W(t)$, $t = (t_1,t_2) \in \mathbb{R}^2_+$ with $EW(t) = 0$ and correlation function $B(t,s) = E(X(t)X(s)) = \min\{s_1,t_1\}\min\{s_2,t_2\}$. Analogously, the $n$-parametric Wiener process is a Gaussian random field $W(t)$, $t \in \mathbb{R}^n_+$ with $EW(t) = 0$ and correlation function $B(t,s) = \prod_{i=1}^n \min(s_i,t_i)$. The multiparametric Wiener process $W(t)$ has independent homogeneous increments. A generalized derivative of the multiparametric Wiener process $W(t)$ is the *Gaussian white noise process* on $\mathbb{R}^n_+$ (Chung and Walsh 2005).

*Poisson random fields* are also reasonable models for many natural phenomena. A Poisson random field is an integer-valued (point) random field where the (random) amount of points that belong to a bounded set from the range of values of the field has a Poisson distribution and the random amounts of points that belong to nonoverlapping sets are mutually independent (Kerstan et al. 1974).

*Markov random field* $X(t)$, $t \in D \subset \mathbb{R}^n$, is a random function that has the Markov property with respect to a fixed system of ordered triples $(S_1,\Gamma,S_2)$ of nonoverlapping subsets from the domain of definition $D$. The Markov property means that for any measurable set $B$ from the range of values of the function $X(t)$ and every $t_0 \in S_2$, the following equality holds true:

$$P\{X(t_0) \in B | X(t), t \in S_1 \cup \Gamma\} = P\{X(t_0) \in B | X(t), t \in \Gamma\}.$$

This means that the future $S_2$ does not depend on the past $S_1$ when the present $\Gamma$ is given. Let, for example, $D = \mathbb{R}^n$, $\{\Gamma\}$ be a family of all spheres in $\mathbb{R}^n$, $S_1$ be the interior of $\Gamma$, and $S_2$ be the exterior of $\Gamma$. A homogeneous and isotropic Gaussian random field $X(t)$, $t \in \mathbb{R}^n$, has the Markov property with respect to the ordered triples $(S_1,\Gamma,S_2)$ if and only if $X(t) = \xi$, where $\xi$ is a random variable. Nontrivial examples of homogeneous and isotropic Markov random fields can be constructed when considering the generalized random fields. Markov random fields are completely described in the class of homogeneous Gaussian random fields on $\mathbb{Z}^n$, in the class of multidimensional homogeneous generalized Gaussian random fields on the space $\mathbb{C}_0^\infty(\mathbb{R}^m)$ and the class of multidimensional homogeneous and isotropic generalized Gaussian random fields (Glimm and Jaffe 1981; Rozanov 1982; Yadrenko 1983).

*Gibbs random fields* form a class of random fields that have extensive applications in solutions of problems in statistical physics. The distribution functions of these

fields are determined by Gibbs distribution (Malyshev and Minlos 1985).

*Homogeneous random field* in the strict sense is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$ (or $t \in \mathbb{Z}^n$), where all its finite-dimensional distributions are invariant under arbitrary translations, that is,

$$F_{t_1+s,\ldots,t_k+s}(x_1,\ldots,x_k) = F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)\ \forall s \in \mathbb{R}^n.$$

*Homogeneous random field* in the wide sense is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$ ($t \in \mathbb{Z}^n$), $E|X(t)|^2 < +\infty$, where $EX(t) = m = \text{const.}$ and the correlation function $EX(t)X(s) = B(t-s)$ depends on the difference $t-s$ of coordinates of points $t$ and $s$.

Homogeneous random field $X(t)$, $t \in \mathbb{R}^n$, $EX(t) = 0$, $E|X(t)|^2 < +\infty$, and its correlation function $B(t) = EX(t+s)X(s)$ admit the spectral representations

$$X(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} Z(d\lambda),$$

$$B(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} F(d\lambda),$$

where $F(d\lambda)$ is a measure on the Borel $\sigma$-algebra $B_n$ of sets from $\mathbb{R}^n$, and $Z(d\lambda)$ is an orthogonal random measure on $B_n$ such that $EZ(S_1)Z(S_2) = F(S_1 \cap S_2)$. The integration range is $\mathbb{R}^n$ in the case of continuous time random field $X(t)$, $t \in \mathbb{R}^n$, and $[-\pi,\pi]^n$ in the case of discrete time random field $X(t)$, $t \in \mathbb{Z}^n$. In the case where the spectral representation of the correlation function is of the form

$$B(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} f(\lambda)d\lambda,$$

the function $f(\lambda)$ is called the spectral density of the field $X(t)$. Based on these spectral representations we can prove, for example, the *law of large numbers* for random field $X(t)$:

The mean square limit

$$\lim_{N\to\infty} \frac{1}{(2N+1)^n} \sum_{|t_i|\le N, i=1,\ldots,n} X(t) = Z\{0\}.$$

This limit is equal to $EX(t) = 0$ if and only if $E|Z\{0\}|^2 = F\{0\}$. In the case where $F\{0\} = 0$ and

$$\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \prod_{i=1}^{n} \log\left|\log\frac{1}{|\lambda_i|}\right| F(d\lambda) < +\infty,$$

the *strong law of large numbers* holds true for the random field $X(t)$.

*Isotropic random field* is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$, $E|X(t)|^2 < +\infty$, where the expectation and the correlation function have properties $EX(t) = EX(gt)$ and $EX(t)X(s) = EX(gt)X(gs)$ for all rotations $g$ around the origin of coordinates. An isotropic random field $X(t)$ admits the decomposition

$$X(t) = \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} X_m^l(r) S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi),$$

where $(r,\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi)$ are spherical coordinates of the point $t \in \mathbb{R}^n$, $S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi)$ are spherical harmonics of the degree $m$, $h(m,n)$ is the amount of such harmonics, $X_m^l(r)$ are uncorrelated stochastic processes such that $EX_m^l(r)X_{m_1}^{l_1}(s) = b_m(r,s)\delta_m^{m_1}\delta_l^{l_1}$, where $\delta_i^j$ is the Kronecker symbol, $b_m(r,s)$ is a sequence of positive definite kernels such that $\sum_{m=0}^{\infty} h(m,n)b_m(r,s) < +\infty$, $b_m(0,s) = 0, m \ne 0$.

Isotropic random field $X(t)$, $t \in \mathbb{R}^2$, on the plane admits the decomposition

$$X(r,\varphi) = \sum_{m=0}^{\infty} \left\{X_m^1(r)\cos(m\varphi) + X_m^2(r)\sin(m\varphi)\right\}.$$

The class of isotropic random fields includes homogeneous and isotropic random fields, multiparametric Brownian motion processes (see ▶Brownian Motion and Diffusions).

*Homogeneous and isotropic random field* is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$, $E|X(t)|^2 < +\infty$, where the expectation $EX(t) = c = \text{const.}$ and the correlation function $EX(t)X(s) = B(|t-s|)$ depends on the distance $|t-s|$ between points $t$ and $s$. Homogeneous and isotropic random field $X(t)$ and its correlation function $B(r)$ admit the spectral representations (Rozanov 1982; Yadrenko 1983; Yaglom 1987)

$$X(t) = c_n \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi) \int_0^{\infty} \frac{J_{m+(n-2)/2}(r\lambda)}{(r\lambda)^{(n-2)/2}} Z_m^l(d\lambda),$$

$$B(r) = \int_0^{\infty} Y_n(r\lambda)d\Phi(\lambda),$$

where

$$Y_n(x) = 2^{(n-2)/2}\Gamma\left(\frac{n}{2}\right)\frac{J_{(n-2)/2}(x)}{x^{(n-2)/2}}$$

is a spherical Bessel function, $\Phi(\lambda)$ is a bounded nondecreasing function called the spectral function of the field $X(t)$, $Z_m^l(d\lambda)$ are random measures with orthogonal values such that $EZ_m^l(S_1)Z_{m_1}^{l_1}(S_2) = \delta_m^{m_1}\delta_l^{l_1}\Phi(S_1 \cap S_2)$, $c_n^2 = 2^{n-1}\Gamma(n/2)\pi^{n/2}$.

Homogeneous and isotropic random field $X(t), t \in \mathbb{R}^2$, on the plane admits the spectral representation

$$X(t,\varphi) = \sum_{m=0}^{\infty} \cos(m\varphi) Y_m(r\lambda) Z_m^1(d\lambda) + \sum_{m=1}^{\infty} \sin(m\varphi) Y_m(r\lambda) Z_m^2(d\lambda).$$

These spectral decompositions of random fields form a power tool for the solution of statistical problems for random fields such as extrapolation, interpolation, filtering, and estimation of parameters of the distribution (Yadrenko 1983; Yaglom 1987a, b).

## About the Author

Dr. Mikhail P. Moklyachuk is a Professor of the Department of Probability Theory, Statistics and Actuarial Mathematics, Kyiv National Taras Shevchenko University, Ukraine. He is the author and coauthor of more than 100 papers and six books, including *Robust estimates for functionals of stochastic processes* (Kyiv University Press, 2008). Professor Moklyachuk has received the Taras Shevchenko prize (Kyiv University best textbook award, 1999) for the textbook *Variational Calculus. Extremum Problems*. He is the editor of the Cooperation Unit of Zentralblatt MATH (Zentralblatt fuer Mathematik/Mathematics Abstracts), coeditor of *Current Index to Statistics*, and member of the editorial board, *Theory of Probability and Mathematical Statistics*.

## Cross References

►Estimation Problems for Random Fields
►Measure Theory in Probability
►Model-Based Geostatistics
►Random Variable
►Spatial Statistics
►Stochastic Processes

## References and Further Reading

Chung KL, Walsh JB (2005) Markov processes, Brownian motion, and time symmetry, 2nd ed. Springer, New York, NY
Glimm J, Jaffe A (1981) Quantum physics: a functional integral point of view. Springer, Berlin/Heidelberg/New York
Kerstan J, Matthes K, Mecke J (1974) Mathematische Lehrbücher und Monographien. II. Abt. Mathematische Monographien. Band XXVII. Akademie, Berlin
Malyshev VA, Minlos RA (1985) Stochastic Gibbs fields. The method of cluster expansions. Nauka, Moskva
Monin AS, Yaglom AM (2007a) Statistical fluid mechanics: mechanics of turbulence, volume I. Edited and with a preface by Lumley JL, Dover, Mineola, NY
Monin AS, Yaglom AM (2007b) Statistical fluid mechanics: mechanics of turbulence, volume II. Edited and with a preface by Lumley JL, Dover, Mineola, NY
Rozanov YuA (1982) Markov random fields. Springer, New York
Yadrenko MI (1983) Spectral theory of random fields. Translation Series in Mathematics and Engineering. Optimization Software, Publications Division, New York; Springer, New York
Yaglom AM (1987a) Correlation theory of stationary and related random functions. volume I. Basic results. Springer Series in Statistics. Springer, New York
Yaglom AM (1987b) Correlation theory of stationary and related random functions, volume II. Supplementary notes and references. Springer Series in Statistics. Springer, New York

# Random Matrix Theory

Jack W. Silverstein
Professor
North Carolina State University, Raleigh, NC, USA

Random matrix theory (RMT) originated from the investigation of energy levels of a large number of particles in quantum mechanics. Many laws were discovered by numerical study in mathematical physics. In the late 1950s, E. P. Wigner formulated the problem in terms of the empirical distribution of a random matrix (Wigner 1955, 1958), which began the investigation into the semicircular law of Gaussian matrices. Since then, RMT has formed an active branch in modern probability theory.

## Basic Concepts

Let $\mathbf{A}$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$. If all $\lambda_j$s are real, then we can construct a 1-dimensional empirical distribution function

$$F^{\mathbf{A}}(x) = \frac{1}{n} \sum_{j=1}^{n} I(\lambda_j \le x),$$

otherwise, we may construct a 2-dimensional empirical distribution function by the real and imaginary parts of $\lambda_j$, i.e.

$$F^{\mathbf{A}}(x,y) = \frac{1}{n} \sum_{j=1}^{n} I(\Re(\lambda_j) \le x; \Im(\lambda_j) \le y).$$

Then, $F^{\mathbf{A}}$ is called the *empirical spectral distribution* (ESD) of $\mathbf{A}$. The main task of RMT is to investigate limiting properties of $F^{\mathbf{A}}$ in the case where $\mathbf{A}$ is random and the order $n$ tends to infinity. If there is a limit distribution $F$, then the limit is called the *limiting spectral distribution* (LSD) of the sequence of the $\mathbf{A}$. Interesting problems include finding the explicit forms of the LSD if it exists and to investigate its properties.

There are two methods used in determining limiting properties of $F^{\mathbf{A}}$ (Bai 1999). One is the *method of moments*, using the fact that the moments of $F^{\mathbf{A}}$ are the scaled traces

of powers of **A**. The other is using *Stieltjes transforms*, defined for any distribution function $F$ as

$$m(z) = \int \frac{1}{x - z} dF(x),$$

for $z \in \mathbb{C}$.

Contrary to the progress made on the eigenvalues of large dimensional random matrices, very few results have been obtained on the limiting properties of the eigenmatrix (i.e., the matrix of the standardized eigenvectors of **A**). Due to its importance in the application to statistics and applied areas, investigation on eigenmatrices is becoming more active.

## Limiting Spectral Distributions

1. **Semicircular Law** A Wigner matrix is defined as a Hermitian (symmetric if real) matrix $\mathbf{W} = (w_{ij})_{n\times n}$ whose entries above or on the diagonal are independent. Then the ESD of $n^{-1/2}\mathbf{W}$ tends to the semicircular law with density

$$p(x) = \frac{1}{2\pi}\sqrt{4 - x^2} I(|x| < 2),$$

if $\mathrm{E}w_{ij} = 0$, $\mathrm{E}|w_{ij}|^2 = 1$ and for any $\delta > 0$,

$$\frac{1}{n^2}\sum_{ij} \mathrm{E}\left|w_{ij}^2\right| I\left(|w_{ij}| \geq \delta\sqrt{n}\right) \to 0.$$

2. **Marčenko–Pastur Law** Let $\mathbf{X} = (x_{ij})_{p\times n}$ whose entries are independent random variables with mean zero and variance 1. If $p/n \to y \in (0, \infty)$ and for any $\delta > 0$,

$$\frac{1}{np}\sum_{ij} \mathrm{E}\left|x_{ij}^2\right| I\left(|x_{ij}| \geq \delta\sqrt{n}\right) \to 0.$$

Then the ESD of $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^*$ (so-called sample covariance matrix) tends to the Marčenko–Pastur law with density

$$\frac{1}{2\pi xy}\sqrt{(b - x)(x - a)} I(a < x < b)$$

where $a = (1 - \sqrt{y})^2$ and $b = (1 + \sqrt{y})^2$. Furthermore, if $y > 1$, the LSD has a point mass $1 - 1/y$ at the origin.

3. **LSD of Products of Random Matrices** Let $\mathbf{T}$ $(p \times p)$ be a Hermitian matrix with LSD H (a probability distribution function) and $\mathbf{S}_n$, $p/n$ satisfy the conditions in item (2). Then the ESD of $\mathbf{S}_n\mathbf{T}$ exists and the Stieltjes transform $m(z)$ is the unique solution on the upper complex plane to the equation

$$m = \int \frac{1}{t(1 - y - yzm) - z} dH(t),$$

where $z$ is complex with positive imaginary part.

## Extreme Eigenvalues and Spectrum Separation

Limits of extreme eigenvalues of large random matrices is one of the important topics. In many cases, under the assumption of finite fourth moment, the extreme eigenvalues almost surely tend to the respective boundaries of the LSD. For the product $\mathbf{S}_n\mathbf{T}$, if the support of the LSD is disconnected, then, under certain conditions, it is proved that there are no eigenvalues among the gaps and the numbers on each side are exactly the same of eigenvalues of $\mathbf{T}$, on the corresponding sides of the interval which determines the gap of the LSD (Bai and Silverstein 1999).

Further deeper investigation into extreme eigenvalues is the Tracy–Widom Law which says that $n^{2/3}$ times the difference of the extreme eigenvalues and the corresponding boundary points tends to the so-called Tracy–Widom law (Tracy and Widom 1994).

## Convergence Rates of Empirical Spectral Distributions

Convergence rates of ESDs of large dimensional random matrices to their corresponding LSDs are important for application of spectral theory of large dimensional matrix. Bai inequality is the basic mathematical tool to establish the convergence rates (Bai 1993a,b). The currently known best rates are that $O(n^{-1/2})$ for the expected ESDs for Wigner matrix and for sample covariance matrix, and $O_p(n^{-2/5})$ and $O_{a.s.}(n^{-2/5+\eta})$ for their ESDs.

The exact rates are still far from known.

## CLT of LSS

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the random matrix **A** and $f$ is a function defined on the space of the eigenvalues, then the LSS (linear spectral statistic) for the random matrix is defined by

$$\frac{1}{n}\sum_{k=1}^{n} f(\lambda_k) = \int f(x) dF^{\mathbf{A}}(x).$$

To investigate the limiting distribution of the LSS, we define $X_n(f) = n\left(\int f(x) d\left(F^{\mathbf{A}}(x) - F(x)\right)\right)$.

Under certain conditions, the normalized LSS, $X_n(f)$, is proved to tend to a normal distribution for the Wigner matrix, the product $\mathbf{S}_n\mathbf{T}$, as well as for the multivariate $F$-matrix, with asymptotic means and variances explicitly expressed by the Stieltjes transforms of the LSDs (Bai and Yao 2005; Bai and Silverstein 2004; Zheng 2010).

These theorems have been found to have important applications to multivariate analysis and many other areas.

## Limiting Properties of Eigenvectors

Work in this area has been primarily done on the matrices in item (2) with **X** containing real entries (Silverstein

1979, 1984, 1990). Write $\mathbf{S}_n = \mathbf{O}\Lambda\mathbf{O}^*$, its spectral decomposition. When the entries of $\mathbf{X}$ are Gaussian, then $\mathbf{S}_n$ is the standard Wishart matrix, with $\mathbf{O}$ Haar-distributed in the group of $p \times p$ orthogonal matrices. The question is to compare the distribution of $\mathbf{O}$ when the entries of $\mathbf{X}$ are not Gaussian to Haar measure when $p$ is large. This has been pursued when $\mathbf{X}$ is made up of iid random variables, by comparing the distribution of $\mathbf{y} = \mathbf{O}^*\mathbf{x}$, where $\mathbf{x}$ is a unit $p$-dimensional vector, to the uniform distribution on the unit sphere in $\mathbb{R}^p$. A stochastic process is defined in terms of the entries of $\mathbf{y}$, which converges weakly to Brownian bridge in the Wishart case. A necessary condition for this process to behave the same way for non Gaussian entries has been shown to be $\mathrm{E}\left(x_{11}^4\right) = 3$, matching the fourth moment of a standardized Gaussian (Silverstein 1984). For certain choices of $\mathbf{x}$ and for symmetrically distributed $x_{11}$, weak convergence to Brownian bridge has been shown in Silverstein (1990).

## About the Author

Professor Silverstein was named IMS Fellow for "seminal contributions to the theory and application of random matrices" (2007). He has (co-)authored over 50 publications, including the book *Spectral Analysis of Large Dimensional Random Matrices* (with Z.D. Bai, 2nd edition, Springer, New York, 2009).

## Cross References

►Eigenvalue, Eigenvector and Eigenspace
►Ergodic Theorem
►Limit Theorems of Probability Theory
►Multivariate Statistical Distributions
►Statistical Inference for Quantum Systems

## References and Further Reading

Bai ZD (1999) Methodologies in spectral analysis of large dimensional random matrices: a review. Stat Sinica 9(3):611–677
Bai ZD (1993a) Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices. Ann Probab 21(2):625–648
Bai ZD (1993b) Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices. Ann Probab 21(2):649–672
Bai ZD, Silverstein JW (2004) CLT for linear spectral statistics of large-dimensional sample covariance matrices. Ann Probab 32(1):553–605
Bai ZD, Silverstein JW (1999) Exact separation of eigenvalues of large dimensional sample covariance matrices. Ann Probab 27(3):1536–1555
Bai ZD, Yao JF (2005) On the convergence of the spectral empirical process of Wigner matrices. Bernoulli 11(6):1059–1092
Silverstein JW (1979) On the randomness of eigenvectors generated from networks with random topologies. SIAM J Appl Math 37:235–245
Silverstein JW (1984) Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. J Multivariate Anal 15(3):295–324
Silverstein JW (1990) Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. Ann Probab 18:1174–1194
Tracy CA, Widom H (1994) Level-spacing distributions and the Airy kernel. Commun Math Phys 159:151–174
Wigner EP (1955) Characteristic vectors bordered matrices with infinite dimensions. Ann Math 62:548–564
Wigner EP (1958) On the distributions of the roots of certain symmetric matrices. Ann Math 67:325–327
Zheng S (2010) Central limit theorem for linear spectral statistics of large dimensional F-Matrix, to appear in Ann Inst Henri Poincaré Probab Stat

# Random Permutations and Partition Models

Peter McCullagh
John D. MacArthur Distinguished Service Professor
University of Chicago, Chicago, IL, USA

## Set Partitions

For $n \geq 1$, a partition $B$ of the finite set $[n] = \{1, \ldots, n\}$ is

- A collection $B = \{b_1, \ldots\}$ of disjoint non-empty subsets, called blocks, whose union is $[n]$
- An equivalence relation or Boolean function $B\colon [n] \times [n] \to \{0,1\}$ that is reflexive, symmetric and transitive
- A symmetric Boolean matrix such that $B_{ij} = 1$ if $i,j$ belong to the same block

These equivalent representations are not distinguished in the notation, so $B$ is a set of subsets, a matrix, a Boolean function, or a subset of $[n]\times[n]$, as the context demands. In practice, a partition is sometimes written in an abbreviated form, such as $B = 2|13$ for a partition of $[3]$. In this notation, the five partitions of $[3]$ are

$$123, \quad 12|3, \quad 13|2, \quad 23|1, \quad 1|2|3.$$

The blocks are unordered, so $2|13$ is the same partition as $13|2$ and $2|31$.

A partition $B$ is a sub-partition of $B^*$ if each block of $B$ is a subset of some block of $B^*$ or, equivalently, if $B_{ij} = 1$ implies $B^*_{ij} = 1$. This relationship is a partial order denoted by $B \leq B^*$, which can be interpreted as $B \subset B^*$ if each partition is regarded as a subset of $[n]^2$. The partition lattice $\mathcal{E}_n$ is the set of partitions of $[n]$ with this partial order. To each pair of partitions $B, B'$ there corresponds a greatest lower bound $B \wedge B'$, which is the set intersection or Hadamard component-wise matrix product. The least upper bound

$B \vee B'$ is the least element that is greater than both, the transitive completion of $B \cup B'$. The least element of $\mathcal{E}_n$ is the partition $\mathbf{0}_n$ with $n$ singleton blocks, and the greatest element is the single-block partition denoted by $\mathbf{1}_n$.

A permutation $\sigma\colon [n] \to [n]$ induces an action $B \mapsto B^\sigma$ by composition such that $B^\sigma(i,j) = B(\sigma(i), \sigma(j))$. In matrix notation, $B^\sigma = \sigma B \sigma^{-1}$, so the action by conjugation permutes both the rows and columns of $B$ in the same way. The block sizes are preserved and are maximally invariant under conjugation. In this way, the 15 partitions of $[4]$ may be grouped into five orbits or equivalence classes as follows:

$$1234, \quad 123|4\,[4], \quad 12|34\,[3], \quad 12|3|4\,[6], \quad 1|2|3|4.$$

Thus, for example, 12|34 is the representative element for one orbit, which also includes 13|24 and 14|23.

The symbol $\#B$ applied to a set denotes the number of elements, so $\#B$ is the number of blocks, and $\#b$ is the size of block $b \in B$. If $\mathcal{E}_n$ is the set of equivalence relations on $[n]$, or the set of partitions of $[n]$, the first few values of $\#\mathcal{E}_n$ are 1, 2, 5, 15, 52, called Bell numbers. More generally, $\#\mathcal{E}_n$ is the $n$th moment of the unit Poisson distribution whose exponential generating function is $\exp(e^t - 1)$. In the discussion of explicit probability models on $\mathcal{E}_n$, it is helpful to use the ascending and descending factorial symbols

$$\alpha^{\uparrow r} = \alpha(\alpha+1)\cdots(\alpha+r-1) = \Gamma(r+\alpha)/\Gamma(\alpha)$$
$$k^{\downarrow r} = k(k-1)\cdots(k-r+1)$$

for integer $r \geq 0$. Note that $k^{\downarrow r} = 0$ for positive integers $r > k$. By convention $\alpha^{\uparrow 0} = 1$.

## Partition Model

The term *partition model* refers to a probability distribution, or family of probability distributions, on the set $\mathcal{E}_n$ of partitions of $[n]$. In some cases, the probability is concentrated on the the subset $\mathcal{E}_n^k \subset \mathcal{E}_n$ of partitions having $k$ or fewer blocks. A distribution on $\mathcal{E}_n$ such that $p_n(B) = p_n(\sigma B \sigma^{-1})$ for every permutation $\sigma\colon [n] \to [n]$ is said to be finitely exchangeable. Equivalently, $p_n$ is exchangeable if $p_n(B)$ depends only on the block sizes of $B$.

Historically, the most important examples are Dirichlet-multinomial random partitions generated for fixed $k$ in three steps as follows.

- First generate the random probability vector $\pi = (\pi_1, \ldots, \pi_k)$ from the Dirichlet distribution with parameter $(\theta_1, \ldots, \theta_k)$.
- Given $\pi$, the sequence $Y_1, \ldots, Y_n, \ldots$ is independent and identically distributed, each component taking values in $\{1, \ldots, k\}$ with probability $\pi$. Each sequence of length $n$ in which the value $r$ occurs $n_r \geq 0$ times has probability
$$\frac{\Gamma(\theta_\cdot)\prod_{j=1}^k \theta_j^{\uparrow n_j}}{\Gamma(n+\theta_\cdot)},$$
where $\theta_\cdot = \sum \theta_j$.
- Now forget the labels $1, \ldots, k$ and consider only the partition $B$ generated by the sequence $Y$, i.e., $B_{ij} = 1$ if $Y_i = Y_j$. The distribution is exchangeable, but an explicit simple formula is available only for the uniform case $\theta_j = \lambda/k$, which is now assumed. The number of sequences generating the same partition $B$ is $k^{\downarrow \#B}$, and these have equal probability in the uniform case. Consequently, the induced partition has probability
$$p_{nk}(B,\lambda) = k^{\downarrow \#B}\frac{\Gamma(\lambda)\prod_{b\in B}(\lambda/k)^{\uparrow \#b}}{\Gamma(n+\lambda)}, \qquad (1)$$
called the uniform Dirichlet-multinomial partition distribution. The factor $k^{\downarrow \#B}$ ensures that partitions having more than $k$ blocks have zero probability.

In the limit as $k \to \infty$, the uniform Dirichlet-multinomial partition becomes

$$p_n(B,\lambda) = \frac{\lambda^{\#B}\prod_{b\in B}\Gamma(\#b)}{\lambda^{\uparrow n}}. \qquad (2)$$

This is the celebrated Ewens distribution, or Ewens sampling formula, which arises in population genetics as the partition generated by allele type in a population evolving according to the Fisher–Wright model by random mutation with no selective advantage of allele types (Ewens 1972). The preceding derivation, a version of which can be found in Chap. 3 of Kingman (1980), goes back to Watterson (1974). The Ewens partition is the same as the partition generated by a sequence drawn according to the Blackwell-McQueen urn scheme (Blackwell and MacQueen 1973).

Although the derivation makes sense only if $k$ is a positve integer, the distribution (1) is well defined for negative values $-\lambda < k < 0$. For a discussion of this and the connection with GEM distributions and Poisson-Dirichlet distributions, see Pitman (2006, Sect. 3.2).

## Partition Processes and Partition Structures

Deletion of element $n$ from the set $[n]$, or deletion of the last row and column from $B \in \mathcal{E}_n$, determines a map $D_n\colon \mathcal{E}_n \to \mathcal{E}_{n-1}$, a projection from the larger to the smaller lattice. These deletion maps make the sets $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ into a projective system

$$\cdots \mathcal{E}_{n+1} \xrightarrow{D_{n+1}} \mathcal{E}_n \xrightarrow{D_n} \mathcal{E}_{n-1} \cdots$$

A family $p = (p_1, p_2, \ldots)$ in which $p_n$ is a probability distribution on $\mathcal{E}_n$ is said to be mutually consistent, or

Kolmogorov-consistent, if each $p_{n-1}$ is the marginal distribution obtained from $p_n$ under deletion of element $n$ from the set $[n]$. In other words, $p_{n-1}(A) = p_n\left(D_n^{-1}A\right)$ for $A \subset \mathcal{E}_{n-1}$. Kolmogorov consistency guarantees the existence of a random partition of the natural numbers whose finite restrictions are distributed as $p_n$. The partition is infinitely exchangeable if each $p_n$ is finitely exchangeable. Some authors, for example Kingman (1980), refer to $p$ as *a partition structure*.

An exchangeable partition process may be generated from an exchangeable sequence $Y_1, Y_2, \ldots$ by the transformation $B_{ij} = 1$ if $Y_i = Y_j$ and zero otherwise. The Dirichlet-multinomial and the Ewens processes are generated in this way. Kingman's (1978) paintbox construction shows that every exchangeable partition process may be generated from an exchangeable sequence in this manner.

Let $B$ be an infinitely exchangeable partition, $B[n] \sim p_n$, let $B^*$ be a fixed partition in $\mathcal{E}_n$, and suppose that the event $B[n] \leq B^*$ occurs. Then $B[n]$ lies in the lattice interval $[\mathbf{0}_n, B^*]$, which means that $B[n] = B[b_1]|B[b_2]|\ldots$ is the concatenation of partitions of the blocks $b \in B^*$. For each block $b \in B^*$, the restriction $B[b]$ is distributed as $p_{\#b}$, so it is natural to ask whether, and under what conditions, the blocks of $B^*$ are partitioned independently given $B[n] \leq B^*$. Conditional independence implies that

$$p_n(B \,|\, B[n] \leq B^*) = \prod_{b \in B^*} p_{\#b}(B[b]), \tag{3}$$

which is a type of non-interference or lack-of-memory property not dissimilar to that of the exponential distribution on the real line. It is straightforward to check that the condition is satisfied by (2) but not by (1). Aldous (1996) shows that conditional independence uniquely characterizes the Ewens family.

## Chinese Restaurant Process

A partition process is a random partition $B \sim p$ of a countably infinite set $\{u_1, u_2, \ldots\}$, and the restriction $B[n]$ of $B$ to $\{u_1, \ldots, u_n\}$ is distributed as $p_n$. The conditional distribution of $B[n+1]$ given $B[n]$ is determined by the probabilities assigned to those events in $\mathcal{E}_{n+1}$ that are consistent with $B[n]$, i.e. the events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$. For the uniform Dirichlet-multinomial model (1), these are

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} (\#b + \lambda/k)/(n+\lambda) & b \in B \\ \lambda(1 - \#B/k)/(n+\lambda) & b = \emptyset. \end{cases} \tag{4}$$

In the limit as $k \to \infty$, we obtain

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} \#b/(n+\lambda) & b \in B \\ \lambda/(n+\lambda) & b = \emptyset, \end{cases} \tag{5}$$

which is the conditional probability for the Ewens process.

To each partition process $p$ there corresponds a sequential description called the Chinese restaurant process, in which $B[n]$ is the arrangement of the first $n$ customers at $\#B$ tables. The placement of the next customer is determined by the conditional distribution $p_{n+1}(B[n+1] \,|\, B[n])$. For the Ewens process, the customer chooses a new table with probability $\lambda/(n+\lambda)$ or one of the occupied tables with probability proportional to the number of occupants. The term, due to Pitman, Dubins and Aldous, is used primarily in connection with the Ewens and Dirichlet-multinomial models.

## Exchangeable Random Permutations

Beginning with the uniform distribution on permutations of $[n]$, the exponential family with canonical parameter $\theta = \log(\lambda)$ and canonical statistic $\#\sigma$ equal to the number of cycles is

$$p_n(\sigma) = \lambda^{\#\sigma}/\lambda^{\uparrow n}.$$

The Stirling number of the first kind, $S_{n,k}$, is the number of permutations of $[n]$ having exactly $k$ cycles, for which $\lambda^{\uparrow n} = \sum_{k=1}^{n} S_{n,k}\lambda^k$ is the generating function. The cycles of the permutation determine a partition of $[n]$ whose distribution is (2), and a partition of the integer $n$ whose distribution is (6). From the cumulant function

$$\log(\lambda^{\uparrow n}) = \sum_{j=0}^{n-1} \log(j+\lambda)$$

it follows that $\#\sigma = X_0 + \cdots + X_{n-1}$ is the sum of independent Bernoulli variables with parameter $E(X_j) = \lambda/(\lambda+j)$, which is evident also from the Chinese restaurant representation. For large $n$, the number of cycles is roughly Poisson with parameter $\lambda \log(n)$, implying that $\hat{\lambda} \simeq \#\sigma/\log(n)$ is a consistent estimate as $n \to \infty$, but practically inconsistent.

A minor modification of the Chinese restaurant process also generates a random permutation by keeping track of the cyclic arrangement of customers at tables. After $n$ customers are seated, the next customer chooses a table with probability (4) or (5), as determined by the partition process. If the table is occupied, the new arrival sits to the left of one customer selected uniformly at random from the table occupants. The random permutation thus generated is $j \mapsto \sigma(j)$ from $j$ to the left neighbour $\sigma(j)$.

Provided that the partition process is consistent and exchangeable, the distributions $p_n$ on permutations of $[n]$ are exchangeable and mutually consistent under the projection $\Pi_n \to \Pi_{n-1}$ on permutations in which element $n$ is deleted from the cyclic representation (Pitman 2006, Sect. 3.1). In this way, every infinitely exchangeable random partition also determines an infinitely exchangeable random permutation $\sigma: \mathbb{N} \to \mathbb{N}$ of the natural numbers. Distributional exchangeability in this context is not to be confused with uniformity on $\Pi_n$.

## On the Number of Unseen Species

A partition of the set $[n]$ is a set of blocks, and the block sizes determine a partition of the integer $n$. For example, the partition 15|23|4 of the set $[5]$ is associated with the integer partition 2 + 2 + 1, one singleton and two doubletons. An integer partition $m = (m_1, \ldots, m_n)$ is a list of multiplicities, also written as $m = 1^{m_1} 2^{m_2} \cdots n^{m_n}$, such that $\sum j m_j = n$. The number of blocks, usually called the number of parts of the integer partition, is the sum of the multiplicities $m_\cdot = \sum m_j$.

Each integer partition is a group orbit in $\mathcal{E}_n$ induced by the action of the symmetric group on the set $[n]$. The multiplicity vector $m$ contains all the information about block sizes, but there is a subtle transfer of emphasis from block sizes to the multiplicities of the parts.

By definition, an exchangeable distribution on set partitions is a function only of the block sizes, so $p_n(B) = q_n(m)$, where $m$ is the integer partition corresponding to $B$. Since there are

$$\frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!}$$

set partitions $B$ corresponding to a given integer partition $m$, to each exchangeable distribution $p_n$ on set partitions there corresponds a marginal distribution

$$q_n(m) \frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!}$$

on integer partitions. For example, the Ewens distribution on integer partitions is

$$\frac{\lambda^{m_\cdot} \Gamma(\lambda) \prod \Gamma(j)^{m_j}}{\Gamma(n+\lambda)} \times \frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!} = \frac{\lambda^{m_\cdot} n! \, \Gamma(\lambda)}{\Gamma(n+\lambda) \prod_j j^{m_j} m_j!}. \tag{6}$$

This version leads naturally to an alternative description as follows. Let $M = M_1, \ldots, M_n$ be independent Poisson random variables with mean $E(M_j) = \lambda \theta^j / j$ for some positive number $\theta$. Then $\sum j M_j$ is sufficient for $\theta$, and the conditional distribution $\mathrm{pr}\left(M = m \mid \sum_{j=1}^{n} j M_j = n\right)$ is the Ewens integer-partition distribution with parameter $\lambda$. This representation leads naturally to a simple method of estimation and testing, using Poisson log-linear models with model formula $1 + j$ and offset $-\log(j)$ for response vectors that are integer partitions.

The problem of estimating the number of unseen species was first tackled in a paper by Fisher (1943), using an approach that appears to be entirely unrelated to partition processes. Specimens from species $i$ occur as a Poisson process (see ▶Poisson Processes) with rate $\rho_i$, the rates for distinct species being independent and identically distributed gamma random variables. The number $N_i \geq 0$ of occurrences of species $i$ in an interval of length $t$ is a negative binomial random variable

$$\mathrm{pr}(N_i = x) = (1-\theta)^{\nu} \theta^x \frac{\Gamma(\nu + x)}{x! \, \Gamma(\nu)}. \tag{7}$$

In this setting, $\theta = t/(1+t)$ is a monotone function of the sampling time, whereas $\nu > 0$ is a fixed number independent of $t$. Specimen counts for distinct species are independent and identically distributed random variables with parameters $\nu > 0$ and $0 < \theta < 1$.

The probability that no specimens from species $i$ occur in the sample is $(1-\theta)^{\nu}$, the same for every species. Most species are unlikely to be observed if either $\theta$ is small, i.e., the time interval is short, or $\nu$ is small.

Let $M_x$ be the number of species occurring $x \geq 0$ times, so that $M_\cdot$ is the unknown total number of species of which $M_\cdot - M_0$ are observed. The approach followed by Fisher is to estimate the parameters $\theta, \nu$ by conditioning on the number of species observed and regarding the observed multiplicities $M_x$ for $x \geq 1$ as multinomial with parameter vector proportional to the negative binomial frequencies (7). For Fisher's entomological examples, this approach pointed to $\nu = 0$, consistent with the Ewens distribution (6), and indicating that the data are consistent with the number of species being infinite. Fisher's approach using a model indexed by species is less direct for ecological purposes than a process indexed by specimens. Nonetheless, subsequent analyses by Good and Toulmin (1956), Holgate (1969) and Efron and Thisted (1976) showed how Fisher's model can be used to make predictions about the likely number of new species in a subsequent temporal extension of the original sample. This amounts to a version of the Chinese restaurant process.

At this point, it is worth clarifying the connection between Fisher's negative binomial formulation and the Ewens partition formulation. The relation between them is the same as the relation between binomial and negative binomial sampling schemes for a Bernoulli process: they are not equivalent, but they are complementary. The partition formulation is an exchangeable process indexed by *specimens*: it gives the distribution of species numbers in a

sample consisting of a fixed number of *specimens*. Fisher's version is also an exchangeable process, in fact an iid process, but this process is indexed by *species*: it gives the distribution of the sample composition for a fixed set of *species* observed over a finite period. In either case, the conditional distribution given a sample containing $k$ species and $n$ specimens is the distribution induced from the uniform distribution on the set of $S_{n,k}$ permutations having $k$ cycles. For the sorts of ecological or literary applications considered by Good and Toulmin (1956) or Efron and Thisted (1976), the partition process indexed by specimens is much more direct than one indexed by species.

Fisher's finding that the multiplicities decay as $E(M_j) \propto \theta^j/j$, proportional to the frequencies in the log-series distribution, is a property of many processes describing population structure, either social structure or genetic structure. It occurs in Kendall's (1975) model for family sizes as measured by surname frequencies. One explanation for universality lies in the nature of the transition rates for Kendall's process, a discussion of which can be found in Sect. 2.4 of Kelly (1978).

## Equivariant Partition Models

A family $p_n(\sigma;\theta)$ of distributions on permutations indexed by a parameter matrix $\theta$, is said to be equivariant under the induced action of the symmetric group if $p_n(\sigma;\theta) = p_n(g\sigma g^{-1}; g\theta g^{-1})$ for all $\sigma, \theta$, and for each group element $g\colon [n] \to [n]$. By definition, the parameter space is closed under conjugation: $\theta \in \Theta$ implies $g\theta g^{-1} \in \Theta$. The same definition applies to partition models. Unlike exchangeability, equivariance is not a property of a distribution, but a property of the family. In this setting, the family associated with $[n]$ is not necessarily the same as the family of marginal distributions induced by deletion from $[n+1]$.

Exponential family models play a major role in both theoretical and applied work, so it is natural to begin with such a family of distributions on permutations of the matrix-exponential type

$$p_n(\sigma;\theta) = \alpha^{\#\sigma} \exp(\mathrm{tr}(\sigma\theta))/M_\alpha(\theta),$$

where $\alpha > 0$ and $\mathrm{tr}(\sigma\theta) = \sum_{j=1}^n \theta_{\sigma(j),j}$ is the trace of the ordinary matrix product. The normalizing constant is the $\alpha$-permanent

$$M_\alpha(\theta) = \mathrm{per}_\alpha(K) = \sum_\sigma \alpha^{\#\sigma} \prod_{j=1}^n K_{\sigma(j),j}$$

where $K_{ij} = \exp(\theta_{ij})$ is the component-wise exponential matrix. This family of distributions on permutations is equivariant.

The limit of the $\alpha$-permanent as $\alpha \to 0$ gives the sum of cyclic permutations

$$\mathrm{cyp}(K) = \lim_{\alpha\to 0} \alpha^{-1} \mathrm{per}_\alpha(K) = \sum_{\sigma:\#\sigma=1} \prod_{j=1}^n K_{\sigma(j),j},$$

giving an alternative expression for the $\alpha$-permanent

$$\mathrm{per}_\alpha(K) = \sum_{B\in\mathcal{E}_n} \alpha^{\#B} \prod_{b\in B} \mathrm{cyp}(K[b])$$

as a sum over partitions. The induced marginal distribution (10) on partitions is of the product-partition type recommended by Hartigan (1990), and is also equivariant. Note that the matrix $\theta$ and its transpose determine the same distribution on partitions, but they do not usually determine the same distribution on permutations.

The $\alpha$-permanent has a less obvious convolution property that helps to explain why this function might be expected to occur in partition models:

$$\sum_{b\subset[n]} \mathrm{per}_\alpha(K[b])\,\mathrm{per}_{\alpha'}(K[\bar{b}]) = \mathrm{per}_{\alpha+\alpha'}(K). \qquad (8)$$

The sum extends over all $2^n$ subsets of $[n]$, and $\bar{b}$ is the complement of $b$ in $[n]$. A derivation can be found in section 2.4 of McCullagh and Møller (2006). If $B$ is a partition of $[n]$, the symbol $K \cdot B = B \cdot K$ denotes the Hadamard component-wise matrix product for which

$$\mathrm{per}_\alpha(K\cdot B) = \prod_{b\in B} \mathrm{per}_\alpha(K[b])$$

is the product over the blocks of $B$ of $\alpha$-permanents restricted to the blocks. Thus the function $B \mapsto \mathrm{per}_\alpha(K\cdot B)$ is of the product-partition type.

With $\alpha, K$ as parameters, we may define a family of probability distributions on $\mathcal{E}_n^k$, i.e., partitions of $[n]$ having $k$ or fewer blocks, as follows:

$$p_{nk}(B) = k^{\downarrow\#B}\,\mathrm{per}_{\alpha/k}(K\cdot B)/\,\mathrm{per}_\alpha(K). \qquad (9)$$

The fact that (9) is a probability distribution on $\mathcal{E}_n$ follows from the convolution property of permanents. The limit as $k \to \infty$

$$p_n(B) = \alpha^{\#B} \prod_{b\in B} \mathrm{cyp}(K[b])/\,\mathrm{per}_\alpha(K), \qquad (10)$$

is a product-partition model satisfying the conditional independence property (3). For $K = \mathbf{1}_n$, the $n \times n$ matrix whose elements are all one, $\mathrm{per}_\alpha(\mathbf{1}_n) = \alpha^{\uparrow n}$ is the ascending factorial function. Thus the uniform Dirichlet-multinomial model (1) and the Ewens model (2) are both obtained by setting $\theta = 0$.

## Leaf-Labelled Trees

Kingman's $[n]$-coalescent is a non-decreasing, $\mathcal{E}_n$-valued Markov process (see ▶Markov Processes) $(B_t)$ in continuous-time starting from the partition $B_0 = \mathbf{0}_n$ with $n$ singleton blocks at time zero. The coalescence intensity is one for each pair of blocks regardless of size, so each coalescence event unites two blocks chosen uniformly at random from the set of pairs. Consequently, the first coalescence occurs after a random time $T_n$ exponentially distributed with rate $\rho(n) = n(n-1)/2$ and mean $1/\rho(n)$. After $k$ coalescences, the partition consists of $n-k$ blocks, and the waiting time $T_k$ for the next subsequent coalescence is exponential with rate $\rho(n-k)$. The time to complete coalescence is the sum of independent exponentials $T = T_n + T_{n-1} + \cdots + T_2$, which is a random variable with mean $2-2/n$ and variance increasing from 1 at $n = 2$ to a little less than 1.16 as $n \to \infty$. In the context of the Fisher–Wright model, the coalescent describes the genealogical relationships among a sample of individuals, and $T$ is the time to the most recent common ancestor of the sample.

The $[n]$-coalescent is exchangeable for each $n$, but the property that makes it interesting mathematically, statistically and genetically is its consistency under selection or sub-sampling (Kingman 1982). If we denote by $p_n$ the distribution on $[n]$-trees implied by the specific Markovian model described above, it can be shown that the embedded tree obtained by deleting element $n$ from the sample $[n]$ is not only Markovian but also distributed as $p_{n-1}$, i.e., the same coalescent rule applied to the subset $[n-1]$. This property is mathematically essential for genealogical trees because the occurrence or non-occurrence of individual $n$ in the sample does not affect the genealogical relationships among the remainder.

A fragmentation $[n]$-tree is a non-increasing $\mathcal{E}_n$-valued Markov process starting from the trivial partition $B_0 = \mathbf{1}_n$ with one block of size $n$ at time $t = 0$. The simplest of these are the consistent binary Gibbs fragmentation trees studied by Aldous (1996), Bertoin (2001, 2006) and McCullagh et al. (2008). The first split into two branches occurs at a random time $T_n$ exponentially distributed with parameter $\rho(n)$. Subsequently, each branch fragments independently according to the same family of distributions with parameter $\rho(\#b)$ for branch $b$, which is a Markovian conditional independence property analogous to (3). Consistency and conditional independence put severe limitations on both the splitting distribution and the rate function $\rho(n)$, so the entire class is essentially one-dimensional.

A rooted leaf-labelled tree $T$ is also a non-negative symmetric matrix. The interpretation of $T_{ij}$ as the distance from the root to the junction at which leaves $i, j$ occur on disjoint branches implies the inequality $T_{ij} \geq \min(T_{ik}, T_{jk})$ for all $i, j, k \in [n]$. The set of $[n]$-trees is a subset of the positive definite symmetric matrices, not a manifold, but a finite union of manifolds of dimension $2n-1$, or $n$ if the diagonal elements are constrained to be equal. Like partitions, rooted trees form a projective system within the positive definite matrices. A fragmentation tree is an infinitely exchangeable random tree, which is also a special type of infinitely exchangeable random matrix.

## Cluster Processes and Classification Models

A $\mathcal{R}^d$-valued cluster process is a pair $(Y, B)$ in which $Y = (Y_1, \ldots)$ is an $\mathcal{R}^d$-valued random sequence and $B$ is a random partition of $\mathbb{N}$. The process is said to be exchangeable if, for each finite sample $[n] \subset \mathbb{N}$, the restricted process $(Y[n], B[n])$ is invariant under permutation $\sigma: [n] \to [n]$ of sample elements.

The Gauss–Ewens process is the simplest non-trivial example for which the distribution for a sample $[n]$ is as follows. First fix the parameter values $\lambda > 0$, and $\Sigma^0, \Sigma^1$ both positive definite of order $d$. In the first step $B$ has the Ewens distribution on $\mathcal{E}_n$ with parameter $\lambda$. Conditionally on $B$, $Y$ is a zero-mean Gaussian matrix of order $n \times d$ with covariance matrix

$$\operatorname{cov}(Y_{ir}, Y_{js} \mid B) = \delta_{ij}\Sigma^0_{rs} + B_{ij}\Sigma^1_{rs},$$

where $\delta_{ij}$ is the Kronecker symbol. A scatterplot color-coded by blocks of the $Y$ values in $\mathcal{R}^2$ shows that the points tend to be clustered, the degree of clustering being governed by the ratio of between to within-cluster variances.

For an equivalent construction we may proceed using a version of the Chinese restaurant process in which tables are numbered in order of occupancy, and $t(i)$ is number of the table at which customer $i$ is seated. In addition, $\epsilon_1, \ldots$ and $\eta_1, \ldots$ are independent Gaussian sequences with independent components $\epsilon_i \sim N_d(0, \Sigma^0)$, and $\eta_i \sim N_d(0, \Sigma^1)$. The sequence $t$ determines $B$, and the value for individual $i$ is a vector $Y_i = \eta_{t(i)} + \epsilon_i$ in $\mathcal{R}^d$, or $Y_i = \mu + \eta_{t(i)} + \epsilon_i$ if a constant non-zero mean vector is included.

Despite the lack of class labels, cluster processes lend themselves naturally to prediction and classification, also called supervised learning. The description that follows is taken from McCullagh and Yang (2006) but, with minor modifications, the same description applies equally to more complicated non-linear versions associated with generalized linear mixed models (Blei et al. 2003). Given the observation $(Y[n], B[n])$ for the 'training sample' $[n]$, together with the feature vector $Y_{n+1}$ for specimen $u_{n+1}$, the conditional distribution of $B[n+1]$ is determined by

those events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$ that are compatible with the observation. The assignment of a positive probability to the event that the new specimen belongs to a previously unobserved class seems highly desirable, even logically necessary, in many applications.

If the classes are tree-structured with two levels, we may generate a sub-partition $B' \leq B$ whose conditional distribution given $B$ is Ewens restricted to the interval $[\mathbf{0}_n, B]$, with parameter $\lambda'$. This sub-partition has the effect of splitting each main clusters randomly into sub-clusters. For the sample $[n]$, let $t'(i)$ be the number of the sub-cluster in which individual $i$ occurs. Given $B, B'$, the Gauss-Ewens two-level tree process is a sum of three independent Gaussian processes $Y_i = \eta_{t(i)} + \eta'_{t'(i)} + \epsilon_i$ for which the conditional distributions may be computed as before. In this situation, however, events that are compatible with the observation $B[n], B'[n]$ are of three types as follows:

$$u_{n+1} \mapsto b' \in B'[n], \quad u_{n+1} \mapsto \emptyset \subset b \in B[n], \quad u_{n+1} \mapsto \emptyset.$$

In all, there are $\#B' + \#B + 1$ disjoint events for which the conditional distribution given $B[n], B'[n], Y[n+1]$ must be computed. An event of the second type is one in which the new specimen belongs to the major class $b \in B$, but not to any of the sub-types previously observed for this class.

## Further Applications of Partition Models

Exchangeable partition models are used to construct nontrivial, exchangeable processes suitable for cluster analysis and density estimation. See Frayley and Raftery (2002) for a review. Here, cluster analysis means cluster detection and cluster counting in the absence of covariate or relational information about the units. In the computer-science literature, cluster detection is also called unsupervised learning. The simplest of these models is the marginal Gauss–Ewens process in which only the sequence $Y$ is observed. The conditional distribution $p_n(B \mid Y)$ on $\mathcal{E}_n$ is the posterior distribution on clusterings or partitions of $[n]$, and $E(B \mid Y)$ is the one-dimensional marginal distribution on pairs of units. In estimating the number of clusters, it is important to distinguish between the sample number $\#B[n]$, which is necessarily finite, and the population number $\#B[\mathbb{N}]$, which could be infinite (McCullagh and Yang 2008).

Exchangeable partition models are also used to provide a Bayesian solution to the multiple comparisons problem. The key idea is to associate with each partition $B$ of $[k]$ a subspace $V_B \subset \mathcal{R}^k$ equal to the span of the columns of $B$. Thus, $V_B$ consists of vectors $x$ such that $x_r = x_s$ if $B_{rs} = 1$. For a treatment factor having $k$ levels $\tau_1, \ldots, \tau_k$, the Gauss–Ewens prior distribution on $R^k$ puts positive mass on the subspaces $V_B$ for each $B \in \mathcal{E}_k$. Likewise, the posterior distribution also puts positive probability on these subspaces, which enables us to compute in a coherent way the posterior probability $\mathrm{pr}(\tau \in V_B)$ or the marginal posterior probability $\mathrm{pr}(\tau_r = \tau_s \mid y)$. For details, see (Gopalan and Berry 1998).

## Acknowledgments

Support for this research was provided in part bt NSF Grant DMS-0906592.

## About the Author

Peter McCullagh is the John D. MacArthur Distinguished Service Professor in the Department of Statistics at the University of Chicago. Professor McCullagh is a past editor of *Bernoulli*, a fellow of the Royal Society of London and of the American Academy of Arts and Sciences. He is co-author with John Nelder of the book *Generalized linear Models* (Chapman and Hall, 1989).

## Cross References

►Cluster Analysis: An Introduction
►Data Mining
►Multivariate Statistical Distributions
►Permanents in Probability Theory

## References and Further Reading

Aldous D (1996) Probability distributions on cladograms. In: Random discrete structures. IMA Vol Appl Math 76. Springer, New York, pp 1–18

Bertoin J (2001) Homogeneous fragmentation processes. Probab Theor Relat Fields 121:301–318

Bertoin J (2006) Random fragmentation and coagulation processes. Cambridge studies in advanced mathematics, vol 102. Cambridge University Press, Cambridge

Blackwell D, MacQueen J (1973) Ferguson distributions via Pólya urn schemes. Ann Stat 1:353–355

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach learn Res 3:993–1022

Efron B, Thisted RA (1976) Estimating the number of unknown species: how many words did Shakespeare know? Biometrika 63:435–447

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42–58

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631

Good IJ, Toulmin GH (1956) The number of new species, and the increase in population coverage when a sample is increased. Biometrika 43:45–63

Gopalan R, Berry DA (1998) Bayesian multiple comparisons using Dirichlet process priors. J Am Stat Assoc 93:1130–1139

Hartigan JA (1990) Partition models. Commun Stat Theor Meth 19:2745–2756

Holgate P (1969) Species frequency distributions. Biometrika 65:651–660
Kelly FP (1978) Reversibility and stochastic networks. Wiley, Chichester
Kendall DG (1975) Some problems in mathematical genealogy. In: Perspectives in probability and statistics: papers in honour of M.S. Bartlett. Academic, London, pp 325–345
Kingman JFC (1975) Random discrete distributions (with discussion). J R Stat Soc B 37:1–22
Kingman JFC (1977) The population structure associated with the Ewens sampling formula. Theor Popul Biol 11:274–283
Kingman JFC (1978) The representation of partition structures. J Lond Math Soc 18:374–380
Kingman JFC (1980) Mathematics of genetic diversity. CBMS-NSF conference series in applied mathematics, 34 SIAM, Philadelphia
Kingman JFC (1982) The coalescent. Stoch Proc Appl 13:235–248
McCullagh P, Møller J (2006) The permanental process. Adv Appl Prob 38:873–888
McCullagh P, Yang J (2006) Stochastic classification models. In: Proceedings of the international congress of mathematicians, 2006, vol 3, pp 669–686
McCullagh P, Yang J (2008) How many clusters? Bayesian Anal 3:1–19
McCullagh P, Pitman J, Winkel M (2008) Gibbs fragmentation trees. Bernoulli 14:988–1002
Pitman J (2006) Combinatorial stochastic processes. Springer, Berlin
Watterson GA (1974) The sampling theory of selectively neutral alleles. Adv Appl Probab 6:217–250

# Random Variable

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

Random variable (r.v.) is a real function defined on the set of outcomes. It reduces the set-theoretical problems in probability to some simpler ones in real analysis. R.v. 's are indispensable in probability computing.

## Motivation

Formal definition of a r.v. is a consequence of some practical and logical needs. Let us start from a measure-theoretic frame $(\Omega, \mathcal{A}, P)$, where $\Omega$ is the set of outcomes, $\mathcal{A}$ is a $\sigma$-algebra of subsets of $\Omega$, serving as random events, and $P$ is a normalized measure on the space $(\Omega, \mathcal{A})$, called probability. Any real function $f = f(\omega)$ transforms the initial probability system $(\Omega, \mathcal{A}, P)$ onto some induced system $(\Omega_f, \mathcal{A}_f, P_f)$, where $\Omega_f$ is the image of $\Omega$ by $f$, $\mathcal{A}_f$ is the $\sigma$-algebra of of subsets $B$ on $\Omega_f$ induced by $f$, while $P_f$ is a probability measure on $(\Omega_f, \mathcal{A}_f)$ defined by

$$P_f(B) = P(\{\omega : f(\omega) \in B\}). \tag{1}$$

For practical reasons we require that the family $\mathcal{A}_f$ includes all intervals $(a; b >$, i.e., that $\mathcal{A}_f \supseteq \mathcal{B}$, where $\mathcal{B}$ means the family of Borel sets in the real line. On the other hand the right side of (1) is well defined, if and only if, $\{\omega : f(\omega) \in B\} \in \mathcal{A}$. Since $\mathcal{A}_f$ is $\sigma$-algebra generated by the intervals, the last condition may be expressed in a more readable form

$$\{\omega : f(\omega) \le c\} \in \mathcal{A} \text{ for any } c \in R. \tag{2}$$

## Formal Definition

Any real function defined on the (measurable) space $(\Omega, \mathcal{A})$ satisfying the condition (2) is said to be a random variable on $(\Omega, \mathcal{A})$. Traditionally, random variables are denoted by capital letters $X(\omega), Y(\omega), Z(\omega)$, or simply $X, Y, Z$. The following exmple shows that not every function of outcome is a random variable.

*Example* 1 Let us set $\Omega = \{1,2,3,4,5\}$, $\mathcal{A} = \{\varnothing, \{1,3,5\}, \{2,4\}, \Omega\}$ and

$$f(\omega) = \begin{cases} 0, & \text{if } \omega \le 2 \\ 1, & \text{if } \omega > 2. \end{cases}$$

By setting in (2) $c = 1$ we get $\{\omega : f(\omega) < 1\} = \{1,2\} \notin \mathcal{A}$. Thus $f$ is not random variable on the space $(\Omega, \mathcal{A})$.

The probability measure $P_X(B) = P(\{\omega : X(\omega) \in B\})$ for $B \in \mathcal{B}$ is said to be distribution of the r.v. $X$. This expression has mainly theoretical sense, because the Borel sets are abstractive objects. More practical in use is so called *cumulative distribution function (c.d.f)* $F$ defined by $F(\alpha) = P(\{\omega : X(\omega) \le \alpha\})$.

*Example* 2 (Two-Dice Game). Here the set of outcomes may be presented in the form $\Omega = \{(i,j) : i,j = 1,2,3,4,5,6\}$, the family of random events $\mathcal{A}$ may be defined as the family of all subsets of $\Omega$, while $X(\omega)$, for any $\omega = (i,j)$ may be defined as $i + j$. Such a r.v. takes the possible values from 2 to 12 with probabilities

$$P_X(k) = \begin{cases} \frac{k-1}{36} & \text{if} \quad k \le 7 \\ \frac{12-k+1}{36} & \text{if} \quad k > 7, \end{cases}$$

while the values of c.d.f. $F_X = F_X(\alpha)$ are collected in the Table 1.

It is worth to add that if $X = X(\omega)$ is a random variable and $f$ is a Borel function, i.e., a real function of a real variable such that $\{x : f(x) \le c\} \in \mathcal{B}$ for all $c \in R$ then the composition $f[X(\omega)]$ is also random variable.

**Random Variable. Table 1** Values of c.d.f. $F_X = F_X(\alpha)$ in example 2

| Interval for $\alpha$ | $(-\infty, 2)$ | $<2,3)$ | $<3,4)$ | $<4,5)$ | $<5,6)$ | $<6,7)$ |
|---|---|---|---|---|---|---|
| $F_X(\alpha)$ | 0 | $\frac{1}{36}$ | $\frac{3}{36}$ | $\frac{6}{36}$ | $\frac{10}{36}$ | $\frac{15}{36}$ |
| Interval for $\alpha$ | $<7,8)$ | $<8,9)$ | $<9,10)$ | $<10,11)$ | $<11,12)$ | $<12,+\infty)$ |
| $F_X(\alpha)$ | $\frac{21}{36}$ | $\frac{26}{36}$ | $\frac{30}{36}$ | $\frac{33}{36}$ | $\frac{35}{36}$ | 1 |

## Classification of r.v. 's

It is well known that any c.d.f. $F$ is continuous, perhaps outside a countable set on the real line. For practical purposes we distinguish two classes of random variables. A r.v. is

1. *Discrete,* if its c.d.f is constant in all intervals designed by the points of its discontinuity
2. *Continuous,* if there exists a nonnegative integrable function $f$ on $R$, called *density,* such that $F(\alpha) = \int_{-\infty}^{+\infty} f(x)dx$ for all $\alpha \in R$.

This classification is fully justified by two different analytical tools used in presentation and calculation of the r.v.'s. Distribution of a discrete r.v. $X$ taking values $x_i$ for some $i \in I$ is given by *probability mass function* $p_i = P(X = x_i)$ and its expectation is calculated by the formula $Ex = \sum_i x_i p_i$. Distribution of a continuous r.v. $X$ is usually given by its *density function* $f$, while its expectation is calculated by the formula $Ex = \int_{-\infty}^{+\infty} xf(x)dx$.

## About the Author

Professor Czesław Stępniak was formerly working at Agricultural University in Lublin, Poland (1972–2001) and heading Department of Statistics and Econometrics, Maria Curie-Skłodowska University in Lublin, Poland (2003–2009). During academic year 1987–1988 he was visiting Mathematical Sciences Institute of Cornell University, Ithaca, NY, as a senior scientist. Jointly with E. Torgersen, C. F. J. Wu and H. Heyer, Czesław Stępniak laid mathematical foundation to comparison of statistical experiments. He was also a recipient of two awards from Ministry of Science and Education in Poland "for a series of papers." Professor Stępniak authored more than 50 articles in peer-reviewed journals. He has two descendents Marek Niezgoda and Zdzisław Otachel.

## Cross References

►Expected Value
►Measure Theory in Probability

## References and Further Reading

Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York
Kolmogorov AN (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin [English translation: Foundations of the theory of probability (2nd edn.), Chelsea, New York, 1956]
Prokhorov YuV (1985) Random variable. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 5, 9–10. Soviet Encyclopedia, Moscow (in Russian)

# Random Walk

Rabi Bhattacharya
Professor of Mathematics
The University of Arizona, Tucson, AZ, USA

The simple random walk $\{S_n : n = 0, 1, \ldots\}$, starting at an integer $x$, is a stochastic process on the integers, given by $S_0 = x$, $S_n = x + X_1 + \ldots + X_n (n \geq 1)$, where $X_n, n \geq 1$, is an independent Bernoulli sequence: $P(X_n = 1) = p$, $P(X_n = -1) = 1 - p = q$, $0 < p < 1$. In the case, $p = q = 1/2$, it is called the *simple symmetric random walk*, while if $p \neq 1/2$, it is *asymmetric*. By the binomial theorem, $P(S_n = y \mid S_0 = 0) = C^n_{\frac{(n+y)}{2}} p^{(n+y)/2} q^{(n-y)/2}$, if $y$ and $n$ are of the same parity, i.e., if either both are odd or both are even. Otherwise, $P(S_n = y \mid S_0 = 0) = 0$. Here $= C^n_m = n!/(m!(n-m)!)$.

For $c \leq x \leq d$ integers, the probability $\pi(x)$ that a simple random walk, starting at $x$, reaches $c$ before $d$ satisfies the equation

$$\pi(x) = p\pi(x+1) + q\pi(x-1) \quad \text{for } c < x < d,\ \pi(c) = 1,\ \pi(d) = 0, \qquad (1)$$

as shown by conditioning on the first step $X_1$. For the symmetric walk, the solution of this equation is $\pi(x) = (d-x)/(d-c)$. Since $\pi(x) \to 1$ as $d \to \infty$, the symmetric walk will reach the state $c$, starting from any state $x > c$, with probability one. By symmetry, it will reach every state with probability one. Iterating this argument one sees that, with probability one, the symmetric random walk visits every state infinitely often. That is, the walk is *recurrent*. For the asymmetric walk, the solution to (1) is $\pi(x) = (1 - (p/q)^{d-x})/(1 - (p/q)^{d-c})$. If $p < 1/2$, then the limit of this is 1 as $d \to \infty$ and, with probability one,

the random walk will visit $c$, starting from $x > c$. On the other hand, if $p > 1/2$, then $\pi(x) \to (q/p)^{x-c} < 1$, as $d \to \infty$. The probability of ever reaching $d$, starting from $x < d$ is obtained by symmetry as 1 if $p > 1/2$ and $(p/q)^{d-x}$ if $p < 1/2$. The asymmetric simple random walk is thus *transient*. Indeed, it follows from the strong law of large numbers (*SLLN*) that if $p > 1/2$, then $S_n \to \infty$ with probability one as $n \to \infty$; and $S_n \to -\infty$, with probability one, if $p < 1/2$. For these and other properties of the random walk, such as those described below, see Feller (1968, Chap. 3), Bhattacharya and Waymire (2009, Chap. 1), or Durrett (1995, Chap. 3). For additional information, refer to Billingsley (1968), and Spitzer (1964).

For computation of various probabilities associated with a simple random walk, the following result proved by D. Andre in 1887 is very useful: Consider the polygonal path of the random walk joining successive points $(j, S_j)$, $(j+1, S_{j+1})$ $(j = 0, 1, \ldots, n-1)$ by line segments. Let $y > 0$. Then (a) the set of paths from $(0,0)$ to $(n, y-1)$ ($n$ and $y-1$ of the same parity) which touch or cross the level $y$, is in one-one correspondence with (b) the set of all paths from $(0,0)$ to $(n, y+1)$ (*Reflection principle*). To prove this, let $\tau$ be the first time a path of the type (a) touches the level $y$ prior to time $n$. Then replace the segment of the path from $(\tau, y)$ to $(n, y-1)$ by its mirror reflection about the level $y$. This gives a path of the type (b). Conversely, given any path of the type (b), reflect about y the segment of the path from $(\tau, y)$ to $(n, y+1)$. This gives a path of the type (a). Here is an application of this principle.

*Example* 1 (First passage time distribution of a simple random walk). Let $y$ be a positive integer, and $F_{n,y}$ the event that the random walk, starting at zero, reaches $y$ for the first time at time $n$, i.e., $F_{n,y} = \{S_j \neq y$, for $0 \le j < n, S_n = y\}$, $n$ and $y$ of the same parity. Altogether there are $C^n_{\frac{n+y}{2}}$ paths from $(0,0)$ to $(n,y)$, each having probability $p^{(n+y)/2} q^{(n-y)/2}$. Of these, the number which cross or touch the level $y$ prior to time n and for which $S_{n-1} = y-1$ is, by the reflection principle, $C^{n-1}_{\frac{n+y}{2}}$. Also the number for which $S_{n-1} = y+1$ is $C^{n-1}_{\frac{n+y}{2}}$. Subtracting these two from the number $C^n_{\frac{n+y}{2}}$ of all paths, one obtains, for all $y \neq 0$ (treating the case $y < 0$ by symmetry),

$$\begin{aligned} P(F_{n,y}) &= \left( C^n_{\frac{n+y}{2}} - 2C^{n-1}_{\frac{n+y}{2}} \right) p^{(n+y)/2} q^{(n-y)/2} \\ &= (|y|/n) C^n_{\frac{n+y}{2}} p^{(n+y)/2} q^{(n-y)/2} \qquad (2) \\ &\quad (n = |y|, |y|+2, |y|+4, \ldots). \end{aligned}$$

One may also consider the simple symmetric random walk $S_0 = x$, $S_n = x + X_1 + \ldots + X_n$ $(n \ge 1)$, in dimension $d \ge 1$, as a stochastic process on the d-dimensional lattice $Z^d$, with $X_n (n \ge 1)$ i.i.d. random vectors, taking values $\pm e_j (j = 1, \ldots, d)$, each with probability $1/2d$. Here $e_j$ is the vector whose $j$-th coordinate is 1 and the remaining $d-1$ coordinates are zero. It was proved by G. Polya in 1921 that this walk is recurrent in dimensions 1,2, and transient in higher dimensions.

De Moivre (1756) obtained the normal approximation to the binomial probability $P(S_n = y \mid S_0 = 0)$, as a combinatorial result. The full potential of this was realized by Laplace (1812) who formulated and derived the far reaching central limit theorem (*CLT*, see ▶Central Limit Theorems). Apparently, Gauss knew about the normal distribution as early as 1794, and assuming this as the distribution of errors of measurement, he obtained his famous method of ▶least squares. Hence the name Gaussian distribution is often used for the normal distribution. The final version of the *CLT* for a *general random walk* $S_n = X_1 + \ldots + X_n$ $(n \ge 1)$, where $X_n$ are arbitrary independent identically distributed (*i.i.d.*) random variables with mean zero and finite variance $\sigma^2 > 0$, was obtained by Lévy (1925): $n^{-1/2}(X_1 + \ldots + X_n)$ converges in distribution to the normal distribution $N(0, \sigma^2)$ with mean zero and variance $\sigma^2$, as $n \to \infty$. In physical terms, this result says the following: if time and length are rescaled so that in one unit of rescaled time there are a large number n of i.i.d. displacements of small rescaled lengths of order $1/\sqrt{n}$, then the random walk displacements over a period of time t will appear as Gaussian with mean zero and variance $t\sigma^2$, the increments over disjoint intervals being independent. That such a Gaussian process exists with continuous sample paths was proved rigorously by N.Wiener in 1923. This process is called the *Brownian motion*, following its implicit use by A. Einstein in 1905–1906 to describe the kinetic motion of colloidal molecules in a liquid, experimentally observed earlier by the botanist R. Brown. Interestingly, even before Einstein, Bachelier (1900) described the random movements of stocks by this Gaussian process. The statement that the rescaled random walk $S_n(n = 0, 1, 2, \ldots)$ converges in distribution to Brownian motion (see ▶Brownian Motion and Diffusions) was proved rigorously by M. Donsker in 1951, and this result is known as the functional central limit theorem (*FCLT*). Both the *CLT* and the *FCLT* extend to arbitrary dimensions d.

As consequences of the *FCLT*, one can derive many asymptotic results for the simple symmetric random walk given by the corresponding result for the limiting Brownian motion. Conversely, by evaluating combinatorially some probability associated with the random walk, one may derive the corresponding probability for the Brownian motion. A Brownian motion with variance parameter $\sigma^2 = 1$ is called a standard Brownian motion, and denoted $\{B_t : t \ge 0\}$ below.

*Example* 2 (Boundary hitting probability of Brownian motion). Let $c \le x \le d$ be arbitrary reals. Then, using the corresponding result for the scaled ▶random walk, one obtains

$$P(\{B_t : t \ge 0\} \text{ reaches c before } d \mid B_0 = x) = (d-x)/(d-c). \quad (3)$$

*Example 3* (Arcsine law). Let $U$ denote the amount of time in $[0,1]$ the Brownian motion spends above zero, i.e., $U$ = Lebesgue measure of the set $\{t : 0 \le t \le 1 : B_t > 0\}$, given $B_0 = 0$. Consider the polygonal path of the simple symmetric random walk $S_j(j = 0, 1, \ldots n)$, starting at zero. By combinatorial arguments, such as the reflection principle, one can calculate exactly the proportion of times the polygonal path lies above zero and, by the *FCLT*, this yields

$$P(U \le x) = (2/\pi) sin^{-1}\sqrt{x}\ (0 \le x \le 1). \quad (4)$$

## Acknowledgments

The author acknowledges support from the NSF grant DMS 0806011.

## About the Author

Rabindranath Bhattacharya received his Ph.D. from the University of Chicago in 1967. He has held regular faculty positions at the University of California at Berkeley, the University of Arizona, and Indiana University, and is currently a Professor of Mathematics at the University of Arizona. He has co-authored a number of graduate texts and monographs: *Normal Approximation and Asymptotic Expansions* (with R. Ranga Rao, John Wiley, 1976), *Stochastic Processes with Applications* (with Edward Waymire, Wiley, 1990), *Asymptotic Statistics* (with M. Denker, Birkhauser, 1990) and, more recently, *A Basic Course in Probability Theory* (with Ed Waymire, Springer, 2007), and *Random Dynamical Systems* (with M. Majumdar, Cambridge University Press, 2007). Among his more than 80 research articles in statistics, probability and mathematics, are Special Invited Papers in the *Annals of Probability* (1977), and the *Annals of Applied Probability* (1999). Professor Bhattacharya was Associate Editor for the following journals: *Annals of Probability* (1976–1981 and 2000–2002), *Annals of Applied Probability* (2006–2009), *Econometric Theory* (1989–1999), *Journal of Multivariate Analysis* (1986–1992), *Journal of Statistical Planning and Inference* (1984–1988, and 2000–2002), and *Statistica Sinica* (2002–2008). Currently he is Associate Editor for *Sankhya* (2009–present). Among his PhD students are several distinguished international scholars such as Ed Waymire (Oregon State University), Vic Patrangenaru (Florida State University), Gopal Basak (Indian Statistical Institute), Oesook Lee (Ewha Woman's University, Seoul, Korea). Rabi Bhattacharya is a Fellow of the IMS, and is a recipient of an Alexander Von Humboldt Forschungspreis, and a Guggenheim Fellowship.

## Cross References

▶Brownian Motion and Diffusions
▶Central Limit Theorems
▶Ergodic Theorem
▶Limit Theorems of Probability Theory
▶Markov Processes
▶Monte Carlo Methods in Statistics
▶Statistical Modeling of Financial Markets
▶Statistical Quality Control
▶Stochastic Modeling Analysis and Applications
▶Stochastic Processes
▶Stochastic Processes: Classification

## References and Further Reading

Bachelier L (1900) Théorie de la speculation. Ann sci école norm sup 17:21–86
Bhattacharya RN, Waymire EC (2009) Stochastic processes with applications. SIAM Classics in Applied Mathematics, vol 61. SIAM, Philadelphia
Billingsley P (1968) Convergence of probability measures. Wiley, New York
De Moivre A (1756) Doctrine of chance. London
Durrett R (1995) Probability: theory and examples. 2nd edn. Duxbury, Belmont
Feller W (1968) An introduction to probability theory and its applications. vol 1. 3rd edn. Wiley, New York
Laplace PS (1812) The theéorie analitique des probabilités. Veuve Courcier, Paris
Lévy P (1925) Calcul des probabilités. Gauthier-Villars, Paris
Spitzer F (1964) Principles of random walk. Van Nostrand. Princeton

# Randomization

Cristiano Ferraz
Associate Professor
Federal University of Pernambuco, Recife, Brazil

Randomization is prescribed in several statistical procedures for reasons related not only to the assurance of scientific objectivity. Randomization, in essence, may be defined as a physical mechanism to assign probabilities to events. In probability sampling, such events are related to the selection of samples from finite populations. Samples are selected according to randomization processes that

guarantee selection probabilities for any specific sample. As a consequence, it also guarantees inclusion probabilities (of first and higher orders) for any element of the population. When a simple random sampling design (without replacement) is employed, for instance, any sample $A$, of size $n$ from a population $\mathcal{U}$, of size $N$ ($n < N$) have the same probability of been selected, and its inclusion probability of first order corresponds to the sample fraction, $n/N$. Restrictions imposed on the randomization lead to different sample designs (e.g., systematic sampling, Bernoulli sampling, Poisson sampling, and stratified sampling) and are responsible for their statistical properties. Similarly, in comparative experiments, the events are related to the assignment of treatments to experimental units. In experiments following the randomization principle, treatments are randomly assigned to available experimental units. This means such an assignment follows a specific randomization protocol. If, for instance, the protocol implies each group of size $r$ from a total $n = tr$ available experimental units have the same probability of receiving a given treatment, the experiment is been conducted to compare $t$ treatments according to a completely randomized design (with $r$ genuine replicates). Once again, restrictions in the randomization lead to different designs (e.g., randomized complete block designs, Latin square designs, and split-plot designs).

Statistical methods of sampling and design of experiments rely on randomization to make valid design-based inferences. In both cases, inferences are supported by real reference distributions, induced by randomization. Its major role may be evident from appropriately derived linear models. A linear model for data from a ▶simple random sample, for instance, may be derived as follows. Let $y_i$ be defined as the $i$th observation of a variable of interest $Y$ under a simple random sample selection scheme (such as the traditional drawing of $n$ balls, one at a time, without replacement, from an urn with $N$ balls labeled from 1 to $N$). Hence, $y_i$ can assume any value $Y_k$ (the value of $Y$ associated to element $k \in \mathcal{U}$). Let also be the following indicator variable defined:

$$\delta_{ik} = \begin{cases} 1, & \text{if } y_i = Y_k \\ 0, & \text{if } y_i \neq Y_k \end{cases}$$

Now, it is possible to write

$$y_i = \sum_{k \in \mathcal{U}} \delta_{ik} Y_k. \tag{1}$$

Let $Y_k$ be rewritten as $\mu + (Y_k - \mu)$, with $\mu = \dfrac{\sum_{k \in \mathcal{U}} Y_k}{N}$. Then, (1) may be rewritten as

$$y_i = \sum_{k \in \mathcal{U}} \delta_{ik} [\mu + (Y_k - \mu)] = \mu + \sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu). \tag{2}$$

Define $\omega_i = \sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu)$ and (2) may be written as

$$y_i = \mu + \omega_i. \tag{3}$$

Expression (3) is the simplest linear model. According to this model, the ith observation of a variable of interest $Y$, observed on a simple random sample, may be regarded as the population mean ($\mu$) plus a random term ($\omega_i$) with statistical properties implied by the randomization scheme. For example, the description of "balls withdrawn from an urn" scheme allows one to write

$$P(\delta_{ik} = 1) = \frac{1}{N}, \text{ for any } k \in \mathcal{U};$$
$$P(\delta_{ik} = 1, \delta_{i'k'} = 1) = 0 \text{ for } k \neq k'; \text{ and,}$$
$$P(\delta_{ik} = 1, \delta_{i'k'} = 1) = \frac{1}{N(N-1)}, \text{ for } i \neq i' \text{ and } k \neq k'.$$

Therefore, the following properties hold:

$$E(\omega_i) = E\left(\sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu)\right) = \frac{1}{N} \sum_{k \in \mathcal{U}} (Y_k - \mu) = 0; \tag{4}$$

$$V(\omega_i) = E(\omega_i^2) = \sum_{k \in \mathcal{U}} \sum_{k' \in \mathcal{U}} (Y_k - \mu)(Y_{k'} - \mu) E(\delta_{ik} \delta_{ik'})$$
$$= \frac{1}{N} \sum_{k \in \mathcal{U}} (Y_k - \mu)^2 = \frac{(N-1)}{N} \sigma^2 \tag{5}$$

where $\sigma^2 = \dfrac{\sum_{k \in \mathcal{U}} (Y_k - \mu)^2}{N - 1}$. It can also be shown that

$$Cov(\omega_i, \omega_{i'}) = -\frac{\sigma^2}{N} \tag{6}$$

Clearly, properties (4), (5), and (6) are consequences of the randomization process. They are not assumptions. Based on them, estimators such as the sample mean can be evaluated and proved unbiased with variances given as stated in many sampling books.

The ideas related to the role of randomization in scientific investigation were originally proposed by Fisher (1925, 1937). Since then, the relevance of the subject motivated works by several authors. Only few of them are referenced here as suggestions for further reading by limitation of space. Hinkelmann and Kempthorne (1994), for instance, explore the role of randomization in designed experiments by deriving linear models and examining in

depth the properties of the induced reference distributions. Särndal et al. (1992) emphasize the fundamental ideas of probability sampling giving attention to unbiased estimation. Finally, Tillé (2006) describes a series of computational algorithms (randomization protocols) to select samples according to the probabilistic method.

## About the Author

Dr. Cristiano Ferraz is the author of the book *Sample design for surveys quality evaluation*, 2008, VDM Verlag Dr Mueller. He has been the director of undergraduate studies in statistics at Federal University of Pernambuco-UFPE, Brazil (2005–2009). Dr Ferraz is currently a faculty member of the UFPE graduate program in statistics, and has been working on sampling and design of experiments.

## Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Causation and Causal Inference
►Clinical Trials, History of
►Confounding and Confounder Control
►Design of Experiments: A Pattern of Progress
►Experimental Design: An Introduction
►Medical Research, Statistics in
►Medical Statistics
►Misuse of Statistics
►Permutation Tests
►Philosophical Foundations of Statistics
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Randomization Tests
►Research Designs
►Statistical Fallacies: Misconceptions, and Myths
►Statistics: An Overview
►Superpopulation Models in Survey Sampling

## References and Further Reading

Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Hinkelmann K, Kempthorne O (1994) Design and Analysis of Experiments, vol 1. Wiley-Interscience, New York
Särndal C-E, Swensson B, Wretmann J (1992) Model assisted survey sampling. Springer, New York
Tillé Y (2006) Sampling algorithms. Springer, New York

# Randomization Tests

Eugene S. Edgington
Professor Emeritus
University of Calgary, Calgary, AB, Canada

A randomization test is a permutation test (see ►Permutation Tests) that is based on randomization (random assignment), where the test is carried out in the following way. A test statistic (such as a difference between means) is computed for the experimental data (measurements or observations). Then the data are repeatedly divided or rearranged in a manner consistent with what the random assignment procedure would have produced if the treatments had no differential effect. The test statistic is computed for each of the resulting data permutations. Those data permutations, including the one for the experimental results, constitute the reference set for determining significance. The proportion of data permutations in the reference set having test statistic values greater than or equal to (or for certain test statistics, less than or equal to) the value for the experimental results is the *p*-value (significance or probability value). Determining significance on the basis of a distribution of test statistics generated by permuting the data is characteristic of all permutation tests; it is when the basis for permuting the data is random assignment (not random sampling) that a permutation test is called a randomization test.

The null hypothesis for a randomization test is that the measurement for each experimental unit (e.g., a subject or a plot of land) is the same under one assignment to treatments as under any alternative assignment. Thus, under the null hypothesis, assignment of experimental units to treatments randomly divides the measurements among the treatments. Each data permutation in the reference set represents the results that, if the null hypothesis is true, would have been obtained for a particular assignment. A randomization test is valid for any kind of sample, no matter how the sample is selected. This is an extremely important property because the use of non-random samples is common in experimentation, and parametric statistical tables (e.g., *t* and *F* tables) are not valid for such samples.

The validity of parametric statistical tables depends on random samples, and the invalidity of application to non-random samples is widely recognized. The random sampling assumption underlying the parametric significance tables is that of a sampling procedure that gives all possible samples of *n* individuals within a specified population

the same probability of being drawn. Arguments regarding the "representativeness" of a non-randomly selected sample are irrelevant to the question of its randomness: a random sample is random because of the sampling procedure used to select it, not because of the composition of the sample. Thus random selection is necessary to ensure that samples are random.

It must be stressed that violation of the random sampling assumption invalidates parametric statistical tables not just for the occasional experiment but for virtually all experiments involving statistical tests. A person conducting a poll may be able to enumerate the population to be sampled and select a random sample by a lottery procedure, but an experimenter would not have enough time, money, or information to take a random sample of the population of the world in order to make statistical inferences about people in general. Not many experiments in biology, education, medicine, psychology, or any other field use randomly selected subjects, and those that do usually concern populations so specific as to be of little interest. For instance, when human subjects for psychological experiments are selected randomly, often they are drawn from a population of students who attend a certain university, are enrolled in a particular class, and are willing to serve as subjects. Biologists and others performing experiments on animals generally do not even pretend to take random samples although they commonly use standard hypothesis testing procedures designed to test null hypotheses about populations. These well-known facts are mentioned here as a reminder of the rareness of random samples in experimentation and of the specificity of the populations on those occasions when random samples are taken.

In most experimentation the concept of population comes into the statistical analysis because it is traditional to discuss the results of a statistical test in terms of inferences about populations, not because the experimenter has sampled randomly some population to which he wishes to generalize. The population of interest to the experiment is likely to be one that cannot be sampled randomly. Random sampling by a lottery procedure, a table of random numbers, or any other device requires sampling a finite population, but experiments of a basic nature are not designed to find out something about a particular finite existing population. For example, with either animals or human subjects the intention is to draw inferences applicable to individuals already dead and individuals not yet born, as well as those who are alive at the present time. If we were concerned only with an existing population, we would have extremely biological laws because every minute some individuals are born and some die, producing a continual change in the existing population. Thus the population of interest in most experiments is not one about which statistical inferences can be made because it cannot be sampled randomly.

A number of desirable properties of randomization tests are a function of their intelligibility. A knowledge of calculus or other aspects of "advanced mathematics" is unnecessary for an experimenter to develop a new randomization test, using only his statistical knowledge of finite statistics involving combinations and permutations. The way in which random assignment is carried out in an experiment permits an experimenter to see whether the method of permuting the data is valid for that experiment for either simple or complex randomization tests. Neither the producer nor the consumer of randomization tests needs to rely on unknown authorities to justify their decision regarding the validity of a randomization test – or of its invalidity. For professors who enjoy making their students think instead of memorize, teaching randomization tests is enjoyable, and the pleasure of reasoning at a level that permits a student to develop new statistical tests that are custom-made to fit a new type of experimental design can appeal to ingenuity of many students.

## About the Author

Dr. Eugene Sinclair Edgington is Emeritus Professor of Psychology at the University of Calgary in Calgary, Alberta, Canada. He received the B.S. (1950) and M.S. (1951) degrees in psychology from Kansas State University, and the Ph.D. (1955) degree in psychology from Michigan State University. He is a member of American Statistical Association and American Psychological Association. Professor Edgington has written numerous papers and is the author of the well known book *Randomization tests* (Marcel Dekker, Inc., 1980; 4th edition with Patrick Onghena, Chapman and Hall/CRC 2007).



## Cross References

►Nonparametric Statistical Inference
►Permutation Tests
►Randomization
►Robust Inference

## References and Further Reading

Edgington ES (1980) Randomization tests. Marcel Dekker, New York
Edgington ES (1987) Randomization tests, 2nd edn. Marcel Dekker, New York
Edgington ES (1995) Randomization tests, 3rd edn. Marcel Dekker, New York
Edgington ES, Onghena P (2007) Randomization tests, 4th edn. Chapman & Hall/CRC, New York

# Rank Transformations

W. J. Conover
Horn Professor of Statistics
Texas Tech University, Lubbock, TX, USA

## Statistics

The Science of Statistics is concerned with the analysis of data. This analysis may be as simple as presenting a graph, or finding an average. In more complex analyzes a statistical model may be assumed, and inferences may be made concerning the more general characteristics of the population of data from which a sample of data was obtained.

## Parametric Versus Nonparametric

If the statistical model involves assumptions regarding the distribution of probabilities that govern the population of data then the resulting statistical methods are usually called "parametric." If the statistical model involves assumptions, but not assumptions regarding the distribution of probabilities governing the population, then the resulting statistical methods are usually called "nonparametric" or "distribution-free." Many of the best nonparametric methods involve the ranks of the data rather than the data itself. By ranks of the data it is meant that the smallest observation in the data set is given rank 1, the second smallest is given rank 2, and so forth.

## Rank Transformation

Parametric methods usually have some optimum property for the parametric model, but are often inferior to the nonparametric method when the parametric model is not appropriate. It is convenient in those cases to "transform" the data to ranks, and to use the parametric method on the ranks instead of the data. These are called "rank transformation methods."

## Example

In some cases the rank transformation method results in a nonparametric method. An early example involves Spearman's rho, published in 1904, which is simply the Pearson product-moment correlation coefficient $r$ calculated on the ranks of the data rather than the data itself. Thus one can test the hypothesis of independence of two variables paired as in $(X, Y)$, without any assumptions regarding the nature of the bivariate distribution from which they came, while the parametric model assumes a bivariate normal distribution. The observations on $X$ are replaced by their ranks from 1 to $n$, the observations on $Y$ are replaced by their ranks from 1 to $n$, and the ranks are placed in the original $n$ pairs where the original data were. The usual correlation coefficient $r$ is calculated on the ranks instead of the data, and the usual hypothesis test is conducted. In the case of independence of $X$ and $Y$ the distribution of Spearman's rho is asymptotically (for large $n$) the same as the distribution of Pearson's $r$. The exact distribution of Spearman's rho can be found, and is given in tables (see Conover 1999, for example), which is useful when $n$ is small.

## RT-1

There are several classifications of the transformation to ranks, as outlined by Conover and Iman (1981). The first type of rank transformation involves ranking all of the data together as one group, from smallest to largest, and then replacing the data in each of the original groups with their ranks. For example, in the case of two independent samples the observations in each sample are replaced with their ranks in the combined sample. Then for a test of equal means the two-sample $t$-test is computed on the ranks instead of the original data, and the test statistic is compared with the $t$-distribution in the usual way as an approximate test. With small sample sizes exact distributions of the test statistic can be obtained. This is equivalent to the *Mann-Whitney Test*, also known as the *two-sample Wilcoxon Test* (see ►Wilcoxon–Mann–Whitney Test). An extension to the case of several independent samples is obvious, with the one-way ►analysis of variance being computed on the ranks instead of the original data, and the $F$-tables being consulted for significance. This is equivalent to the *Kruskal-Wallis Test*. Details of these tests may be found in Conover (1999).

## RT-2

A second type of rank transformation involves subsets of the data being ranked separately from other subsets, as in the correlation case mentioned earlier where the observations on $X$ were ranked among themselves, and the observations on $Y$ were ranked among themselves. The original data are replaced by their resulting ranks, and the statistic computed on the ranks. If independence is of interest, $r$ is calculated, resulting in Spearman's rho, as mentioned earlier.

Another example of this second type of rank transformation is in the two-way layout, where the observations in each row are ranked among themselves only, and a two-way analysis of variance is computed on these ranks to see if there is a significant difference in the column means.

Thus we obtain a form of the *Friedman Test*. In the case of only two columns we obtain a form of the ►*Sign Test*.

### RT-3

This brings us to a third type of rank transformation, where the ranks are determined after an appropriate re-expression of the data. Again consider pairs of data, $n$ observations on a bivariate $(X, Y)$ random variable. If the null hypothesis is equal means rather than independence, as was the case in the previous paragraph, the differences $X - Y$ are first computed, and then these differences are ranked on the basis of their absolute values $|X - Y|$ with the smallest absolute difference getting rank 1, and so on. Then the signs of the difference are applied to the ranks, and the one-sample t-test is computed on these signed ranks.

### RT-4

The fourth type of rank transformation is an extension of the second type and third type combined. That is, subgroups of data are re-expressed, such as by subtracting a covariate or dividing by the consumer price index. Then each re-expressed subgroup is ranked by itself, and the standard parametric test is applied to the ranks. This could lead to an ►analysis of covariance by testing equality of means on the ranks of the re-expressed groups.

Another example of this fourth type of rank transformation is a nonparametric test for equal slopes presented by Hogg and Randles (1975). Several groups of paired data $(X, Y)$ are first combined to find the least squares regression estimate $y = a+bx$. Then the residuals $(Y-y)$ from this model are ranked overall, and compared with the ranks of the $X's$ as described in their paper in a rank version of the parametric test for the same hypothesis.

### Discussion

The rank transformation may result in a nonparametric test as indicated in the examples above, or the result may be a robust test such as when the first type of rank transform is applied to a two-way layout, or it may result in a test that is not even always robust such as applying the first type of rank transformation to a two-way layout with several observations per cell and trying to test for interaction.

### About the Author

Dr. William Jay Conover received his Bachelor of Science in Mathematics from Iowa State University and his Master of Arts and Ph.D. in Mathematical Statistics from Catholic University of America. He is Paul Whitfield Horn Professor of Statistics, area of Information Systems and Quantitative Sciences, College of Business Administration, Texas Tech University. He was Elected Fellow of the American Statistical Association "for significant contributions to nonparametric statistics, for wide ranging and effective statistical consulting, and for excellence as a teacher and administrator" (1979). Among many his awards, Professor Conover was awarded the 1986 Don Owen Award and 1999 Wilks Medal. He has (co-)authored about 40 refereed papers and several books, including highly-regarded text *Practical Nonparametric Statistics* (3rd edition, John Wiley & Sons, 1999). His publications have been cited approximately 400 times each year for the last two decades, as reported by the Science Citation Index and the Social Sciences Citation Index. He is listed in in Who's Who in America (since 1987), and in Who's Who in the World, (since 1990–1991).

### Cross References

►Measures of Dependence
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Scales of Measurement and Choice of Statistical Methods
►Statistical Fallacies: Misconceptions, and Myths
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

### References and Further Reading

Conover WJ (1999) Practical nonparametric statistics, 3rd edn. Wiley, New York
Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 33:124–129
Hogg RV, Randles RH (1975) Adaptive distribution-free regression methods and their application. Technometrics 17:399–407

## Ranked Set Sampling

Douglas A. Wolfe
Professor and Chair
The Ohio State University, Columbus, OH, USA

### Introduction

In experimental settings where data are collected with the goal of making inferences about some aspects of an underlying population it is always important to design the study in such a way as to obtain as much useful information as possible while minimizing the overall cost of the experiment. This is particularly true when the initial step in collecting these data is to select the particular units from the finite or infinite population on which measurements

are to be taken. In this context, the goal of minimizing experimental cost is most often equivalent to minimizing the sample size while still achieving the desired accuracy of the inferences that follow.

The most commonly used approach for collecting data from a population is that of a simple random sample (SRS). If the population is infinite, the observations in a SRS are independent and identically distributed random variables. Even if the population is finite and sampling is done without replacement so that the sample observations are no longer independent, there is still a probabilistic guarantee that each measurement in the SRS can be considered as representative of the population. Despite this assurance, there is a distinct possibility that a specific SRS might not provide a truly representative picture of the complete population and larger sample sizes might be required to guard against such atypical samples.

Statisticians have, of course, developed a number of ways to guard against such unrepresentative samples without resorting to unduly large sample sizes. Sampling designs such as stratified sampling, probability sampling, and ►cluster sampling all provide additional structure on the sampling process that improves the likelihood that a collected sample provides a good representation of the population while trying to control the sampling costs involved in both the selection of the units to include in the sample and the cost of making the actual measurements on the selected units.

A novel sampling approach with this goal in mind was introduced by McIntyre (1952, reprinted in 2005) for situations where taking the actual measurements for sample units is difficult (e.g., costly, destructive, time-consuming, etc.), but there are inexpensive mechanisms readily available for either informally or formally ranking a set of sample units. Sample data collected via such a preliminary ranking scheme are known in the literature as ranked set sample (RSS) data.

## Balanced Ranked Set Samples

To obtain a RSS of $k$ observations from a population, we first select a SRS of $k^2$ units from the population and randomly divide them into $k$ subsets of $k$ units each. Within each of these subsets, the $k$ units are rank ordered (least to greatest) by some informative mechanism (such as visual comparisons, expert opinion, or through the use of auxiliary variables) that does not involve actual measurements on the attribute of interest for the sample units. The unit that is judged to be the smallest in the first of these rank ordered subsets is then included in the RSS and the attribute of interest is formally measured for this unit. This measurement is denoted by $X_{[1]}$, where [1] is used instead of the usual round bracket (1) for the smallest order statistic because $X_{[1]}$ is only judgment ranked to be the smallest among the $k$ units in the first subset; it may or may not actually have the smallest measurement among the $k$ units.

The same ranking process is used to judgment rank the second subset of $k$ units and the item ranked as the second smallest of the $k$ units is selected and its attribute measurement, $X_{[2]}$, is obtained and added to the RSS. From the third subset of size $k$ we select the unit judgment ranked to be the third smallest and add its attribute measurement, $X_{[3]}$, to the RSS. This process continues until we add the attribute measurement for the unit ranked to the largest of the k units in the final subset of size $k$, denoted by $X_{[k]}$, to the RSS.

The resulting collection of $k$ measurements $X_{[1]}, \ldots, X_{[k]}$ is called a *balanced ranked set sample of size k*, where the term balanced refers to the fact that we have collected one judgment order statistics for each of the ranks $1, 2, \ldots, k$. This entire process is called a *cycle* and $k$ is the *set size*. To obtain a balanced RSS with a desired total number of measured observations (i.e., total sample size) $n = kq$, we repeat the entire process for $q$ independent cycles, yielding the balanced RSS of size $n$ : $X_{[i]j}$, $i = 1, \ldots, k$; $j = 1, \ldots, q$.

To illustrate the advantages of RSS over SRS, we consider the problem of estimation of a population mean. Let $X_1, \ldots, X_n$ be a SRS of size $n$ from a distribution with mean $\mu$ and finite variance $\sigma^2$. Let $X_{[1]}, \ldots, X_{[n]}$ be the judgment order statistics for a balanced RSS from this distribution based on a single cycle with set size $n$. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \overline{X}^* = \frac{1}{n}\sum_{j=1}^{n} X_{[j]}$$

be the corresponding SRS and RSS sample means, respectively. It is well known that the SRS mean $\overline{X}$ is unbiased for $\mu$ and that $Var(\overline{X}) = \frac{\sigma^2}{n}$. Dell and Clutter (1972) and Takahasi and Wakimoto (1968) showed that the RSS sample mean $\overline{X}^*$ is also an unbiased estimator of $\mu$, and this is true even when there are errors in the ranking mechanism used to obtain the RSS data. Moreover, they provided an explicit formula for the variance of $\overline{X}^*$, namely,

$$\begin{aligned} Var(\overline{X}^*) &= \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2 \\ &= Var(\overline{X}) - \frac{1}{n^2}\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2, \end{aligned} \tag{1}$$

where $\mu^*_{[j]} = E(X_{[j]})$, for $j = 1, \ldots, n$.

Since $\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2 \geq 0$, it follows from Eq. (1) that the variance of the SRS mean $\overline{X}$ is always at least as large as the

variance of the RSS mean $\overline{X}^*$, regardless of the accuracy of the ranking process. Thus the RSS mean $\overline{X}^*$ is a more precise estimator of the population mean $\mu$ than the SRS mean $\overline{X}$ based on the same number of measured observations. The gain in precision is a monotonically increasing function of the quantity $\sum_{j=1}^{n}\left(\mu^*_{[j]}-\mu\right)^2$, which is itself an increasing function of the accuracy of the judgment rankings. The more reliable the judgment ranking process, the more separated will be the judgment order statistic expectations, $\mu^*_{[r]}, r = 1,\ldots,n$, and the more improvement we can expect from using RSS instead of SRS. The worst-case scenario where there is no gain from using RSS occurs when $\mu^*_{[1]} = \mu^*_{[2]} = \cdots = \mu^*_{[n]} = \mu$, which corresponds to no information in our ranking process and thus completely random rankings.

## Unbalanced Ranked Set Samples

For most settings, balanced RSS is the natural and preferred approach. There are, however, settings where measuring different numbers of the various judgment order statistics (unbalanced RSS) can lead to improved RSS procedures. This is the case, for example, when we are estimating the location parameter $\theta$ for a unimodal, symmetric distribution. In that setting when the ranking process is reasonably accurate, the optimal RSS would be to measure the sample median from each of the $k$ sets, resulting in an extremely unbalanced RSS, and then estimate $\theta$ by the average of these $k$ set sample medians. Stokes (1995), Bhoj (1997), and Kaur et al. (1997) were the first to point out the optimality of unbalanced RSS under appropriate conditions.

## Other Factors Affecting RSS

Properties of procedures based on RSS data are affected by a number of factors that are unique to this sampling approach. First, and probably foremost, is the accuracy of the ranking process. While the balanced RSS sample mean is always as efficient as the SRS sample mean based on the same number of measured observations, this gain in efficiency can be minimal if the ranking process is not reasonably accurate. Moreover, the SRS sample mean can even be more efficient than estimators based on unbalanced RSS data when the ranking process is not reliable. There have been a number of approaches in the literature to modeling this degree of imperfection in the rankings. Frey (2007) provides a general discussion of these approaches and presents a broad class of imperfect ranking models that can be used to assess the effect of imperfect ranking on RSS procedures. A second factor that affects the properties of RSS procedures is the set size. Generally speaking, the effectiveness of RSS procedures improves with increasing set size but this is counterbalanced by the fact that the ranking accuracy generally decreases with increased set size. Finally, the relative costs of sampling, ranking, and measuring units can be an important factor to consider in evaluating RSS versus SRS competitors.

## Resources

The original paper by McIntyre (1952, 2005) is a good place to start with understanding the motivation behind the RSS sampling approach. Kaur et al. (1995) and Patil (1995) provide general reviews of the research and applications involving RSS data and Wolfe (2004) gives a general introduction to RSS methodology with a special emphasis on nonparametric procedures. Cheng et al. (2004) have the only monograph/textbook on the subject.

## About the Author

Douglas Wolfe is Professor and Chair, Department of Statistics at The Ohio State University in Columbus. He is a co-author of two well known texts: *Nonparametric Statistical Methods* (with Myles Hollander, Wiley-Interscience, 2nd edition 1999) and *Introduction to the Theory of Nonparametric Statistics* (with Ronald Randles, Wiley, 1979). Professor Wolfe is a two-time recipient of the Ohio State University Alumni Distinguished Teaching Award (1973–1974) and (1988–89). He was a member of the Noether Award Selection Committee, (2007–2010) and was Chair, ASA Nonparametric Statistics Section (2009).

## Cross References

►Handling with Missing Observations in Simple Random Sampling and Ranked Set Sampling
►Order Statistics
►Ordered Statistical Data: Recent Developments
►Ranks
►Simple Random Sample

## References and Further Reading

Bhoj DS (1997) New parametric ranked set sampling. J Appl Stat Sci 6:275–289
Cheng Z, Bai ZD, Sinha BK (2004) Ranked set sampling: theory and applications. Springer, New York
Dell TR, Clutter JL (1972) Ranked set sampling theory with order statistics background. Biometrics 28:545–555
Frey J (2007) New imperfect rankings models for ranked set sampling. J Stat Planning Infer 137:1433–1445
Kaur A, Patil GP, Sinha AK, Taillie C (1995) Ranked set sampling: an annotated bibliography. Environ Ecol Stat 2:25–54
Kaur A, Patil GP, Taillie C (1997) Unequal allocation models for ranked set sampling with skew distributions. Biometrics 53: 123–130

McIntyre GA (1952, 2005) A method for unbiased sampling, using ranked sets. Aust J Agri Res 3:385–390. Reprinted in The American Statistician 59(3):230–232
Patil GP (1995) Editorial: ranked set sampling. Environ Ecol Stat 2:271–285
Stokes SL (1995) Parametric ranked set sampling. Ann Inst Stat Math 47:465–482
Takahasi K, Wakimoto K (1968) On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann Inst Stat Math 20:1–31
Wolfe DA (2004) Ranked set sampling: an approach to more efficient data collection. Stat Sci 19(4):636–643

# Ranking and Selection Procedures and Related Inference Problems

S. Panchapakesan
Professor Emeritus
Southern Illinois University, Carbondale, IL, USA

## Introduction

A statistical ranking or selection procedure is typically called for when the experimenter (the decision-maker) is faced with the problem of comparing a certain number $k$ of populations in order to make a decision about preferences among them.

Consider $k$ populations, each characterized by the value of a parameter $\theta$. In an agricultural experiment, for example, the different populations may represent different varieties of wheat and the parameter $\theta$ may be the average yield of a variety. The classical approach in this situation is to test the so-called homogeneity hypothesis $H_0$ that $\theta_1 = \cdots = \theta_k$, where the $\theta_i$ are the unknown values of the parameter for the $k$ populations. In the case of the familiar one-way classification model, the populations are assumed to be normal with unknown means $\theta_1, \ldots, \theta_k$, and a common unknown variance $\sigma^2$. The homogeneity hypothesis $H_0$ is tested using Fisher's ▶analysis of variance (ANOVA) technique. However, this usually does not solve the real problem of the experimenter, which is not simply to accept or reject the homogeneity hypothesis. The real goal is often to choose the best population (the variety with the largest average yield). The inadequacy of the ANOVA is not in the design aspects of the procedure; it rather lies in the types of decisions that are made on the basis of the data. The attempts to formulate the decision problem in order to achieve this realistic goal of selecting the best treatment mark the beginnings of ranking and selection theory.

The formulation of a $k$-sample problem as a multiple decision problem enables one to answer the natural questions regarding the best populations. The formulation of multiple decision procedures in the framework of what has now come to be known as ranking and selection procedures began with the now-classic paper by Bechhofer (1954).

## Basic Formulations of the Ranking and Selection Problem

We have $k$ populations, $\Pi_1, \ldots, \Pi_k$, each indexed by a parameter $\theta$, where the cumulative distribution function (cdf) of $\Pi_i$ is $F(x;\theta_i)$ for $i = 1, 2, \ldots, k$. We assume that the family $\{F(x;\theta)\}$ is stochastically increasing in $\theta$, i.e., $F(x;\theta_1) \geq F(x;\theta_2)$ for $\theta_1 \leq \theta_2$ for all $x$, and that the parameters can be ordered from the smallest to the largest. Denote the true ordered $\theta$-values by $\theta_{[1]} \leq \theta_{[2]} \leq \ldots \leq \theta_{[k]}$. To fix ideas, we assume that larger the value of $\theta$, more preferable is the population. Hence, the population associated with $\theta_{[k]}$ is called the *best* population. We assume that there is no prior information as to the correspondence between the ordered and the unordered $\theta_i$. Ranking and selection problems have generally been formulated adopting one of two main approaches known as the *indifference-zone formulation* and the *subset selection formulation*.

In the indifference-zone formulation due to Bechhofer (1954), the goal is to select a fixed number of populations. Consider the basic goal of selecting the one best population. Based on samples of size n taken from each population, we seek a procedure to select one of the populations as the best. The natural procedure would be to compute estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ from each sample and claim that the population that yielded the largest $\hat{\theta}_i$ is the best population. Here, a correct selection occurs when the selected population is the best. We require a guaranteed minimum probability of a correct selection (PCS), denoted by $P^*$, whenever $\theta_{[k]}$ is sufficiently larger than $\theta_{[k-1]}$. Let $\delta = \delta(\theta_{[k]}, \theta_{[k-1]})$ denote a suitably defined measure of the separation between the populations associated with $\theta_{[k]}$ and $\theta_{[k-1]}$. Let $\Omega = \{\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)\}$. Define $\Omega(\delta^*) = \{\vec{\theta} \,|\, \delta(\theta_{[k]}, \theta_{[k-1]}) \geq \delta^* > 0\}$. For specified $\delta^*$ and $P^*(1/k < P^* < 1)$, it is required that

$$PCS \geq P^* \text{ whenever } \vec{\theta} \in \Omega(\delta^*). \tag{1}$$

To be meaningful, we choose $P^* > 1/k$; otherwise, the requirement (1) can be met by randomly choosing one of the populations as the best. The region $\Omega(\delta^*)$ of the parameter space $\Omega$ is called the *preference-zone* (PZ) as this is where we have strong preference for a correct selection.

The complement of the PZ is known as the *indifference-zone* (IZ), a region where we do not require a guaranteed PCS. The PCS in the PZ depends, in general, on the configuration of $\vec{\theta}$. In many cases, there is a *least favorable configuration* (LFC) of $\vec{\theta}$ for which the PCS attains a minimum over the PZ for any sample size. If we can make the PCS at the LFC equal to $P^*$, then the probability requirement (1) will be satisfied. The usual choices for $\delta = \delta(\theta_{[k]}, \theta_{[k-1]})$ are $\delta = \theta_{[k]} - \theta_{[k-1]}$ in the case of a location parameter and $\delta = \theta_{[k]}/\theta_{[k-1]}$ in the case of a scale parameter. In the case of nonnegative $\theta$ which is not a scale parameter, one may choose either of these two special forms depending on other aspects of the problem.

Bechhofer (1954) introduced the IZ formulation by considering $k$ normal populations with means $\theta_1, \ldots, \theta_k$, and a common known variance $\sigma^2$. Here, $\delta = \theta_{[k]} - \theta_{[k-1]}$. Based on samples of size $n$ from these normal populations, he proposed the natural selection procedure, say $R_1$, which selects the population that yielded the largest sample mean. The LFC for $R_1$ is $\theta_{[1]} = \cdots = \theta_{[k-1]} = \theta_{[k]} - \delta^*$. For a specified $(\delta^*, P^*)$, the minimum sample size needed to meet the probability requirement (1) is given by

$$n = \left\langle 2\left(\sigma H/\delta^*\right)^2 \right\rangle, \tag{2}$$

where $\langle b \rangle$ stands for the smallest integer greater than or equal to $b$, $H$ satisfies

$$\Pr\{Z_1 \le H, \ldots, Z_{k-1} \le H\} = P^*, \tag{3}$$

and the $Z_i$ are standard normal variables with equal correlation $\rho = 0.5$.

Some generalized goals that have been considered are: (I) Selecting the $t$ best populations for $t \ge 2$, (a) in an ordered manner or (b) in an unordered manner, and (II) Selecting a fixed subset of size $m$ that will contain at least $s$ of the $t$ best populations.

The first of these itself is a special case of the general ranking goal of Bechhofer (1954), which is to partition the set of $k$ populations into $s$ nonempty subsets $I_1, I_2, \ldots, I_s$ consisting of $k_1, k_2, \ldots, k_s$ $(k_1 + k_2 + \cdots + k_s = k)$ populations, respectively, such that for $\Pi_i \in I_\alpha$, $\Pi_j \in I_\beta$, $1 \le \alpha < \beta \le s$, we have $\theta_i < \theta_j$.

In the above general ranking problem, Fabian (1962) introduced the idea of $\Delta$–correct ranking. Roughly speaking, a ranking decision is $\Delta$–correct if wrongly classified populations are not too much apart. The special case of $s = 2$ and $k_1 = k - 1$ for a location parameter family is of interest. In this case, a $\Delta$-correct ranking is equivalent to selecting one population $\Pi_i$ for which $\theta_i > \theta_{[k]} - \Delta$, $\Delta > 0$; such a population is called a *good* population. The goal of selecting a good population has been considered by several subsequent authors.

In the normal means selection problem of Bechhofer (1954) mentioned previously, if the common variance $\sigma^2$ is unknown, a single-sample procedure does not exist. It can be seen from (2) that the minimum sample size $n$ needed in order to satisfy the probability requirement (1) cannot be determined without the knowledge of the variance. In this case, a two-stage selection procedure is necessary to control the PCS. The first two-stage procedure for this problem was studied by Bechhofer et al. (1954). This procedure uses the first stage samples to obtain an estimate of $\sigma^2$.

In the subset selection formulation for selecting the best (i.e., the population associated with $\theta_{[k]}$), we seek a rule which will select a nonempty subset of random size that includes the best population. Here no assertion is made about which population in the selected subset is the best. The size $S$ of the selected subset is determined by the sample data. In contrast with the IZ formulation, there is no specification of a PZ (or an IZ). The experimenter specifies $P^*$, the minimum PCS to be guaranteed no matter what the unknown values of the $\theta_i$ are. The selection rule is based on the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$. The expected size of the selected subset is a performance characteristic of a procedure.

In the case of normal means problem, assuming a common known variance $\sigma^2$, Gupta (1954, 1965) proposed a procedure based on a sample of size $n$ from each population. This rule, say $R_2$, selects population $\Pi_i$ if the sample mean $\overline{X}_i$ from it satisfies:

$$\overline{X}_i \ge \overline{X}_{[k]} - d\sigma/\sqrt{n}, \tag{4}$$

where $d$ is a positive constant to be chosen so that the minimum PCS is guaranteed. The LFC in this case is given by $\theta_1 = \theta_2 = \cdots = \theta_k$. By equating the PCS at the LFC to $P^*$, we get $d = \sqrt{2H}$, where $H$ is given by (3).

When $\sigma^2$ is unknown, Gupta (1954) proposed the rule $R_3$ which is $R_2$ with $\sigma^2$ replaced by the pooled sample variance $s^2$ based on $\nu = k(n-1)$ degrees of freedom and a different constant $d'$, which turns out to be the one-sided upper $(1 - P^*)$ equicoordinate point of the equicorrelated $(k-1)$–variate central $t$ distribution with equal correlation $\rho = 0.5$ and the associated degrees of freedom $\nu$.

Seal (1955) proposed a class of procedures that included Gupta's maximum-type procedure and an alternative (average-type) procedure that Seal advocated using. The superiority of Gupta's procedure under certain slippage configurations and with regard to certain optimality properties and its comparative ease in handling theoretical details accelerated the growth of the subset selection literature.

Subset selection can be thought of as a screening procedure towards selecting one population as the best. The IZ approach has no requirements regarding correct selection when the true parametric configuration lies in the IZ, whereas the (random size) subset selection formulation does not control the size of the selected subset. A modified formulation, called the *restricted subset selection*, puts an upper bound for the (random) size of the selected subset (see Santner 1975). Using the restricted subset selection formulation for the goal of selecting a subset of the $k$ populations whose size does not exceed $m(1 \le m \le k)$ so that the selected subset includes at least one of the $t(1 \le t \le k-1)$ best with a guaranteed probability, Panchapakesan (2005) has provided a fresh look at the salient features of the IZ and subset selection approaches.

Over the last almost six decades, several aspects of selection and ranking have been investigated. Substantial accomplishments have been made concerning procedures for specific univariate and multivariate parametric families, conditional procedures, nonparametric procedures, sequential and multistage procedures, procedures for restricted families such as the IFR (increasing failure rate) and IFRA (increasing failure rate on the average) distributions, decision-theoretic developments, and Bayes and empirical Bayes procedures. For detailed accounts of these, see Gupta and Panchapakesan (1979, 1985, 1996), Panchapakesan (2006) and the references contained therein.

## Inference Problems Associated with Ranking and Selection

One related inference problem is the point and interval estimation of the ordered parameters, $\theta_{[1]}, \ldots, \theta_{[k]}$. Some attempts have been made to combine selecting the population associated with $\theta_{[k]}$ and estimating $\theta_{[k]}$ with simultaneous probability control; see, for example, Rizvi and Lal Saxena (1974). Another related inference problem is the estimation of the PCS for a selection procedure; see, for example, Gupta et al. (1990). Another interesting problem is known as *estimation after selection* in which the interest is to estimate the parameter of the selected population in the case of a procedure for selecting one population, or to estimate a known function such as the average of the parameters of the selected populations in the case of subset selection. Here the object of inference depends on the sample data that are to be used in the procedure. Such a statistical procedure has been called a *selective inference procedure*. This is different from a nonselective inference procedure in which the identity of the object of inference is fixed and is determined before the data were obtained. For references to several papers dealing with this, see Gupta and Panchapakesan (1996) and Panchapakesan (2006).

In a given situation, we may use a natural rule to select the best population and may want to simultaneously test if the selected population is uniquely the best. Such a problem was first considered by Gutmann and Maymin (1987). Recently, a few papers have appeared dealing with location and scale parameter cases and selecting the best multinomial cell using inverse sampling. For a discussion of these, see Cheng and Panchapakesan (2009) and the references given therein.

## Concluding Remarks

Our aim here is to give a brief introduction to ranking and selection procedures. As such, we have given only a few important references. Gupta and Panchapakesan (1979) provide a comprehensive survey of the literature up to the date with a bibliography of some 600 main references. For references to later developments and other books on the subject see Gupta and Panchapakesan (1985, 1996) and Panchapakesan (2006).

## About the Author

Subrahmanian Panchapakesan is Professor Emeritus of Mathematics at the Southern Illinois University at Carbondale. He has published close to 90 journal articles, book chapters and reports, mostly on ranking and selection. He is a member of the ASA, IMS, and the International Statistical Institute (elected). Professor Panchapakesan is a co-author (with Shanti S. Gupta) of the well known text *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations* (John Wiley and Sons, 1979) and a co-editor (with N. Balakrishnan) of *Advances in Statistical Decision Theory and Applications* (Boston: Birkhäuser, 1997). He received the 2003 Thomas L. Saaty Prize for Applied Advances in the Mathematical and Management Sciences (awarded by the *American Journal of Mathematical and Management Sciences*). He is currently Associate Editor of *Communications in Statistics: Theory and Methods* and *Communications in Statistics: Simulation and Computation* (2002–), and an Editorial Board Member of *American Journal of Mathematical and Management Sciences* (1993–).

## Cross References

►Analysis of Variance
►Explaining Paradoxes in Nonparametric Statistics
►Multiple Statistical Decision Theory
►Sequential Sampling

## References and Further Reading

Bechhofer RE (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann Math Stat 25:16–39

Bechhofer RE, Dunnett CW, Sobel M (1954) A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika 41:170–176
Cheng S-R, Panchapakesan S (2009) Is the selected population the best?: location and scale parameter cases. Comm Stat Theor Meth 38:1553–1560
Fabian V (1962) On multiple decision methods for ranking population means. Ann Math Stat 33:248–254
Gupta SS (1956) On a decision rule for a problem in ranking means. PhD thesis (Mimeo Ser 150), University of North Carolina, Chapel Hill
Gupta SS (1965) On some multiple decision (selection and ranking) rules. Technometrics 7:225–245
Gupta SS, Leu L-Y, Liang T (1990) On lower confidence bounds for PCS in truncated location parameter models. Comm Stat Theor Meth 19:527–546
Gupta SS, Panchapakesan S (1979) Multiple decision procedures: theory and methodology of selecting and ranking populations. Wiley, New York (Reprinted by Society for Industrial and Applied Mathematics, Philadelphia, 2002)
Gupta SS, Panchapakesan S (1985) Subset selection procedures: review and assessment. Am J Math Manage Sci 5: 235–311
Gupta SS, Panchapakesan S (1996) Design of experiments with selection and ranking goals. In: Ghosh S, Rao CR (eds) Design and analysis of experiments. Elsevier, Amsterdam, pp 555–585
Gutmann S, Maymin Z (1987) Is the selected population the best? Ann Stat 15:456–461
Panchapakesan S (2005) Restricted subset selection procedures for normal means: a brief review with a fresh look at the classical formulations of Bechhofer and Gupta. Comm Stat Theor Meth 34:1265–1273
Panchapakesan S (2006) Ranking and selection procedures. In: Balakrishnan N, Read C, Vidakovic B (eds) Encyclopedia of statistical sciences, vol 10, 2nd edn. pp 6907–6915
Rizvi MH, Lal Saxena KM (1974) On interval estimation and simultaneous selection of ordered location or scale parameters. Ann Stat 2:1340–1345
Santner TJ (1975) A restricted subset selection approach to ranking and selection problems. Ann Stat 3:334–349
Seal KC (1955) On a class of decision procedures for ranking means of normal populations. Ann Math Stat 26:387–398

# Ranks

Elisabeth Svensson
Professor Emerita
Swedish Business School at Örebro University, Örebro, Sweden

## Uni-variate Ranking

A common approach in nonparametric statistical method is to transform data to ranks. A ranking of n ordered observations $\{x_1 < x_2 <, \ldots, < x_n\}$ will be a set of $n$ ranks $\{r_1 <, \ldots, < r_n\}$ where $x_i$ is represented by the rank $r_i$ in calculations. The rank sum is $\frac{1}{2}n(n+1)$, and $\sum r^2 = \frac{n^3 - n}{12}$. Hence the mean rank is $\frac{1}{2}(n+1)$ and the variance $\frac{1}{12}(n^2-1)$ assuming uniform distribution of all possible rankings. For untied observations the rank $r_i$ equals the number of observations less than $x_{i+1}$, $i = 1, \ldots, n$.

Assessments on scales with a limited number of categories will produce groups of observations that are tied to the same category, which means that these observations will share the same rank value. The midrank of an observation in the $i^{th}$ category, $i = 1, \ldots, m$, is then

$$\bar{r}_i = \sum_{v=1}^{i-1} x_v + \tfrac{1}{2}(x_i + 1),$$

where $x_v$ denotes the $v$th category frequency, $v = 1, \ldots, m$. Then $\sum r^2 < \frac{1}{12}(n^3 - n)$ and the variance is decreased, the correction term being

$$t^{(X)} = \sum_{v=1}^{m} \left(x_v^3 - x_v\right).$$

The mid-ranks of the marginal distribution $X$ of the Fig. 1 are (2, 9, 23, 41).

The calculations of the Wilcoxon–Mann–Whitney test statistics (see ►Wilcoxon–Mann–Whitney Test) of difference between two independent groups of data and of the Spearman rank-order correlation coefficient are based on this type of rank transformations.

## Augmented Ranking

In the evaluation of paired assessments made on rating scales regarding reliability of inter- or intra-rater agreement but also regarding change in outcome, the pairs of data can be transformed to ranks taking account of the information given by the pairs of data. In this *augmented ranking approach*, (aug-rank), by Svensson, the ranks are tied to the pairs of data $(X, Y)$, i.e., to the observations in the cells of a square contingency table alternatively to

| X / Y | $C_1$ | $C_2$ | $C_3$ | $C_4$ | Total |
|---|---|---|---|---|---|
| $C_4$ | | | 1 (31; 49) | 1 (50; 50) | 2 |
| $C_3$ | | 2 (13.5; 31.5) | 2 (29.5; 33.5) | 14 (42.5; 41.5) | 18 |
| $C_2$ | 1 (3; 15) | 1 (12; 16) | 11 (23; 22) | 3 (34; 29) | 16 |
| $C_1$ | 2 (1.5; 1.5) | 8 (7.5; 6.5) | 3 (16; 12) | 1 (32; 14) | 14 |
| Total | 3 | 11 | 17 | 19 | 50 |

**Ranks. Fig. 1** Example of a frequency distribution of paired assessments of a four-point rating scale, and the pairs of augmented ranks $\left(\bar{R}_{ij}^{(X)}; \bar{R}_{ij}^{(Y)}\right)$

| X / Y | $C_1$ | $C_2$ | $C_3$ | $C_4$ | Total |
|---|---|---|---|---|---|
| $C_4$ | | | | 2 | 2 |
| $C_3$ | | | 1 | 17 | 18 |
| $C_2$ | | | 16 | | 16 |
| $C_1$ | 3 | 11 | | | 14 |
| Total | 3 | 11 | 17 | 19 | 50 |

**Ranks. Fig. 2** The rank-transformable pattern of agreement, RTPA, uniquely defined by the two sets of marginal distributions

the points of a scatter plot of data from visual analogue scale assessments. This means that the augmented rank of the assessments $X$ depends on the pairing with $Y$.

The mean augmented rank according to the assessments $X$ is

$$\overline{R}_{ij}^{(X)} = \sum_{k=1}^{i-1}\sum_{l=1}^{m} x_{kl} + \sum_{l=1}^{j-1} x_{il} + \frac{1}{2}(1 + x_{ij})$$

for $1 \le i,j \le m$, where $x_{ij}$ is the *ijth* cell frequency, $i$ and $j = 1, \ldots, m$ and $m$ is the number of categories. The augmented mean rank of the observations in the *ijth* cell according to assessments $Y$, $\overline{R}_{ij}^{(Y)}$, is defined correspondingly, see Fig. 1. This aug-rank approach makes it possible to identify and separately analyse a possible systematic component of observed disagreement from the occasional, noise, variability, (see ►Measures of Agreement). A complete agreement in all pairs of aug-ranks, $\overline{R}_{ij}^{(X)} = \overline{R}_{ij}^{(Y)}$, for all $i$ and $j = 1, \ldots, m$ defines *the rank-transformable pattern of agreement* (RTPA), which is uniquely related to the two marginal distributions, see Fig. 2.

## About the Author

Professor Svensson is Past President Swedish Society for Medical Statistics (2000–2002). She is an Elected member of the International Statistical Institute (2002), member of the International Association for Statistical Education (2000), International Society for Clinical Biostatistics; (Member of Executive Committee 2001–2004), and Elected member of the scientific board of Statistics Sweden (2003–).

## Cross References

►Measures of Agreement
►Measures of Dependence
►Nonparametric Rank Tests
►Rank Transformations
►Ranking and Selection Procedures and Related Inference Problems
►Wilcoxon–Mann–Whitney Test

## References and Further Reading

Gibbons JD, Chakraborty S (2003) Nonparametric statistical inference, 4th edn (revised and expanded). Marcel Dekker, New York

Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw Hill, New York

Svensson E (1997) A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical J 39:643–657

# Rao–Blackwell Theorem

Arthur Cohen
Professor
Rutgers University, Piscataway, NJ, USA

The Rao–Blackwell Theorem (RB Theorem) attributed to C.R. Rao and David Blackwell links the notions of sufficient statistics and unbiased estimation. Let $\mathbf{X}$, a random vector represent the data. Assume the distribution of $\mathbf{X}$ depends on a parameter $\boldsymbol{\theta}$. A statistic $S(\mathbf{X})$ is said to be sufficient if the conditional distribution of $\mathbf{X}$ given $S$ does not depend on $\boldsymbol{\theta}$. A statistic $T(\mathbf{X})$ is said to be an unbiased estimator of $g(\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$, if $E_{\boldsymbol{\theta}} T(\mathbf{X}) = g(\boldsymbol{\theta})$ where $E$ stands for expected value. The RB Theorem, which is constructive says the following:

Let $U(\mathbf{X})$ be any unbiased estimator of $g(\boldsymbol{\theta})$ and let $\sigma_U^2$ be the variance of $U$. Let

$$W(\mathbf{X}) = E[U(\mathbf{X})|S(\mathbf{X})].$$

That is, $W(\mathbf{X})$ is the conditional expected value of $U(\mathbf{X})$ given $S(\mathbf{X})$. Then $W(\mathbf{X})$ is unbiased and $\sigma_U^2 \ge \sigma_W^2$, where $\sigma_W^2$ is the variance of W.

Evaluating unbiased estimators by their variance clearly corresponds to evaluating estimators using a squared error loss function. A well known extension of the RB Theorem is achieved by replacing a squared error loss function with any convex loss function.

The utility of the theorem is highlighted in situations where it is easy to find a simple unbiased estimator of $g(\boldsymbol{\theta})$. Sometimes this can be done using only a subset of the data and then the construction typically yields an excellent unbiased estimator. We proceed with some applications and an extension of the theorem.

One issue in quality control is to estimate the proportion of items produced whose measurements do not

meet specifications i.e. fall outside a given interval (L, U). Assuming measurements are normal with mean $\mu$ and variance $\sigma^2$ the quantity to estimate is

$$p = 1 - \int_L^U \varphi(z; \mu, \sigma^2) dz,$$

where $\varphi(z; \mu, \sigma^2)$ is the normal density. Based on a sample of size $n$, labeled $x_1, \ldots, x_n$ sufficient statistics are $\bar{x} = \sum x_i/n$, $s^2 = \sum(x - \bar{x})^2/(n-1)$. A simple unbiased estimator of p is

$$\begin{aligned} \hat{p} &= 0 \quad \text{if } L \le x_1 \le U. \\ &= 1 \quad \text{otherwise} \end{aligned}$$

Lieberman and Resnikoff (1955) Rao–Blackwellize $\hat{p}$ by deriving $E(\hat{p}|\bar{x}, s^2)$ resulting in what turns out to be the minimum variance unbiased estimator of $p$.

Cohen and Sackrowitz (1974) consider a common mean model. That is, consider two independent random samples, $x_1, \ldots, x_m$ from $N(\mu, \sigma_x^2)$ and $y_1, \ldots, y_n$ from $N(\mu, \sigma_y^2)$. In the course of estimating the common mean $\mu$ it was desired to seek an unbiased estimator of $\gamma = \sigma_x^2/(\sigma_x^2 + \sigma_y^2)$. For both $m$ and $n$ greater than or equal to 5, the sample could be split up in such a way to quickly find an unbiased estimator of $\gamma$. Then the simple estimator could be Rao–Blackwellized. This type of application is also suitable to find a good unbiased estimator of a correlation coefficient or intraclass correlation coefficient as was done in Olkin and Pratt (1958).

Cohen et al. (1985) consider the problem of estimating a quantile of a symmetric distribution. The cases of known and unknown centers of symmetry are studied. Convex combinations of a pair of ►order statistics from the sample are intuitive simple estimators of a quantile that exceeds 0.5. That is, suppose $X_{(1)} \le \ldots \le X_{(n)}$ are the order statistics from a population whose center of symmetry is known to be $\theta_0$. Then the ordered values of $Y_j = |X_j - \theta_0|$ are sufficient statistics. The Rao–Blackwellized version of the convex combination then is a superior estimator of the quantile in terms of mean squared error.

Given two independent samples from populations with distributions characterized by parameters $\theta_1$ and $\theta_2$ respectively, suppose population $i$, $i = 1$ *or* $2$ is selected if $\bar{X}_i$ is the larger sample mean. Suppose we wish to estimate the mean of the selected population. Note such a mean is a random variable. An estimator of a selected mean is said to be unbiased if its expected value equals the expected value of the selected mean. By taking an additional single observation from the selected population, and Rao–Blackwellizing it, using all the data Cohen and Sackrowitz (1989) display an estimator of the selected mean that is minimum variance conditionally unbiased under some assumptions regarding the underlying distributions.

An extension of the RB Theorem and the construction aspect of it appears in Brown et al. (1976). The extension gives a construction based on a conditional expectation of a decision procedure given the sufficient statistic that leads to a better procedure for all bowl shaped loss functions simultaneously even those that are not convex. Furthermore the construction preserves the property of median unbiasedness of any estimator.

## About the Author

Arthur Cohen is a Professor of Statistics at Rutgers University, New Jersey, USA. He served as chairperson from 1968–1977. He is a Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. He served as Editor of the *Annals of Statistics* from 1989–1991. He served as one of five editors of the *Journal of Multivariate Analysis* from 1978–1989. He also was an Associate editor of the *Journal of the American Statistical Association* and the *Journal of Statistical Planning and Inference*. He is the author of over 130 papers appearing in statistical journals.

## Cross References

►Adaptive Sampling
►Bivariate Distributions
►Estimation
►Loss Function
►Minimum Variance Unbiased
►Properties of Estimators
►Statistical Quality Control
►Sufficient Statistics
►Unbiased Estimators and Their Applications

## References and Further Reading

Brown LD, Cohen A, Strawderman WE (1976) A complete class theorem for Strict monotone likelihood ratio with applications. Ann Stat 4:712–722

Cohen A, Sackrowitz HB (1989) Two stage conditionally unbiased estimators of the selected mean. Stat Probab Lett 8: 273–278

Cohen A, Sackrowitz HB (1974) On estimating the common mean of two normal distributions. Ann Stat 2:1274–1282

Cohen A, Lo SH, Singh K (1985) Estimating a quantile of a symmetric distribution. Ann Stat 13:1114–1128

Lieberman GJ, Resnikoff GJ (1955) Sampling plans for inspection by variables. J Am Stat Assoc 50:457–516

Olkin I, Pratt J (1958) Unbiased estimation of certain correlation coefficients. Ann Math Stat 29:201–211

# Rating Scales

Elisabeth Svensson
Professor Emerita
Swedish Business School at Örebro University, Örebro, Sweden

Rating scales have been used in psychology and psychophysics for over 100 years, and the use of rating scales and other kinds of ordered classifications is nowadays inter-disciplinary and unlimited. As there are no standardised rules for the operational definitions of qualitative variables, a considerable variety in types of rating scales for the same variable, in different applications, is common. A *rating scale* consists of a number of ordered categorical recordings of an item.

*The verbal descriptive scale (VDS), also called the verbal rating scale*, consists of a discrete number of verbally described ordered response categories, or description of criteria, grading the level of responses. The set of categories can also refer to a time scaling, also called a frequency-of-use scale, like "often, seldom, never," Fig. 1.

A *Likert scale* is a type of VDS, the descriptive categories being agreement levels to statements. Figure 2 shows two different operational definitions of perceived health, a VDS-5 scale and a Likert scale from the same ▶questionnaire, The Short-Form 36 (SF-36). The categories of the VDS represent five levels of perceived health, from excellent to poor. The Likert scale has one level of the variable, in this case "excellent health," and five levels of agreement with the statement of excellent health. The response categories except for a complete agreement with the statement (definitely true) contain no information about other levels of health. Hence the Likert scale assessments are just comparable with the binary responses: yes my health is excellent, no my health is not excellent.

*A numerical rating scale (NRS)* consists of a range of numerals indicating the ordered response levels without any description of the categories, except from the end points. Figure 3 shows a seven-point NRS of pain.

*A visual analog scale (VAS)* consists of a straight line anchored by the extremes of the variable being measured. The variable can be measured by a bipolar construct of the VAS, the anchors being opposing adjectives, or by a mono-polar scale, the anchors being "no sign at all" to "the most extreme alternative." A rating method that combines the verbal descriptor scale and the VAS, called the *graphic rating scale (GRS)* consists of a line with no breaks or divisions. There should be three to five discrete categories beneath the horizontal line, and the extreme categories should not be worded such that they are never employed, see Fig. 4.

A *pictogram* is a visual scale, the categories being faces or other pictures with different expressions illustrating the variable of interest.

| VDS-6 intensity scale | VDS-4 grading of symptom | VDS-5 time scale |
|---|---|---|
| **How much. . . ?** | [ ] no evidence | **How often. . . ?** |
| [ ] Extremely high | [ ] slight signs | [ ] None of the time |
| [ ] Very high | [ ] moderate signs | [ ] A little of the time |
| [ ] Moderate | [ ] considerable signs | [ ] Some of the time |
| [ ] Slight | | [ ] Most of the time |
| [ ] Very low | | [ ] All of the time |
| [ ] non-existing | | |

**Rating Scales. Fig. 1** Examples of verbal descriptive scale categories

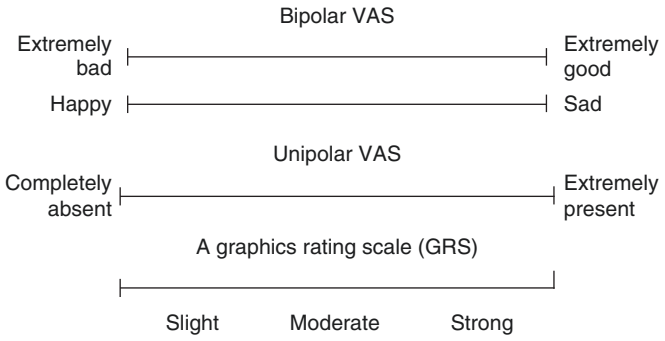
| VDS-5 scale of health | Likert scale of excellent health |
|---|---|
| In general, would you say your health is | My health is excellent |
| [ ] Excellent | [ ] Definitely true |
| [ ] Very good | [ ] Mostly true |
| [ ] Good | [ ] Don't know |
| [ ] Fair | [ ] Mostly false |
| [ ] Poor | [ ] Definitely false |

**Rating Scales. Fig. 2** Examples of two different types of scales for assessment of health from the Short-Form-36 (SF-36), item 1 and item 11d, respectively

"How is your pain now?"

| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| no pain at all | | | | | | unbearable pain |

**Rating Scales. Fig. 3** Example of a numeric rating scale (NRS-7) of pain




**Rating Scales. Fig. 4** Examples of Visual analogue scales (VAS) and a Graphic rating scale (GRS)

A *transitional scale*, the categories being *completely disappeared, much better, somewhat better, unchanged, somewhat worse, much worse* is useful when patient's perceived change after treatment is evaluated.

Assessments on rating scales, of any kind, produce ordinal data, the responses indicating only an ordering, although the use of numerical labelling could give a false impression of mathematical values. These so-called rank-invariant properties of ordinal data are well recognized, and several authors have stressed the fact that arithmetic operations are not appropriate for such data, therefore rank based statistical methods are recommended for analysis of data from rating scales.

## About the Author

For biography *see* the entry ►Ranks.

## Cross References

►Measures of Agreement
►Scales of Measurement
►Validity of Scales
►Variables

## References and Further Reading

Teeling Smith G (ed) (1988) Measuring health: a practical approach. Wiley, Chichester

Svensson E (2000a) Concordance between ratings using different scales for the same variable. Stat Med 19(24):3483–3496

Svensson E (2000b) Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biomet J 42: 417–434

# Record Statistics

Mohammad Ahsanullah[1], Valery B. Nevzorov[2]
[1]Professor
Rider University, Lawrenceville, NJ, USA
[2]Professor
St. Petersburg State University, St. Petersburg, Russia

In 1952, Chandler defined the so-called record times and record values and gave groundwork for a mathematical theory of records. For six decades beginning his pioneering work, about 500 papers and some monographs devoted to different aspects of the theory of records appeared. This theory relies largely on the theory of ►order statistics and is especially closely connected to extreme order statistics. Records are very popular because they arise naturally in many fields of studies such as climatology, sports, medicine, traffic, industry and so on. Such records are memorials of their time. The annals of records reflect the progress in science and technology and enable us to study the evaluation of mankind on the basic of record achievements in various areas of its activity. A large number of record data saved for a long time inspirited the appearance of different mathematical models reflecting the corresponding record processes and forecasting the future record results.

## Definitions of Records

Let $X_1, X_2, \ldots$ be a sequence of random variables and $X_{1,n} \le X_{2,n} \le \ldots \le X_{n,n}$, $n = 1, 2, \ldots$, be the corresponding

order statistics. For any $n = 1, 2, \ldots$ denote also $M(n) = X_{n,n} = \max\{X_1, X_2, \ldots, X_n\}$ and $m(n) = X_{1,n} = \min\{X_1, X_2, \ldots, X_n\}$. Now one can define the classical upper record times $L(n)$ and upper record values $X(n)$ as follows:

$$L(1) = 1,\ X(1) = X_1 \text{ and then}$$

$$L(n+1) = \min\{j : X_j > X(n)\}, \quad X(n+1) = X_{L(n+1)},$$

$$n = 1, 2, \ldots \tag{1}$$

One can use the following alternative definitions:

$$L(1) = 1, \quad L(n+1) = \min\{j : X_j > M(L(n))\}, \ldots$$
$$n = 1, 2, \ldots, \tag{2}$$

and

$$X(n) = M(L(n)), \quad n = 1, 2, \ldots$$

If we replace the sign $>$ and $M(L(n))$ in (2) by $<$ and $m(L(n))$, respectively, we obtain the definitions of the lower record times and the lower record values. Indeed, the theories of upper and lower records coincide practically in all their details, since the lower records for the sequence $X_1, X_2, \ldots$ correspond to the upper records for the sequence $-X_1, -X_2, \ldots$ Using the sign $\geq$ in (1) instead of $>$ we introduce the so-called weak upper records, when any repetition of the previous record value is also considered as a new record. Analogically, the sign $\leq$ in (1) gives the opportunity to define the weak lower record times and the weak lower record values. Note that the theory of weak records has practical meaning only for sequences of the initial $X'$s, which have discrete distributions.

The so-called $k$th records are a natural extension of records. The $k$th record times $L(n,k)$ and the $k$th record values $X(n,k)$ for any $k = 1, 2, \ldots$ are defined as follows:

$$L(1,k) = k, \quad L(n+1,k) = \min\{j > L(n,k) : X_j$$
$$> X_{j-k,j-1}\},\ n = 1, 2, \ldots, \tag{3}$$

and

$$X(n,k) = X_{L(n,k)-k+1,L(n,k)}, \quad n = 1, 2, \ldots \tag{4}$$

To be precise, (3) and (4) define the $k$th upper record times and the $k$th upper record values respectively. One can also introduce the $k$th lower record times and the $k$th lower record values changing the event $X_j > X_{j-k,j-1}$ in (3) by $X_j < X_{k,j-1}$ and replacing $X_{L(n,k)-k+1,L(n,k)}$ in (4) by $X_{k,L(n)}$. If $k = 1$ then definitions of $k$th record values $X(n,k)$ and $k$th record times $L(n,k)$ coincide with the definitions of $X(n)$ and $L(n)$ given in (1).

We will use $N(n)$ to denote the number of records among random variables $X_1, X_2, \ldots, X_n$, $n = 1, 2, \ldots$, and $N(n,k)$ will denote the number of $k$th records in a sequence $X_1, X_2, \ldots, X_n$ respectively.

## Sequential Ranks and Record Indicators

The essential role in the theory of records play the sequential ranks $R(n)$, the record indicators $\xi_j$, $j = 1, 2, \ldots$, and the indicators of the $k$*th* records $\xi_j(k)$, $j = 1, 2, \ldots$, $k = 1, 2, \ldots$. The sequential rank $R(n)$ denotes the rank of $X_n$ among $X_1, X_2, \ldots, X_n$, i.e.,

$$X_n = X_{R(n),n},\ n = 1, 2, \ldots$$

The record indicator $\xi_j$ is defined as follows: $\xi_j = 1$ if $X_j$ is a record value and $\xi_j = 0$ otherwise. Analogically, the indicator $\xi_j(k)$ can be defined for any $k = 1, 2, \ldots$ and $j \geq k$:

$\xi_j(k) = 1$, if $X_j$ is the $k$th record value and $\xi_j(k) = 0$ otherwise. There are some useful relations between record indicators and sequential ranks. We will formulate some simple equalities for indicators of the upper records. Indeed, analogical results are also valid for the lower records. Note, that

$$\{\xi_j = 1\} = \{R(j) = j\} = \{X_j = M(j)\} = \{X_j = X_{j,j}\}$$
$$= \{X_j \text{ is a record value}\}$$

and for any $k = 1, 2, \ldots$ and $j \geq k$

$$\{\xi_j(k) = 1\} = \{R(j) \geq j - k + 1\} = \{X_{j-k+1,j} \geq X_{j-k,j-1}\}$$
$$= \{X_j > X_{j-k,j-1}\} = \{X_j \text{ is a } k\textit{th} \text{ record}\}.$$

The record indicators allow us to give convenient relations for the numbers of records $N(n)$ and $N(n,k)$:

$$\{N(n) = m\} = \{\xi_1 + \ldots + \xi_n = m\},$$
$$n = 1, 2, \ldots,\ m = 1, 2, \ldots, n; \tag{5}$$
$$\{N(n,k) = m\} = \{\xi_k(k) + \ldots + \xi_n(k) = m\},$$
$$n = k, k+1, \ldots,\ m = 1, 2, \ldots, n-k+1. \tag{6}$$

The classical theory of records is connected with the situation when the initial sequence of random variables $X_1, X_2, \ldots$ is a sequence of independent random variables with a common continuous distribution function. In this case, sequential ranks and record indicators have a number of useful and rather convenient properties.

**Theorem 1** For independent identically distributed random variables $X_1, X_2, \ldots$ with a continuous distribution function $F$ the sequential ranks $R(1), R(2), \ldots$ are independent and $P\{R(n) = m\} = 1/n$, $m = 1, 2, \ldots, n$, $n = 1, 2, \ldots$

**Theorem 2** Under conditions of Theorem 1, for any fixed $k = 1, 2, \ldots$, indicators $\xi_k(k), \xi_{k+1}(k), \ldots$ are independent and $P\{\xi_j(k) = 1\} = k/j$, $j = k, k+1, \ldots$

As a partial case of Theorem 2 it follows that under conditions of Theorem 1 record indicators $\xi_1, \xi_2, \ldots$ are independent and $P\{\xi_j = 1\} = 1 - P\{\xi_j = 0\} = 1/j$, $j = 1, 2, \ldots$

These results together with equalities (5) and (6) allow us to find that the distributions of $N(n)$ and $N(n,k)$ are expressed as distributions of sums of independent random variables. One can also see that under conditions of Theorem 1, there are some convenient relations for the *kth* record times $L(n,k)$ and, in particular, for the record times $L(n)$ :

$$P\{L(n,k) > m\} = P\{N(m,k) < n\} = P\{\xi_k(k) + \xi_{k+1}(k) + \ldots + \xi_m(k) < n\} \quad (7)$$

and

$$P\{L(n) > m\} = P\{N(m) < n\} = P\{\xi_1 + \xi_2 + \ldots + \xi_m < n\}. \quad (8)$$

It follows from relations (7) and (8) that if $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with any continuous distribution function $F$, then distributions of record times and numbers of records do not depend on $F$.

One more important situation in the classical record theory is connected with sequences of independent identically distributed random variables having a discrete distribution. Without loss of generality, we can suppose that $X'$s take nonnegative integer values. For discrete distributions we introduce another type of record indicators. Let $\eta_n = 1$ if $n$ is a record value in the sequence $X_1, X_2, \ldots$, that is there exists such $m = 1, 2, \ldots$ that $X(m) = n$, and $\eta_n = 0$ otherwise (compare with indicators $\xi_n$!). Analogously, for any $k = 1, 2, \ldots$ we can introduce indicators $\eta_n(k)$ for $k$th record values : $\eta_n(k) = 1$, if $n$ is a $k$th record value in the sequence $X_1, X_2, \ldots$, and $\eta_n(k) = 0$ otherwise. The following results are valid for such type of indicators.

**Theorem 3** Let $X, X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables taking values $0, 1, 2, \ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0, 1, 2, \ldots$ Then for any fixed $k = 1, 2, \ldots$ indicators $\eta_n(k)$, $n = 0, 1, 2, \ldots$, are independent and

$$P\{\eta_n(k) = 1\} = 1 - P\{\eta_n(k) = 0\} = (p_n / P\{X \geq n\})^k, \quad n = 0, 1, 2, \ldots$$

Indeed, under $k = 1$, as a partial case of this theorem, one gets that record indicators $\eta_0, \eta_1, \eta_2, \ldots$ are independent and $P\{\eta_n = 1\} = 1 - P\{\eta_n = 0\} = p_n / P\{X \geq n\}$, $n = 0, 1, 2, \ldots$

It is easy to see that under conditions of Theorem 3, we can express distributions of *kth* record values for discrete random variables via distributions of sums of independent indicators:

$$P\{X(n,k) > m\} = P\{\eta_0(k) + \ldots + \eta_m(k) < n\}, \quad m = 0, 1, 2, \ldots, \ n = 1, 2, \ldots, \quad (9)$$

and, in particular, under $k = 1$ one has equality

$$P\{X(n) > m\} = P\{\eta_0 + \ldots + \eta_m < n\}, \ m = 0, 1, 2, \ldots, \ n = 1, 2, \ldots \quad (10)$$

As an example, we can consider the case, when $X'$s have the geometric distribution with some parameter $0 < p < 1$, that is $P\{X_j = n\} = (1-p)p^n$, $n = 0, 1, 2, \ldots$, for $j = 1, 2, \ldots$. In this situation, $P\{\eta_n(k) = 1\} = (1-p)^k$ and $P\{\eta_n(k) = 0\} = 1 - (1-p)^k$, for any $n = 0, 1, 2, \ldots$. It means that the sum $\eta_0(k) + \ldots + \eta_m(k)$ has the binomial $B(m+1, q)$ distribution with a parameter $q = (1-p)^k$. Hence,

$$P\{X(n,k) > m\} = \sum_{j=0}^{n-1} ((m+1)!/j! \ (m+1-j)!) q^j (1-q)^{m+1-j}, \quad \text{if } 1 \leq n \leq m+1,$$

and

$$P\{X(n,k) > m\} = 1, \text{ if } n > m+1.$$

It was mentioned above that for discrete distributions it is useful to introduce weak records together with classical (strong) record values. Weak records may arise, for example, in some sports competitions where any athlete who repeats the record achievement is also declared as a record-holder. If we consider $X'$s having a common discrete distribution, it is useful to introduce one more type of record statistics. Let conditions of Theorem 3 be valid. We define random variables $\mu_0, \mu_1, \mu_2, \ldots$, where $\mu_n$ denotes the number of those weak records in the sequence $X_1, X_2, \ldots$ that are equal to $n$. The following result is valid.

**Theorem 4** Let $X, X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables taking values $0, 1, 2, \ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0, 1, 2, \ldots$ Then for any fixed $k = 1, 2, \ldots$, random variables $\mu_0, \mu_1, \mu_2, \ldots$ are independent and

$$P\{\mu_n = m\} = (1 - r(n))(r(n))^m, \ n = 0, 1, 2, \ldots; \ m = 0, 1, 2, \ldots,$$

where

$$r(n) = p_n / P\{X \geq n\}.$$

Let $X_\omega(1), X_\omega(2), \ldots$ denote the weak (upper) record values in the sequence $X_1, X_2, \ldots$ Then for any $n = 1, 2, \ldots$

and $m = 0, 1, 2, \ldots$ the following relation is valid:

$$P\{X_\omega(n) > m\} = P\{\mu_0 + \mu_1 + \ldots + \mu_m < n\}. \quad (11)$$

Thus, we see that there are some convenient representations of record values, record times, numbers of records ((5–11) among them), which allow us to impress these record statistics in terms of sums of independent random variables.

## Distributions of Record Times

Let us consider the classical case, when $X_1, X_2, \ldots$ are independent and have a continuous distribution function $F$. Using the independence of the corresponding record indicators $\xi_1, \xi_2, \ldots$ one gets for any $n = 1, 2, \ldots$ and $1 < j(1) < j(2) < \ldots < j(n)$ that

$$\begin{aligned} &P\{L(1) = 1, L(2) = j(2), \ldots, L(n) = j(n)\} = \\ &\quad P\{\xi_1 = 1, \xi_2 = 0, \ldots, \xi_{j(2)-1} = 0, \xi_{j(2)} = 1, \\ &\quad \xi_{j(2)+1} = 0, \ldots, \xi_{j(3)-1} = 0, \\ &\xi_{j(3)} = 1, \ldots, \xi_{j(n)-1} = 0, \xi_{j(n)} = 1\} = \\ &\quad P\{\xi_1 = 1\}P\{\xi_2 = 0\} \ldots P\{\xi_{j(2)-1} = 0\} \\ &\quad P\{\xi_{j(2)} = 1\}P\{\xi_{j(2)+1} = 0\} \ldots \\ &\quad P\{\xi_{j(3)-1} = 0\}P\{\xi_{j(3)} = 1\} \ldots \\ &\quad P\{\xi_{j(n)-1} = 0\}P\{\xi_{j(n)} = 1\} = \\ &\quad 1/(j(2) - 1)(j(3) - 1) \ldots (j(n) - 1)j(n). \end{aligned} \quad (12)$$

Note that here the joint distribution of record times does not depend on $F$. Now one can see from (12) that

$$\begin{aligned} P\{L(n) = m\} = \sum 1/(j(2) - 1)(j(3) - 1) \\ \ldots (j(n-1) - 1)(m - 1)m, \end{aligned}$$

where the sum is taken over all $j(2), j(3), \ldots, j(n-1)$, such that $1 < j(2) < j(3) < \ldots < j(n-1) < m$.

It follows from (12) also that

$$\begin{aligned} P\{L(n) = j(n) | L(n-1) = j(n-1), L(n-2) \\ = j(n-2), \ldots, L(2) = j(2), L(1) = 1\} \\ = j(n-1)/j(n)(j(n) - 1) \end{aligned}$$

and

$$P\{L(n) = j | L(n-1) = i\} = i/j(j-1), \quad n = 2, 3, \ldots, j > i.$$

Hence, we see that the sequence of record times $L(1), L(2), \ldots$ in the announced situation forms a Markov chain (see ▶Markov Chains).

It was mentioned above that record times are closely related to the random variables $N(n)$, since for any $n = 1, 2, \ldots$ and $m = 1, 2, \ldots$ the following equalities are valid:

$$P\{L(n) > m\} = P\{N(m) < n\}$$

and

$$P\{L(n) = m\} = P(N(m-1) = n-1,\ N(m) = n\}. \quad (13)$$

Equalities (5) and (13) allow us to express the distributions of $L(n)$ in terms of independent record indicators:

$$\begin{aligned} P\{L(n) = m\} &= P\{N(m-1) = n-1, \xi_m = 1\} \\ &= P\{N(m-1) = n-1\}/m \\ &= P\{\xi_1 + \xi_2 + \ldots \xi_m - 1 \\ &= n-1\}/m. \end{aligned} \quad (14)$$

Representations (5) and (14) of $N(m)$ and $L(n)$ give a possibility to find the distributions of these record statistics via the Stirling numbers of the first kind $S_n(k)$, which are defined by equalities

$$x(x-1)\ldots(x-n+1) = \sum_{k \geq 0} S_n(k)x^k.$$

It appears that

$$P\{N(m) = k\} = (-1)^k S_m(k)/m! = |S_m(k)|/m!,$$

$$k = 1, 2, \ldots, m,$$

and

$$P\{L(n) = m\} = |S_{m-1}(n-1)|/m!,\ m = n,\ n+1, \ldots$$

Representations (5) and (14) give the following formulas for the corresponding generating functions:

$$P_m(s) = Es^{N(m)} = s(1+s)(2+s)\ldots(m-1+s)/m!$$

and

$$\begin{aligned} Q_n(s) = Es^{L(n)} &= 1 - (1-s)\sum_{k=0}^{n-1}(-\log(1-s))^k/k! \\ &= \int_0^{-\log(1-s)} v^{n-1}\exp(-v)dv/(n-1)! \end{aligned} \quad (15)$$

## Distributions of Record Values

Let us again consider the case when $X_1, X_2, \ldots$ are independent and have a continuous distribution function $F$. The record value $X(n)$ can be presented as $X_{L(n)}$, where $L(n)$ is

the corresponding record time with a generating function (15). One can see that then

$$P\{X(n) < x\} = P\{X_{L(n)} < x\} = P\{M(L(n)) < x\}, \quad (16)$$

where $M(n) = \max\{X_1, X_2, \ldots, X_n\}$ and $P\{M(n) < x\} = F^n(x)$. It follows from (15) and (16) that

$$\begin{aligned} P\{X(n) < x\} &= P\{M(L(n)) < x\} \\ &= \sum_{m \geq 0} P\{M(m) < x\} P\{L(n) = m\} \\ &= \sum_{m \geq 0} F^m(x) P\{L(n) = m\} = Q_n(F(x)) \\ &= \int_0^{-\log(1-F(x))} v^{n-1} \exp(-v) dv/(n-1)! \end{aligned}$$

The following result is valid for distributions of record values.

**Theorem 5** Let $X(1) < X(2) < \ldots$ be the record values in a sequence of independent random variables having a common continuous distribution function $F$, and let $U(1) < U(2) < \ldots$ be the record values related to the uniform distribution on the interval $[0,1]$. Then for any $n = 1,2,\ldots$ the random vector $(F(X(1)), \ldots, F(X(n)))$ has the same distribution as $(U(1), \ldots, U(n))$.

**Corollary 1** Let $X(1) < X(2) < \ldots$ and $Y(1) < Y(2) < \ldots$ be, respectively, the record values in a sequence of independent random variables $X_1, X_2, \ldots$ having a common continuous distribution function $F_1$ and in a sequence of independent identically distributed random variables $Y_1, Y_2, \ldots$ with a continuous distribution function $F_2$. Then for any $n = 1,2,\ldots$ the vector $(Y(1), Y(2), \ldots, Y(n))$ has the same distribution as the vector $(H_2(X(1)), H_2(X(2)), \ldots, H_2(X(n)))$, where $H_2(x) = G_2(F_1(x))$ and $G_2$ is the inverse function to $F_2$. Analogously, the vectors $(X(1), X(2), \ldots, X(n))$ and $(H_1(Y(1)), H_1(Y(2)), \ldots, H_1(Y(n)))$, where $H_1(x) = G_1(F_2(x))$ and $G_1$ is the inverse function to $F_1$, are identically distributed.

Let us consider the partial case of record values $Z(1) < Z(2) < \ldots$ related to the standard exponential $E(1)$ distribution (the case, when $F(x) = 1 - \exp(-x)$, $x > 0$). We get

$$P\{Z(n) < x\} = \int_0^x v^{n-1} \exp(-v) dv/(n-1)!,$$

that is, in this situation $Z(n)$ has the gamma-distribution with parameter $n$. It means that $Z(n)$ has the same distribution as the sum $v_1 + v_2 + \ldots + v_n$ of independent $E(1)$-distributed random variables $v_1, v_2, \ldots$ Moreover, for any $n = 1,2,\ldots$ the vector $(Z(1), Z(2), \ldots, Z(n))$ has the same distribution as the vector $(v_1, v_1 + v_2, \ldots, v_1 + v_2 + \ldots + v_n)$. It means that the vector $(Z(1), Z(2) - Z(1), \ldots, Z(n) - Z(n-1))$ consists of independent elements and each of these elements has the standard exponential $E(1)$ distribution.

Combining the previous results, we can get the following representation for record values $X(1) < X(2) < \ldots$ related to any continuous distribution function $F$. Let $G$ below denote the inverse function to $F$.

**Representation 1** For any $n = 1,2,\ldots$

$$(X(1), X(2), \ldots, X(n)) \stackrel{d}{=} (H(v_1), H(v_1 + v_2), \ldots, H(v_1 + v_2 + \ldots + v_n)),$$

where $v_1, v_2, \ldots$ are independent random variables having the exponential $E(1)$ distribution and $H(x) = G(1 - \exp(-x))$.

Taking into account the property of the exponential records it is not difficult to obtain the joint density function $f_n(x_1, x_2, \ldots, x_n)$ of the record values $Z(1), Z(2), \ldots, Z(n)$. It appears that

$$f_n(x_1, x_2, \ldots, x_n) = \exp(-x_n), \quad \text{if } 0 < x_1 < x_2 < \ldots < x_n, \text{ and } f_n(x_1, x_2, \ldots, x_n) = 0, \quad \textit{otherwise.}$$

In the general case, when $X_1, X_2, \ldots$ have a distribution function $F$ and a density function $f$, the joint density function of the record values $X(1), X(2), \ldots, X(n)$ is given by the formula

$$f_n(x_1, x_2, \ldots, x_n) = r(x_1) r(x_2) \ldots r(x_n) (1 - F(x_n)), \quad x_1 < x_2 < \ldots < x_n,$$

where $r(x) = f(x)/(1 - F(x))$.

Now we consider the conditional distributions

$$\varphi(x | x_1, x_2, \ldots, x_n) = P\{X(n+1) > x | X(1) = x_1, X(2) = x_2, \ldots, X(n) = x_n\}, \; x_1 < x_2 < \ldots < x_n < x,$$

for record values $X(1) < X(2) < \ldots < X(n) < X(n+1)$. It appears that

$$\begin{aligned} \varphi(x | x_1, x_2, \ldots, x_n) &= P\{X(n+1) > x | X(n) = x_n\} \\ &= (1 - F(x))/(1 - F(x_n)), x > x_n. \end{aligned} \quad (17)$$

It is interesting that equality (17) does not need the continuity of the distribution function $F$. It follows from (17) that record values $X(1), X(2), \ldots$ form a Markov chain.

If we now consider discrete $X$'s taking values $0, 1, 2, \ldots$ then (17) can be rewritten in the form

$$P\{X(n+1) > j | X(n) = m\} = P\{X > j\}/P\{X \geq m+1\}, \quad j > m \geq n-1.$$

It follows from the latter equality that in this case

$$P\{X(1) = j_1, X(2) = j_2, \ldots, X(n) = j_n\} = P\{X = j_n\}\omega(j_1)\omega(j_2)\ldots\omega(j_{n-1}), \quad 0 \le j_1 < j_2 < \ldots < j_{n-1} < j_n,$$

where $\omega(j) = P\{X = j\}/P\{X > j\}$.

The simplest discrete case is presented by the geometric distribution. The following result is valid.

**Theorem 6** Let $X, X_1, X_2, \ldots$ be independent identically distributed random variables such that

$$P\{X = j\} = (1-p)p^{j-1},\ j = 1, 2, \ldots;\ 0 < p < 1,$$

and $X(1) < X(2) < \ldots$ be the record values in a sequence $X_1, X_2, \ldots$ Then the interrecord values $X(1), X(2) - X(1), X(3) - X(2), \ldots$ are independent and have the same geometric distribution as $X$.

## Distributions of *k*th Record Values

The $k$th record values $X(n,k)$ are a natural extension of records $X(n)$. It is interesting that distributions of the *kth* records can be expressed via distributions of the classical record values. Really, together with a sequence of independent random variables $X_1, X_2, \ldots$ having a common distribution function $F$, let us consider one more sequence

$$Y_1 = \min\{X_1, X_2, \ldots, X_k\},$$
$$Y_2 = \min\{X_{k+1}, X_{k+2}, \ldots, X_{2k}\}, \ldots$$

Now let $X(n,k)$ be the $k$th record value based on $X$'s and $Y(n)$ be the classical records based on the sequence $Y_1, Y_2, \ldots$ It appears that for any fixed $k = 2, 3, \ldots$ and any $n = 1, 2, \ldots$ the vector $(X(1,k), X(2,k), \ldots, X(n,k))$ has the same distribution as the vector $(Y(1), Y(2), \ldots, Y(n))$.

Note that this result is valid for discrete $X$'s as well. One can immediately obtain some important results for the *kth* records taking into account the analogous results for the classical record values. For example, if $Z(n,k)$, $n = 1, 2, \ldots$, denote the *kth* records for the standard exponential distribution, then the vector $(Z(1,k), Z(2,k), \ldots, Z(n,k))$ has the same distribution as the vector $(v_1/k, (v_1 + v_2)/k, \ldots, (v_1 + v_2 + \ldots + v_n)/k)$, where $v_1, v_2, \ldots$ are the independent exponentially $E(1)$ distributed random variables. Hence, the following relation is valid for $k$th records related to a sequence of $X_1, X_2, \ldots$ with a continuous distribution function $F$.

**Representation 2** For any $n = 1, 2, \ldots$

$$(X(1,k), X(2,k), \ldots, X(n,k)) \stackrel{d}{=} (H(v_1/k), H((v_1 + v_2)/k), \ldots, H((v_1 + v_2 + \ldots + v_n)/k),$$

where $v_1, v_2, \ldots$ are independent random variables having the exponential $E(1)$ distribution, $H(x) = G(1 - \exp(-x))$ and $G$ is the inverse function to $F$.

Some useful results for the $k$th records follow immediately from representation 2 and analogous results for the classical records. Say, one gets that

$$P\{X(n,k) < x\} = \int_0^{-k\log(1-F(x))} v^{n-1}\exp(-v)dv/(n-1)!$$

and this equality is valid for any $k = 1, 2, \ldots$ and any continuous distribution function $F$.

**Theorem 7** For any $k = 1, 2, \ldots$ the sequence $X(1,k)$, $X(2,k), \ldots$ forms a Markov chain and

$$P\{X(n+1,k) > x | X(n,k) = u\} = ((1 - F(x))/(1 - F(u)))^k, x > u.$$

More complete theory of records is given in monographs (Ahsanullah 1988, 1995, 2004; Ahsanullah and Nevzorov 2001; Arnold et al. 1998; Nevzorov 2000). Different results for record values can be found in references (Adke 1993; Ahsanullah 1978, 1979, 1981, 1987, 1988, 1995, 2004; Ahsanullah and Nevzorov 2001, 2004, 2005; Akhundov and Nevzorov 2008; Akhundov et al. 2007; Andel 1990; Arnold et al. 1998; Bairamov 2000; Balakrishnan and Nevzorov 2006; Ballerini and Resnick 1985, 1987; Berred et al. 2005; Biondini and Siddiqui 1975; Chandler 1952; Deheuvels 1984; Deheuvels and Nevzorov 1994; Dziubdziela and Kopocinsky 1976; Foster and Stuart 1954; Gulati and Padgett 2003; Gupta 1984; Haiman 1987; Houchens 1984; Nagaraja 1978, 1982; Nevzorov 1984, 1987, 1990, 1992, 1995, 2000; Nevzorov and Balakrishnan 1998; Nevzorov et al. 2003; Nevzorova et al. 1997; Pfeifer 1982, 1984, 1991; Renyi A 1962; Resnick 1973; Shorrock 1972a, b; Siddiqui and Biondini 1975; Smith 1988; Smith and Miller 1986; Stepanov 1992; Tata 1969; Vervaat 1973; Williams 1973; Yang 1975).

## Acknowledgments

The work was partially supported by RFBR grant 09-01-00808.

## Cross References

►Markov Chains
►Order Statistics
►Ordered Statistical Data: Recent Developments
►Sequential Ranks

## References and Further Reading


Adke SR (1993) Records generated by Markov sequences. Stat Probab Lett 18:257–263

Ahsanullah M (1978) Record values and the exponential distribution. Ann Inst Stat Math 30A:429–433

Ahsanullah M (1979) Characterizations of the exponential distribution by record values. Sankhya B 41:116–121

Ahsanullah M (1981) On a characterization of the exponential distribution by weak homoscedasticity of record values. Biometr J 23:715–717

Ahsanullah M (1987) Two characterizations of the exponential distribution. Commun Stat Theory Meth 16:375–381

Ahsanullah M (1988) Introduction to record statistics. Ginn, Needham Heights

Ahsanullah M (1995) Record statistics. Nova Science, Commack

Ahsanullah M (2004) Record values – theory and applications. University Press of America, Lanham

Ahsanullah M, Nevzorov VB (2001) Ordered random variables. Nova Science, New York

Ahsanullah M, Nevzorov VB (2004) Characterizations of distributions by regressional properties of records. J Appl Stat Sci 13:33–39

Ahsanullah M, Nevzorov VB (2005) Order statistics. Examples and exercises. Nova Science, New York

Akhundov I, Nevzorov VB (2008) Characterizations of distributions via bivariate regression on differences of records. In: Records and branching processes. Nova Science, New York, pp 27–35

Akhundov I, Berred A, Nevzorov VB (2007) On the influence of record terms in the addition of independent random variables. Commun Stat Theory Meth 36:1291–1303

Andel J (1990) Records in an $AR(1)$ process. Ricerche Mat 39: 327–332

Arnold BC, Balakrishnan N, Nagaraja HN (1998) Records. Wiley, New York

Bairamov IG (2000) On the characteristic properties of exponential distribution. Ann Inst Stat Math 52:448–452

Balakrishnan N, Nevzorov VB (2006) Record values and record statistics. In: Encyclopedia of statistical sciences, 2nd edn. Wiley, 10, 6995–7006

Ballerini R, Resnick S (1985) Records from improving populations. J Appl Probab 22:487–502

Ballerini R, Resnick S (1987) Records in the presence of a linear trend. Adv Appl Probab 19:801–828

Berred A, Nevzorov VB, Wey S (2005) Normalizing constants for record values in Archimedean copula processes. J Stat Plan Infer 133:159–172

Biondini R, Siddiqui MM (1975) Record values in Markov sequences. In: Statistical inferences and related topics –2 Academic, New York, pp 291–352

Chandler KN (1952) The distribution and frequency of record values. J R Stat Soc Ser B 14:220–228

Deheuvels P (1984) The characterization of distributions by order statistics and record values – a unified approach. J Appl Probab 21:326–334 (Correction, 22, 997)

Deheuvels P, Nevzorov VB (1994) Limit laws for $k$-record times. J Stat Plan Infer 38:279–308

Dziubdziela W, Kopocinsky B (1976) Limiting properties of the $k$th record values. Zastos Mat 15:187–190

Foster FG, Stuart A (1954) Distribution free tests in time-series band on the breaking of records. J R Stat Soc B16:1–22

Gulati S, Padgett WJ (2003) Parametric and nonparametric inference from record breaking data. Springer, London

Gupta RC (1984) Relationships between order statistics and record values and some characterization results. J Appl Probab 21: 425–430

Haiman G (1987) Almost sure asymptotic behavior of the record and record time sequences of a stationary Gaussian process. In: Mathematical statistics and probability theory, vol A. D. Reidel, Dordrecht, pp 105–120

Houchens RL (1984) Record value theory and inference. PhD thesis, University of California, Riverside

Nagaraja HN (1978) On the expected values of record values. Aust J Stat 20:176–182

Nagaraja HN (1982) Record values and extreme value distributions. J Appl Probab 19:233–239

Nevzorov VB (1984) Record times in the case of nonidentically distributed random variables. Theory Probab Appl 29:808–809

Nevzorov VB (1987) Records. Theory Probab Appl 32:201–228

Nevzorov VB (1990) Generating functions for the $k$th record values – a martingale approach. Zap Nauchn Semin LOMI 184:208–214 (in Russian). Translated version in J Soviet Math 44: 510–515

Nevzorov VB (1992) A characterization of exponential distributions by correlation between records. Math Meth Stat 1:49–54

Nevzorov VB (1995) Asymptotic distributions of records in nonstationary schemes. J Stat Plan Infer 44:261–273

Nevzorov VB (2000) Records: mathematical theory. Translations of mathematical monographs, vol 194. Am Math Soc

Nevzorov VB, Balakrishnan N (1998) Record of records. In: Handbook of statistics, vol 16. Elsevier, Amsterdam, pp 515–570

Nevzorov VB, Balakrishnan N, Ahsanullah M (2003) Simple characterization of Student's $t_2$-distribution. Stat 52(part 3):395–400

Nevzorova LN, Nevzorov VB, Balakrishnan N (1997) Characterizations of distributions by extremes and records in Archimedean copula process. In: Advances in the theory and practice of statistics: a volume in honor of Samuel Kotz. Wiley, New York, pp 469–478

Pfeifer D (1982) Characterizations of exponential distributions by independent nonstationary record increments. J Appl Probab 19:127–135. (Correction, 19, 906)

Pfeifer D (1984) Limit laws for inter-record times from nonhomogeneous record values. J Organ Behav Stat 1:69–74

Pfeifer D (1991) Some remarks on Nevzorov's record model. Adv Appl Probab 23:823–834

Renyi A (1962) Theorie des elements saillants d'une suite d'observations. Colloquim on combinatorial methods in probability theory. Math. Inst., Aarhus Univ., Aarhus, Denmark, 1–10 August 1962, pp 104–117. See also: Selected papers of Alfred Renyi, vol. 3 (1976), Akademiai Kiado, Budapest, pp 50–65

Resnick SI (1973) Limit laws for record values. Stoch Proc Appl 1: 67–82

Shorrock RW (1972a) A limit theorem for inter-record times. J Appl Probab 9:219–223

Shorrock RW (1972b) On record values and record times. J Appl Probab 9:316–326

Siddiqui MM, Biondini RW (1975) The joint distribution of record values and inter-record times. Ann Probab 3:1012–1013

Smith RL (1988) Forecasting records by maximum likelihood. J Am Stat Assoc 83:331–338

Smith RL, Miller JE (1986) A non-Gaussian state space model and application in prediction of records. J R Stat Soc Ser B 48:79–88

Stepanov AV (1992) Limit theorems for weak records. Theor Probab Appl 37:586–590
Tata MN (1969) On outstanding values in a sequence of random variables. Z Wahrscheinlichkeitsth verw Geb 12:9–20
Vervaat W (1973) Limit theorems for records from discrete distributions. Stochast Proc Appl 1:317–334
Williams D (1973) On Renyi's record problem and Engel's series. Bull Lond Math Soc 5:235–237
Yang MCK (1975) On the distribution of the inter-record times in an increasing population. J Appl Probab 12:148–154

# Recursive Partitioning

Hugh A. Chipman
Canada Research Chair in Mathematical Modelling, Professor
Acadia University, Wolfville, NS, Canada

## Introduction

Recursive partition (RP) models are a flexible method for specifying the conditional distribution of a variable $y$, given a vector of predictor values $x$. Such models use a tree structure to recursively partition the predictor space into subsets where the distribution of $y$ is successively more homogeneous. The terminal nodes of the tree correspond to the distinct regions of the partition, and the partition is determined by splitting rules associated with each of the internal nodes. By moving from the root node through to the terminal node of the tree, each observation is then assigned to a unique terminal node where the conditional distribution of $y$ is determined. The two most common response types are continuous and categorical, with corresponding tasks often known as regression and classification.

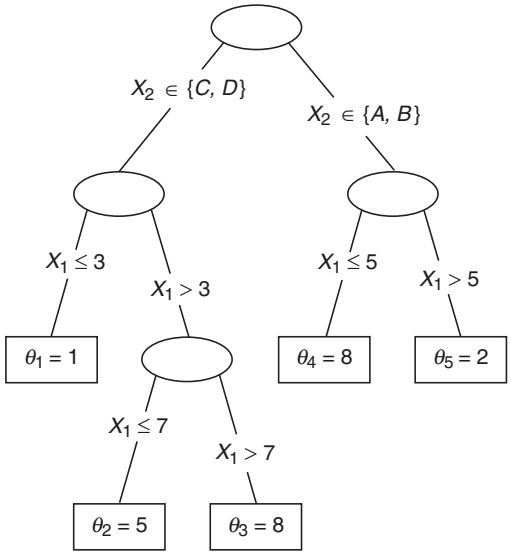
Given a data set, a common strategy for finding a good tree is to use a greedy algorithm to grow a tree and then to prune it back to avoid overfitting. Such greedy algorithms typically grow a tree by sequentially choosing splitting rules for nodes on the basis of maximizing some fitting criterion. This generates a sequence of trees each of which is an extension of the previous tree. A single tree is then selected by pruning the largest tree according to a model choice criterion such as cost-complexity pruning, cross-validation, or hypothesis tests of whether two adjoining nodes should be collapsed into a single node.

Early work in RP models includes Morgan and Sonquist (1963), who developed a recursive partitioning strategy (AID – Automatic Interaction Detection) for a continuous response. There were many offshoots of this work, including Kass (1980) and Hawkins and Kass (1982). Recursive partitioning models were popularized in the statistical community by the book "Classification and Regression Trees" by Breiman et al. (1984). RP models have also been developed in the machine learning community, with work by Quinlan on the ID3 (1986 and references therein) and C4.5 (1993) algorithms being among the most widely recognized.

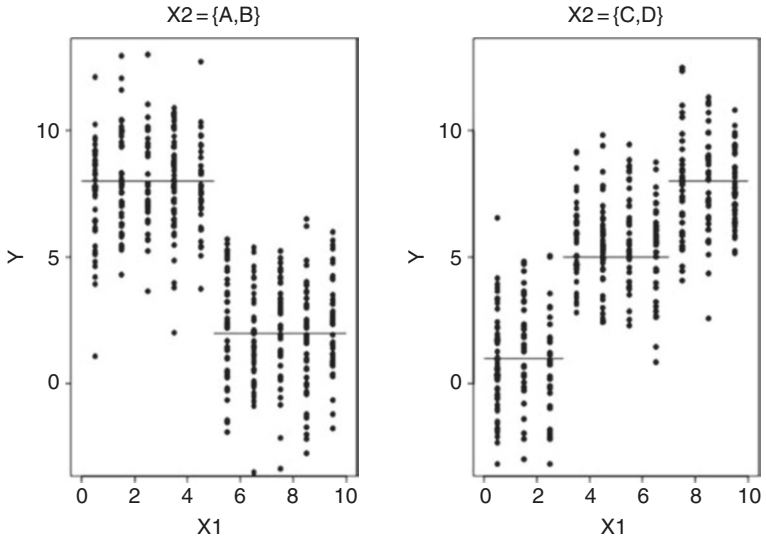
## Structure of a RP model

A RP model describes the conditional distribution of $y$ given a vector of predictors $x = (x_1, x_2, \ldots, x_p)$. This model has two main components: a tree $T$ with $b$ terminal nodes, and a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_b)$ which associates the (possibly vector-valued) parameter $\theta_j$ with the $j$th terminal node. If $x$ lies in the region corresponding to the $j$th terminal node then $y|x$ has distribution $f(y|\theta_j)$, where we use $f$ to represent a parametric family indexed by $\theta_j$. The model is called a regression tree or a classification tree according to whether the response $y$ is quantitative or qualitative, respectively. An example of a RP model with binary splits is displayed in Fig. 1, and data sampled from its induced partition is displayed in Fig. 2.

Before describing the example tree, we discuss the general structure of a RP model for the case of a binary tree. A binary tree $T$ subdivides the predictor space as follows: Each internal node has an associated splitting rule which uses a predictor to assign observations to either its left or




**Recursive Partitioning. Fig. 1** A regression tree where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$




**Recursive Partitioning. Fig. 2** A realization of 800 observations sampled from the tree model depicted in Fig. 1

right child node. The terminal nodes thus identify a partition of the predictor space according to the subdivision defined by the splitting rules. For quantitative predictors, the splitting rule is based on a split value $s$, and assigns observations for which $\{x_i \leq s\}$ or $\{x_i > s\}$ to the left or right child node respectively. For qualitative predictors, the splitting rule is based on a category subset $C$, and assigns observations for which $\{x_i \in C\}$ or $\{x_i \notin C\}$ to the left or right child node respectively.

Several assumptions have been made to simplify exposition. First, splitting rules are assumed to subdivide a region into two sub-regions, giving a binary tree. Second, only one predictor variable is assumed to be used for each splitting rule. Both these restrictions can be relaxed.

For illustration, Fig. 1 depicts a regression tree model where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$. $x_1$ is a quantitative predictor taking values in [0,10], and $x_2$ is a qualitative predictor with categories $(A, B, C, D)$. The binary tree has nine nodes of which $b = 5$ are terminal nodes. The terminal nodes subdivide the $x$ space into five nonoverlapping regions. The splitting variable and rule are displayed at each internal node. For example, the leftmost terminal node corresponds to $x_1 \leq 3.0$ and $x_2 \in \{C, D\}$. The $\theta_i$ value which identifies the mean of $y$ given $x$ is displayed at each terminal node. Note that $\theta_i$ decreases in $x_1$ when $x_2 \in \{A, B\}$, but increases in $x_1$ when $x_2 \in \{C, D\}$. A realization of 800 observations sampled from this model is displayed in Fig. 2.

If $y$ were a qualitative variable, a classification tree model would be obtained by using an appropriate categorical distribution at each terminal node. For example, if $y$ was binary with categories $C_1$ or $C_2$, one might consider the Bernoulli model $P(y \in C_1) = \theta = 1 - P(y \in C_2)$ with a possibly different value of $\theta$ at each terminal node. A standard classification rule for this model would then classify $y$ into the category yielding the smallest expected misclassification cost. When all misclassification costs are equal, this would be the category with largest probability.

## Learning the RP Model

To learn or estimate a RP model, we assume that a training sample consisting of tuples $(x_i, y_i), i = 1, \ldots, n$ is available. Both the tree $T$ and the terminal node parameters $\Theta$ must be estimated using the training data.

For a fixed $T$, a common assumption is that the response values are i.i.d. within each terminal node. The data in each terminal node can be considered a separate sample, and conventional estimation techniques (e.g., maximum likelihood) yield familiar node parameter estimates $\hat{\theta}_j$ such as the sample mean for a continuous normal response and sample proportions for a categorical multinomial response.

Armed with a recipe for estimating $\Theta$ given $T$, we can now consider estimation of $T$. First, an objective function must be specified, providing a mechanism to assess the quality of a particular tree $T$. The log-likelihood of the training data is one such criterion. For a normal response model, the corresponding criterion would be the minimization of a residual sum of squares. For a multinomial response, the multinomial log-likelihood would be used. Ciampi (1991) was one of the first to develop a likelihood-based approach to RP models. Other criteria have been

proposed for specific response classes, such as the Gini index (Breiman et al. 1984) for a categorical response.

With an objective function quantifying the quality of a tree, the estimation problem becomes a search over all possible trees to optimize the objective. Although splitting rules for continuous $x$ are real-valued, the objective function will only change when training points are moved among terminal nodes of the tree. Thus it is common to consider only splitting rules defined at data points, and require that each terminal node contain at least one training point. The search over the set of trees is thus a combinatorial search over a finite but very large discrete space.

The most common search algorithm is a greedy forward search, in which all training observations are initially grouped into a single node. The algorithm considers splitting into two child nodes, examining all possible splits on all possible variables. The splitting rule yielding the best value of the objective function (e.g., the smallest residual sum of squares when summed over the two child nodes) is selected. The procedure is repeated in each child node recursively until a large tree is grown.

Several strategies can be employed to decide how large a tree to grow. In the CHAID algorithm of Kass (1980), hypothesis tests were used to decide when to stop subdividing, yielding a final tree. Breiman et al. (1984) suggest growing a maximal tree, and then pruning away sibling nodes that do not significantly improve the objective function over the value assigned to their parent node. Their reasoning was that the forward greedy search might sometimes stop early, missing significant effects. For example, in the tree displayed earlier, no initial split leads to a large reduction in residual sum of squares because of the interaction pattern. Their backward pruning was facilitated by the idea of cost-complexity pruning, in which a modified objective function was minimized:

$$\text{Loss}(T;\alpha) = \text{RSS}(T) + \alpha|T|, \tag{1}$$

where $|T|$ represents the number of terminal nodes of the tree. Penalty parameter $\alpha \geq 0$ controls the trade-off between tree size and accuracy. Breiman et al. showed that (1) can be minimized as $\alpha$ increases from 0 to $\infty$ by considering a nested sequence of pruned trees, starting with the largest tree identified. The optimal $\alpha$ and a corresponding tree are selected so as to minimize a cross-validated estimate of the objective function.

While other methods for identifying the best tree have been proposed, the greedy forward search is quick and can be quite effective.

## Strengths and Weaknesses of RP Models

The structure of RP models enables them to identify *interactions*. For instance, in Figs. 1 and 2, we see an interaction effect between $X_1$ and $X_2$: If $X_2 = \{A, B\}$ then response $y$ decreases with increasing $X_1$. If $X_2 = \{C, D\}$ then response $y$ increases with increasing $X_1$. This is perhaps the greatest strength of RP models, and one of the reasons they are used for exploratory data analysis.

This strength is also a weakness. If the relation between predictors and response is *additive*, very large trees will be needed to capture this relationship. For instance, if

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \text{error},$$

then a tree with 32 terminal nodes will be required to even approximate this function with a single step along each of the five predictor axes.

Trees are popular among practitioners because of their *interpretability*. It is natural to interpret the sequence of conditions leading to a terminal node of a tree. Care must be taken with such interpretations, especially if dependencies exist among predictors. In such cases, multiple trees with different splits on different variables may fit the data equally well.

In addition to dealing with mixed predictor types, RP models can handle missing values of predictors via several strategies. For missing predictor values in the training data, one could (i) treat "missing" as a new category for a categorical predictor, or (ii) identify surrogate splitting variables that produce splits similar to a missing predictor. If predictor values are missing when making predictions for new observations, either of these strategies may be employed, or one may terminate the branching process when a missing value is needed in a branch, and base predictions on the interior node.

The most common form of RP models utilize a single variable for each splitting rule. This *axis alignment* aids in interpretability, but can be a weakness if variation in the response occurs along a linear combination of predictors, rather than along the axes. The additive function of five variables mentioned above is an example of this.

By virtue of subdividing the data into smaller subgroups, an RP model can suffer from *sparsity*, especially if more complex statistical models are utilized in the terminal nodes. For instance, a significant challenge in modifying RP models for survival data with censoring (LeBlanc and Crowley 1993) is the pooled nature of Kaplan–Meier estimates (see ►Kaplan-Meier Estimator) of the survival curve. This data sparsity is one of the primary reasons for the use of simple models in terminal nodes.

A weakness of RP models is *sensitivity* of results to small data perturbations. Breiman (1996) demonstrated that when RP models were fit to bootstrap samples of the data, there could be substantial variation in tree structure. While this would seem to be a weakness, Breiman leveraged this idea to produce Ensemble methods discussed below in section ►"Ensembles of Trees".

Because of the greedy nature of the search over the space of trees, inference for the resultant model is difficult. Although confidence intervals and hypothesis tests can easily be constructed conditional on a specific tree $T$, the adaptive nature of the learning algorithm means that the statistical properties of estimators, intervals and tests will be seriously undermined. Methods that take account of the search include adjustments for multiple testing (Hawkins and Kass 1982) and Bayesian approaches (Chipman et al. 1998; Denison Mallick and Smith 1998).

## Extensions

The popularity of RP models has lead to a number of extensions and the development of related methods.

A variety of search strategies have been proposed as alternatives to the greedy forward stepwise approach. These include the use of stochastic search optimizers such as genetic algorithms (Fan and Gray 2005) and simulated annealing (Sutton 1991; Lutsko and Kuijpers 1994) and MCMC (Chipman et al. 1998; Denison et al. 1998). Tibshirani and Knight (1999) used the bootstrap to perturb data before executing a greedy search.

Variations on the tree structure have also been considered, including splitting rules based on linear combinations of real-valued predictors (Loh and Vanichsetakul 1988). Some RP algorithms (e.g., AID) allow nodes to have more than two child nodes, complicating the search but sometimes making interpretation clearer. Quinlan's C4.5 splits categorical predictors by generating a different child node for each categorical level of the corresponding predictor.

The statistical model in terminal nodes has also been extended to richer models, such as linear regression (Alexander and Grimshaw 1996; Chipman et al. 2002), ►generalized linear models (Chipman et al. 2003), and Gaussian process models (Gramacy and Lee 2008).

## Ensembles of Trees

RP models have been used as a "base learner" in a number of algorithms that seek to achieve greater predictive accuracy by combining together multiple instances of a model.

In noticing the sensitivity of trees to small perturbations, Breiman (1996) developed a strategy known as bootstrap aggregation or "Bagging" for generating multiple trees and combining them to achieve greater prediction accuracy. For instance, with a continuous response, each bootstrap tree would be used to generate predictions at a particular test point, and these predictions would be averaged to form an ensemble prediction.

A further enhancement led to Random Forests (Breiman 2001). Additional variation in the search algorithm was introduced by randomizing the choice of predictor in splitting rules. This led to a richer set of trees, and could further improve predictive accuracy.

Another form of ensemble model using RP models is boosting (Freund and Schapire 1997). In this algorithm, a sequence of RP models are learned, each depending on those already identified via data weights that depend on predictive accuracy of earlier RP models. These weights encourage the next RP model to better fit those observations that have been incorrectly classified. At the end of the boosting sequence, an ensemble prediction is generated by a weighted combination of predictions from each learner in the ensemble.

Although neither boosting or random forests require that the base learner be a RP model, these have yielded the most popular and successful form of ensemble model.

## Related Work

A model closely related to RP models is the hierarchical mixture of experts model (Jordan and Jacobs 1994). In this model, a different logistic function of the predictors is used in each interior node to probabilistically assign data points to the left and right children. In doing so, the hard boundaries associated with splitting rules are replaced with soft decisions indexed by continuous parameters. In terminal nodes, predictions are given by ►logistic regression. Tree size and topology is typically fixed in advance, and the tree learning algorithm becomes a continuous optimization problem.

## About the Author

Hugh A. Chipman is Professor and Canada Research Chair in Mathematical Modelling, Acadia University Department of Mathematics and Statistics. He is Editor-Elect (2010) and will be Editor (2011–2014), *Technometrics*. He was elected as a Fellow of the American Statistical Association (2008) and received the CRM-SSC award (Canada, 2009).

## Cross References

►Data Mining
►Exploratory Data Analysis
►Interaction
►Kaplan-Meier Estimator
►Logistic Regression

## References and Further Reading


Alexander WP, Grimshaw SD (1996) Treed regression. J Comput Graph Stat 5:156–175
Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
Breiman L (2001) Random forests. Mach Learn 45:5–32
Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees, Wadsworth, Belmont
Chipman HA, George EI, McCulloch RE (1998) Bayesian CART model search. J Am Stat Assoc 93:935–948
Chipman HA, George EI, McCulloch RE (2002) Bayesian treed models. Mach Learn 48:299–320
Chipman HA, George EI, McCulloch RE (2003) Bayesian treed generalized linear models. In: Bernardo JM, Bayarri M, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian statistics vol 7. Oxford University Press, Oxford
Ciampi A (1991) Generalized regression trees. Comput Stat Data Anal 12:57–78
Denison D, Mallick B, Smith AFM (1998) A Bayesian CART algorithm. Biometrika 85:363–377
Fan G, Gray JB (2005) Regression analysis using TARGET. J Comput Graph Stat 14:206–218
Gramacy RB, Lee HKH (2008) Bayesian treed Gaussian process models with an application to computer modeling. J Am Stat Assoc 103:1119–1130
Hawkins DM, Kass GV (1982) Automatic interaction detection. In: Hawkins DM (ed) Topics in applied multivariate analysis. Cambridge University Press, Cambridge
Jordan MI, Jacobs RA (1994) Mixtures of experts and the EM algorithm. Neural Comput 6:181–214
Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Appl Stat 29:119–127
LeBlanc M, Crowley J (1993) Survival trees by goodness of split. J Am Stat Assoc 88:457–467
Loh W-Y, Vanichsetakul N (1988) Tree-structured classification via generalized discriminant analysis. J Am Stat Assoc 83: 715–725
Lutsko JF, Kuijpers B (1994) Simulated annealing in the construction of near-optimal decision trees. In: Cheeseman P, Oldford RW (eds) Selecting models from data: AI and statistics IV. Springer, New York, pp 453–462
Morgan JA, Sonquist JN (1963) Problems in the analysis of survey data and a proposal. J Am Stat Assoc 58:415–434
Quinlan JR (1986) Induction of decision trees. Mach Learn 1: 81–106
Quinlan JR (1993) C4.5: tools for machine learning. Morgan Kauffman, San Mateo
Sutton C (1991) Improving classification trees with simulated annealing. In: Keramidas E (ed) Proceedings of the 23rd symposium on the interface, Interface Foundation of North America
Tibshirani R, Knight K (1999) Model search by bootstrap 'bumping'. J Comput Graph Stat 8:671–686


# Regression Diagnostics


Shuangzhe Liu[1], Alan H. Welsh[2]
[1]Associate Professor, Faculty of Information Sciences and Engineering
University of Canberra, Canberra, ACT, Australia
[2]E.J. Hannan Professor of Statistics
Australian National University, Canberra, ACT, Australia

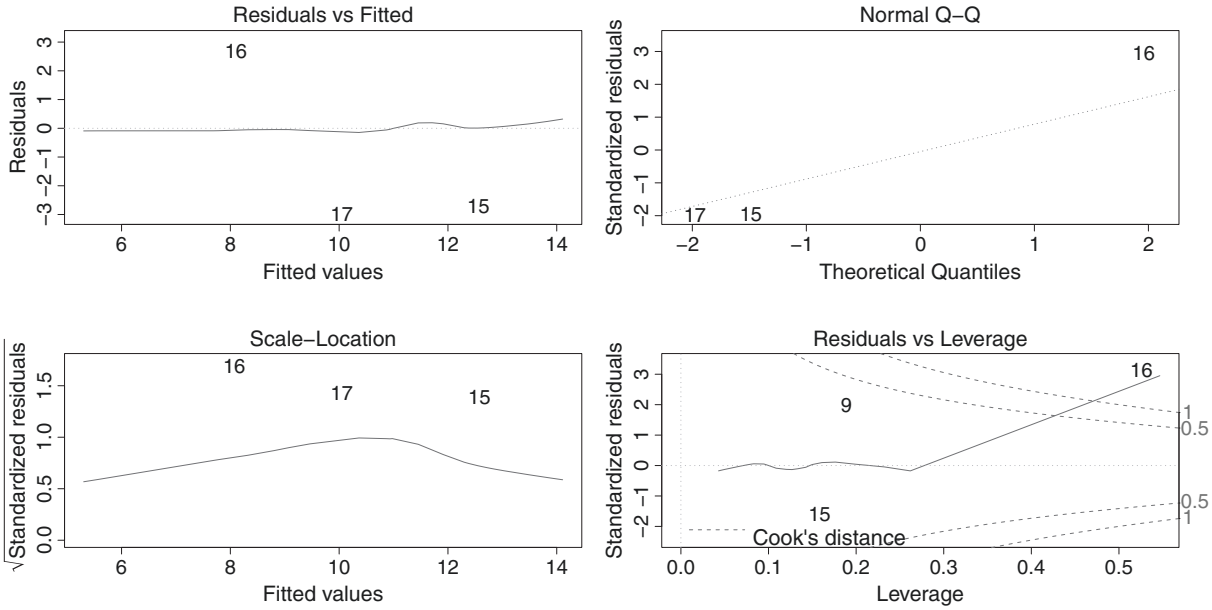

Regression diagnostics are a set of mostly graphical methods which are used to check empirically the reasonableness of the basic assumptions made in the model. These informal methods are an important part of regression modelling: many formal conclusions and inferences (including confidence intervals, statistical tests, prediction etc.) derived from a fitted model only make sense if the assumptions of the model hold. If the regression assumptions are violated, any application of results obtained from the model can be very misleading.

For a data set of $n$ observations of a response variable $y$ and $k$ explanatory variables $x_j$ $(j = 1, \ldots, k)$, the standard linear regression model (see ►Linear Regression Models) for the relationship between the response and the explanatory variables can be written in matrix notation as

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

where $y = (y_i)$ is an $n$-vector of observations, $X = (x_{ij})$ is an $n \times k$ matrix of independent variables, $\beta = (\beta_j)$ is a $k$-vector of unknown parameters and $\epsilon = (\epsilon_i)$ is an $n$-vector of unobserved random variables, often called errors. The basic assumptions of the model are that the relationship between $y$ and $X$ is linear, the $\epsilon_i$ are independent, have constant variance and are normally distributed.

The basic quantities on which diagnostics are based are the residuals and fitted values. For any estimator $\hat{\beta}$ of $\beta$, the fitted values are $\hat{y} = X\hat{\beta}$ and the residuals are $\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta}$. The residuals provide information about the errors in the model so are fundamental in diagnostics. Various forms of standardized residuals can also be calculated. If $X$ is of full column rank so $\hat{\beta} = (X'X)^{-1}X'y$ is the least squares estimator of $\beta$, the fitted values can be written as $\hat{y} = Hy$, where $H = X(X'X)^{-1}X' = (h_{ij})$ is the hat matrix and the $i$th diagonal element $h_{ii}$ is called the leverage of the $i$th observation. The residuals can be standardized as $\hat{\epsilon}_i/s$, where $s^2 = (n-k)^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$, as $\hat{\epsilon}_i/s(1-h_{ii})^{1/2}$ (internally Studentized) or as $\hat{\epsilon}_i/s_{(i)}(1-h_{ii})^{1/2}$ (externally Studentized), where $s^2_{(i)} =$




**Regression Diagnostics. Fig. 1** Diagnostic plots based on the least squares fit of a linear regression model to the salinity data of Ruppert and Carroll (1980)

$(n - k - 1)^{-1} \sum_{j \neq i}^{n} \hat{\epsilon}_{(j)}^2$ and $\hat{\epsilon}_{(j)}$ is the residual for the $j$th observation calculated from the $n - 1$ observations after excluding the $j$th observation. Other useful quantities include ▶Cook's Distance which is a measure of influence involving the square of the Studentized residual and the potential function $h_{ii}/(1 - h_{ii})$.

The most widely used diagnostic plots are residual plots which plot the residuals against the fitted values (checking for linearity, constant variance and ▶outliers), spread plots which plot the square root of the Studentized residuals against the fitted values (checking for constant variance, outliers), QQ-plots which plot the ordered residuals against their expected values under normality (normality, outliers), and leverage plots which plot Studentized residuals against the leverage (checking for ▶influential observations). These four plots are illustrated in Fig. 1 for the salinity data (Ruppert and Carroll 1980) which have 28 observations and 3 explanatory variables. The plots are supplemented by lines and curves which aid in their interpretation. The most interesting features are the departure from normality in the upper tail (observation 16) shown in the QQ-plot and the confirmation that this observation is influential in the leverage plot. In general, outliers may be difficult to find without the use of robust method: using robust methods, Ruppert and Carroll also identified observations 15 and 17 as outliers in these data. Other useful plots include added-variable plots (examining the relationship between $y$ and $x_j$ after adjusting for the other explanatory variables) and partial-residual plots (checking for linearity). In addition, there are a number of specialized plots which can be used to check for dependence: these include various time series and spatial plots, correlograms (ACF, PACF), variograms and spectrum plots. These methods are extensively documented in the statistical literature.

See for example the list of references at the end of this entry.

Graphical methods are preferred in diagnostics because they are more informative than numerical ones and often suggest ways in which deficiencies in a model can be rectified. A good illustration is Anscombe's (1973) set of 4 different datasets with the same summary statistics but four distinct regression relationships between the response and explanatory variables.

Diagnostic methods are important in all statistical modelling including generalised linear models (de Jong and Heller 2008), time series analysis (Li 2003) etc.

## About the Authors

Shuangzhe Liu is Associate Professor in the Discipline of Mathematics and Statistics in the Faculty of Information Sciences and Engineering at the University of Canberra.He

holds a PhD in Econometrics from the Tinbergen Institute, University of Amsterdam. He is a member, Statistical Society of Australia (2010–), and Australian Mathematical Sciences Institute (2007–). He is a Contributing Editor, *Current Index to Statistics* (2000–), and an Associate Editor, *Chilean Journal of Statistics* (2009–).

Alan Welsh is the E.J. Hannan Professor of Statistics and the Head of the Centre for Mathematics and its Applications at the Australian National University. He is a fellow of the Australian Academy of Science, the Institute for Mathematical Statistics and the American Statistical Association. He is currently Applications Editor of the *Australian and New Zealand Journal of Statistics* and an Associate Editor of the *Journal of the American Statistical Association*. He has published over 95 papers and a book on statistical inference.

## Cross References

►Cook's Distance
►Influential Observations
►Linear Regression Models
►Outliers
►Residuals
►Robust Regression Estimation in Generalized Linear Models
►Simple Linear Regression

## References and Further Reading

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27: 17–21
Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York
Belsley DA, Kuh E, Welsch RE (2004) Regression diagnostics: identifying influential data and sources of collinearity, 2nd edn. Wiley, New York
Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall/CRC, New York
Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York
de Jong P, Heller GZ (2008) Generalized linear models for insurance data. Cambridge University Press, Cambridge
Fox J (1991) Regression diagnostics: an introduction. Sage, New York
Fox J (2008) Applied regression analysis and generalized linear models, 2nd edn. Sage, New York
Li WK (2003) Diagnostic checks in time series. Chapman & Hall/CRC, New York
Ruppert D, Carroll RJ (1980) Trimmed least squares in the linear model. J Am Stat Assoc 75:828–838
Wheeler D (2009) Spatially varying coefficient regression models: diagnostic and remedial method for collinearity. Vdm Verlag Dr. Müller, p 132. ISBN 3-63911437-X

# Regression Models with Increasing Numbers of Unknown Parameters

Asaf Hajiyev
Professor, Chair
Baku State University, Baku, Azerbaijan

## Introduction

Consider the regression model

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, 2, \ldots, N \tag{1}$$

where $x_i$ is the point of observation, $y_i$ an observable value, $\varepsilon_i$ a random error at the point $x_i$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)^T$ is the vector of unknown parameters. Let us suppose that the number of unknown parameters $m$ depends on the number of observations $N$ and $m$ may increase, when $N$ becomes larger. Such regressions are called models with increasing number of unknown parameters. The variances of observation error are unknown and may be different. At each point $x_i$ there is only one observable value, $y_i$, that does not allow estimation of the variance.

Regression models with an increasing number of unknown parameters and with unknown and different variances of observation error are of interest in important applications. This is because, with an increased number of unknown parameters, the unknown function can be approximated more accurately in experiments. Moreover, in some applications, repeated tests at a single point are costly (financially and technically), which hampers the estimation of the unknown error variance, which is different at different observation points.

Regression models have been widely addressed in numerous publications (Demidenko 1989; Huet et al. 1996; Sen and Srivastava 1997), but models with an increasing number of unknown parameters have received little attention, which motivates our interest in this subject. The main aims of our investigations are

- Direct estimation (without estimation of a variance) of the elements of the covariance matrix of the vector $\sqrt{N}(\theta^* - \theta)$, where $\theta^*$ is the least square estimator (l.s.e.).
- Construction of a confidence band for the unknown function $f(x, \theta)$.

## Linear Regression Models

Let us assume that

$$f(x, \theta) = \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \ldots + \theta_{m(N)} \phi_{m(N)}(x), \tag{2}$$

where $\phi_1(x), \phi_2(x), \ldots, \phi_{m(N)}(x)$ is a system of linearly independent and bounded functions. Expression (1) can be rewritten in a vector form as

$$Y = X\theta + \varepsilon, \tag{3}$$

where $Y$ is the vector of observable values, $X$ the design matrix, defined as $X = //x_{ij}//, x_{ij} = \phi_i(x_j)$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; with $\theta = (\theta_1, \theta_2, \ldots, \theta_{m(N)})^T$ being the vector of unknown parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$ denoting the error-vector. The number of unknown parameters depends on the number of observations and moreover

$$m(N)/N \to 0, \ as \ n \to \infty. \tag{4}$$

The sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ is assumed to have uniformly bounded and independent random variables with $E\varepsilon_i = 0, E\varepsilon_i^2 = \sigma_i{}^2$ being unknown, different, and

$$0 < (\sigma_*)^2 \le \sigma_i{}^2 \le (\sigma^*)^2 < \infty.$$

Let $\theta^* = (X^TX)^{-1}X^TY^Y$ be the l.s.e and $trA = \sum_{i=1}^{n} a_{ii}$ be the trace of the matrix $A$ with elements $a_{ij}$.

**Definition 1** The vector $\theta = (\theta_1, \theta_2, \ldots, \theta_{m(N)})^T$ with random elements and increasing dimension converges to zero in probability $\theta \xrightarrow{P} 0$, if $\sum_{i=1}^{m(N)} \theta_i^2 \xrightarrow{P} 0$ as $N \to \infty$.

Let $0 < \lambda_1(N) \le \lambda_2(N) \le \ldots \le \lambda_m(N)$ be eigenvalues of the matrix $(X^TX)/N$.

**Theorem 1** Let the conditions (2)–(4) be true. Then $(\theta^* - \theta) \xrightarrow{P} 0$, if and only if $(1/N)tr(X^TX/N) \xrightarrow{P} 0$ as $N \to \infty$.

**Definition 2** The vector $\theta^P = (\theta_1{}^P, \theta_2{}^P, \ldots, \theta_{m(N)}{}^P, 0, \ldots, 0)^T$ is called $m$-finite and $p$-consistent estimator of the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_N)^T$, if

$$\forall \delta > 0 \, P\left\{\sum_{i=1}^{m(N)} \left(\theta_i{}^P - \theta_i\right)^2 < \delta\right\} \ge p \text{ holds true, as } N \to \infty.$$

*Example* 1 Consider $f(x, \theta) = \sum_{i=1}^{\infty} \theta_i \phi_i < \infty, |\phi_i(x)| \le 1$, $\sum_{i=1}^{\infty} \theta_i^2 < \infty$. The problem is to find such $m(p, N, \delta)$ ($\forall \delta > 0$ and given $0 < p < 1$, $N > 0$), for which $P\left\{\sum_{i=1}^{m(N)} \left(\theta_i^* - \theta_i\right)^2 < \delta\right\} \ge p$ holds true, where $\theta_i^*$ is the l.s.e. on $N$ observations. For simplicity, we assume $E\varepsilon_i = 0, E\varepsilon_i^2 \le 1$.

Consider $y_i = \sum_{i=1}^{m(N)} \theta_i \phi_i(x) + \delta_i$, where

$$\delta_i = \sum_{j=m(N)+1}^{\infty} \theta_i \phi_i(x) + \varepsilon_i, \ \delta_i$$
$$= \sum_{j=m(N)+1}^{\infty} \theta_i \phi_i(x) \to 0 \text{ as } N \to \infty.$$

Assuming that for large values of $N$ $P\left\{\sum_{i=1}^{\infty}\left(\theta_i{}^* - \theta_i\right)^2 < \delta\right\} \approx P\left\{\sum_{i=1}^{m(N)} \left(\theta_i{}^* - \theta_i\right)^2 < \delta\right\}$. If the conditions of the Theorem 1 hold true, then we obtain

$$P\left\{\sum_{i=1}^{m(N)} \left(\theta_i{}^* - \theta_i\right)^2 < \delta\right\} \ge 1 - (m+1)/(N\lambda_1(N)\delta)$$

from Chebyshev inequality. Taking $p = 1 - (m+1)/(N\lambda_1(N)\delta)$, we get $m = (1-p)(N\lambda_1(N)\delta) - 1$. Now in the capacity of a consistent estimator of the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_N)^T$, we can take the vector $\theta^P = \left(\theta_1{}^P, \theta_2{}^P, \ldots, \theta_{m(N)}{}^P, 0, \ldots, 0\right)^T$, where $m$ was found, above. According to the Theorem 1 the vector $\theta^*$ is a consistent estimator of $\theta$.

## Estimation of Covariance Matrix

Denote

$$D_N = E(\theta^* - \theta)(\theta^* - \theta)^T$$
$$= (1/N)(X^TX/N)^{-1}[X^T(E\varepsilon\varepsilon^T)X/N](X^TX/N)^{-1}$$
$$= (1/N)(X^TX/N)^{-1}[X^TI(\sigma^2)X/N](X^TX/N)$$

where $I(\sigma^2) = //z_{ij}//$ is an unknown matrix, $//z_{ij}// = \sigma_i\sigma_j\delta_{ij}$, $\delta_{ij}(i, j = 1, 2, \ldots, N)$ is Kroneker symbol

$$C_N = X^TI(\sigma^2)X/N, \ C_N = //c_{kl}//, \ k, l = 1, 2, \ldots, m;$$
$$y^* = X\theta^*,$$

$$I_{kl}(x) = //a_{ij}{}^{kl}//, \ i, j = 1, 2, \ldots, m;$$

$$//a_{ij}{}^{kl}// = \phi_k(x_j)\phi_l(x_j)\delta_{ij}, \ c_{kl}{}^*$$
$$= (1/N)(y^* - y)I_{kl}(x)(y^* - y),$$
$$C_N{}^* = //c_{kl}{}^*//, \ k/l = 1, 2, ..., m.$$

**Theorem 2** Let $E\varepsilon_i{}^4 < \infty$ and $\left(m\sqrt{m}\right)/(N\lambda_1(N)) \to 0$, as $N \to \infty$, then

$$\left(c_{ij}{}^* - c_{ij}\right) \xrightarrow{P} 0, \quad E\left(c_{ij}{}^* - c_{ij}\right) \to 0 \ as \ N \to \infty.$$

*Remark* 1 In Theorem 2 we do not need the existence of the limit $c_{kl}^*$ and $c_{kl}$. This is because the difference between them converges to zero, in probability.

**Theorem 3** If $\left(m\sqrt{m}\right)/(N\lambda_1(N))$ will be bounded, then $\sqrt{N}(\theta^* - \theta) \Rightarrow N(0, D_N)$ as $N \to \infty$, where $\Rightarrow N(0, D_N)$ means a convergence in probability to the normal distribution with the covariance matrix $D_N$.

Different approaches for estimating the elements of covariance matrix were suggested in Belyaev and Hajiyev (1979), Hajiyev (2004), and Wu (1986).

## Nonlinear Regression Models

Let us assume that $f(x_i, \theta)$ in (1) is a nonlinear function and

$$f(x_i, \theta),\ \partial f(x, \theta)/\partial\theta,\ \partial^2 f(x, \theta)/\partial\theta_i\partial\theta_j, (i, j = 1, 2, \ldots, m)$$

are bounded and continuous functions of $(x, \theta)$, $\theta \in \Theta$ is a compact set. Denote

$$f_{ij}(\theta) = \partial f(x_j, \theta)/\partial\theta_i;$$

$F_N(\theta)$ is the matrix with elements $f_{ij}(\theta)$, $0 < \mu_1{}^N(\theta) \le \ldots \le \mu_m{}^N(\theta)$ eigenvalues of the matrix $\left[F_N{}^T(\theta)F_N(\theta)/N\right]$ and $B(r)$ be the sphere of the radius $r > 0$ centered at the point $\theta^*$. A least squares estimator of $\theta$ is constructed by the iterative process

$$\theta_N(s+1) = \theta_N(s) + \left[F_N{}^T(\theta_N(s))F_N(\theta_N(s))\right]^{-1} F_N{}^T(\theta_N(s))(y - f(x, \theta_N(s))). \quad (5)$$

The question arises as to whether the iterative process (5) converges or not. Relation (5) can be represented as

$$\theta_N(s+1) = u(\theta_N(s)) = \theta_N(s) + A_N(\theta_N(s))\delta_N(\theta_N(s)),$$

where

$$A_N(\theta_N(s)) = \left[F_N{}^T(\theta_N(s))F_N(\theta_N(s))/N\right]^{-1} F_N{}^T(\theta_N(s))$$

$$\delta_N(\theta_N(s)) = y - f(x, \theta_N(s)),\ \delta_N{}^*(\theta_N(s)) = y - f(x, \theta^*).$$

Define

$$\zeta_{N,r}{}^p(\theta) = m(\partial A_N(\theta)/\partial\theta_p)\varepsilon, p = 1, 2, .., m; \theta \in B(r),$$
$$L_p = \partial u_N(\theta)/\partial\theta_p,$$
$$\tau_N(r) = \max_{p=1,2,\ldots,m}\sup_{\theta\in\Theta}//L_p//.$$

Below, the convergence of random variables is understood as convergence in probability.

**Theorem 4** If there exists such $N$ that $m(N)^5/\left[N\left(\lambda_1{}^N(\theta)\right)^4\right] \to 0, r \to 0$, then

$$m(N)\tau_N(r) \to 0 \text{ and for any } p,\ \zeta_{N,r}{}^p(\theta) \to 0, r \to 0.$$

Introduce $\rho_N(\theta) = u_N(\theta) - \theta, \rho^* = \rho(\theta^*)$.

**Theorem 5** Let $\theta(0) \in B(r)$ and $\tau_N(r) + (//\rho^*//)/r < 1$. Then under the conditions of Theorem 4, there exists a random variable $\theta_N$ such that

$$\sqrt{N}(\theta_N - \theta^*) \Rightarrow N[0, \sum(\theta^*)] \text{ as } N \to \infty,$$

where

$$\sum(\theta^*) = \left[F_N{}^T(\theta^*)F_N(\theta^*)/N\right]^{-1} \left[F_N{}^T(\theta^*)I(\sigma^2)F_N(\theta^*)/N\right]\left[F_N{}^T(\theta^*)F_N(\theta^*)/N\right]^{-1},$$

that is, $\theta_N$ is a $\sqrt{N}$ consistent estimator and $\theta_N$ can be used as l.s.e. on $N$ observations. Using the approach suggested in Hajiyev and Hajiyev (2009), (similarly as for linear models) the elements of a covariance matrix can be estimated.

## The Construction of Asymptotic Confidence Bands

Consider the quadratic form

$$(\theta^* - \theta)^T(D_N)^{-1}(\theta^* - \theta) \le \chi_\gamma{}^2(m)/N. \quad (6)$$

According to Theorem 5, the left side of (6) has asymptotically chi-square distribution random with degrees of freedom $m$. In (6) instead of $D_N^{-1}$ (according to the Theorem 2) can be used estimates (Hajiyev and Hajiyev 2009) the matrix $D_N^{-1}$ elements. For the construction of a confidence band for $f(x, \theta)$, it is necessary to find $\inf f(x, \theta)$ and $\sup f(x, \theta)$, $\theta \in \varepsilon_\gamma(\theta)$, which are lower and upper boundaries of a confidence band and

$$\varepsilon_\gamma(\theta) = \left[\theta : (\theta^* - \theta)^T\left(D_N{}^{-1}\right)(\theta^* - \theta) \le \chi_\gamma^2(m)/N\right]$$

is the confidence ellipsoid, $\chi_\gamma{}^2(m)$ is the $\gamma > 0$ level quantile of the ▶chi-square distribution with $m$ degrees of freedom.

## About the Author

Dr. Asaf Hajiyev is a Professor and Chair, Department of Queuing Systems, Institute of Cybernetics, the Azerbaijan National Academy of Sciences (ANAS) and Head, Department of Probability and Mathematical Statistics, Baku State University. In 1980, he was awarded the Azerbaijan Lenin Komsomol Prize in Science and Technology. He is a Member of Bernoulli Society for Mathematical Statistics and Probability (1986), Elected member of the International Statistical Institute (2000), Elected correspondent-member Azerbaian National Academy of Science (2001), Elected member of the Third World Academy of Sciences (TWAS), Italy (2004). In 2008, he was elected a Member of the Mongolian National Academy of Sciences. Professor Hajiyev is Head of the Coordinating Council of the Azerbaian National Academy of Science in Mathematics. He

was a Member of the Azerbaijan Parliament: first (1995–2000), second (2000–2005), and third (2005–2010) convocations. During 2004–2006, he was elected Vice President of the Parliamentary Assembly of the Black Sea Economy Cooperation. He has authored more than 100 scientific papers and two books, including *Encyclopedia in Theory of Probability and Mathematical Statistics*.

## Cross References

►Convergence of Random Variables
►Eigenvalue, Eigenvector and Eigenspace
►Least Squares
►Linear Regression Models
►Nonlinear Regression

## References and Further Reading

Belyaev YuK, Hajiyev AH (1979) Sov J Comput Syst Sci 4:79–83
Demidenko EZ (1989) Optimization and regression. Nauka, Moscow. In Russian
Hajiyev AH (2004) Linear regression models with increasing numbers of unknown parameters. Doklady Mathematics 70(3): 887–891
Hajiyev AH, Hajiyev VG (2009) Nonlinear regression models with increasing numbers of unknown parameters. Doklady, Math 426(2):166–169
Huet S, Bouvier A, Gruey MA, Jolivet E (1996) Statistical tools for nonlinear regression. Springer, New-York
Sen A, Srivastava M (1997) Regression analysis: theory, methods and applications. Springer, Berlin
Wu CFJ (1986) Jackknife, bootstrap, and other resampling methods in regression analysis. Ann Stat 14:1261–1350

# Regression Models with Symmetrical Errors

Francisco José A. Cysneiros
Associate Professor
CCEN-UFPE - Cidade Universitária - Recife, Recife, Brazil

The normality assumption is a very attractive option for the errors of regression models with continuous response variables. However, when it is not satisfied, some transformation can be adopted for the response variable to obtain, at least, the symmetry property. It is known that the estimates of the coefficients in normal regression models are sensitive to extreme observations. Alternatives to the assumption of normal errors have been proposed in the literature. One of those alternatives is to consider that the errors have distributions with heavier tails than the normal distribution, in order to reduce the influence of outlier observations. In this context, Lange et al. (1989) proposed the Student $t$ model with unknown $\nu$ degrees of freedom. In the last decade, several results appeared as alternatives to modeling distributions other than the normal errors as, for instance, the symmetrical (or elliptical) distributions. Some of these results can be found in Fang et al. (1990), and Fang and Anderson (1990).

## Symmetrical Nonlinear Models

Consider the symmetrical nonlinear model

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $y_1, \ldots, y_n$ are the observed responses, $\mu_i = \mu_i(\boldsymbol{\beta}; \mathbf{x})$ is an injective and at least twice differentiable function with respect to $\boldsymbol{\beta}$. In addition, we suppose that the derivative matrix $\mathbf{D}_\beta = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ has rank $p(p < n)$ for all $\boldsymbol{\beta} \in \Omega_\beta \subset I\!R^p$, where $\Omega_\beta$ is a compact set with interior points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the parameter vector of interest, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of explanatory variable values and $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with the symmetrical density function $f_{\epsilon_i}(\epsilon) = g(\epsilon^2/\phi)/\sqrt{\phi}$, $y \in I\!R$, where $g : I\!R \to [0, \infty)$ is such that $\int_0^\infty g(u)du < \infty$. The function $g(\cdot)$ is typically known as the density generator. We will denote $\epsilon_i \sim S(0, \phi, g)$. The symmetrical class includes all symmetrical continuous distributions with heavier and lighter tails than the normal ones. When they exist, $E(Y_i) = \mu_i$ and $Var(Y_i) = \xi\phi$, where $\xi > 0$ is a constant that may be obtained from the expected value of the radial variable or from the derivative of the characteristic function (see, for instance, Fang et al. 1990). The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$ is given by $L(\boldsymbol{\theta}) = -n/2 \log \phi + \sum_{i=1}^n log\{g(u_i)\}$, where $u_i = \phi^{-1}\{y_i - \mu_i\}^2$. The score functions for $\boldsymbol{\beta}$ and $\phi$ take, respectively, the forms

$$\mathbf{U}_\beta(\boldsymbol{\theta}) = \frac{1}{\phi}\mathbf{D}_\beta^T\mathbf{V}(\mathbf{y} - \boldsymbol{\mu}) \ \ and$$
$$\mathbf{U}_\phi(\boldsymbol{\theta}) = 2\phi^{-1}\{Q_V(\boldsymbol{\beta}, \phi)/\phi - n\},$$

where $\mathbf{V} = diag\{v_1, \ldots, v_n\}$ with $v_i = -2W_g(u_i)$, $W_g(u) = \frac{g'(u)}{g(u)}$, $g'(u) = \frac{dg(u)}{du}$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $Q_V(\boldsymbol{\beta}, \phi) = \{(\mathbf{y} - \boldsymbol{\mu})^t\mathbf{V}(\mathbf{y} - \boldsymbol{\mu})\}$. The Fisher information matrix for $\boldsymbol{\theta}$ can be expressed as $\mathbf{K}_{\theta\theta} = diag\{\mathbf{K}_{\beta\beta}, K_{\phi\phi}\}$, where $\mathbf{K}_{\beta\beta} = 4d_g\phi^{-1}\mathbf{D}_\beta^T\mathbf{D}_\beta$ and $K_{\phi\phi} = n(4\phi^2)^{-1}(4f_g - 1)$ with $d_g = E\left\{W_g^2(U^2)U^2\right\}$, $f_g = E\left\{W_g^2(U^2)U^4\right\}$ and $U \sim S(0, 1, g)$. Thus, $\boldsymbol{\beta}$ and $\phi$ are orthogonal. Due to the similarity between the inference for elliptical and normal models, it is reasonable to expect that for large $n$ and under suitable regularity conditions, the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are approximately normal of means $\boldsymbol{\beta}$ and $\phi$ and variance–covariance matrices $\mathbf{K}_{\beta\beta}^{-1}$ and $\mathbf{K}_{\phi\phi}^{-1}$, respectively. General expressions for $W_g(u)$, $W_g'(u)$, $d_g, f_g$, and $\xi$ may be found, for instance, in Cysneiros and Paula (2005).

## Parameter Estimation

Some iterative procedures such as Newton–Raphson, BFGS, and Fisher scoring method can be used. Fisher scoring method can be easily applied to obtain $\hat{\boldsymbol{\theta}}$, where the iterative process can be interpreted as a modified least squares. The iterative process for $\hat{\boldsymbol{\theta}}$ take the form

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= \left\{\mathbf{D}_{\beta}^{T(m)}\mathbf{D}_{\beta}^{(m)}\right\}^{-1}\mathbf{D}_{\beta}^{T(m)}\mathbf{Z}^{(m)},\\ \phi^{(m+1)} &= \frac{1}{n}Q_V(\boldsymbol{\beta}^{(m+1)},\phi^{(m)}), \quad m=0,1,2,\ldots, \quad (2)\end{aligned}$$

where $\mathbf{Z} = \mathbf{D}_{\beta}\boldsymbol{\beta} + \frac{1}{4d_g}\mathbf{V}(\mathbf{y}-\boldsymbol{\mu})$. In linear case, we have $\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{y}$. Starting values should be given for $\boldsymbol{\beta}$ and $\phi$, for example, least square estimates. As we can see from the iterative process (2), the observations with small value for $v_i$ are down weighted for estimating $\boldsymbol{\beta}$. In particular, for the normal model, we have $v_i = 1$, $\forall i$. For the Student $t$ model with $\nu$ degrees of freedom, power exponential with shape parameter $k$, and logistic type II distributions, the values of $v_i$ are given in the Table 1.

It may be showed for the Student $t$ and logistic type II distributions, $v_i$ is inversely proportional to $u_i$. This property also follows for the power exponential distribution when $0 < k \leq 1$. Then, robustness aspects of $\hat{\beta}$ against outlying observations appear in these three heavy-tailed error distributions. In general, when the errors of the model have distribution with heavier tails than normal, the values of the weights $v_i$ have small values for $u_i$ large. Thus, models where the distribution of error have heavy tails can reduce the influence of extreme observations, while in the normal nonlinear regression model the weights are equal for all observations. In consequence, estimates in symmetrical regression models are less sensitive to the extreme observations than normal regression models. Extensions in the area of heteroscedastic symmetrical regression models can be found in Cysneiros et al. (2007, 2010) and codes in S-Plus and R to fit symmetrical regression models can be obtained in the Web page www.de.ufpe.br/~cysneiros/elliptical/elliptical.html.

**Regression Models with Symmetrical Errors. Table 1**
Expression of $v_i$ for some symmetrical distributions

| Distribution | $v_i$ |
|---|---|
| Normal | 1 |
| Student-t | $\frac{\nu+1}{(\nu+u_i)}$ |
| Logistic-II | $\frac{2exp(-\sqrt{u_i})-1}{(-2\sqrt{u_i})[1+exp(-\sqrt{u_i})]}$ |
| Power exponential | $\frac{1}{(1+k)u_i^{k/(k+1)}}$ |

## About the Author

Francisco Cysneiros is Associate Professor and Vice-Director of graduate studies (statistics graduate program) of Department of Statistics at Federal University of Pernambuco, Brazil. He is also Vice-Head of Department of Statistics (2005–2009). Professor Cysneiros is currently a member of the Advisory Board of Biometric Brazilian Journal (2008–2010) and he is an Associate Editor of the *Chilean Journal of Statistics*. He has served as a member of the Exact Science Committee – FACEPE (Research Foundation of Pernambuco, Brazil) and he is a fellowship researcher of the CNPq/Brazil since 2006.

## Cross References

►Heavy-Tailed Distributions
►Logistic Regression
►Nonlinear Models
►Nonlinear Regression
►Student's t-Distribution

## References and Further Reading

Cysneiros FJA, Paula GA (2005) Restricted methods in symmetrical linear regression models. Comput Stat Data Anal 49:689–708
Cysneiros FJA, Paula GA, Galea M (2007) Heteroscedastic symmetrical linear models. Stat Probab Lett 77:1084–1090
Cysneiros FJA, Cordeiro GM, Cysneiros AHMA (2010) Corrected maximum likelihood estimators in heteroscedastic symmetric nonlinear models. J Stat Comput Sim 80:451–461
Fang KT, Anderson TW (1990) Statistical inference in elliptical contoured and related distributions. Allerton Press, New York
Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman & Hall, London
Lange KL, Little RJ, Taylor J (1989) Robust statistical modelling using the t-distribution. J Am Stat Assoc 84:881–896

# Relationship Between Statistical and Engineering Process Control

Alberto Luceño
Professor
University of Cantabria, Santander, Spain

## Introduction

Many industrial processes must be adjusted from time to time to continuously maintain their outputs close to target. The reason for this is that such processes may be affected by disturbances produced, for example, by machines loosing their adjustment, components wearing out, and varying feed stock characteristics. Industrial control is a continual endeavor to keep measures of quality as close as possible to their target values for indefinite periods of time.

This may be attained using process monitoring and process adjustment tools. Monitoring implies continually checking the desired state of the process to detect and eliminate assignable causes of variation that can send the process out of control. Adjustment implies forecasting future deviations and taking corrective actions by feedback and/or feedforward. Process control can potentially benefit by using complementary tools of process monitoring and process adjustment within the same application.

## Process Monitoring Techniques

Process monitoring, or process surveillance, is a part of Statistical Process Control (SPC) that is used when the process can be brought to a satisfactory state of statistical control by systematically applying standardization of criteria, materials, methods, practices and processes.

Process monitoring is usually implemented by means of ►control charts, such as the Shewhart charts, the CUmulative SUM (CUSUM) charts, or the Exponentially Weighted Moving Average (EWMA) charts, among others.

The purpose of such methods is to continually check, or supervise, the state of the process in order to detect any conceivable out of control situation as soon as possible while simultaneously minimizing the rate of false alarms (i.e., alarms that eventually turn out to have no special cause).

When an alarm is triggered, a search for the special and potentially assignable cause of variation that presumably produced the alarm should be started. This search should end with the detection of such assignable cause and its permanent removal from the system. If the search fails, so that no special cause is eventually found, the alarm should be counted as a false alarm.

## Process Adjustment Techniques

Process adjustment is often considered as a part of Engineering Process Control (EPC) and is used when the process cannot be brought to a satisfactory state of statistical control, even after systematic application of standardization techniques. Much efforts have recently been dedicated, however, to bring some important features of process adjustment to the attention of the statistical community (e.g., see Box and Kramer 1992; Box and Luceño 1997a,b, 2002; Box et al. 2009; Luceño 2003, or Montgomery et al. 1994).

Process adjustment is often implemented by first using forecasting tools to estimate future deviations from target and subsequently modifying, or adjusting, an input compensatory variable so as to make those predicted deviations equal to zero (or to an appropriate small value in asymmetric situations). A process adjustment scheme may use feedback adjustments, feedforward adjustments, or a combination of both. Some types of feedback adjustment schemes are repeated adjustment schemes, constrained adjustment schemes, Proportional Integral Derivative (PID) control schemes, bounded adjustment schemes, among other.

The purpose of these methods is to indicate when and by how much the process has to be sampled and adjusted to keep it close to target. The only actions called for are to sample and to adjust the process when and as indicated by the adjustment scheme. The objective may be to minimize the output variance (or the mean squared error at the output) without any additional constraints, or to minimize the output variance constrained by a bound on the input variance, or by a bound on the frequency of adjustment, or on the frequency of sampling, or on the amount of each adjustment, among other possibilities.

## Conclusion

One can tentatively conclude that declarations of alarms and searches for special and potentially assignable causes of variation are not called for in the context of process adjustment techniques, but in the context of process monitoring techniques. By the same token, process adjustments are not called for in the context of process monitoring techniques, but much more appropriately in the context of process adjustment techniques.

Nevertheless, the appropriate combination of process monitoring and process adjustment tools, and their complementary use in SPC, is the subject of controversy within the statistical community. Further information can be found in the bibliography that follows, as well as in many documents produced by the International Organization for Standardization (ISO) and related organizations (e.g., ANSI, DIN, BSI, CEN).

## About the Author

For biography *see* the entry ►Control Charts.

## Cross References

►Control Charts
►Industrial Statistics
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

## References and Further Reading

Box GEP, Kramer T (1992) Statistical process monitoring and feedback adjustment. A discussion. Technometrics 34:251–267
Box GEP, Luceño A (1997a) Statistical control by monitoring and feedback adjustment. Wiley, New York
Box GEP, Luceño A (1997b) Discrete proportional-integral adjustment and statistical process control. J Qual Technol 29:248–260

Box GEP, Luceño A (2002) Feedforward as a supplement to feedback adjustment in allowing for feedstock changes. J Appl Stat 29:1241–1254
Box GEP, Luceño A, Paniagua-Quiñones MA (2009) Statistical control by monitoring and adjustment, 2nd edn. Wiley, New York
Luceño A (2003) Dead-band adjustment schemes for on-line feedback quality control. In: Khattree R, Rao CR (eds) Handbook of statistics: statistics in industry, vol 22. Elsevier, Amsterdam, pp 695–727
Montgomery DC (2005) Introduction to statistical quality control, 5th edn. Wiley, New York
Montgomery DC, Keats BJ, Runger GC, Messina WS (1994) Integrating statistical process control and engineering process control. J Qual Technol 26:79–87
NIST/SEMATECH (2009) e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/
Ruggery F, Kenetts RS, Faltin FW (eds) (2007) Encyclopedia of statistics in quality and reliability. Wiley, New York

# Relationships Among Univariate Statistical Distributions

Lawrence M. Leemis
Professor
The College of William & Mary, Williamsburg, VA, USA

Certain statistical distributions occur so often in applications that they are named. Examples include the binomial, exponential, normal, and uniform distributions. These distributions typically have *parameters* that allow for a certain degree of flexibility for modeling. Two important applications of these common statistical distributions are: $(a)$ to provide a probability model the outcome of a random experiment, and $(b)$ to provide a reasonable approximation to a data set.

Statistical distributions are traditionally introduced in separate sections in introductory probability texts, which obscures the fact that there are relationships between these distributions. The purpose of this section is to overview the various types of relationships between these common univariate distributions.

Common distributions and their relationships are presented in the encyclopedic work of Johnson et al. (1994, 1995) and Johnson et al. (2005). More concise treatments are given in Balakrishnan and Nevzorov (2003), Evans et al. (2000), Patel et al. (1976), Patil et al. (1985a, b), and Shapiro and Gross (1981). Figures that highlight the relationships between distributions are given in Casella and Berger (2002), Leemis and McQueston (2008), Marshall and Olkin (1985), Morris and Lock (2009), Nakagawa and Yoda (1977), Song (2005), and Taha (1982).

Since there are well over 100 named distributions used by probabilists and statisticians, the next three sections simply classify and illustrate some of the relationships.

## Special Cases

The first type of relationship between statistical distributions is known as a *special case*, which occurs when one distribution collapses to a second distribution for certain settings of its parameters. Two well-known examples are:

- A ►gamma distribution collapses to the exponential distribution when its shape parameter equals 1.
- A normal distribution with mean $\mu$ and variance $\sigma^2$ collapses to a standard normal distribution when $\mu = 0$ and $\sigma = 1$.

There are also certain special cases in which two statistical distributions overlap for a single setting of their parameters. Examples include $(a)$ the exponential distribution with a mean of two and the ►chi-square distribution with two degrees of freedom, $(b)$ the chi-square distribution with an even number of degrees of freedom and the Erlang distribution with scale parameter two, and $(c)$ the Kolmogorov–Smirnov distribution (all parameters known case) for a sample of size $n = 1$ and the $U(1/2,1)$ distribution, where $U$ denotes the uniform distribution (see ►Uniform Distribution in Statistics).

## Transformations

The second type of relationship between statistical distributions is known as a *transformation*. The term "transformation" is used rather loosely here, to include the distribution of an order statistic, truncating a random variable, or taking a mixture of random variables. Some well-known examples include:

- The random variable $(X - \mu)/\sigma \sim N(0,1)$ when $X \sim N(\mu, \sigma^2)$, where $N$ denotes the normal distribution.
- An Erlang random variable is the sum of mutually independent and identically distributed exponential random variables.
- The natural logarithm of a log normal random variable has the normal distribution.
- A hyperexponential random variable is the mixture of mutually independent exponential random variables.
- An order statistic taken from a sample of mutually independent $U(0,1)$ random variables has the ►beta distribution.

- A geometric random variable is the floor of an exponential random variable.
- If $X$ has the $F$ distribution with parameters $n_1$ and $n_2$, then $\left(1+(n_1/n_2)X\right)^{-1}$ has the beta distribution.
- If $X \sim U(0,1)$ then the floor of $10^X$ has the Benford distribution (Benford 1938).

It is also the case that two random variables from different statistical families can be combined via a transformation to form another common distribution, for example,

$$\frac{Z}{\sqrt{Y/n}} \sim t(n)$$

where $t(n)$ is the $t$ distribution with $n$ degrees of freedom, $Z$ is a standard normal random variable, and $Y$ is a chi-square random variable with $n$ degrees of freedom that is independent of $Z$.

## Limiting Relationships

The third type of relationship between statistical distributions is known as a *limiting* or *asymptotic* relationship, which is typically formulated in the limit as one or more parameters approach the boundary of the parameter space. Three well-known examples are:

- A standard normal distribution is the limit of a $t$ distribution as its degrees of freedom parameter approaches infinity.
- If $X_1, X_2, \ldots, X_n$ are mutually independent $U(0,1)$ random variables, then

$$n(1-\max\{X_1, X_2, \ldots, X_n\})$$

  approaches an exponential random variable in the limit as $n \to \infty$.
- The gamma distribution approaches the normal distribution as its shape parameter approaches infinity.

## Bayesian Relationships

The fourth type of relationship between statistical distributions is known as a *Bayesian* or *stochastic parameters* relationship, in which one or more of the parameters of a distribution are considered to be random variables rather than fixed constants. Two well-known examples are:

- If a random variable has a ▶binomial distribution with fixed parameter $n$ and random parameter $p$ which has the beta distribution, then the resulting random variable has the beta–binomial distribution.
- If a random variable has a negative binomial distribution with fixed parameter $n$ and random parameter $p$ which has the beta distribution, then the resulting random variable has the beta-negative binomial distribution.

## Internal Properties

The fifth and last type of relationship between statistical distributions is actually a relationship between a statistical distribution and itself. There are occasions when a particular operation on one or more random variables from a certain statistical family result in a new random variable that remains in that family. These are best thought of as properties of a statistical distribution rather than relationships between statistical distributions. Some well-known examples include:

- The *linear combination* property indicates that linear combinations of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ for $i = 1, 2, \ldots, n$; $a_1, a_2, \ldots, a_n$ are real constants, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

- The *convolution* property indicates that sums of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim \chi^2(n_i)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

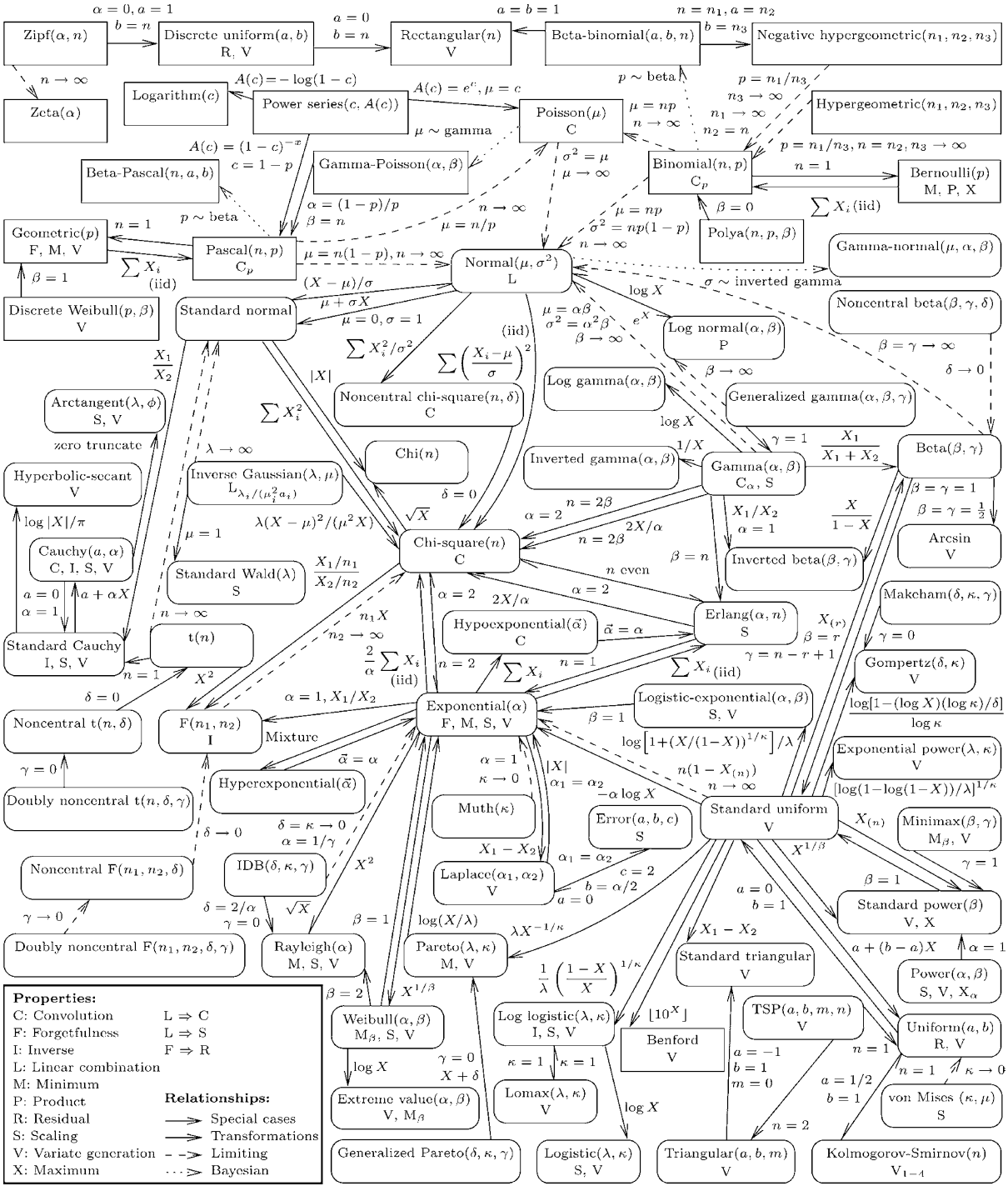
$$\sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} n_i\right),$$

  where $\chi^2$ denotes the chi-square distribution. The convolution property is a special case of the linear combination property.
- The *scaling* property implies that any positive real constant times a random variable having this distribution comes from the same distribution family. For example, if $X \sim Weibull(\alpha, \beta)$ and $k$ is a positive, real constant, then

$$kX \sim Weibull(\alpha k^{\beta}, \beta).$$

- The *product* property indicates that products of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim \text{log normal}\left(\mu_i, \sigma_i^2\right)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\prod_{i=1}^{n} X_i \sim log\ normal\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$




**Relationships Among Univariate Statistical Distributions. Fig. 1** Univariate distribution relationships

- The *inverse* property indicates that the reciprocal of a random variable of this type comes from the same distribution family. For example, if $X \sim F(n_1, n_2)$ then

$$\frac{1}{X} \sim F(n_2, n_1),$$

where $F$ denotes the $F$ distribution.

- The *minimum* property indicates that the smallest of mutually independent and identically distributed random variables from a distribution comes from the same distribution family. For example, if $X_i \sim exponential(\lambda_i)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\min\{X_1, X_2, \ldots, X_n\} \sim exponential\left(\sum_{i=1}^{n} \lambda_i\right),$$

  where the exponential parameter is a rate.
- The *residual* property indicates that the conditional distribution of a random variable left-truncated at a value in its support belongs to the same distribution family as the unconditional distribution. For example, if $X \sim U(a, b)$, and $k$ is a real constant satisfying $a < k < b$, then the conditional distribution of $X$ given $X > k$ belongs to the uniform family.

Many of the relationships described here are contained in Fig. 1 from Leemis and McQueston (2008), which is reprinted with permission from *The American Statistician*.

## About the Author

Dr. Lawrence Leemis is a Professor and former Department Chair, Department of Mathematics, The College of William & Mary in Virginia, U.S.A. He has authored or co-authored more than 100 articles, book chapters, and reviews. He has published three books: *Reliability: Probabilistic Models and Statistical Methods* (Prentice-Hall, 1995), *Discrete-Event Simulation: A First Course*, with Steve Park, (Prentice-Hall, 2006), and *Computational Probability: Algorithms and Applications in the Mathematical Sciences*, with John Drew, Diane Evans, and Andy Glen (Springer, 2008). He has won more than ten research and teaching awards, including the INFORMS Computing Society Prize in 2006. He is currently an associate editor for *Naval Research Logistics*, and has previously been an associate Editor for *IEEE Transactions on Reliability* and a Book Reviews Editor for the *Journal of Quality Technology*.

## Cross References

►Beta Distribution
►Binomial Distribution
►Chi-Square Distribution
►F Distribution
►Gamma Distribution
►Geometric and Negative Binomial Distributions
►Normal Distribution, Univariate
►Statistical Distributions: An Overview
►Testing Exponentiality of Distribution
►Uniform Distribution in Statistics
►Univariate Discrete Distributions: An Overview

## References and Further Reading

Balakrishnan N, Nevzorov VB (2003) A primer on statistical distributions. Wiley, Hoboken
Benford F (1938) The law of anomalous numbers. Proc Am Phil Soc 78:551–572
Casella G, Berger R (2002) Statistical inference, 2nd edn. Duxbury, Belmont
Evans M, Hastings N, Peacock B (2000) Statistical distributions, 3rd edn. Wiley, New York
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol I, 2nd edn. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol II, 2nd edn. Wiley, New York
Leemis LM, McQueston JT (2008) Univariate distribution relationships. Am Stat 62(1):43–53
Marshall AW, Olkin I (1985) A family of bivariate distributions generated by the bivariate Bernoulli distribution. J Am Stat Assoc 80:332–338
Morris CN, Lock KF (2009) Unifying the named natural exponential families and their relatives. Am Stat 63(3):247–253
Nakagawa T, Yoda H (1977) Relationships among distributions. IEEE Trans Reliab 26(5):352–353
Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York
Patil GP, Boswell MT, Joshi SW, Ratnaparkhi MV (1985a) Discrete models. International Co-operative Publishing House, Burtonsville
Patil GP, Boswell MT, Ratnaparkhi MV (1985b) Univariate continuous models. International Co-operative Publishing House, Burtonsville
Shapiro SS, Gross AJ (1981) Statistical modeling techniques. Marcel Dekker, New York
Song WT (2005) Relationships among some univariate distributions. IIE Trans 37:651–656
Taha HA (1982) Operations research: an introduction, 3rd edn. Macmillan, New York



# Renewal Processes

Kosto V. Mitov[1], Michael A. Zazanis[2]
[1]Professor, Faculty of Aviation
National Military University, Pleven, D. Mitropolia, Bulgaria
[2]Professor
Athens University of Economics and Business, Athens, Greece

Let $\{X_n; n \in \mathbb{N}\}$ be a sequence of independent, identically distributed random variables with values in $\mathbb{R}^+$ and distribution function $F$. The process $\{S_n; n \in \mathbb{N}_0\}$ defined

by means of $S_0 := 0, S_n := S_{n-1} + X_n, n = 1, 2, \ldots$ is called an *ordinary renewal process*. The non-negative random variables $X_n$ are called increments or, in many applications, inter-event times. In connection with the sequence of random points in time, $\{S_n\}$, one can define the *counting process* $N_t = \sum_{n=0}^{\infty} 1(S_n \le t), t \in \mathbb{R}^+$, where $1(A)$ designates the *indicator function* of the event $A$ (which is 1 if $A$ occurs and 0 otherwise). The *renewal function* associated with a renewal process is the increasing, right-continuous function $U(t) := EN_t = \sum_{n=0}^{\infty} F^{*n}(t)$ where $F^{*n}(t)$ denotes the $n$-fold *convolution* of the distribution function $F$ with itself (hence $F^{*n}(t) = P(S_n \le t)$).

Renewal processes are intimately related to the theory of the so-called *renewal equation* which is a linear integral equation of the form

$$Z(x) = z(x) + \int_0^x Z(x-y)F(dy) \tag{1}$$

where $z : \mathbb{R}^+ \to \mathbb{R}$ is a Borel function, bounded on finite intervals, and $F$ a probability distribution on $\mathbb{R}^+$. $F$ and $z$ are assumed to be given and the object is to determine the (unique) solution $Z$ which is bounded on finite intervals, and study its asymptotic behavior as $x \to \infty$. Its solution is given, in terms of the renewal function by the convolution $Z(x) = \int_0^x z(x-y)U(dy)$.

Renewal processes are important as special cases of random ▶point processes. In this respect the Poisson process (see ▶Poisson Processes) on the real line is the simplest and most important renewal process. They occur naturally in the theory of replacement of industrial equipment, the theory of queues, in branching processes, and in many other applications. In the framework of perpetual replacement of a single item, $X_n$ is the life of the $n$th such item which, as soon as it fails, is replaced by a new one with independent duration distributed according to $F$. Then $N_t$ is the number of items used in the time interval $[0, t]$ and $S_{N_t}$ is the time of the last replacement before $t$. We define three additional processes $\{A_t; t \ge 0\}$, $\{B_t; t \ge 0\}$, and $\{C_t; t \ge 0\}$ as follows: $A_t := t - S_{N_t - 1}$ is the *age*, $B_t := S_{N_t} - t$ is the *remaining life*, and $C_t := A_t + B_t = X_{N_t}$ is the total life duration of the item currently in use. (The age and remaining life are also known as the backward and forward recurrence times.) The statistics of these processes can be described by means of appropriate renewal equations. For instance, if $W_x(t) := P(A_t \le x)$ then conditioning on $S_1$ (using the so-called "renewal argument") we obtain

$$W_x(t) = (1 - F(t))1(t \le x) + \int_0^t W_x(t-s)dF(s). \tag{2}$$

If we allow the first increment to have a different distribution from all the others, i.e. if we set $S_0 = X_0$ and $S_n = S_{n-1} + X_n, n = 1, 2, \ldots$ where $X_0$ is independent of the $\{X_n\}$ and, unlike them, has distribution $F_0$, different from $F$, we obtain a *delayed renewal process*. This type of process is important because it provides additional flexibility in accommodating different initial conditions. Of course, its limiting properties are not affected by this modification. Of particular importance, assuming the mean $m$ to be finite, is the choice $F_0 = F_I$, given by

$$F_I(x) := \frac{1}{m} \int_0^x (1 - F(y))dy. \tag{3}$$

With this choice, $\{S_n\}$ becomes a *stationary point process*. $F_I$ is called the *integrated tail distribution* associated with the distribution $F$.

Of fundamental importance are the limit theorems related to renewal processes. If $m := \int_0^{\infty} xdF(x)$ denotes the mean of the increments, then the *Elementary Renewal Theorem* states that $\lim_{t\to\infty} t^{-1}U(t) = m^{-1}$. (The result holds also in the case $m = \infty$ provided that we interpret $m^{-1}$ as 0.) A refinement is possible if the increments have finite second moment, in which case $\lim_{t\to\infty} (U(t) - t/m) = EX_1^2/(2m^2)$. An analogous bound, due to Lorden (1970), also holds for all $t \ge 0$: $U(t) \le t/m + EX_1^2/m^2$. When the second moment exists we also have a Central Limit Theorem for the number of events up to time $t$: As $t \to \infty$, $\frac{N_t - t/m}{\sigma\sqrt{t/m^3}} \xrightarrow{d} Z$ where $Z$ is a standard Normal random variable and $\sigma^2 = Var(X_1)$.

Much deeper is *Blackwell's Theorem* which states that, if $F$ in non-lattice and the mean $m$ is finite then

$$\lim_{t\to\infty} (U(t+h) - U(t)) = h/m \qquad \text{for all } h > 0. \tag{4}$$

(A distribution $F$ on $\mathbb{R}^+$ is *lattice* with lattice size $\delta$ if there exists $\delta > 0$ such that the support of $F$ is a subset of $\{n\delta; n = 0, 1, 2, \ldots\}$ and $\delta$ is the largest such number.) If $F$ is lattice $(\delta)$ then (4) still holds, provided that $h$ is an integer multiple of $\delta$. Also, if $m = \infty$ the theorem still holds with $m^{-1} = 0$. Blackwell's original proof (1948) of (4) depended on harmonic analysis techniques. In the 1970s with the widespread use of coupling techniques simpler probabilistic proofs of the renewal theorem became available. (See Lindvall [1992] for a complete account.) An integral version of Blackwell's theorem, the *Key Renewal Theorem*, states that, if $z$ is directly Riemann integrable then the limit $\lim_{x\to\infty} \int_0^x z(x-y)dU(y)$ exists and equals $m^{-1} \int_0^{\infty} z(x)dx$. This then gives the limiting behavior of any function which satisfies a renewal equation (1). (Direct Riemann integrability is a direct extension of the Riemann

integral from bounded intervals to unbounded ones: Fix $h > 0$ and let $\overline{\gamma}_n(h) = \sup_{nh \le x < (n+1)h} z(x)$, $\underline{\gamma}_n(h) = \inf_{nh \le x < (n+1)h} z(x)$. Set $\overline{I}(h) := \sum_{n=0}^{\infty} h\overline{\gamma}_n(h)$ and $\underline{I}(h) := \sum_{n=0}^{\infty} h\underline{\gamma}_n(h)$. Clearly, if $h_1 > h_2 > 0$ then $\underline{I}(h_1) \le \underline{I}(h_2) \le \overline{I}(h_2) \le \overline{I}(h_1)$, though these quantities may not necessarily be finite. If $\lim_{h\to 0} \underline{I}(h)$ and $\lim_{h\to 0} \overline{I}(h)$ exist and are equal then $z$ is directly Riemann integrable. It should be noted that the direct Riemann integral is more restrictive than either the improper Riemann integral or the Lebesgue integral.)

The discrete version of the renewal theorem is simpler but not elementary. Suppose we are given a probability distribution $\{f_n; n = 1, 2, \ldots\}$ which is *non–arithmetic*, i.e. g.c.d.$\{n : f_n > 0\} = 1$ and has mean $m = \sum_{n=1}^{\infty} nf_n$, and define the renewal sequence $\{u_n; n = 0, 1, 2, \ldots\}$ via $u_0 = 1$, $u_n = f_n + f_{n-1}u_1 + \cdots + f_1 u_{n-1}$. Then $\lim_{n\to\infty} u_n = m^{-1}$ (interpreted as 0 when $m = \infty$). This is the celebrated Erdös–Feller–Pollard (1948) renewal theorem (see Feller [1968, 1971, Vol. 1, Chap. 13]) which marks the beginning of modern renewal theory and played a central rôle in the treatment of ▶Markov chains with countable state space. Interesting behavior arises if the non-arithmetic distribution function $\{f_n\}$ has infinite mean: Suppose that $\sum_{k=n+1}^{\infty} f_k = L(n)n^{-\alpha}$ where $0 < \alpha < 1$ and $L(n)$ is a *slowly varying* function. (A real function $L$ is said to be slowly varying if it is positive, measurable, and for every $\lambda > 0$, $L(\lambda x)/L(x) \to 1$ as $x \to \infty$.) Then (Garsia and Lamperti 1962) $\lim_{n\to\infty} n^{1-\alpha} L(n)u_n = \pi^{-1} \sin \pi\alpha$. If $1/2 < \alpha < 1$, this can be sharpened to $\lim_{n\to\infty} n^{1-\alpha} L(n)u_n = \pi^{-1} \sin \pi\alpha$. Analogous results in continuous time are also proved. Suppose that $F(.)$ is continuous, $F(0+) = 0$, $F(\infty) = 1$, $m = \infty$, and

$$1 - F(t) \sim \frac{t^{-\alpha} L(t)}{\Gamma(1-\alpha)} \Leftrightarrow m(t) := \int_0^t (1 - F(u))du \sim \frac{t^{1-\alpha} L(t)}{\Gamma(2-\alpha)}, \; t \to \infty, \qquad (5)$$

where $\alpha \in [0, 1)$ and $L(\cdot)$ is a slowly varying function at infinity. Under these conditions the growth rate of $U(t)$ is given by (see e.g. Bingham et al. [1987, Chap. 8]),

$$U(t) \sim C_\alpha t/m(t), \text{ as } t \to \infty, \text{ where } C_\alpha = [\Gamma(\alpha+1)\Gamma(2-\alpha)]^{-1}.$$

Erickson (1970) proved a version of Blackwell's theorem in the infinite mean cycle case. It states that if in (5), $\alpha \in \left(\frac{1}{2}, 1\right]$, then for any fixed $h > 0$

$$\lim_{t\to\infty} m(t)[U(t) - U(t-h)] = C_\alpha h.$$

If $\alpha \in \left(0, \frac{1}{2}\right]$, then lim has to be replaced by lim. Several versions of the Key Renewal Theorem in the infinite mean cycle case are also proved in Teugels (1968), Erickson (1970), and Anderson and Athreya (1987).

Using the Key Renewal Theorem one can obtain the asymptotic behavior of the age and the current and residual life. If $Y$ is a random variable with distribution $P(Y \le y) = \frac{1}{m}\int_0^y x dF(x)$ and $V$ is uniform in $[0, 1]$ and independent of $Y$, then

$$(A_t, B_t, C_t) \xrightarrow{d} (VY, (1-V)Y, Y) \text{ as } t \to \infty.$$

In particular the limiting marginal distribution of the age (which is the same as that of the residual life) is

$$\lim_{t\to\infty} P(A_t \le x) = F_I(x),$$

the integrated tail distribution given in (3). The limiting behavior of these processes gives rise to the so called "renewal paradox." For instance, the limiting distribution of the item currently in use is

$$\lim_{t\to\infty} P(C_t \le x) = \frac{1}{m}\int_0^x y dF(y)$$

with corresponding mean, provided that the second moment of $F$ exists, given by $m + \sigma^2/m$. Hence if we inspect such a process a long time after it has started operating (and is therefore in equilibrium) the part we are going to see will have longer life duration than average. Of course this is simply an instance of *length-biased sampling* and its effects are more pronounced when the variability of the distribution $F$ around its mean is large.

In the infinite mean cycle case the life time processes $A_t$ and $B_t$ have a linear growth to infinity, i.e. the normalized processes $A_t/t$ and $B_t/t$ have non-degenerate limit laws, jointly or separately. This result is usually called the Dynkin–Lamperti theorem (Dynkin 1955; Lamperti 1962). (See also Bingham et al. [1987, Chap. 8]). The theorem states that the condition (5) with $\alpha \in (0, 1)$ is necessary and sufficient for the existence of non-degenerate limit laws for $A_t/t$, $B_t/t$,

$$\lim_{t\to\infty} P(A_t/t \le x) = \pi^{-1} \sin \pi\alpha \int_0^x u^{-\alpha}(1-u)^{\alpha-1}, \quad 0 < x < 1,$$

$$\lim_{t\to\infty} P(B_t/t \le x) = \pi^{-1} \sin \pi\alpha \int_0^x u^{-\alpha}(1+u)^{-1} du, \quad x > 0.$$

An important and immediate generalization of the renewal equation (1) is to allow $F$ to be a general positive finite measure on $\mathbb{R}^+$. Setting $\|F\| := F(\mathbb{R}^+)$ one distinguishes the *excessive* case where $\|F\| > 1$, the *defective* case where $\|F\| < 1$, and the *proper* case we have already discussed, where $\|F\| = 1$. In the excessive case one can always find

a (unique) $\beta > 0$ such that $\int_0^\infty e^{-\beta x} dF(x) = 1$. One can define then a probability distribution function $F^\#$ via the relationship $dF^\#(x) = e^{-\beta x} dF(x)$, $x \geq 0$. Multiplying both sides of (1) by $e^{-\beta x}$ and setting $z^\#(x) = e^{-\beta x} z(x)$, $Z^\#(x) = e^{-\beta x} Z(x)$, the proper renewal equation $Z^\#(x) = z^\#(x) + \int_0^x z^\#(x-y) dF^\#(y)$ is obtained. The Key Renewal Theorem then yields

$$\lim_{x\to\infty} e^{-\beta x} Z(x) = \frac{1}{m^\#} \int_0^\infty z^\#(y) dy,$$

which establishes that, asymptotically, $Z$ grows exponentially with rate $\beta$. We should point out that the defective case is not entirely similar. While formally one again tries to identify $\beta > 0$ so that $\int_0^\infty e^{\beta x} dF(x) = 1$, this may or may not be possible according to whether the distribution function $\frac{1}{\|F\|} F(x)$ is *light-tailed* or *heavy-tailed*. In the former case one proceeds just as in the excessive case. (For more details see Feller [1968, 1971, Vol. 2, Chap. 11]). This type of analysis is characteristic of the applications of renewal theory to areas such as population dynamics, the theory of collective insurance risk, and to the economic theory or replacement and depreciation (Jorgenson 1974; Feldstein and Rothchild 1974).

Alternating renewal processes arise in a natural way in many situations, like queueing systems and reliability of industrial equipment, where working(busy) periods ($X$) interchange with idle periods ($T$). Consider a sequence of random vectors with non-negative coordinates $(T_i, X_i)$, $i = 1, 2, \ldots$. It defines an *alternating renewal sequence* $(S_n, S'_{n+1})$ as follows $S_0 = 0$, $S'_n = S_{n-1} + T_n$, $S_n = S'_n + X_n = S_{n-1} + (T_n + X_n)$, $n = 1, 2, \ldots$. An interpretation in terms of the reliability theory is the following. There are two types of renewal events: $S_n$ is the moment when the installation of a new element begins (The installation takes time $T_n$); $S'_{n+1}$ is the moment when the installation ends and the new element starts working. (The working period has length $X_n$). The renewal process $N(t) = \sup\{n : S_n \leq t\}$ counts the pairs of renewal events in the interval $[0, t]$. The processes $\sigma_t = \max\{0, t - S'_{N(t)+1}\}$ – *spent working time* and $\tau_t = \min\{S_{N(t)+1} - t, X_{N(t)+1}\}$ – *residual working time* generalize the lifetime processes $A_t$ and $B_t$. Their properties are derived in Mitov and Yanev (2001) in the infinite mean cycle case.

The central place that renewal theory holds in the analysis of stochastic systems is due to the concept of *regeneration*. Let $\{X_t; t \in \mathbb{R}^+\}$ be a process with values in $\mathcal{S}$ (e.g. a Euclidean space $\mathbb{R}^d$) and sample paths that are càdlàg (right-continuous with left-hand limits) a.s.. Such a process is called *regenerative* with respect to a (possibly delayed) renewal process $\{S_n\}$, defined on the same probability space, if, for each $n \in \mathbb{N}$ the *post* $S_n$ *process* $(\{X_{S_n+t}\}_{t\geq 0}, \{S_{n+k} - S_n\}_{k\in\mathbb{N}})$ is independent of $\{S_0, S_1, \ldots, S_n\}$ and its distribution does not depend on $n$, i.e. $(\{X_{S_n+t}\}_{t\geq 0}, \{S_{n+k} - S_n\}_{k\in\mathbb{N}}) \stackrel{d}{=} (\{X_{S_0+t}\}_{t\geq 0}, \{S_k - S_0\}_{k\in\mathbb{N}})$ for all $n$. The existence of an embedded, non-lattice renewal process with respect to which the process $\{X_t\}$ is regenerative, together with the finiteness of the mean $m := E[S_1 - S_0]$ is enough to guarantee the existence of a "stationary version," say $\{\tilde{X}_t\}$, to which $\{X_t\}$ converges as $t$ goes to infinity. The statistics of $\{\tilde{X}_t\}$ can be determined by analyzing the behavior of $\{X_t\}$ over any *regenerative cycle*, i.e. a random time interval of the form $[S_n, S_{n+1})$. If $k \in \mathbb{N}$, $t_i \in \mathbb{R}^+$, $i = 1, 2, \ldots, k$, and $f : \mathcal{S}^k \to \mathbb{R}$ is any bounded, continuous function then

$$Ef(\tilde{X}_{t_1}, \ldots, \tilde{X}_{t_k}) = \frac{1}{m} E \int_{S_0}^{S_1} f(X_{t_1+t}, \ldots, X_{t_k+t}) dt.$$

Nowadays, our view of whole areas of probability, including parts of the theory of ▶Markov processes is influenced by renewal theoretic tools and related concepts of regeneration. The analysis of many stochastic models is greatly facilitated if one identifies certain embedded points in time that occur according to a renewal process and with respect to which the process is regenerative. The fact that these regeneration cycles are independent, identically distributed, also facilitates the statistical analysis of the simulation output of regenerative systems.

A detailed representation of the renewal theory and its applications could be found, for instance, in the following books Asmussen (2003), Bingham et al. (1987), Feller (1968, 1971), and Resnick (1992).

## Cross References


▶Point Processes
▶Poisson Processes
▶Queueing Theory
▶Statistical Inference for Stochastic Processes
▶Stochastic Processes: Classification


## References and Further Reading


Asmussen S (2003) Applied probability and queues, 2 edn. Springer, New York

Anderson KK, Athreya KB (1987) A renewal theorem in the infinite mean case. Ann Probab 15:388–393

Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Cambridge University Press, Cambridge

Dynkin EB (1955) Limit theorems for sums of independent random quantities. Izves Akad Nauk U.S.S.R 19:247–266

Erickson KB (1970) Strong renewal theorems with infinite mean. Trans Am Math Soc 151:263–291

Feldshtein M, Rotschild M (1974) Toward an economic theory of replacement investment, Econometrica 42(3):393–423
Feller W (1968, 1971) An introduction to probability theory and its applications, vol 1 and 2. Wiley, New York
Garsia A, Lamperti J (1962) A discrete renewal theorem with infinite mean. Comment Math Helv 37:221–234
Jorgenson DW (1974) Investment and production: a review. In: Intriligator M, Kendrick D (eds) Frontiers of quantitative economics, vol 2. Amsterdam, North-Holland, pp 341–366
Lamperti J (1962) An invariance principle in renewal theory. Ann Math Stat 33:685–696
Lindvall T (1992) Lectures on the coupling method. Wiley, New York
Lorden G (1970) On the excess over the boundary. Ann Math Stat 41:520–527
Mitov KV, Yanev NM (2001) Limit theorems for alternating renewal processes in the infinite mean case. Adv Appl Probab 33:896–911
Resnick S (1992) Adventures in stochastic processes. Birkhäuser, Boston

# Repeated Measures

Geert Molenberghs
Professor
Universiteit Hasselt & Katholieke Universiteit Leuven, Leuven, Belgium

Repeated measures are obtained whenever a specific response is measured repeatedly in a set of units. Examples are hearing thresholds measured on both ears of a set of subjects, birth weights of all litter members in a toxicological animal experiment, or weekly blood pressure measurements in a group of treated patients. The last example is different from the first two examples in the sense that the time dimension puts a strict ordering on the obtained measurements within subjects. The resulting data are therefore often called longitudinal data. Obviously, a correct statistical analysis of repeated measures or longitudinal data can only be based on models which explicitly take into account the clustered nature of the data. More specifically, valid models should account for the fact that repeated measures within subjects are allowed to be correlated. For this reason, classical (generalized) linear regression models are not applicable in this context. An additional complication arises from the highly unbalanced structure of many data sets encountered in practice. Indeed, the number of available measurements per unit is often very different between units, and, in the case of longitudinal data, measurements may have been taken at arbitrary time points, or subjects may have left the study prematurely, for a number of reasons (sometimes known but mostly unknown). A large number of models have been proposed in the statistical literature, during the last few decades. Overviews are given in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005).

## About the Author

For biography *see* the entry ▶Linear Mixed Models.

## Cross References

▶Nonlinear Mixed Effects Models
▶Research Designs
▶Sample Survey Methods
▶Statistical Analysis of Longitudinal and Correlated Data

## References and Further Reading

Brown H, Prescott R (1999) Applied mixed models in medicine. Wiley, New York
Crowder MJ, Hand DJ (1990) Analysis of repeated measures. Chapman & Hall, London
Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Chapman & Hall, London
Davis CS (2002) Statistical methods for the analysis of repeated measurements. Springer, New York
Demidenko E (2004) Mixed models: theory and applications. Wiley, New York
Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford
Fahrmeir L, Tutz G (2002) Multivariate statistical modelling based on generalized linear models, 2nd edn. Springer, New York
Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs G (2009) Longitudinal data analysis. Handbook. Wiley, Hoboken
Goldstein H (1995) Multilevel statistical models. Edward Arnold, London
Hand DJ, Crowder MJ (1995) Practical longitudinal data analysis. Chapman & Hall, London
Hedeker D, Gibbons RD (2006) Longitudinal data analysis. Wiley, New York
Kshirsagar AM, Smith WB (1995) Growth curves. Marcel Dekker, New York
Leyland AH, Goldstein H (2001) Multilevel modelling of health statistics. Wiley, Chichester
Lindsey JK (1993) Models for repeated measurements. Oxford University Press, Oxford
Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2005) SAS for mixed models, 2nd edn. SAS Press, Cary
Longford NT (1993) Random coefficient models. Oxford University Press, Oxford
Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer, New York
Pinheiro JC, Bates DM (2000) Mixed effects models in S and S-Plus. Springer, New York
Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer Series in Statistics. Springer, New York
Vonesh EF, Chinchilli VM (1997) Linear and non-linear models for the analysis of repeated measurements. Marcel Dekker, Basel

Weiss RE (2005) Modeling longitudinal data. Springer, New York
West BT, Welch KB, Gałecki AT (2007) Linear mixed models: a practical guide using statistical software. Chapman & Hall/CRC Press, Boca Raton
Wu H, Zhang J-T (2006) Nonparametric regression methods for longitudinal data analysis. Wiley, New York
Wu L (2010) Mixed effects models for complex data. Chapman & Hall/CRC Press, Boca Raton

# Representative Samples

Ksenija Dumicic
Full Professor, Head of Department of Statistics, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

According to Lavrakas (2008), a representative sample is one that ensures external validity in relationship to the population of interest the sample is meant to represent. In addition, it should be said that a representative sample is a probability sample, so, *sampling errors* for estimates can be calculated and the *estimates* from the sample survey can be generalized with confidence to the sampling population.

Random selection, i.e., being objective and unbiased, is an essential element of *survey sampling*. There are many factors that affect the representativeness of a sample, though traditionally attention has mostly been paid to sample design and coverage. More recently, the focus has extended to the nonresponse issues.

Zarkovic (1956), Kruskal and Mosteller (1980), Bellhouse (1988), Kish (2002), and Rao (2005) wrote histories of random sampling methods as representative methods. The statistical literature gives examples of all the meanings for "representative sampling," such as: general acceptance for data; absence of selective forces; miniature of the population; typical or ideal cases; and proper coverage of the population. Kruskal and Mosteller (1979) added the following new meanings: representative sampling as a specific sampling method; representative sampling as permitting unbiased estimation; and representative sampling as sufficient to serve a particular purpose. Occasionally, it is also determined as a vague term.

The development of modern *sampling theory* started in around 1895 when the Norwegian statistician A.N. Kiaer, the first director of Statistics Norway, published his *Representative Method* and was the first to promote "the representative method" over the ►*census* as a complete enumeration. Kiaer stated that if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to other survey variables. For Kiaer, a representative sample is a "miniature" of the actual population, though the selection of units is based on *purposive selection*, according to a rational scheme based on general results of previous investigations. He presented his thoughts at a meeting of the International Statistical Institute in Bern in 1895. Many famous statisticians did not agree on Kiaer's new approach, as no measure of the accuracy of the estimates could be obtained. A basic problem was that the representative method lacked a formal theory of inference.

It was Sir A.L. Bowley, an English statistician, who pioneered the use of *simple random sampling*, for which the accuracy measures of estimates could be computed. He introduced *stratified random sampling* with *proportional allocation*, leading to a representative sample with equal *inclusion probabilities*. By the 1920s, the representative method was widely used. In 1924, the International Statistical Institute played a prominent role with its formation of a committee to report on the representative method. In 1935, Polish scientist J. Neyman published his now famous paper (Neyman 1934). He developed a new theory and laid the theoretical foundations for design-based sampling or the probability sampling approach to inference from survey samples. He showed that stratified random sampling is preferable to ►balanced sampling and introduced the *optimal allocation* of units based on efficiency in his theory of stratified random sampling without replacement, by relaxing the condition of equal inclusion probabilities for sampling units. In 1943, M.H. Hansen and W.N. Hurwitz published their theory of *multistage cluster samples*. In 1944, W.G. Madow and L.H. Madow conducted the first theoretical study of the precision of *systematic sampling*. The classical theory of survey sampling was more or less completed in 1952. Horvitz and Thompson (1952) completed the classical theory, and the *random sampling* approach was almost unanimously accepted. Most of the classical books on *sampling* were also published by then: W.G. Cochran in 1953 (see the last edition: Cochran 1977), Deming (1950), Hansen et al. (1953a, b), and Yates (1949). Later, a great contribution to probability sampling was given by Kish (1965).

The representative method is applied in *survey research*, *both social and business*, in *official statistics*, for *public opinion polling*, in *market research*, etc. It is also applied for *audit sampling* and *statistical quality control*.

The sampling technique used by G. Gallup was *quota sampling* for opinion polling. Gallup's approach was in great contrast with that of *Literary Digest* magazine, the

leading polling organization at that time. This magazine conducted regular "America Speaks" polls with a *convenient sample* of the sample size near to two million people. The presidential election of 1936 turned out to be decisive for both approaches (Rao 2005). Gallup correctly predicted using a sample size of 3,000 that the candidate Alf Landon would beat Franklin Roosevelt. It seemed to be strange how could a prediction based on such a large sample be wrong? The explanation was the fatal flaw in the *Literary Digest*'s sampling mechanism. The automobile registration lists and telephone directories applied were not representative samples. In the 1930s, cars and telephones were typically owned by the middle and upper classes. More well-to-do Americans tended to vote Republican and the less well-to-do were inclined to vote Democrat. Therefore, Republicans were over represented in the *Literary Digest* sample. As a result of this famous historical error, opinion researchers learned that the *manner of selection* of a sample is more important than the *sample size*. Also, among nonprobability sample designs, a *quota sample*, if well designed, would be the most similar to a probability sample as a representative one.

In a sample survey, researchers must judge whether the sample is actually representative of the target population. The best way of ensuring a representative sample is to have a complete sampling frame (i.e., directory, list or map) covering all the elements in the population, and to know that each and every element (e.g., person, household, enterprise, etc.) on the list has a nonzero probability (equal or unequal) of being included in the sample. Furthermore, it is necessary to use random selection to draw elements from the sampling frame into the sample based either on a random number generator or on systematic selection procedure. Also, it is essential to collect complete data from every single sampled element.

Completely up-to-dated sampling frames of the populations of interest are very rare. If there are elements in the target population with a zero probability of selection, sample estimates cannot be generalized to these elements. For example, if unemployed persons belong to the population of interest, but were not registered as unemployed, then they would have a zero probability of inclusion in the sample. Further, very modern Internet surveys, Tele-Voting and Push Polling samples are not based on solid sampling frames and instead use non-representative sample designs. As such, the results in such surveys cannot be generalized to the overall population and users should be aware that these are nothing more than amusement techniques.

First, to judge the representativeness of a sample, the use of some prior knowledge about the population main variables structures is recommended, for comparison with the sample structures. Occasionally, an extra random sample is helpful. Further, to correct for biases, survey researchers apply post-stratification. Post-stratification is the process of weighting some of the respondents in the responding sample relative to others, so that the characteristics of the responding sample are essentially equal to those of the target population for those characteristics that can be controlled to complete coverage data (e.g., age, gender, educational level, geography, etc.). Applying post-stratification adjustments reduces the bias due to *noncoverage and nonresponse*. And finally, it is necessary to limit the conclusions to those elements in the sampling frame to only those with a nonzero probability of inclusion. In other words, to avoid biases, researchers need to estimate coverage, and both unit and item-nonresponse.

## About the Author

Dr. Ksenija Dumicic is Professor, and, since 2006 Head of Department of Statistics, Faculty of Economics and Business, University of Zagreb, Croatia. She is founder of the postgraduate studies, Statistical Methods for Economic Analysis and Forecasting. She has been a member of Editorial Boards for several journals in Croatia, Serbia and Bosnia and Herzegovina. She has authored and coauthored more than 80 papers and two books in statistical research methodology. She specializes in statistical sample survey methods and has supervised over 10 postgraduate students.

## Cross References

►Balanced Sampling
►Non-probability Sampling Survey Methods
►Nonresponse in Surveys
►Nonresponse in Web Surveys
►Sample Survey Methods
►Telephone Sampling: Frames and Selection Techniques

## References and Further Reading

Bellhouse DR (1988) A brief history of random sampling methods. In: Krishnaiah PR, Rao CR (eds) Handbook of statistics: sampling. Elsevier, Amsterdam, pp 1–14

Cochran WG (1977) Sampling tehniques, 3rd edn. Wiley, New York

Deming WE (1950) Some theory of sampling. Wiley, New York

Hansen MH, Hurwitz WN, Madow WG (1953a) Sample survey methods and theory, volume I. Wiley, New York

Hansen MH, Hurwitz WN, Madow WG (1953b) Sample survey methods and theory, volume II. Wiley, New York

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685

Kish L (1965) Survey sampling. Wiley, New York

Kish L (2002) New paradigms (models) for probability sampling. In: Survey methodology, vol 28(1), Statistics Canada, Catalogue No. 12001XIE, pp 31–34

Kruskal W, Mosteller F (1979) Representative sampling, III: the current statistical literature. Int Stat Rev 47:245–265

Kruskal WH, Mosteller F (1980) Representative sampling IV: the history of the concept in statistics, 1895–1939. Int Stat Rev 48:169–195
Lavrakas PJ (2008) Encyclopedia of survey research methods, volume 2. SAGE Publications, California
Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J Roy Stat Soc 97:558–606
Rao JNK (2005) Interplay between sample survey theory and practice: an appraisal. In: Survey methodology, vol 31(2), Statistics Canada, Catalogue No. 12001XPB, pp 117–138
Yates F (1949) Sampling methods for censuses and surveys. Griffin, London
Zarkovic SS (1956) Note on the history of sampling methods in Russia. J Roy Stat Soc A 119:336–338

# Research Designs

Roxana Toma, G. David Garson
North Carolina State University, Raleigh, NC, USA

## Introduction

Research design is a term broadly referring to any plan for gathering data systematically in such a way as to be able to arrive at conclusions. On the subject selection dimension, designs may be experimental, quasi-experimental, or non-experimental. On the measurement dimension, designs may be between-subjects or within-subjects.

Experimental studies are characterized by ►randomization of subjects into treatment and control groups. Control groups may receive no treatment or some standard treatment, where "treatment" is exposure to some stimulus. Randomization serves to control for variables which are not included explicitly in the study. In quasi-experimental designs, treatment and comparison groups are not composed of randomized subjects, even if data are gathered through random sampling. In the absence of randomization, control for confounding variables must be accomplished explicitly through statistical techniques. Finally, a design is non-experimental if there is systematic collection of data with respect to topics of interest but there is no randomization of subjects as in experimental studies nor statistical controls as in quasi-experimental designs. Most case study designs exemplify this category.

By type of measurement, between-subjects designs are the most common. The researcher is comparing between subjects who experience different treatments. There are different subjects for each level of the independent variable(s). Any given subject is exposed to only one level and comparisons are made between subjects' reactions or effects. In contrast, in within-subjects designs the same subjects are measured for each level of the independent variable, as in before-after studies or panel studies. Similar subjects, as in matched pair's designs, are of the same type. When subjects are measured more than once, within-subjects designs are also called repeated measures designs. Since the subjects are the same for all levels of the independent variable(s), they are their own controls (i.e., subject variables are controlled). However, there is greater danger to validity in the form of carryover effects due to exposure to earlier levels in the treatment sequence (e.g., practice, fatigue, attention) and there is danger of attrition in the sample. Counterbalancing is a common but not foolproof strategy to address carryover effects: e.g., half the subjects get treatment *A* first, then *B*, while the other half get B first, then *A*.

## Factorial and Block Designs

Factorial designs use categorical independent variables to establish groups. For instance, in a two factor design, the independent variables might be information type (fiction, non-fiction) and media type (television, print, Internet), generating two times three = six categories. A factorial design is "fully crossed" if there is a group for every possible combination of factors (independent variables). An "incomplete" factorial design, leaving out some of the groups, may be preferred if some combinations of values of factors are nonsensical or of no theoretical interest. In experimental designs, an equal number of subjects are assigned randomly to each of the six possible groups (e.g., to the fiction-television group). The researcher might then measure subjects on information retention. A null outcome would be indicated by the average retention score being the same for all six groups of the factorial design. Unequal mean retention scores would indicate a main effect of information type or media type, and/or an interaction effect of both. Quasi-experimental designs may also be factorial, but groups are established by stratified random sampling, not randomization of subjects, entailing the need for more complex and explicit statistical controls and possibly less conclusive results.

**Balanced designs** are simply factorial designs where there are equal numbers of cases in each subgroup (cell) of the design, assuring that the factors are independent of one another (but not necessarily the covariates). Unbalanced designs have unequal n's in the cells formed by the intersection of the factors.

**Randomized block designs** stratify the subjects and for each strata, a factorial design is run. This is typically done when the researcher is aware of nuisance factors that

need to be controlled (e.g., there might be an air conditioned room stratum and a no air conditioning stratum) or if there were other mitigating structural factors known in advance (e.g., strata might be different cities). That is, the blocking variables which stratify the sample are factors which are considered to be control variables, not independent variables as they would be in a simple factorial design. Randomized block designs seek to control for the effects of main factors and their interactions, controlling for the blocking variable(s).

**Nested designs** have two or more factors, but the levels of one factor are never repeated as levels in the other factor(s). This happens in hierarchical designs, for instance, when a forester samples trees, then samples seedlings of each sampled tree for survival rates. The seedlings are unique to each tree and represent a random factor. Likewise, a researcher could sample drug companies, then could sample drug products for quality within each sampled company. This contrasts with crossed designs of ordinary two-way (or higher) analysis of variance, in which the levels of one factor appear as levels in another factor (e.g., tests may appear as levels across schools). We can get the mean of different tests by averaging across schools, but we cannot get the mean survival rate of different seedlings across trees because each tree has its own unique seedlings. Likewise, we cannot compute the mean quality rating for a drug product across companies because each company has its own unique set of products. Latin square and Graeco-Latin square designs discussed below are nested designs.

## Random Versus Fixed Effects Designs

Most designs are fixed effects models, meaning that data are collected on all categories of the independent variables. In random effects models (also called random factors models), in contrast, data are collected only for a sample of categories. There is replaceability, meaning that the levels of the factor are randomly or arbitrarily selected and could be replaced by other, equally acceptable levels. The purpose of random effects modeling is generalizability – the researcher wishes to generalize findings beyond the particular, randomly or arbitrarily selected levels in the study. For instance, a researcher may study the effect of item order in a ►questionnaire. Six items could be ordered 720 ways. However, the researcher may limit him- or herself to the study of a sample of six of these 720 ways. The random effects model in this case would test the null hypothesis that the effects of ordering are zero. Note that "mixed factorial design" is also possible simply by having a random effects model with a fixed factor and a random factor.

**Treatment by replication design** is a common random effects model. The treatment is a fixed factor, such as exposure to different types of public advertising, while the replication factor is represented by the particular respondents who are treated. Sometimes it is possible and advisable to simplify analysis from a hierarchical design to a simple treatment by replication design by shifting the unit of analysis. An example would be to use class averages rather than student averages in a design in which students represent a random factor nested within teachers as another random factor (the shift drops the student random factor from analysis). Note also that the greater the variance of the random effect variable, the more levels are needed (e.g., more subjects in replication) to test the fixed (treatment) factor at a given alpha level of significance.

## Common Experimental Designs

A very large number of research designs have been devised for experimental design. Though not exclusive to experimental design, the most prevalent examples are outlined below.

**Completely randomized designs** assign an equal number of subjects randomly to each of the cells formed by the factors. In the quasi-experimental mode, where there is no control by randomization, the researcher must measure and employ controls explicitly by using covariates.

**Latin square designs** extend block designs to control for two categorical variables. This design requires that the researcher assume all interaction effects are zero. Normally, if one had three variables, each of which could assume four values, then one would need $4^3 = 64$ observations just to have one observation for every possible combination. Under Latin square design, however, the number of necessary observations is reduced to $4^2 = 16$ because the third variable is nested. For instance, suppose there are 4 teachers, 4 classes, and 4 textbooks. The 16 groups in the design would be the 16 different class-textbook pairs. Each teacher would teach in each of the four classes, using a different text each time. Each class would be taught by each of the four different teachers, using a different text each time. However, only 16 of the 64 possible teacher-class-textbook combinations would be represented in the design because textbooks represent a nested factor, with each class and each teacher being exposed to a given textbook only once. Eliminating all but 16 cells from the complete (crossed) design requires the researcher to assume that there are no significant teacher-textbook or class-textbook interaction effects, only the main effects for teacher, class, and textbook.

**Graeco-Latin square designs** extend Latin square block designs to control for three categorical variables.

**Split-plot designs**. Like randomized complete block designs, in split plot designs there is still a blocking factor but each block is split into two segments and segments are assigned to the blocks in random order. Within any segment, treatments are assigned in random order. For instance, in a study of health improvement effects, the blocking factor might be in the form of three age groups, treatment in the form of three levels of dosage of medicine, with the segmentation variable being two brands of medicine. Splitting the three age blocks yields six segments, with each age group having a Brand A and Brand B segment. Each of the six segments is homogenous by brand. Split-plot designs are used when homogeneity rather than randomization within blocks is required (in agriculture, for instance, equipment considerations could dictate that any given plot segment only receive one brand of fertilizer).

**Split-plot repeated measures designs** can be used when the same subjects are measured more than once. In a typical split-plot repeated measures design, subjects are measured on some variable over a number of trials. Subjects are also split by some grouping variable. In this design, the between-subjects factor is the group (treatment or control) and the repeated measure is, for example, the test scores for two trials. The resulting statistical output will include a main treatment effect (reflecting being in the control or treatment group) and a group-by-trials interaction effect (reflecting treatment effect on posttest scores, taking pretest scores into account).

## Common Quasi-Experimental Designs

As with experimental designs, numerous types of quasi-experimental designs exist, many enumerated in the classic work of Cook and Campbell (1979).

**One-group posttest-only design**. This design lacks a pretest baseline or a comparison group, making it impossible to come to reliable conclusions about a treatment effect.

**Posttest-only design with nonequivalent comparison groups**. In this common social science design, it is also impossible to come to reliable conclusions about treatment effect based solely on posttest information on two nonequivalent groups since effects may be due to treatment or to nonequivalencies between the groups.

**Posttest-only design with predicted higher-order interactions**. The presence of an interaction effect creates two or more expectations compared to the one-expectation one-group posttest-only design. Because there are more expectations, there is greater verification of the treatment effect but the explanation accounting for the interaction is more complex and therefore may be less reliable.

**One-group pretest-posttest design**. This is a common but flawed design subject to such threats to validity as history (events intervening between pretest and posttest), maturation (changes in the subjects that would have occurred anyway), regression toward the mean (the tendency of extremes to revert toward averages), testing (the learning effect on the posttest of having taken the pretest), and the like.

**Two-group pretest-posttest design using an untreated control group**. If a comparison group which does not receive treatment is added to what otherwise would be a one-group pretest-posttest design, threats to validity are greatly reduced. This is the classic experimental design but in quasi-experimental design, since the groups are not equivalent, there is still the possibility of selection bias.

**Double pretest design**. The researcher can strengthen pretest-posttest designs by having two (or more) pretest measures to establish if there is a trend in the data independent of the treatment effect measured by the posttest.

**Regression-discontinuity design**. If there is a treatment effect, then the slope of the regression line relating scores before and after treatment would be the same, but there would be a discontinuous jump in the intercept (and possibly also change in slope) following treatment. This design is extended in the simple interrupted time series design in which there are multiple pretests and posttests. The trend found in multiple pretests can be compared to the trend found in multiple posttests to assess whether apparent post-treatment improvement may simply be an extrapolation of a maturation effect which was leading toward improvement anyway.

**Regression point displacement design**. In this design there is a treatment group (e.g., a county) and a large number of comparison groups (e.g., other counties in the state). For instance, in a study of the effect of an after-school intervention on juvenile crime, the researcher might regress juvenile crime rates on median income in the pretest condition and note the position of the test county in the regression scattergram. In the posttest condition, the regression is re-run and the location of the test county is noted. If displaced on the scattergram, the researcher concludes that the intervention had an effect. This type of design assumes no misspecification of the model and assumes an invariant relation of independents to dependents between pretest and posttest.

**Other designs**. Cook and Campbell (1979) discussed other research designs for which space does not permit discussion here. These include nonequivalent dependent variables pretest-posttest designs, removed-treatment pretest-posttest designs, repeated-treatment designs, switching replications designs, reversed-treatment pretest-posttest nonequivalent comparison groups designs,

cohort designs with cyclical turnover, four-group designs with pretest-posttest and posttest-only groups, interrupted time series designs with a nonequivalent no-treatment comparison group, interrupted time series designs with nonequivalent dependent variables, interrupted time series designs with removed treatment, interrupted time series designs with multiple replications, and interrupted time series designs with switching replications.

## About the Author

G. David Garson is a Professor of public administration at North Carolina State University, where he teaches graduate research methodology. He is Editor of the *Social Science Computer Review* and is author of *Statnotes: Topics in Multivariate Analysis* (http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm), an online text utilized by over a million scholars and researchers each year. His recently authored or edited books include *Handbook of Research on Public Information Technology* (2008; 3rd ed. 2010), *Patriotic Information Systems: Privacy, Access, and Security Issues of Bush Information Policy* (2008), *Modern Public Information Technology Systems* (2007), and *Public Information Technology and E-Governance: Managing the Virtual State* (2006; 2nd ed. 2010). He may be contacted at david_garson@ncsu.edu.

## Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Clinical Trials: An Overview
►Design of Experiments: A Pattern of Progress
►Designs for Generalized Linear Models
►Experimental Design: An Introduction
►Factorial Experiments
►Farmer Participatory Research Designs
►Graphical Analysis of Variance
►Incomplete Block Designs
►Interaction
►Medical Research, Statistics in
►Multilevel Analysis
►Optimum Experimental Design
►Randomization
►Repeated Measures
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistical Design of Experiments (DOE)
►Student's t-Tests
►Uniform Experimental Design

## References and Further Reading

Bordens K, Abbott BB (2008) Research design and methods: a process approach, 7th edn. McGraw-Hill, New York
Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis issues for field settings. Houghton-Mifflin, Boston
Creswell JW (2008) Research design: qualitative, quantitative, and mixed methods approaches, 3rd edn. Sage Publications, Thousand Oaks, CA
Leedy P, Ormrod JE (2009) Practical research: planning and design, 9th edn. Prentice-Hall, Upper Saddle River, NJ
Levin IP (1999) Relating statistics and experimental design. Sage Publications, Thousand Oaks, CA. Quantitative Applications in the Social Sciences series #125
Pedhazur EJ, Schmelkin LP (1991) Measurement, design, and analysis: an integrated approach. Lawrence Erlbaum Assoc, Mahwah, NJ
Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin, Boston

# Residuals

Samprit Chatterjee
Professor Emeritus of Statistics
Graduate School of Business Administration
Professor
New York University, New York, NY, USA

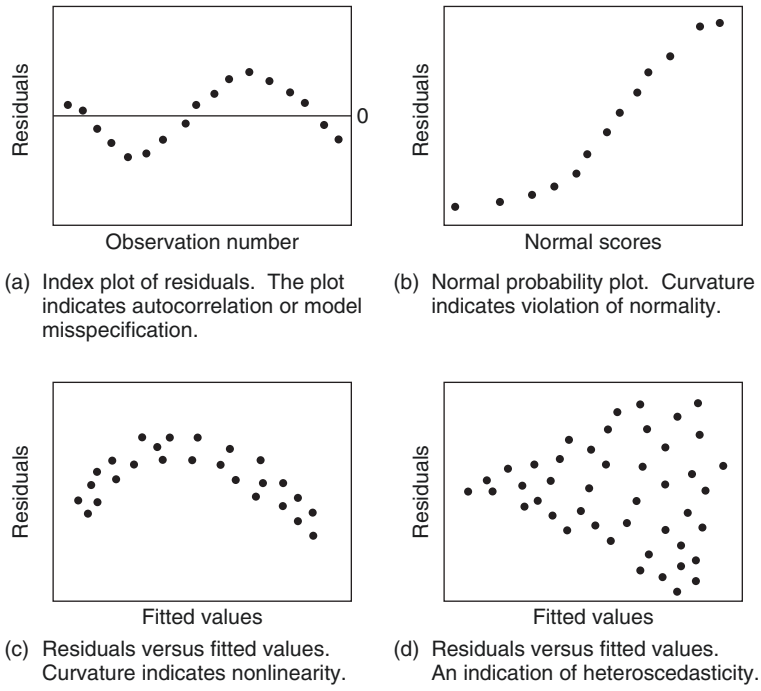
Residual is an important concept in statistical model building. Residual is defined as the difference between an observed value $(Y)$ and the value fitted by a statistical model $\hat{Y}$

$$\text{Residual}_i = Y_i - \hat{Y}_i$$

A large value of the residual (positive or negative) shows the model does not fit the particular data point. The pattern of the residuals will often reveal the inadequacy of the fitted model. A plot of the residuals, often called the residual plot, is an important tool in regression model building. In this article we will concentrate on the role of residuals in regression analysis.

With $n$ observations in a regression data set there will be $n$ residuals. The sum of the $n$ residuals from a least squares fit will be zero. Instead of working with the residuals as defined above we usually work with the standardized residuals. The residuals are scaled so they have unit standard deviation. We will call the standardized residuals for brevity residual. We will be talking about residuals obtained from least squares fit in our discussion.

The magnitude and the pattern of the distribution of residuals will reveal a great deal about the adequacy of the model describing the data. For moderate sized data the residuals can be thought of as normal deviates (mean 0, and Standard deviation 1). Large residuals, with values greater than 2 or 2.5, are called ►outliers. The data points




(a) Index plot of residuals. The plot indicates autocorrelation or model misspecification.

(b) Normal probability plot. Curvature indicates violation of normality.

(c) Residuals versus fitted values. Curvature indicates nonlinearity.

(d) Residuals versus fitted values. An indication of heteroscedasticity.

**Residuals. Fig. 1** Several configurations illustrating possible violations of model assumptions

corresponding to the outliers are not well fitted by the model. These points should be examined carefully, as they often represent transcription errors or contamination of the data. By contamination of data we mean an observation which comes from another population. As an example consider a data set of weekly production of a factory. Most weeks have 5 workdays, but there may be few weeks with only 4 workdays. The weeks with 4 workdays will not be well fitted by the model, and those weeks will be outliers.

The residual plot should show no structure. The distribution should appear random. Instead of trying to describe random structure (an impossible task!) we provide examples of some commonly observed structures of residual plots, and indicate the model deficiencies they indicate.

The four graphs depicted in Fig. 1 give some of the commonly observed pattern of residual plots which indicate model deficiencies. Plot (c) indicates the data has a nonlinear component which is not included in the specified mode. Inclusion of a squared term in the model will remove the structure from the residuals.

The graph (a) indicates that the successive values are correlated, a common feature of time series data. This pattern may also arise if the model is misspecified. Methods for removing auto regression have to be adopted. Working with successive differences is a good first step.

The graph (d) indicates that the error variance is not constant and increases with size. The data is often classified as heteroscedastic. This is often referred to as the size effect. To account for the size effect, sometimes we work with logarithms of the data, use weighted least squares, or introduce a variable which reflects size.

The graph (b) is the normal plot of the residuals and is used to assess the normality of the residuals. This plot is not very effective for small sample sizes.

The residual plots are one of the most effective diagnostic plots for model fitting. No regression analysis is complete without a residual plot analysis.

## About the Author

Samprit Chatterjee received his B.A. in 1958 from Calcutta University, and in 1967 he received a PhD from Harvard University. He was Research Assistant to Professor W.G. Cochran of Harvard University. He has been a visiting professor at Stanford, Sloan School of Management, Harvard School of Public Health, ETH (Zurich), University of Tampere, and University of Auckland (New Zealand). Since 1973, he has been Professor of Statistics, Graduate School of Business Administration, New York University. Professor Chatterjee has (co-)authored about 60 publications, including successful textbook *Regression*

*Analysis by Example* (with Ali Hadi, 4th edition, John Wiley & Sons, 2006).

## Cross References


►Absolute Penalty Estimation
►Bootstrap Methods
►Gauss-Markov Theorem
►Generalized Linear Models
►Graphical Analysis of Variance
►Heteroscedasticity
►Influential Observations
►Jarque-Bera Test
►Least Absolute Residuals Procedure
►Linear Regression Models
►Outliers
►Regression Diagnostics
►Simple Linear Regression
►Structural Time Series Models
►Time Series Regression
►Vector Autoregressive Models


## References and Further Reading


Chatterjee S, Hadi A (1988) Sensitivity analysis in linear regression. Wiley

Chatterjee S, Hadi A (2006) Regression analysis by example, 4th edn. Wiley


# Response Surface Methodology


André I. Khuri
Professor Emeritus
University of Florida, Gainesville, FL, USA


*Response surface methodology* (RSM) is an area of statistics that incorporates the use of design and analysis of experiments along with model fitting of a response of interest denoted by $y$. One of the main objectives of RSM is the determination of operating conditions on a group of control (or input) variables that yield optimal response values over a certain region of interest denoted by $\mathcal{R}$.

In a typical response surface (RS) investigation, several factors are first identified by the experimenter as having possible effects on the response $y$. In some experiments, the number of such factors may be large. In this case, factor screening is carried out in order to eliminate factors deemed to be unimportant. This represents the first stage in the RS investigation. The execution of this stage requires the use of an initial design which consists of a number of specified settings of the control variables. Each set of such settings is used to produce a value on the response $y$. A low-degree polynomial model, usually chosen to be of the first degree, is then fitted to the resulting data set. Following factor screening, additional experiments are carried out which lead to a new region of experimentation where the actual exploration of the response will take place. By this we mean fitting a suitable polynomial model of degree higher than the one used in the initial screening stage. Such a model can be expressed as

$$y = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$ is a vector of $k$ of control variables representing the levels of the factors that were retained after the initial screening, $\boldsymbol{f}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$ whose elements consist of powers and cross products of powers of $x_1, x_2, \ldots, x_k$ up to a certain degree denoted by $d$. Typically, $d = 2$ or higher depending on the adequacy of model (1). Furthermore, $\boldsymbol{\beta}$ is a vector of $p$ unknown parameters and $\epsilon$ is a random experimental error term assumed to have a zero mean. A commonly used form of model (1) is the second-degree model,

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum\sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \epsilon. \tag{2}$$

In this case, the elements of $\boldsymbol{\beta}$ consist of $\beta_0$, the $\beta_i$'s, $\beta_{ij}$'s, and $\beta_{ii}$'s $(i, j = 1, 2, \ldots, k,\ i < j)$. The quantity $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}$ in model (1) is called the mean response at $\boldsymbol{x}$ and is denoted by $\eta(\boldsymbol{x})$. Thus,

$$\eta(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}. \tag{3}$$

In order to estimate the parameter vector $\boldsymbol{\beta}$, a series of $n$ experiments is carried out in each of which the response $y$ is measured at specified settings of $x_1, x_2, \ldots, x_k$. Let $\boldsymbol{x}_u = (x_{u1}, x_{u2}, \ldots, x_{uk})'$, where $x_{ui}$ is the setting of $x_i$ at the $u$th experimental run, and let $y_u$ denote the corresponding response value $(i = 1, 2, \ldots, k,\ u = 1, 2, \ldots, n)$. From model (1) we then have

$$y_u = \boldsymbol{f}'(\boldsymbol{x}_u)\boldsymbol{\beta} + \epsilon_u,\ \ u = 1, 2, \ldots, n. \tag{4}$$

Model (4) can be expressed in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{X}$ is a matrix of order $n \times p$ whose $u$th row is $\boldsymbol{f}'(\boldsymbol{x}_u)$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. The first

column of $\boldsymbol{X}$ is $\mathbf{1}_n$, the column of $n$ ones. The $n \times k$ matrix,

$$\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \tag{6}$$

whose rows consist of the settings of $x_1, x_2, \dots, x_k$ used at the $n$ experimental runs is called the *design matrix*. If $\boldsymbol{\epsilon}$ is assumed to have a zero mean and a variance–covariance matrix $\sigma^2 \boldsymbol{I}_n$, where $\sigma^2$ is an unknown variance component and $\boldsymbol{I}_n$ is the matrix of ones of order $n \times n$, then $\boldsymbol{\beta}$ is estimated by the ordinary least-squares estimator,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{7}$$

The variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2. \tag{8}$$

Using formula (3), an estimate of the mean response, $\eta(\boldsymbol{x})$, at a point $\boldsymbol{x}$ in the region of interest, $\mathcal{R}$, is given by

$$\hat{\eta}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\beta}},$$

which is also known as the *predicted response* at $\boldsymbol{x}$ and is denoted by $\hat{y}(\boldsymbol{x})$. Thus,

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\beta}}. \tag{9}$$

The variance of $\hat{y}(\boldsymbol{x})$ is then of the form

$$\mathrm{Var}[\,\hat{y}(\boldsymbol{x})] = \sigma^2 \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{f}(\boldsymbol{x}). \tag{10}$$

This is called the *prediction variance*. The process of selecting the design matrix $\boldsymbol{D}$ and the subsequent fitting of model (1) represents the second stage of the RS investigation.

The third stage involves the determination of optimum operating conditions on the control variables, $x_1, x_2, \dots, x_k$, that yield either maximum or minimum values of $\hat{y}(\boldsymbol{x})$ over the region $\mathcal{R}$. This is a very important stage since it amounts to determining the settings of the control variables that should be used in order to obtain "best" values for the response. For example, if $y$ represents the yield of some chemical product, and if the corresponding control variables consist of $x_1$ = reaction temperature and $x_2$ = length of time of the reaction, then it would be of interest to determine the settings of $x_1$ and $x_2$ that result in a maximum yield.

The proper choice of the design matrix $\boldsymbol{D}$ given in formula (6) is very important. This is true because $\boldsymbol{D}$ is used to predict the response and determine its prediction variance (see formula (10)). The size of the latter quantity has to be small in order to get good quality predictions. This is particularly true since the optimization of $\hat{y}(\boldsymbol{x})$ in formula (9) leads to the determination of optimum operating conditions on $x_1, x_2, \dots, x_k$ in the third stage of a RS investigation.

Several criteria are available for the choice of the design $\boldsymbol{D}$. Some of these criteria pertain to the prediction variance, such as *D-optimality* and *G-optimality*. A review of such criteria can be found in several textbooks such as Khuri and Cornell (1996, Chap. 12), Atkinson and Donev (1992, Chap. 10) and Myers and Montgomery (1995, Sect. 8.2.1), among others. Other design criteria deal with the minimization of the bias caused by fitting the wrong model as explained in Box and Draper (1959, 1963).

If model (1) is of the first degree, that is,

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon, \tag{11}$$

then common designs for fitting the model are $2^k$ factorial, Plackett-Burman, and simplex designs. These are referred to as *first-order designs*. A coverage of such designs can be found in, for example, Khuri and Cornell (1996, Chap. 3) and Myers and Montgomery (1995, Chaps. 3, 4, and 7). On the other hand, if model (1) is of the second degree, as shown in formula (2), then common second-order designs include $3^k$ factorial, the central composite design, and the Box-Behnken design. A coverage of these designs can be found in, for example, Khuri and Cornell (1996, Chap. 4), which also includes reference to other lesser-known second-order designs (see also Myers and Montgomery 1995, Sect. 7.4).

Methods for the determination of optimum conditions on the control variables depend on the nature of the fitted model in (1). If it is of the second degree, as in model (2), then the method of ridge analysis can be used to optimize $\hat{y}(\boldsymbol{x})$ in (9). This method was introduced by Hoerl (1959) and later formalized by Draper (1963) (see also Khuri and Cornell 1996, Chap. 5). A more recent modification of this method that takes into account the size of the prediction variance was given by Khuri and Myers (1979).

Historically, the development of RSM was initiated by the work of Box and Wilson (1951) which introduced the sequential approach in a RS investigation. The article by Box and Hunter (1957) is considered to be a key paper, which along with the one by Box and Wilson (1951), provided an outline of the basic principles of RSM. Several review articles were also written about RSM. These include those by Mead and Pike (1975), Myers

et al. (1989, 2004), and Myers (1999). In addition, a comprehensive coverage of RSM can be found in the books by Khuri and Cornell (1996), Myers and Montgomery (1995), and Box and Draper (2007).

New developments and modeling trends were introduced into the RSM literature in the late 1970s. They provided further extensions of the classical techniques used in RSM. Some of these developments include the *analysis of multiresponse experiments*, which deals with several response variables that are measured for each setting of a group of control variables (see Khuri 1996a), the *response surface approach to robust parameter design* (see, for example, Myers et al. 1992), *response surface models with random effects* (see Khuri 1996b, 2006). Furthermore, in the design area, several graphical techniques were introduced for comparing response surface designs. These include the use of *variance dispersion graphs*, as in Giovannitti-Jensen and Myers (1989), the *quantile plots* of the prediction variance, as in Khuri et al. (1996), and the *fraction of design space plots* by Zahran et al. (2003). The main advantage of the graphical approach is its ability to explore the prediction capability of a response surface design throughout the region of interest, $\mathcal{R}$. By contrast, standard design optimality criteria, such as *D*- or *G-optimality*, use single-valued criteria functions to evaluate a given design. This does not give adequate information about the design's performance at various locations inside the region $\mathcal{R}$.

## About the Author

Dr. André I. Khuri is Professor Emeritus, Department of Statistics, University of Florida, Gainesville, Florida, USA. He is a Fellow of the American Statistical Association (since 1992) and an Elected Member of the International Statistical Institute (since 1989). He holds two PhDs, one in mathematics (University of Florida, 1969) and one in statistics (Virginia Tech, 1976). He has published more than 100 papers in statistics journals, in addition to 5 books, including *Response Surfaces* (1987, Dekker; 2nd edition, 1996) with John Cornell, *Statistical Tests for Mixed Linear Models* (1998, Wiley) with Thomas Mathew and Bimal Sinha, *Advanced Calculus with Applications in Statistics* (1993, Wiley; 2nd edition, 2003), *Response Surface Methodology and Related Topics* (2006, an edited book, World Scientific), and *Linear Model Methodology* (2009, Chapman & Hall/CRC). Professor Khuri was Associate Editor of *Technometrics* (1983–1992) and *Journal of Statistical Planning and Inference* (1995–2003). Currently, he is Editorial Advisor for *Journal of Probability and Statistical Science* (since 2003).

## Cross References

►Optimal Designs for Estimating Slopes
►Optimum Experimental Design
►Statistical Design of Experiments (DOE)

## References and Further Reading

Atkinson AC, Donev AN (1992) Optimun experimental designs. Oxford University Press, New York
Box GEP, Draper NR (1959) A basis for the selection of a response surface design. J Am Stat Assoc 54:622–654
Box GEP, Draper NR (1963) The choice of a second order rotatable design. Biometrika 50:335–352
Box GEP, Draper NR (2007) Response surfaces, mixtures, and ridge analyses, 2nd edn. Wiley, Hoboken
Box GEP, Hunter JS (1957) Multifactor experimental designs for exploring response surfaces. Ann Math Stat 28: 195–241
Box GEP, Wilson KB (1951) On the experimental attainment of optimum conditions (with discussion). J R Stat Soc B13: 1–45
Draper NR (1963) Ridge analysis of response surfaces. Technometrics 5:469–479
Giovannitti-Jensen A, Myers RH (1989) Graphical assessment of the prediction capability of response surface designs. Technometrics 31:159–171
Hoerl AE (1959) Optimum solution of many variables equations. Chem Eng Prog 55:69–78
Khuri AI (1996a) Multiresponse surface methodology. In: Ghosh S, Rao CR (eds) Handbook of statistics, vol 13. Elsevier Science B. V., Amsterdam, pp 377–406
Khuri AI (1996b) Response surface models with mixed effects. J Qual Technol 28:177–186
Khuri AI (2006) Mixed response surface models with heterogeneous within-block error variances. Technometrics 48: 206–218
Khuri AI, Cornell JA (1996) Response surfaces, 2nd edn. Dekker, New York
Khuri AI, Myers RH (1979) Modified ridge analysis. Technometrics 21:467–473
Khuri AI, Kim HJ, Um Y (1996) Quantile plots of the prediction variance for response surface designs. Comput Stat Data Anal 22:395–407
Mead R, Pike DJ (1975) A review of response surface methodology from a biometric viewpoint. Biometrics 31:803–851
Myers RH (1999) Response surface methodology – current status and future directions. J Qual Technol 31:30–44
Myers RH, Montgomery DC (1995) Response surface methodology. Wiley, New York
Myers RH, Khuri AI, Carter WH (1989) Response surface methodology: 1966–1988. Technometrics 31:137–157
Myers RH, Khuri AI, Vining GG (1992) Response surface alternatives to the Taguchi robust parameter design approach. Am Stat 46:131–139
Myers RH, Montgomery DC, Vining GG, Borror CM, Kowalski SM (2004) Response surface methodology: a retrospective and literature survey. J Qual Technol 36:53–77
Zahran A, Anderson-Cook CM, Myers RH (2003) Fraction of the design space to assess prediction capability of response surface designs. J Qual Technol 35:377–386

# Ridge and Surrogate Ridge Regressions

Ali S. Hadi
Professor and Vice Provost
The American University in Cairo, Cairo, Egypt
Emeritus Professor
Cornell University, Ithaca, NY, USA

Ridge regression is a method for the estimation of the parameters of a linear regression model (see ▶Linear Regression Models) which is useful when the predictor variables are highly *collinear*, that is, when there is a strong linear relationship among the predictor variables. Hoerl (1959) named the method *ridge regression* because of its similarity to ridge analysis used in his earlier work to study second-order response surfaces in many variables. Some standard references for ridge regression are Hoerl and Kennard (1970, 1976), Belsley et al. (1980), and Chatterjee and Hadi (2006).

The standard linear regression model can be written as

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

where $\mathbf{Y}$ is an $n \times 1$ vector of observations on the response variable, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is an $n \times p$ matrix of $n$ observations on $p$ predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. It is usual to assume that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}_n$, where $\sigma^2$ is unknown constant and $\mathbf{I}_n$ is the identity matrix of order $n$.

Without loss of generality, we also assume that $\mathbf{Y}$ and the columns of $\mathbf{X}$ are centered and scaled to have unit length so that $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ are matrices of correlation coefficients. If a variable $\mathbf{V}$ is not centered or scaled, it's $i$-th element, $v_i$, can be replaced by $(v_i - \bar{v})/\sqrt{\sum_{i=1}^n (v_i - \bar{v})^2}$.

An estimate for $\boldsymbol{\beta}$ is obtained by minimizing

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean (or $L_2$) norm. The minimization of this ordinary least squares (OLS) problem leads to the so-called system of normal equations,

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}. \tag{3}$$

Provided that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists, the solution of this system of linear equations is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4}$$

with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

If collinearity is present, the linear system in (3) is said to be *ill-conditioned* and $\hat{\boldsymbol{\beta}}$ in (4) can be unstable, that is, a slight change in the data can result in a substantial change in the values of the estimated regression coefficients. Furthermore, collinearity usually inflates the variance of $\hat{\boldsymbol{\beta}}$ and this in turn deflates the $t$-statistics for testing the significance of the regression coefficients, which can lead to the wrong conclusion that the coefficients of some important predictors are statistically insignificant.

A measure for assessing the condition of the linear system in (3) is the condition number of $\mathbf{X}$, which is defined as $\kappa = \sqrt{\lambda_1/\lambda_p}$, where $\lambda_1$ and $\lambda_p$ are the largest and smallest eigenvalues of $\mathbf{X}^T\mathbf{X}$, respectively. Large values of the condition number indicate ill-conditioned system. A measure for assessing the effect of collinearity on variance inflation is the *variance inflation factor* (VIF). For the $j$-th predictor variable $X_j$, the VIF is the $j$-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. It can be shown that $\mathrm{VIF}_j = 1/\left(1 - R_j^2\right)$, where $R_j^2$ is the multiple correlation coefficient when $X_j$ is regressed on all other predictor variables. When $X_j$ has a strong linear relationship with all other predictors, $R_j^2$ would be very large (close to 1), causing $\mathrm{VIF}_j$ to be very large. As a rule of thumb, values of variance inflation factors greater than 10 are indicative of the presence of collinearity.

To obtain a stable (*regularized*) solution, we replace the problem in (2) by minimizing

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2, \tag{5}$$

for some value of $k > 0$, suitably chosen by the user. The explicit solution of the problem in (5) is

$$\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}, \tag{6}$$

where $\mathbf{I}_p$ is the identity matrix of order $p$. The expected value and variance of $\hat{\boldsymbol{\beta}}(k)$ are

$$E(\hat{\boldsymbol{\beta}}(k)) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{7}$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + k\,\mathbf{I}_p)^{-1}. \tag{8}$$

In statistics, the solution $\hat{\boldsymbol{\beta}}(k)$ is known as *ridge regression* (see Hoerl 1962) and the ridge parameter $k$ is a penalizing factor. But more generally, it is known as the Tikhonov regularization (TR) method (Tikhonov 1943) and the factor $k$ is known as the Tikhonov factor.

By comparing (4) and (6), one can see that the ridge estimator is obtained by adding a small positive quantity $k$ to each of the diagonal elements of the matrix $\mathbf{X}^T\mathbf{X}$. Clearly, when $k = 0$, the ridge estimator in (6) becomes the OLS estimator in (4). It is clear from (7) that for $k > 0$, ridge estimators are biased for $\boldsymbol{\beta}$.

The variance inflation factors as a function of $k$ are the diagonal elements of the matrix $(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}$. Hoerl and Kennard (1970) show that there exists a value of $k > 0$ such that

$$E[(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta})] < E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})], \quad (9)$$

which means that the total mean square error of the ridge estimators are less than that of the OLS.

The choice of the ridge parameter $k$ is therefore important. The optimal value of $k$ is difficult to find, but there exists several alternative methods for estimating $k$. First, an appropriate value of $k$ can be found graphically by examining the *ridge trace*, which is a simultaneous plot of the elements of $\hat{\boldsymbol{\beta}}(k)$ versus $k$ (usually between 0 and 1). The smallest value of $k$, for which (a) the estimated vector of regression coefficients, $\hat{\boldsymbol{\beta}}(k)$, is stable, (b) the variance inflation factors are less than 10 (close to 1), and (c) the residual sum of squares is close to its minimum value, is chosen and used in (6) to obtain the ridge estimators.

Second, numerical methods for estimating $k$ have been proposed. For example, Hoerl et al. (1975) suggest estimating $k$ by $\hat{k} = p\hat{\sigma}^2/(\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}})$, where $\hat{\sigma}^2 = SSE/(n-p)$ and $SSE$ is the OLS residual sum of squares. Other numerical methods have also been suggested; see, for example, Lawless and Wang (1976), Wahba et al. (1979), Hoerl and Kennard (1981), Masuo (1988), Khalaf and Shukur (2005), and Dorugade and Kashid (2010).

Forms of ridge regression other than (6) are possible. For example the ridge parameter $k$ (which is a scalar) can be replaced by a diagonal matrix with possibly different diagonal elements, or even with a full $p \times p$ matrix, but these alternatives are less common in practice.

More recently, Jensen and Ramirez (2008) cast some doubt about the ability of ridge estimators to actually improve the condition of an ill-conditioned linear system and provide stable estimated regression coefficients and smaller variance inflation factors. Note that the ridge estimator in (6) is the solution of the linear system

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}, \quad (10)$$

which replaces the system of normal equations in (3). The condition number of the matrix $(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)$ on the left-hand side of (10) is $\sqrt{(\lambda_1 + k)/(\lambda_p + k)}$, which is smaller than the condition number of $(\mathbf{X}^T\mathbf{X})$, which is $\sqrt{\lambda_1/\lambda_p}$. Thus adding $k$ to each of the diagonal elements of $\mathbf{X}^T\mathbf{X}$ improves its condition. But the matrix $\mathbf{X}$ on the right-hand side of (10) remains ill-conditioned. To also improve the condition of the right-hand side of (10), Jensen and Ramirez (2008) propose replacing the ill-conditioned regression model in (1) by the *surrogate* but less ill-conditioned model

$$\mathbf{Y} = \mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (11)$$

where $\mathbf{X}_k = \mathbf{U}(\boldsymbol{\Lambda} + k\mathbf{I}_p)^{1/2}\mathbf{V}^T$, the matrices $\mathbf{U}$ and $\mathbf{V}$ are obtained from the *singular-value decomposition* of $\mathbf{X} = \mathbf{UDV}^T$ (see, e.g., Golub and van Loan 1989) with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, and $\mathbf{D}$ is a diagonal matrix containing the corresponding ordered singular values of $\mathbf{X}$. Note that the square of the singular values of $\mathbf{X}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$, that is, $\mathbf{D}^2 = \boldsymbol{\Lambda}$. Because $\mathbf{X}_k^T\mathbf{X}_k = \mathbf{X}^T\mathbf{X} + k\mathbf{I}_p$, the least squares estimator of the regression coefficients in (11) is the solution of the linear system

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}_k^T\mathbf{Y}, \quad (12)$$

which is given by

$$\hat{\boldsymbol{\beta}}_s(k) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}_k^T\mathbf{Y}. \quad (13)$$

Jensen and Ramirez (2008) study the properties of the surrogate ridge regression estimator, $\hat{\boldsymbol{\beta}}_s(k)$, in (13), and using a case study they demonstrate that the surrogate estimator is more conditioned than the classical ridge estimator, $\hat{\boldsymbol{\beta}}(k)$, in (6). For example, they observe that (a) the condition of the variance of $||\hat{\boldsymbol{\beta}}_s(k)||$ is monotonically increasing in $k$ and (b) the maximum variance inflation factor is monotonically decreasing in $k$. These properties do not hold for the classical ridge estimator, $\hat{\boldsymbol{\beta}}(k)$.

## About the Author

Professor Hadi is the Vice Provost and the Director of Graduate Studies and Research, the American University in Cairo (AUC). He is also a Stephen H. Weiss Presidential Fellow and Professor Emeritus, Cornell University, USA. He is the Founding Director of the Actuarial Science Program at AUC (2004–present). Dr. Hadi is the Editor-in-Chief, *International Statistical Review* (2009–present) and Co-founding Editor, *Journal of Economic and Social Research* (1998–present). He is an Elected Fellow of the American Statistical Association (1997) and Elected Member of the International Statistical Institute (1998). He has authored/coauthored nearly 100 articles in international refereed journals, and has published 5 books including, *Regression Analysis by Example* (with Samprit Chatterjee, Wiley, 4th edition, 2006).

## Cross References

►Absolute Penalty Estimation
►Linear Regression Models
►Multicollinearity
►Multivariate Statistical Analysis
►Partial Least Squares Regression Versus Other Methods
►Properties of Estimators

## References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Chatterjee S, Hadi AS (2006) Regression analysis by example, 4th edn. Wiley, New York

Dorugade AV, Kashid DN (2010) Alternative method for choosing ridge parameter for regression. Appl Math Sci 4:447–456

Golub GH, van Loan C (1989) Matrix computations. John Hopkins, Baltimore

Hoerl AE (1959) Optimum solution of many variables. Chem Eng Prog 55:69–78

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

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–68

Hoerl AE, Kennard RW (1976) Ridge regression: iterative estimation of the biasing parameter. Commun Stat, Theory Methods A5:77–88

Hoerl AE, Kennard RW (1981) Ridge regression – 1980: advances, algorithms, and applications. Am J Math Manag Sci 1:5–83

Hoerl AE, Kennard RW, Baldwin KF (1975) Ridge regression: some simulations. Commun Stat, Theory Methods 4:105–123

Jensen DR, Ramirez DE (2008) Anomalies in the foundations of ridge regression. Int Stat Rev 76:89–105

Khalaf G, Shukur G (2005) Choosing ridge parameter for regression problem. Commun Stat, Theory Methods 34:1177–1182

Lawless JF, Wang P (1976) A simulation study of ridge and other regression estimators. Commun Stat, Theory Methods 14:1589–1604

Masuo N (1988) On the almost unbiased ridge regression estimation. Commun Stat, Simul 17:729–743

Tikhonov AN (1943) On the stability problems. Dokl Akad Nauk SSSR 39:195–198

Wahba G, Golub GH, Health CG (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21:215–223

# Rise of Statistics in the Twenty First Century

Jon R. Kettenring
Drew University, Madison, NJ, USA

## Introduction

As the end of the first decade of the twenty first century approaches, it is fair to say that statistics as a profession is on the rebound in ways that matter. In recent years there have been recurring laments about missed opportunities and lack of respect for statistics and statisticians. Yet, statisticians are on the move, identifying and embracing new opportunities, and being increasingly expansive in defining the scope of their field and its relationship to the world at large.

A sign of progress is the growing recognition by outsiders of the importance of the statistics discipline. For example, a recent editorial in *Science* magazine (Long and Alpern 2009), based on a new report, "Scientific Foundations for Future Physicians," argues that "students should arrive at medical school prepared in the sciences, including some areas not currently required, such as statistics and biochemistry."

Even more recently, this dramatic headline appeared on the front page of *The New York Times*: "For today's graduate, just one word: statistics." The accompanying article (Lohr 2009) talked about "the rising stature of statisticians," "the new breed of statisticians" who analyze "vast troves of data," and how the "data explosion" is "open[ing] up new frontiers" for statistics.

The goal in this essay is to discuss a few disparate factors that are relevant to my thesis about the current rise of and strong future for statistics and statisticians.

## Renewal

One of the great qualities of the statistics profession, and an important reason for optimism about the future, is its tradition of introspection, self-assessment, and adjustment, leading to renewal of how we view our field, how we teach our subject, and how we interact with others. In part, this reflects the healthy questioning that statisticians engage in whenever confronted with a new problem. We want to know the whole story. We are politely but usefully skeptical of everything we are told. We examine all assumptions. We strive for quality data and supportable conclusions drawn from sound analyses. We know how to unlock underlying truths from complex and noisy circumstances. We recognize our limitations but also are able to develop new methods as needed.

## Education

►Statistics education has received considerable attention and improvement during the last 50 years. It is even an accepted area for research – and we've learned a lot about how to teach students more effectively. Statistics courses are now frequently found in high schools in the U.S. Over 100,000 high school students take the Advanced Placement Examination in Statistics annually. Introductory courses for college students are often taught by specialists who have devoted careers to perfecting ways of breaking novices gently and carefully into the wonders of statistical thinking. The wide availability of statistical software has allowed students to experience first hand what it is like to analyze real data.

Consulting courses have been a popular way to expose graduate students to current data analysis and modeling

problems. Going a step further, there are now success stories that involve the orchestrated merging of statistics education, consulting, and research. A well-developed example of this synergistic approach, from the University of California at Riverside, is described by Jeske et al. (2007). As a result of such efforts, students are entering the workforce with wider experiences and broader skill sets.

Another plus has been the evolution of postdoctoral programs in statistics. Once rarities, they are increasingly common in academia, government, industry and research institutes such as the National Institute of Statistical Sciences (NISS) in the U.S.A. In the past 18 years, NISS has engaged more than 60 postdoctoral students in cross-disciplinary research projects.

Still, the potential for *substantial* improvements remains, and a few specific ones are spelled out in Lindsay et al. (2004). The recent eye-catching proposal by Brown and Kass (2009), based on their experience working together in neuroscience, calls for strong reforms based on "deepening cross-disciplinary involvement" and "a broad vision of the discipline of statistics." The aggressive changes that they propose amount to a culture change (comment by Gibbs and Reid in the discussion) and won't come easily (comment by Johnstone) but are necessary to keep up with "big science" (comment by Nolan and Temple Lang).

## Cross-Disciplinary Research

Statistics has always been driven in part by applications. It is only in the last few decades that full-blown cross-disciplinary research involving statistics has become widely accepted within the profession. Now it is not only part of the culture but it is also spurring growth as statisticians respond to new data problems such as those posed by neuroscience. Other fruitful fields are easy to tick off as well, e.g., ▶bioinformatics, healthcare, life sciences, climate change, the environment, manufacturing, business strategy, privacy and confidentiality, bioterrorism, and national defense. Increasingly, the associated problems involve the analysis of massive datasets, i.e., ones of extraordinary size and complexity. When confronted with such challenges, teamwork is perhaps the most effective strategy and is very likely to trump purely statistical ones (Kettenring 2009).

A variety of other examples of cross-disciplinary work are listed in Lindsay et al. (2004), under the heading of statistics in science and industry, to illustrate the now common "interplay between statistics and other scientific disciplines." It is also worth noting that funding opportunities for such crossover activities have been on the rise.

## Credentials

Within the American Statistical Association (ASA) there have been strong debates for at least 15 years about the need for credentialing (as in accreditation based on experience or certification based on testing) of professional statisticians. In a perfect world there would be no need for more than a suitable academic degree, but ours is not so neat and tidy. Practitioners with no degrees in statistics but excellent records of accomplishment often have difficulty achieving the stature that they deserve or require for success in their careers. Similar problems are encountered by those who work in fringe areas or for small employers where the professionalism of statisticians is misunderstood or underappreciated. We also face the unfortunate companion situation of practitioners who claim competency in statistics but lack it. This can result in damaging malpractice.

Similar concerns were no doubt behind earlier movements by the Royal Statistical Society, the Statistical Society of Canada, and the Statistical Society of Australia Inc. to provide accreditation programs that help to differentiate practitioners who are good at statistics from those who only claim to be. Motivated in part by the apparent success of these ventures, the ASA has taken several steps along the same path, culminating in approval in August 2009 of a plan (Bock et al. 2009) to launch its own *voluntary* accreditation program for professional statisticians.

In a randomized survey of ASA members, 41% reported that they would apply to such a program were it offered because it would provide evidence of competency and a credential useful for employment, among other factors. This change in thinking from "why do we need such a thing" to "maybe the time is right" illustrates that the field is evolving not only in theoretical and philosophical ways but also in very pragmatic ones aimed at meeting the needs of practitioners operating in very competitive environments.

Time will tell, but if these credentialing programs fulfill their potential, they will have served a very useful purpose by helping to legitimize and support a broad class of professionals who are highly qualified to practice statistics.

## Journals

Journals are an essential component of the statistics infrastructure. They serve as a collective record of historical and current developments in the field. JSTOR, e.g., provides a very important centralized archive and point of access for more than 40 of the leading ones in statistics and probability. It includes journals such as *Biometrika*, the *Journal of the American Statistical Association (JASA)*, the *Journals of the Royal Statistical Society (JRSS)*, and *Technometrics*. Yet there are many more. The Current Index

to Statistics lists over 160 "core" journals on its website, www.statindex.org/CIS/news/CIS_core_journals.pdf. As of April 2009 the website, http://w4.stern.nyu.edu/ioms/research.cfm?doc_id=3532, hosted by the Stern School at New York University, included over 200 names under the heading of statistics and probability. Its breadth reflects the soft boundary view of statistics. Examples include *Analytical Chemistry, Econometrica*, the *Journal of Machine Learning Research*, and *Water Resources Research*.

Comparing journals across fields is dicey business, but it is tempting to see where one stands. In Lindsay et al. (2004) it was observed that *JASA* was "far and away the most cited mathematical science journal" for the period 1991–2001. I've also taken note of more recent data on the website www.eigenfactor.org, where 68 journals in probability and statistics are ranked based on cross-journal citation patterns, along with nearly 8,000 others. Citations of a journal to itself are not counted. The journals are quantified by their "Eigenfactor Score," which measures "the journal's total importance to the scientific community", and the "Article influence Score," which measures "the average influence, per article, of the papers in a journal" (Bergstrom et al. 2008). For the most recent year available, 2007, based on citations to the previous five years, the top four in the probability and statistics category are *JASA*, *Statistics in Medicine*, the *Annals of Statistics*, and *JRSS Series B* by the first measure, and *JRSS Series B*, *JASA*, the *Annals of Statistics*, and *Biostatistics* by the second. (The last title is a convenient reminder of the enormous success story of ►biostatistics as a subfield that has led the way in growth and career opportunities.)

The median Article Influence Score is 0.45 across all journals and 0.70 for those in the probability and statistics category. In comparison, the medians are 0.72 for neuroscience, 0.62 for mathematics, 0.60 for psychology, 0.58 for economics, and 0.59 for physics. The point is that statistics journals are publishing articles of relatively broad interest and high influence, at least by this measure. The vitality of our better journals provides a strong backbone for the future of statistics research, practice, and education.

## Holistic Statistics

In Kettenring (1997), under the heading of "holistic statistics," I asked whether statistics in the twenty first century should be equated so strongly to the more traditional core topics of statistics as it had been in the past and followed with these points:

- There is a natural tension between narrowly focused pursuits of science vs. broader ones that favor synthesis and interdisciplinarity.
- The core of statistics should be nourished by surrounding itself with vigorous areas of application.
- Broad-minded statisticians are needed to work across boundaries and operate in fast-paced environments.
- A more inclusive definition of statistics would better reflect its strong interdisciplinary character.
- Such an inclusive interpretation of statistics is where the future lies.

In similar spirit, Hand (2009) talks about the importance of "greater statistics" (Chambers 1993) as an overarching discipline that deals with (quoting Chambers) "everything related to learning from data." It is this expansive view of statistics that I intended in the title of this essay and what Hand has in mind when he talks about the "magic" of modern statistics.

## Wrap Up

Taking a bit of license with a popular adage, we can safely say that the future of statistics isn't what it used to be. These are meant to be encouraging words for students looking for a field of study that is full of life and much needed in a modern information age that is swamped with data and in need of help on what to do about it. Or, as the distinguished economist Hal Varian put it in Lohr (2009), "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."

## Acknowledgments

The author was President of the American Statistical Association in 1997. He thanks Daniel Jeske, Alan Karr, and David Morganstein for their suggestions.

## About the Author

Dr. Jon R. Kettenring is Past President of the American Statistical Association (1997). He has been Director of RISE since 2008. Previously, he was Executive Director of the Mathematical Sciences Research Center at Telcordia Technologies. Throughout his career, he has maintained a strong interest in statistics research and its application to solve real problems in the telecommunications industry. He is a Fellow of ASA and AAAS and an Elected Member of the International Statistical Institute. The National Institute of Statistical Sciences sponsored The Future of Data Analysis Conference in his honor at Avaya Labs Research, Basking Ridge, NJ (September 30, 2005).

## Cross References

►Careers in Statistics
►Online Statistics Education
►Role of Statistics

►Role of Statistics in Advancing Quantitative Education
►Role of Statistics: Developing Country Perspective
►Statistics Education
►Statistics: An Overview

## References and Further Reading

Bergstrom CT, West JD, Wiseman MA (2008) The eigenfactor metrics. J Neurosci 28:11433–11434
Bock ME, Hoerl R, Kettenring J, Kirkendall N, Mason R, Morganstein D, Nair V, O'Neill R, Oppenheimer L, Wasserstein R (2009) Report to the ASA board of directors by the individual accreditation proposal review group. www.amstat.org/news/pdfs/Kettenring_AccreditationReport.pdf
Brown EM, Kass RE (2009) What is statistics? Am Stat 63: 105–123
Chambers JM (1993) Greater or lesser statistics: a choice for future research. Stat Comput 3:182–184
Hand DJ (2009) Modern statistics: the myth and the magic (with discussion). J R Stat Soc A 172:287–306
Jeske DR, Lesch SM, Deng H (2007) The merging of statistics education, consulting and research: a case study. J Statist Educ 15. www.amstat.org/publications/jse/v16n3/jeske.html
Kettenring JR (1997) Shaping statistics for success in the 21st century. J Am Stat Assoc 92:1229–1234
Kettenring JR (2009) Massive datasets. WIRES Comput Stat 1:25–32
Lindsay BG, Kettenring JR, Siegmund DO (2004) A report on the future of statistics (with discussion). Stat Sci 19:387–412
Lohr S (2009) For today's graduate, just one word: statistics. The New York Times A1 and A3
Long S, Alpern R (2009) Science for future physicians. Science 324:1241

# Risk Analysis

Michael R. Greenberg
Professor
Rutgers University, New Brunswick, NJ, USA

## Introduction

Risk analysis originated in safety and systems engineering and following the events at the Three Mile Island nuclear facilities in the United States developed into an interdisciplinary approach to better understand and manage hazards. It has been applied to many human and ecological health issues, such as air borne spread of biological agents, destruction of the United States chemical weapons stockpile, cyber attacks, facility safety, food contamination, hazardous waste management, medical decision-making, nuclear power and waste management, and natural hazards such as earthquakes, floods, and tornadoes. Several of the ideas and models can be extended to economic, social, and even political risk.

Risk analysis is divided into risk assessment and risk management, although feedback loops exist among the stages. To provide continuity to this entry, the author uses the example of a terrorist planning to kill bus riders.

## Risk Assessment

Risk assessors try to answer three questions.

1. What can go wrong?
2. What are the chances that something with serious consequences will go wrong?
3. What are the consequences if something does go wrong?

This so-called "triplet" of questions (Kaplan and Garrick 1981; Garrick 1984) can be written as follows:

$$R = (S, P, C)$$

where $R$ is the risk; $S$ is a hypothesized risk scenario event of what can go wrong; $P$ is the probability of that scenario occurring; and $C$ are the consequences.

### Scenarios

Analysts create risk scenarios. For example, the terrorist could board a bus and detonate a bomb or leave a bomb near a stop and detonate it remotely. When there are thousands of potential triggering scenarios, analysts identify the worst consequences and then they work backward to scenarios that could produce them.

Analysts use fault trees or event trees to build out risk assessment scenarios. Event trees start with an event and follow it through branches. Some of the branches lead to insignificant consequences, while others end in serious outcomes. Fault tree analyses begin with the end state and work backwards to identify the event or sequence of events that will trigger it. They also are used together.

### Likelihood

Quantification of the ►likelihood of events was the major improvement introduced by risk assessment to safety analysis. Analysts develop a probability distribution of the likelihood of events, and then apply them to the trees. This step, the author believes, is the most difficult challenge faced by risk assessors (Hora 2007; Aven and Renn 2009; Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council 2007; Cox 2009; Dillon et al. 2009).

When there is no deliberate intent driving an event, for instance, a bus has a flat tire and crashes, then estimating probabilities from historical records and experiences make sense. The problem is how to estimate the likelihood

of a deliberate attack. Do we assume that the terrorists are intent as well as capable and will adjust to countermeasures? If a terrorist is assumed to achieve optimal or near optimal success, then we need to game theoretic or more generally agent-based modeling to support these estimates.

Eliciting likelihood estimates from experts is part science and part art. Expert opinion can be obtained through surveys. The results may not translate directly into an absolute measurement of likelihood, but perhaps an ordinal scale that can be used to set priorities. The key is training the experts to participate in the survey (Hora 2007; Aven and Renn 2009; Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council 2007).

### Consequences

Estimating consequences is part three of the risk assessment process. In the case of the bus passengers, the worst outcomes would be death of all the riders and for a transit system it would also be public fear of using the bus. When all the branches are followed, the results include estimates of the number of deaths, injuries, physical damage to assets, environmental effects, and local/regional economic impacts. These consequences are ranked with regard to severity.

Summarizing, risk assessments produce a list of risks. What to do about them is the responsibility of risk managers.

## Risk Management

The author offers the following three questions for risk managers:

1. How can consequences be prevented or reduced?
2. How can recovery be enhanced, if the scenario occurs?
3. How can key local officials, expert staff, and the public be informed to reduce concern and increase trust and confidence?

### Prevention

Risk managers try to contain risk within an acceptable (typically regulatory-based) level by implementing mitigation measures. These options are engineered and behavioral, and collectively they can be seen as the essence of a multi-criteria decision-making model (MCDM) (Chankong and Haimes 2008), where prevention options $O_i(i = 1, 2, \ldots, m)$ are defined and the decision-maker(s) evaluate them based on selected criteria $C_j(j = 1, 2, \ldots, n)$. In the case of our at risk bus passengers, the company can monitor the bus stops, and train the drivers can look for suspicious behavior (they are already trained to deal with rowdy and intoxicated riders).

Prevention is partly based on engineering options, life cycle cost, union contracts, ethical and other considerations, and requires blending of quantitative and qualitative data into a multi-criteria decision making framework. The audience for this volume recognizes the problems of scarcity of data, lack of knowledge, and subjectivity. Various mathematical and statistical techniques are available in the literature to deal with decision-making with uncertainties (Chankong and Haimes 2008; Heinz Center 2000; Skidmore and Toya 2002; Greenberg et al. 2007; Bedford and Cooke 2001; Edwards et al. 2007).

Two difficult challenges for risk managers are the time and space dimensions. In fact, we do not know a great deal about the geographical and temporal impacts of risk-related events (Zinn 2009). The direct consequences of a bus explosion event include impacts on the passengers and the driver. But indirect effects could include fear of using the bus, leading to a loss of revenue for the system and for businesses that depend on it, and to induced income effects caused by job losses. That is, when people lose their jobs, they begin to reduce their purchases.

The local impact is the area directly impacted by the event. Regional impacts occur in surrounding areas that are affected by direct losses. State, national and international impacts are felt as economic consequence ripples across the landscape. Some of these impacts are felt immediately or within a month or two of the event. Others are intermediate in length and measured in months and even a year or two out from the events. If the event is large enough there will also be long-term impacts that can be measured for many years. For example, the author has studied the impact of large scale loss of energy supply, leading to loss of confidence in the region and relocation out of the area. Yet we also have learned that a devastated region will receive funds from insurance companies, not-for-profits, and government agencies (Singpurwalla 2006), so negative consequences may be less than had been anticipated.

Economic impact tools allow us to estimate some of the consequences of such risk management decisions, albeit each of these has important data requirements, limitations, and capabilities (Modarres et al. 2010). The key to successfully using the economic models is the willingness of analysts to probe deeply into the events and through the stages that follow. Using sophisticated models without first understanding the event is a waste of time and money.

### Recovery

Even the best efforts to prevent risk events sometimes fail. Every risk manager needs a plan to respond to events. Like

the risk mitigation response, these options are both human and engineered. The response options $R_i (i = 1, 2, \ldots, m)$ are defined and the decision-maker(s) evaluate them based on selected criteria $C_j (j = 1, 2, \ldots, n)$. In the case of our bus passengers, for example, police would cordon off the area, ambulances would arrive, and the injured would be moved to a nearest health care facility that is able to treat those that are alive.

### Communications

At some major risk-related events, there are misunderstandings about who is in charge, who should perform what function, and sometimes the results are tragic. Firemen and police should be fully equipped and trained to deal with hazards. Transit workers and upper-level managers should know how to cope with passengers who show signs of suspicious behavior, and how to prevent panic rather than contribute to it. A good deal of research is focusing on crisis communications specifically and risk communications more generally, and principles have been articulated about how to manage risk events. However, it will take a systematic and ongoing effort to diffuse these suggestions to managers and to front-line employees.

## Summary

Risk analysis is a multidisciplinary field that includes researchers trained in physics, chemistry, biology, engineering, mathematics, economics, psychology, geography, sociology, communications, political science and others. This essay touches on the six key questions that risk analysts try to answer. Some good books are available (Bedford and Cooke 2001; Edwards et al. 2007; Zinn 2009; Singpurwalla 2006; Modarres et al. 2010). However, in a rapidly moving field like this, most books are out of date quickly. The author recommends consulting two journals: *Risk Analysis: an International Journal* and the *Journal of Risk Research*.

## Acknowledgments

I would like to thank colleagues of many years for helping me learn about risk analysis, especially Vicky Bier, Tony Cox, John Garrick, Bernard Goldstein, Ortwin Renn, Paul Slovic, and Yacov Haimes.

## About the Author

Dr. Michael R. Greenberg is Professor and Associate Dean of the Faculty of the Edward J. Bloustein School of Planning and Public Policy of Rutgers University. He is Director, National Center for Neighborhood and Brownfields, and Director, Rutgers National Transportation Security Center of Excellence. He has been a member of National Research Council Committees that focus on waste management, such as the destruction of the U.S. chemical weapons stockpile and nuclear weapons. Professor Greenberg has contributed more than 500 publications to scientific journals like *Cancer Research*, *American Journal of Epidemiology*, *Risk Analysis*, *American Journal of Public Health*, and public interest ones like *Urban Affairs Review*, *Housing Policy Debate*, *Society*, *the Sciences*, and *Public Interest*. He has received awards for research from the United States Environmental Protection Agency, the Society for Professional Journalists, the Public Health Association, the Association of American Geographers, and Society for Risk Analysis. In 2003, he received the Distinguished Career Achievement Award, International Society for Risk Analysis. He has supervised about 70 PhD students. Currently, Professor Greenberg is the Editor-in-Chief, *Risk Analysis: An International Journal*.

## Cross References

►Actuarial Methods
►Banking, Statistics in
►Bias Analysis
►Insurance, Statistics in
►Likelihood
►Quantitative Risk Management
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

## References and Further Reading

Aven T, Renn O (2009) The role of quantitative risk assessment for characterizing risk and uncertainty and delineating appropriate risk management options, special emphasis on terrorism risk. Risk Anal 29(4):587–599

Bedford T, Cooke R (2001) Probabilistic risk analysis: foundations and methods. Cambridge University Press, Cambridge, UK

Chankong V, Haimes Y (2008) Multiobjective decision making: theory and methodology. Dover, New York

Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council (2007) Interim report on methodological improvements to the Department of Homeland Security's biological agent risk analysis. National Academy, Washington, DC

Cox LA Jr (2009) Improving risk-based decision-making for terrorism applications. Risk Anal 29(3):336–341

Dillon R, Liebe R, Bestafka T (2009) Risk-based decision-making for terrorism applications. Risk Anal 29(3):321–335

Edwards W, Miles R Jr, von Winterfeldt D (2007) Advances in Decision Analysis. Cambridge University Press, Cambridge, UK

Garrick BJ (1984) Recent case studies and advances in probabilistic risk assessments. Risk Anal 4:262–279

Greenberg M, Lahr M, Mantell N, Felder N (2007) Understanding the economic costs and benefits of catastrophes and their aftermath: a review and suggestions for the as-federal government. Risk Anal 27(1):83–96

Heinz Center for Science, Economics and the Environment (2000) The hidden costs of coastal hazards: implications for risk assessment and mitigation. Island Press, Washington, DC
Hora S (2007) Eliciting probabilities from experts. In: Edwards W, Miles R, von Winterfeldt D (eds) Advances in decision analysis. Cambridge University Press, Cambridge, UK, pp 129–153
Kaplan S, Garrick BJ (1981) On the quantitative definition of risk. Risk Anal 1(1):11–27
Modarres M, Kaminskiy M, Krivtsov V (2010) Reliability Engineering and Risk Analysis. Taylor & Francis Group, Boca Raton, Florida, FL
Singpurwalla N (2006) Reliability and risk. Wiley, New Jersey
Skidmore M, Toya H (2002) Do natural disasters promote long-run growth? Economic Inquiry 40:664–687
Zinn J (2009) Social Theories of Risk and Uncertainty: An Introduction. Blackwell, Oxford, UK

# Robust Inference

Elvezio Ronchetti
Professor
University of Geneva, Geneva, Switzerland

►Robust statistics deals with deviations from ideal parametric models and their dangers for the statistical procedures derived under the assumed model. Its primary goal is the development of procedures which are still reliable and reasonably efficient under small deviations from the model, i.e., when the underlying distribution lies in a neighborhood of the assumed model. Robust statistics is then an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. The field is now some 50 years old. Indeed one can consider Tukey (1960), Huber (1964), and Hampel (1968) the fundamental papers which laid the foundations of modern robust statistics. Book-length expositions can be found in Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel et al. (1986), Maronna et al. (2006).

More specifically, in robust testing one would like the level of a test to be stable under small, arbitrary departures from the distribution at the null hypothesis (*robustness of validity*). Moreover, the test should still have good power under small arbitrary departures from specified alternatives (*robustness of efficiency*). For confidence intervals, these criteria correspond to stable coverage probability and length of the confidence interval.

Many classical tests do not satisfy these criteria. An extreme case of nonrobustness is the F-test for comparing two variances. Box (1953) showed that the level of this test becomes large in the presence of tiny deviations from the normality assumption (see Hampel et al. 1986; 188–189). Well known classical tests exhibit robustness problems too. The classical t-test and F-test for linear models are relatively robust with respect to the level, but they lack robustness of efficiency with respect to small departures from the normality assumption on the errors (cf. Hampel 1973; Schrader and Hettmansperger 1980; Ronchetti 1982; Heritier et al. 2009: 35). Nonparametric tests are attractive since they have an exact level under symmetric distributions and good robustness of efficiency. However, the distribution free property of their level is affected by asymmetric contamination (cf. Hampel et al. 1986: 201). Even ►randomization tests which keep an exact level, are not robust with respect to the power if they are based on a non-robust test statistic like the mean.

The first approach to formalize the robustness problem was Huber's (1964, 1981) minimax theory, where the statistical problem is viewed as a game between the Nature (which chooses a distribution in the neighborhood of the model) and the statistician (who chooses a statistical procedure in a given class). The statistician achieves robustness by constructing a minimax procedure which minimizes a loss criterion at the worst possible distribution in the neighborhood. More specifically, in the problem of testing a simple hypothesis against a simple alternative, Huber (1965, 1981) found the test which maximizes the minimum power over a neighborhood of the alternative, under the side condition that the maximum level over a neighborhood of the hypothesis is bounded. The solution to this problem which is an extension of ►Neyman-Pearson Lemma, is the censored likelihood ratio test. It can be interpreted in the framework of capacities (Huber and Strassen 1973) and it leads to exact finite sample minimax confidence intervals for a location parameter (Huber 1968). While Huber's minimax theory is one of the key ideas in robust statistics and leads to elegant and exact finite sample results, it seems difficult to extend it to general parametric models, when no invariance structure is available.

The infinitesimal approach introduced in Hampel (1968) in the framework of estimation, offers an alternative for more complex models. The idea is to view the quantities of interest (for instance the bias or the variance of an estimator) as functionals of the underlying distribution and to use their linear approximations to study their behavior in a neighborhood of the ideal model. A key tool is a derivative of such a functional, the influence function (Hampel 1974) which describes the local stability of the functional.

To illustrate the idea in the framework of testing, consider a parametric model $\{F_\theta\}$, where $\theta$ is a real parameter and a test statistic $T_n$ which can be written (at least asymptotically) as a functional $T(F_n)$ of the empirical distribution function $F_n$. Let $H_0 : \theta = \theta_0$ be the null hypothesis and $\theta_n = \theta_0 + \Delta/\sqrt{n}$ a sequence of alternatives. We consider a neighborhood of distributions $F_{\epsilon,\theta,n} = (1-\epsilon/\sqrt{n})F_\theta + (\epsilon/\sqrt{n})G$, where $G$ is an arbitrary distribution and we can view the asymptotic level $\alpha$ of the test as a functional of a distribution in the neighborhood. Then by a von Mises expansion of $\alpha$ around $F_{\theta_0}$, where $\alpha(F_{\theta_0}) = \alpha_0$, the nominal level of the test, the asymptotic level and (similarly) the asymptotic power under contamination can be expressed as

$$\lim_{n\to\infty} \alpha(F_{\epsilon,\theta_0,n}) = \alpha_0 + \epsilon \int IF(x;\alpha,F_{\theta_0})dG(x) + o(\epsilon), \tag{1}$$

$$\lim_{n\to\infty} \beta(F_{\epsilon,\theta_n,n}) = \beta_0 + \epsilon \int IF(x;\beta,F_{\theta_0})dG(x) + o(\epsilon), \tag{2}$$

where

$$IF(x;\alpha,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0))IF(x;T,F_{\theta_0}) \Big/ [V(F_{\theta_0},T)]^{1/2},$$

$$IF(x;\beta,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})IF(x;T,F_{\theta_0}) \Big/ [V(F_{\theta_0},T)]^{1/2},$$

$\alpha_0 = \alpha(F_{\theta_0})$ is the nominal asymptotic level, $\beta_0 = 1 - \Phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})$ is the nominal asymptotic power, $E = [\xi'(\theta_0)]^2 \Big/ V(F_{\theta_0},T)$ is Pitman's efficacy of the test, $\xi(\theta) = T(F_\theta)$, $V(F_{\theta_0},T) = \int IF(x;T,F_{\theta_0})^2 dF_{\theta_0}(x)$ is the asymptotic variance of T, and $\Phi^{-1}(1-\alpha_0)$ is the $1-\alpha_0$ quantile of the standard normal distribution $\Phi$ and $\phi$ is its density (see Ronchetti 1979; Rousseeuw and Ronchetti 1979). More details can be found in Markatou and Ronchetti (1997) and Huber and Ronchetti (2009, Chap. 13).

Therefore, bounding the influence function of the the test statistic $T$ from *above* will ensure *robustness of validity* and bounding it from *below* will ensure *robustness of efficiency*. This is in agreement with the exact finite sample result about the structure of the censored likelihood ratio test obtained using the minimax approach.

In the multivariate case and for general parametric models, the classical theory provides three asymptotically equivalent tests, Wald, score, and likelihood ratio test, which are asymptotically uniformly most powerful with respect to a sequence of contiguous alternatives. If the parameter of the model is estimated by a robust estimator such as an $M$-estimator $T_n$ defined by the estimating equation $\sum_{i=1}^n \psi(x_i;T_n) = 0$, natural extensions of the three classical tests can be constructed by replacing the score function of the model by the function $\psi$. This leads to formulas similar to (1) and (2) and to optimal bounded influence tests (see Heritier and Ronchetti 1994).

## About the Author

Professor Elvezio Ronchetti is Past Vice-President of the Swiss Statistical Association (1988–1991). He was Chair, Department of Econometrics, University of Geneva (2001–2007). He is an Elected Fellow of the American Statistical Association (2001) and of the International Statistical Institute (2008). Currently he is an Associate Editor, *Journal of the American Statistical Association* (2005–present) and Director of the Master of Science and PhD Program in Statistics, University of Geneva (2009–present). He is the co-author (with F.R. Hampel, P.J. Rousseeuw, and W.A. Stahel) of the well known text *Robust Statistics : The Approach Based on Influence Functions* (Wiley, New York, 1986, translated also into Russian), and of the 2nd edition of Huber's classic *Robust Statistics* (with P. J. Huber, Wiley 2009.)

## Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Confidence Interval
►Multivariate Technique: Robustness
►Neyman-Pearson Lemma
►Nonparametric Statistical Inference
►Power Analysis
►Randomization Tests
►Robust Regression Estimation in Generalized Linear Models
►Robust Statistical Methods
►Robust Statistics
►Statistical Inference
►Statistical Inference: An Overview
►Student's t-Tests
►Tests for Homogeneity of Variance

## References and Further Reading

Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335
Hampel FR (1968) Contribution to the theory of robust estimation. PhD Thesis, University of California, Berkeley
Hampel FR (1973) Robust estimation: a condensed partial survey. Z Wahrsch Verwandte Geb 27:87–104
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Heritier S, Ronchetti E (1994) Robust bounded-influence tests in general parametric models. J Am Stat Assoc 89:897–904
Heritier S, Cantoni E, Copt S, Victoria-Feser M-P (2009) Robust methods in biostatistics. Wiley, Chichester
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
Huber PJ (1965) A robust version of the probability ratio test. Ann Math Stat 36:1753–1758
Huber PJ (1968) Robust confidence limits. Z Wahrsch Verwandte Geb 10:269–278
Huber PJ (1981) Robust statistics. Wiley, New York
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, New York
Huber PJ, Strassen V (1973) Minimax tests and the Neyman-Pearson lemma for capacities. Ann Stat 1:251–263, 2:223–224
Markatou M, Ronchetti E (1997) Robust inference: the approach based on influence functions. In: Maddala GS, Rao CR (eds) Handbook of Statistics, vol 15. North Holland, Amsterdam, pp 49–75
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York
Ronchetti E (1979) Robustheitseigenschaften von Tests. Diploma Thesis, ETH Zürich, Switzerland
Ronchetti E (1982) Robust testing in linear models: The infinitesimal approach. PhD Thesis, ETH Zürich, Switzerland
Rousseeuw PJ, Ronchetti E (1979) The influence curve for tests. Research report 21. Fachgruppe für Statistik, ETH Zürich, Switzerland
Schrader RM, Hettmansperger TP (1980) Robust analysis of variance based upon a likelihood ratio criterion. Biometrika 67:93–101
Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I (ed) Contributions to probability and statistics. Stanford University Press, Stanford, pp 448–485

# Robust Regression Estimation in Generalized Linear Models

Nor Aishah Hamzah[1], Mohammed Nasser[2]
[1]Professor
University of Malaya, Kuala Lumpur, Malaysia
[2]Professor
University of Rajshahi, Rajshahi, Bangladesh

The idea of ▶generalized linear models (GLM) generated by Nelder and Wedderburn (1972) seeks to extend the domain of applicability of the linear model by relaxing the normality assumption. In particular, GLM can be used to model the relationship between the explanatory variable, $X$, and a function of the mean, $\mu_i$, of a continuous or discrete responses. More precisely, GLM assumes that $g(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the $p$-vector of unknown parameters and $g(\cdot)$ is the link function that determines the scale on which linearity is assumed. Models of this type include logistic and probit regression, Poisson regression, linear regression with known variance, and certain models for lifetime data.

Specifically, let $Y_1, Y_2, \ldots, Y_n$, be $n$ independent random variables drawn from the exponential family with density (or probability function)

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \qquad (1)$$

for some specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$. Here, $E(Y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$ with usual notation of derivative.

The most common method of estimating the unknown parameter, $\beta$, is that of maximum likelihood estimation (MLE) or quasi-likelihood methods (QMLE), which are equivalent if $g(\cdot)$ is the canonical link such as the logit function for the ▶logistic regression, the log function for ▶Poisson regression, or the identity function for the Normal regression. That is, when $g(\mu_i) = \theta_i$, the MLE and QMLE estimator of $\beta$ are the solutions of the $p$-system of equations:

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, \ldots, p. \qquad (2)$$

The estimator defined by (2) can be viewed as an $M$-estimator with score function

$$\psi(y_i; \beta) = (y_i - \mu_i)x_i \qquad (3)$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Since the score function defined by (3) is proportional to $x$ and $y$, the maximum possible influence in both the $x$ and $y$ spaces are unbounded. When $y$ is categorical, the problem of unbounded influence in $x$ remains and in addition, the breakdown possibility by inliers arises (Albert and Anderson 1984). As such, the corresponding estimator of $\beta$ based on (2) is therefore non-robust. Any attempt to improve the estimation of such $\beta$ should limit such influences. Two basic approaches are usually employed in order to address the problems stated above, that is: (a) diagnostics and (b) robust estimation.

## Diagnostic Measures

In most diagnostics approaches, the MLE is first employed and subsequently diagnostics tools are used to identify potential influential observations. For details on diagnostic measures, readers are referred to the published works of Pregibon (1981, 1982), McCullagh and Nelder (1989), Johnson (1985), Williams (1987), Pierce and Schafer (1986), Thomas and Cook (1990), and Adimari and Ventura (2001).

While these techniques have been quite successful in identifying individual influential points, its generalization to jointly influential points cannot guarantee success. The development of a robust method in the early 1980s provides an option that offers automatic protection against anomalous data. A recent trend in diagnostic research is (a) to detect wild observations by using the classical diagnostic method after initially deploying the robust method (Imon and Hadi 2008) or (b) to use robust method in any case (Cantoni and Ronchetti 2001; Serigne and Ronchetti 2009).

## Robust Estimation

Since the score function in (3) is subject to influence of outlying observation, both in the $X$ and $y$, appropriate robust estimations are those of the GM-estimates. These include the Mallows-type (Pregibon 1979) and Schweppe-type (Stefanski et al. 1986; Künsch et al. 1989). The proposed methods are discussed here. Let

$$\ell(\theta_i, y_i) = \log f(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \quad (4)$$

and define the $i$-th deviance as $d_i = d_i(\theta_i) = 2\{\ell(\tilde{\theta}_i, y_i) - \ell(\theta_i, y_i)\}$, where $\tilde{\theta}_i$ is the MLE based on observation $y_i$ alone, that is, $\tilde{\theta}_i = (b')^{-1}(y_i)$. The deviance $d_i$ can be interpreted as a measure of disagreement of the $i$-th observation and the fitted model. Thus, MLE that aims at maximizing the likelihood function also aims at minimizing the deviances, specifically minimizing $M(\beta) = \sum_{i=1}^{n} d_i(\theta)$.

In an attempt to robustify the MLE, the first modification of the MLE introduced by Pregibon (1979) is to replace the minimization criterion with $M(\beta) = \sum_{i=1}^{n} \rho(d_i)$.

The function $\rho(\cdot)$ acts as a filter that limits the contribution of extreme observations in determining the fits to the data. Minimizing the criterion above can be obtained by finding the root solutions to the following score function

$$\sum_{i=1}^{n} \psi(d_i) = \sum_{i=1}^{n} w_i s_i x_{ij} = 0, \quad j = 1, \ldots, p, \quad (5)$$

with $s_i = \partial\ell(\theta_i, y_i)/\partial\eta_i$, and $w_i (0 \le w_i \le 1)$ given by $w(d_i) = \partial\rho(d_i)/\partial d_i$. Note that this is simply the weighted version of the maximum likelihood score equations with data-dependent weights.

## Mallows-Type GM Estimate

Based on Huber's loss function, the corresponding weight function $w_i = \min\{1, (H/d_i)^{1/2}\}$ with adjustable tuning constant $H$, which aims at achieving some specified efficiency, can be used (Pregibon 1982). By solving (5), one can obtain a class of Mallows M-estimates. This type of estimation is resistant to poorly fitted data, but not to extreme observations in the covariate space that may exert undue influence on the fit.

## Schweppe-Type GM Estimate

Extending the results obtained by Krasker and Welsch (1982) and Stefanski et al. (1986), Künsch et al. (1989) proposed bounded influence estimators that are also conditionally Fisher-consistent. Subject to a bound $b$ on the measure of sensitivity $\gamma_\psi (\gamma_\psi \le b < \infty)$, the following modification to the score function was proposed:

$$\psi_{BI} = \left\{ y - \mu - c\left( x^T\beta, \frac{b}{(x^T B^{-1} x)^{1/2}} \right) \right\} w_b(|r(y, x, \beta, B)|(x^T B^{-1} x)^{1/2}) x^T$$

where $c(\cdot, \cdot)$ and $B$ are the respective bias-correction term and dispersion matrix chosen so that the estimates are conditionally Fisher-consistent with bounded influence, with weight function of the form $w_b(a) = \min\{1, b/a\}$ based on Huber's loss function. As in Schweppe-type GM estimates, $w_b(\cdot)$ downweight observations with a high product of corrected residuals and leverage. Details on the terms used here can be found elsewhere (see, e.g., Huber (1981) on infinitesimal sensitivity).

Besides the general approach in robust estimation in GLM several researchers put forward various other estimators for specific case of GLM. For example, when $y$ follows a Gamma distribution with log link function, Bianco et al. (2005) considered redescending M-estimators and showed that the estimators are Fisher-consistent without any correction term. In the logistic model, Carrol and Pederson (1993) proposed weighted MLE to robustify estimators, Bianco and Yohai (1996) extended the work of Morgenthaler (1992) and Pregibon (1982) on M-estimators while Croux and Haesbroeck (2003) developed a fast algorithm to execute Bianco–Yohai estimators. Gervini (2005) presented robust adaptive estimators and recently Hobza et al. (2008) opened a new line proposing robust median estimators in ▶logistic regression (see also Hamzah 1995). The robust Poisson regression model (RPR) (see ▶Poisson Regression) was proposed by Tsou (2006) for the inference about regression parameters for more general count data; here one need not worry about the correctness of the Poisson assumption.

## Cross References

▶Generalized Linear Models
▶Influential Observations
▶Outliers

►Regression Diagnostics
►Robust Statistics

## References and Further Reading


Adimari G, Ventura L (2001) Robust inference for generalized linear models with application to logistic regression. Stat Probab Lett 55:413–419

Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10

Bianco A, Yohai V (1996) Robust estimation in the logistic regression model. In: Rieder H (ed) Robust statistics, data analysis, and computer intensive methods. Lecture notes in statistics, vol 109, Springer, New York, pp 17–34

Bianco AM, Garcia Ben M, Yohai VJ (2005) Robust estimation for linear regression with asymmetric error. Can J Stat 33: 511–528

Carrol RJ, Pederson S (1993) On robustness in logistic regression model. J Roy Stat Soc B 55:693–706

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Stat Assoc 96:1022–1030

Croux C, Haesbroeck G (2003) Implementing the Bianco and Yohai estimator for logistic regression. Comput Stat Data Anal 44:273–295

Gervini D (2005) Robust adaptive estimators for binary regression models. J Stat Plan Infer 131:297–311

Hobza T, Pardo L, Vajda I (2008) Robust median estimator in logistic regression. J Stat Plan Infer 138:3822–3840

Hamzah NA (1995) Robust regression estimation in generalized linear models, University of Bristol, Ph.D. thesis

Huber PJ (1981) Robust Statistics. Wiley, New York

Imon AHMR, Hadi AS (2008) Identification of multiple outliers in logistic regression. Commun Stat Theory Meth 37(11): 1697–1709

Johnson W (1985) Influence measures for logistic regression: Another point of view. Biometrika 72:59–65

Krasker WS, Welsch RE (1982) Efficient bounded-influence regression estimation. J Am Stat Assoc 77:595–604

Künsch H, Stefanski L, Carroll RJ (1989) Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. J Am Stat Assoc 84:460–466

McCullagh P, Nelder JA (1989) Generalized Linear Models. Chapman and Hall, London

Morgenthaler S (1992) Least-absolute-deviations fits for generalized linear models. Biometrika 79:747–754

Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. J Roy Stat Soc A 135:370–384

Pierce DA, Schafer DW (1986) Residual in generalized linear model. J Am Stat Assoc 81:977–990

Pregibon D (1979) Data analytic methods for generalized linear models. University of Toronto, Ph. D. thesis

Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9:705–724

Pregibon D (1982) Resistant fits for some commonly used logistic models with medical applications. Biometrics 38: 485–498

Serigne NL, Ronchetti E (2009) Robust and accurate inference for generalized linear models. J Multivariate Anal 100:2126–2136

Stefanski L, Carroll RJ, Ruppert D (1986) Optimally bounded score functions for generalized linear models, with applications to logistic regression. Biometrika 73:413–425

Thomas W, Cook RD (1990) Assessing influence on predictions from generalized linear models. Technometrics 32:59–65

Tsou T-S, Poisson R (2006) regression. Journal of Statistical Planning and Inference 136:3173–3186

Williams DA (1987) Generalized linear model diagnostics using the deviance and single case deletions. Appl Stat 36:181–191


# Robust Statistical Methods


Ricardo Maronna
Professor
University of La Plata and C.I.C.P.B.A., La Plata, Buenos Aires, Argentina


## Outliers

The following Table (Hand et al. 1994: 278) contains 20 measurements of the speed of light in suitable units (km/s minus 299000) from the classical experiments performed by Michelson and Morley in 1887.

| | | | | |
|---|---|---|---|---|
| 880 | 880 | 880 | 860 | 720 |
| 720 | 620 | 860 | 970 | 950 |
| 880 | 910 | 850 | 870 | 840 |
| 840 | 850 | 840 | 840 | 840 |

We may represent our data as

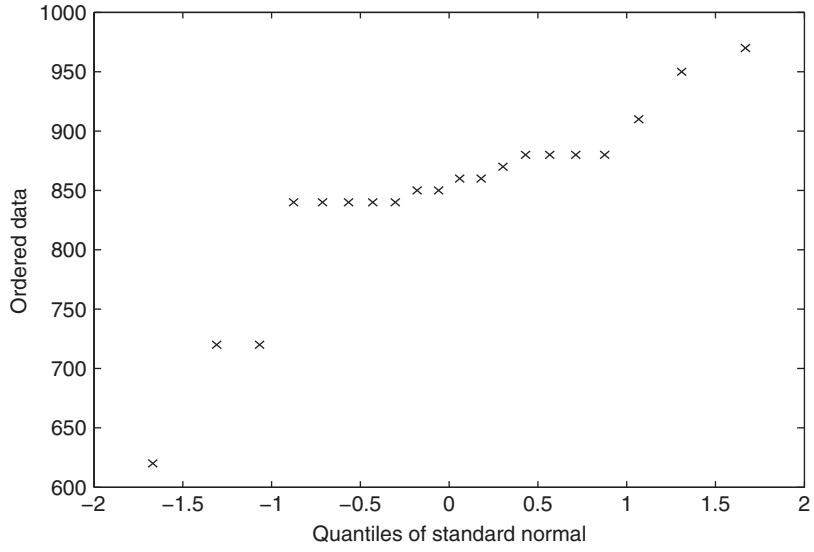
$$x_i = \mu + u_i,\ i = 1, \ldots, n \tag{1}$$

where $n = 20$, $\mu$ is the true (unknown) speed value and $u_i$ are random observation errors. We want a point estimate $\widehat{\mu}$ and a ►confidence interval for $\mu$.

Figure 1 is the normal QQ-plot of the data. The three smallest observations clearly stand out from the rest. The central part of the plot is approximately linear, and therefore we may say that the data are "approximately normal."

The left-hand half of the following Table shows the sample mean and standard deviation (SD) of the complete data and also of the data without the three smallest observations (the right-hand half will be described below).

| | Mean | SD | Median | MADN |
|---|---|---|---|---|
| Complete data | 845.0 | 79.1 | 855.0 | 29.6 |
| 3 obs. omitted | 872.9 | 38.5 | 860.0 | 29.6 |




**Robust Statistical Methods. Fig. 1** Speed of light: normal QQ-plot of data

We see that these three observations inflate the SD and diminish the mean.

The confidence intervals with level 0.95 for the mean with the complete data and with the three outliers removed are respectively [807.94, 882.06] and [853.14 892.74].

We see that even data from a carefully controlled experiment may contain atypical observations ("outliers") which may overly influence the conclusions from the experiment. Although the proportion of outliers is low (3/20=15%) they have a serious influence.

The oldest approach to deal with this problem is to employ some diagnostic tool to detect ▶outliers, delete them, and then recompute the statistics of interest. Barnett and Lewis (1998) is a useful source of methods for outlier detection.

Using a good outlier diagnostic is clearly better than doing nothing, but has its drawbacks:

- Deletion requires a subjective decision. When is an observation "outlying enough" to be deleted?
- The user or the author of the data may feel uneasy about deleting observations
- There is a risk of deleting "good" observations, which results in underestimating data variability
- Since the results depend on the user's subjective decisions, it is difficult to determine the statistical behavior of the complete procedure.

Robust statistical methods are procedures that require no subjective decisions from the user, and that

- give approximately the same results as classical methods when there are no atypical observations, and
- are only slightly affected by a small or moderate proportion of atypical observations.

The sample median $\mathrm{Med}(\mathbf{x})$ is a robust alternative to the median. The median absolute deviation from the median $\mathrm{MAD}(\mathbf{x}) = \mathrm{Med}(|\mathbf{x}\text{-}\mathrm{Med}(\mathbf{x})|)$ is a robust dispersion estimate. The normalized MAD: $\mathrm{MADN}(\mathbf{x}) = \mathrm{MAD}(\mathbf{x})/0.675$ is a robust alternative to the SD; for large normal samples MADN and SD are approximately equal. The right-hand half of the Table above shows the sample median and MADN for the complete data and the data with the three smallest observations omitted. We see that the median has only a small change, and that MADN remains the same.

Then, why not always use the median instead of the mean? To answer this question we have to analyze the behavior of the estimates at a given model. Assume that $u_i$ are normal: $\mathrm{N}(0, \sigma^2)$. Then the sample mean has variance $\mathrm{Var}(\bar{x}) = \sigma^2/n$, while for large $n$ the sample median has $\mathrm{Var}(\mathrm{Med}(\mathbf{x})) \approx 1.571\sigma^2/n$ (proofs for all results can be found in Maronna et al. 2006). We say that the sample median has asymptotic efficiency 1/1.571 = 0.636 at the normal. This means that we have to pay a high price for the median's robustness. We may make requirement (1) above more precise by stating that we want an estimate with a high efficiency at the normal, while keeping condition (2). We now consider two approaches to attain this goal.

## M Estimates

Let $u_i$ have a positive density function $f$, so that $x_i$ in (1) has density $f(x-\mu)$. Then the maximum likelihood estimate (MLE) of $\mu$ is the solution of

$$\prod_{i=1}^{n} f(x_i-\mu) = \max.$$

Taking logs we get

$$\sum_{i=1}^{n} \rho(x_i-\mu) = \min \tag{2}$$

where $\rho = -\log(x)$. If $f$ $\mathrm{N}(0,1)$ we have $\rho(x) = \left(x^2+\log(2\pi)\right)/2$. Note that using this $\rho$ is equivalent to using $\rho(x) = x^2$, which yields $\widehat{\mu} = \bar{x}$. If $f$ is the double exponential density $f(x) = 0.5\exp(-|x|)$ we get likewise $\rho(x) = |x|$, which yields $\widehat{\mu} = \mathrm{Med}(\mathbf{x})$.

An M estimate is defined through (2) where $\rho(x)$ is a given function (which does not necessarily correspond to a MLE). To fulfill (1) it has to be approximately quadratic for small $x$; to fulfill (2) it must increase more slowly than $x^2$ for large $x$. An important case is the Huber $\rho$-function

$$\rho(x) = \begin{cases} x^2 & \text{for } |x| \le k \\ 2k|x| - k^2 & \text{for } |x1 > k. \end{cases}$$

Figure 2 plots $\rho$ for $k = 2$.

The limit cases $k \to \infty$ and $k \to 0$ correspond respectively to $x^2$ and $|x|$, and therefore the estimate is an intermediate between the mean and the median. Differentiating (2) we get that $\widehat{\mu}$ is a solution to the *estimating equation*

$$\sum_{i=1}^{n} \psi(x_i-\widehat{\mu}) = 0 \tag{3}$$

where $\psi = \rho'$. For the Huber function we have that (up to a constant)

$$\psi(x) = \begin{cases} -k & \text{for } x < -k \\ x & \text{for } |x1 \le k. \\ k & \text{for } x > k \end{cases}$$

Figure 3 displays $\psi$ for $k = 2$.

The boundedness of $\psi$ makes the estimate robust.

It can be shown that for any symmetric distribution of the $u_i$, for large $n$ the distribution of $\widehat{\mu}$ is approximately $\mathrm{N}(\mu, v/n)$ where the asymptotic variance $v$ is given by

$$v = \frac{\mathrm{E}\psi(x-\mu)^2}{[\mathrm{E}\psi'(x-\mu)]^2}.$$

The following Table gives the normal efficiencies of the Huber estimate for different values of $k$.

| $k$ | Efficiency |
|---|---|
| 0 | 0.64 |
| 1.0 | 0.90 |
| 1.4 | 0.95 |
| $\infty$ | 1.00 |

It is seen that $k = 1.4$ yields a high efficiency.

Define now the *weight function* $W$ as $W(x) = \psi(x)/x$. For the Huber function we have

$$W(x) = \begin{cases} 1 & \text{for } |x| \le k \\ k/|x| & \text{for } |x| > k. \end{cases}$$

Figure 4 plots Huber's $W$ :

We may rewrite (3) as

$$\sum_{i=1}^{n} W(x_i-\widehat{\mu})(x_i-\widehat{\mu}) = 0$$

and therefore

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

where $w_i = W(x_i-\widehat{\mu})$. This shows that a location M estimate can be thought of as a weighted mean with weights $w_i$, where observations distant from the "bulk" of the data receive smaller weights.

Note however that (4) is not an explicit formula for $\widehat{\mu}$, since $w_i$ depends on both $x_i$ and $\widehat{\mu}$. It can however be used as a basis for the iterative numerical computing of $\widehat{\mu}$.

## L Estimates

A different approach to robust location estimates is based on the ordered observations $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ ("order statistics"). The simplest is the *α-trimmed mean*. For $\alpha \in [0, 0.5]$ let $m = [\alpha(n-1)]$ where $[.]$ stands for the integer part. Then the $\alpha$-trimmed mean is defined as

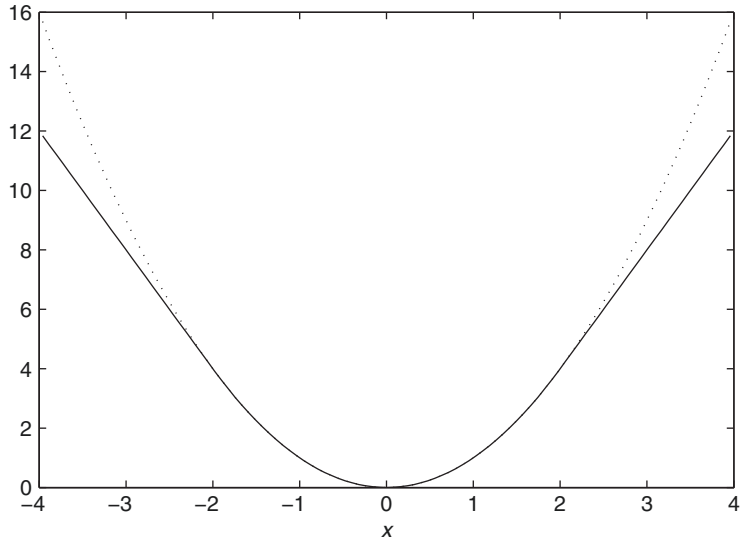
$$\bar{x}_\alpha = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)}. \tag{5}$$

That is, a proportion $\alpha$ of the largest and smallest observations are deleted. It can be shown that for $\alpha = 0.25$ the efficiency of $\bar{x}_\alpha$ is 0.83, although it seems that we are "deleting" half of the sample!. The reason is that $\bar{x}_\alpha$ is actually a function of *all* observations, even of those that do not appear in (5).

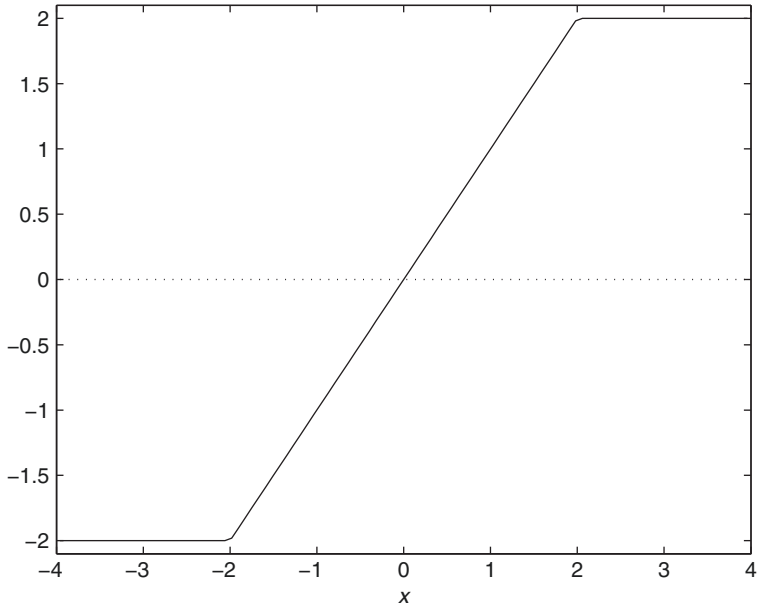
In general, L estimates are linear combinations of ▶order statistics:

$$\widehat{\mu} = \sum_{i=1}^{n} a_i x_{(i)}$$

where the $a_i$ are constants such that $a_i = a_{n-1+i}$ and $\sum_{i=1}^{n} a_i = 1$.




**Robust Statistical Methods. Fig. 2** Huber $\rho(x)$ with $K = 2$ (full line) and $x^2$ (dotted line)



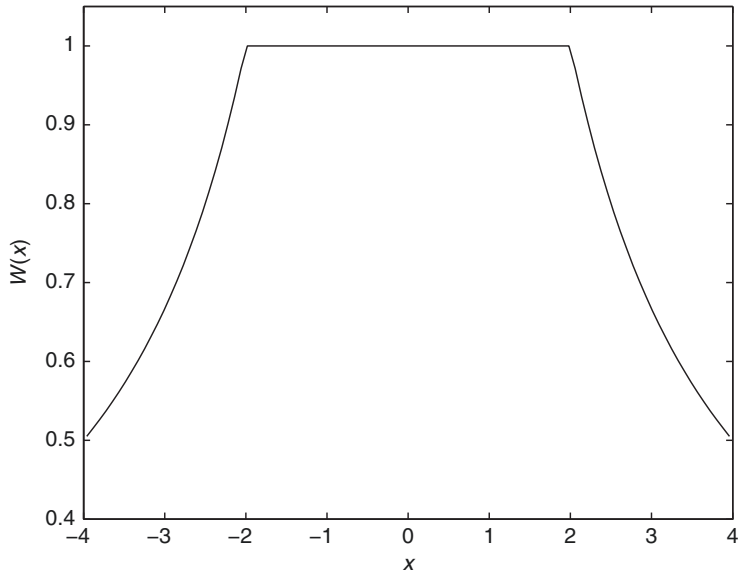

**Robust Statistical Methods. Fig. 3** Huber's $\psi$ for $k = 2$

Although L estimates seem simpler than M estimates, they are difficult to generalize to regression and multivariate analysis. On the other hand, M estimates can be generalized to more complex situations.

## General Considerations

The present exposition attempts to give the reader a flavor of what robust methods are, through the incomplete treatment of a very simple situation. It is based on the author's experience and personal preferences.

The book by Maronna et al. (2006) contains a general and up to date account of robust methods. The classic book by Huber (1981) and the recent one by Jurecková and Picek (2006) contain more theoretical material. Hampel et al. (1986) gives a particular approach to robustness. Rousseeuw and Leroy (1987) deal (although not exclusively) with an important approach to robust regression.




**Robust Statistical Methods. Fig. 4** Huber's weight function for $k = 2$

## About the Author

Dr. Ricardo Maronna is Consulting Professor, University of La Plata, and Researcher at C.I.C.P.B.A., both in La Plata, Argentina. He has been three times Head of the Mathematics Department of the Faculty of Exact Sciences of the University of La Plata. He is Past President of the Statistical Society of Argentina (2003–2004). He is the author or co-author of 40 papers on statistical methods and their applications, and of the highly praised book *Robust Statistics: Theory and Methods* (with R.D. Martin and V.J. Yohai, John Wiley and Sons, 2006). "This book belongs on the desk of every statistician working in robust statistics, and the authors are to be congratulated for providing the profession with a much-needed and valuable resource for teaching and research." (*Tyler, David E. (2008), Journal of the American Statistical Association*, **103:** June 2008, p. 889.)

## Cross References

►Adaptive Linear Regression
►Adaptive Methods
►Mean Median and Mode
►Multivariate Outliers
►Multivariate Technique: Robustness
►Order Statistics
►Outliers
►Robust Inference
►Robust Statistics

## References and Further Reading

Barnett V, Lewis T (1998) Outliers in statistical data, 3rd edn. Wiley, New York

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

Hand DJ, Daly F, Dunn AD, McConway KJ, Ostrowski E (1994) A handbook of small data sets. Chapman & Hall, London

Huber PJ (1981) Robust statistics. Wiley, New York

Jurecková J, Picek J (2006) Robust statistical methods with R. Chapman & Hall, London

Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

# Robust Statistics

Peter J. Huber
Klosters, Switzerland

## Introduction

The term "robust" was introduced into the statistical literature by Box (1953). By then, robust methods such as trimmed means, had been in sporadic use for well over a century, see for example Anonymous (1821). However, Tukey (1960) was the first person to recognize the

extreme sensitivity of some conventional statistical procedures to seemingly minor deviations from the assumptions, and to give an eye-opening example. His example, and his realization that statistical methods optimized for the conventional Gaussian model are unstable under small perturbations were crucial for the subsequent theoretical developments initiated by Huber (1964) and Hampel (1968).

In the 1960s robust methods still were considered "dirty" by most. Therefore, to promote their reception in the statistical community it was crucial to mathematize the approach: one had to prove optimality properties, as was done by Huber's minimax results (1964, 1965, 1968), and to give a formal definition of qualitative robustness in topological terms, as was done by Hampel (1968, 1971). The first book-length treatment of theoretical robustness was that by Huber (1981, 2nd edition by Huber and Ronchetti 2009).

## *M*-Estimates and Influence Functions

With Huber (1964) we may formalize a robust estimation problem as a game between the Statistician and Nature. Nature can choose any distribution within some uncertainty region, say an $\varepsilon$-contamination neighborhood of the Gaussian distribution (i.e., a fraction $\varepsilon$ of the observations comes from an arbitrary distribution). The Statistician can choose any $M$-estimate, that is, an estimate defined as the solution $\hat{\theta}$ of an equation of the form

$$\sum \psi(x_i, \theta) = 0, \tag{1}$$

where $\psi$ is an arbitrary function. If $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$ is the logarithmic derivative of a probability density, then $\hat{\theta}$ is the maximum likelihood estimate. The Statistician aims to minimize the worst-case asymptotic variance of the estimate.

It can be seen from (1) that in large samples the influence of the $i$th observation toward the value of $\hat{\theta}$ is proportional to $\psi(x_i, \theta)$. Hampel (1968; 1974, see also Hampel et al. 1986) generalized this notion through his *influence curve* (or *influence function*) to more general types of estimators. In the case of $M$-estimates the influence function is proportional to $\psi(x, \theta)$. Arguably, the influence function is the most useful heuristic tool of robustness. To limit the influence of gross errors, the influence function should be bounded, and a simple method for constructing a robust $M$-estimate is to choose for $\psi$ a truncated version of the logarithmic derivative of the idealized model density.

In simple cases, in particular the estimation of a one-dimensional location parameter, the game between the Statistician and Nature has an explicit asymptotic minimax solution: find the *least favorable* distribution (i.e., minimizing Fisher information) within the uncertainty region. This is the minimax strategy for Nature. The asymptotic minimax strategy for the Statistician then is the maximum likelihood estimate for the least favorable distribution. In fact, error distributions occurring in practice are well modeled by least favorable distributions corresponding to contamination rates between 1% and 10%, better than by the Gaussian model itself.

Note that bounding the influence provides safety not only against ▶outliers ("gross errors"), but also against all other types of contamination. All three approaches: the simple-minded truncation of the logarithmic derivative, the asymptotic minimax solution, and the finite sample minimax solution (see below) lead to qualitatively identical $\psi$-functions.

## Robustness, Large Deviations and Diagnostics

By 1970 John Tukey's interests had changed their focus, he scorned models, and for him, robust methods now were supposed to have a good performance for the widest possible variety of (mostly longtailed) distributions. His shift of the meaning of the word "robust" inevitably created some confusion. I still hold (with Tukey 1960, Huber 1964 and Hampel 1968) that robust statistics should be classified with parametric statistics, and that robustness primarily should be concerned with safeguarding against ill effects caused by finite but small deviations from an idealized model, with emphasis on the words *small* and *model*. Interpretation of the results in terms of a model becomes difficult if one leaves the neighborhood of that model. Good properties far away from the model should be regarded as a bonus rather than as a must.

The concern with large deviations (see ▶Large Deviations and Applications) has caused a concomitant confusion between the complementary roles of diagnostics and robustness. The purpose of robustness is to *safeguard* against deviations from the assumptions, while the purpose of diagnostics is to *identify* and *interpret* such deviations. Robustness is concerned in particular with deviations that are near or below the limits of detectability. Safeguards against those can be achieved in a mechanical, almost blind fashion, even if the sparsity of data may prevent you from going beyond. Diagnostics on the other hand comes into play with larger deviations; it is an art, requiring insight into the processes generating the data.

## The Breakdown Point

The standard interpretation of contamination models is that a dominant fraction $1 - \varepsilon$ of the data consists of "good"

observations that follow the idealized model, while a small fraction $\varepsilon$ of "bad" observations does not.

The breakdown point $\varepsilon^*$ is the smallest fraction $\varepsilon$ of bad observations that may cause an estimator to take on arbitrarily large aberrant values. This concept is a very simple, but extremely useful global characteristic of a robust procedure. Hampel (1968) had given it an asymptotic definition, but actually, it is most useful in small sample situations (Donoho and Huber 1983).

Robust statistical procedures should have a reasonably high breakdown point (i.e., at least in the range of 10% to 25%). A higher values is desirable – if it comes for free and does not unduly impair performance at the model. Indeed, robust $M$-estimates of one-dimensional parameters in large samples typically approach the maximum possible breakdown point of 50%. This is not so in higher dimensions: $M$-estimates of $d$-dimensional location parameters and covariance matrices have a disappointingly low breakdown point $\varepsilon^* \leq 1/(d+1)$, see Maronna (1976). For a while this limit was thought to hold generally for all affine equivariant estimators, but then it was found that with the help of projection pursuit methods it is possible to construct estimators approaching an asymptotic breakdown point of 50%, see Donoho and Huber (1983). However, these estimators are overly pessimistic by having a low efficiency at the model, and they are very computer intensive.

Over the years it has become fashionable to strive for the highest possible breakdown point, particularly in regression situations, where observations that are influential through their position in factor space (the so-called "leverage points") present peculiar problems. While a proof that the theoretical maximum of 50% can be attained is interesting and theoretically important, the corresponding procedures in my opinion suffer from what I have called the Souped-up Car Syndrome (Huber 2009): they optimize one aspect to the detriment of others. For example, the high breakdown point $S$-estimators of regression even lack the crucial stability attribute of robust procedures (Davies 1993, Section 1.6).

With high values of $\varepsilon$, alternative interpretations of contamination models become important, transcending the ubiquitous presence of a small fraction of gross errors. The data may be a mixture of two or more sets with different (e.g., ethnic) origins, and the task no longer is to ignore a small discordant minority of gross errors (a robustness problem), but to disentangle larger mixture components (a diagnostic problem). High breakdown point procedures can be used for diagnostic purposes, namely to identify a dominant mixture component, but they need not provide the best possible approach.

## Bayesian Robustness

The term "robust" had been coined by a Bayesian (Box 1953). Ironically, while there is a fairly large literature in the form of journal articles – see, for example, Berger's (1994) overview – Bayesianism never quite assimilated the concept. The reason seems to be that for an orthodox Bayesian statistician probabilities exist only in his mind, and that he therefore cannot separate the model (i.e., his belief) from the true underlying situation. For a pragmatic Bayesian like Box, robustness was a property of the model (which he was willing to adjust in order to achieve robustness), while for a pragmatic frequentist like Tukey, it was a property of the procedure (and he would tamper with the data by trimming or weighting them to achieve robustness). To me as a decision theorist, the dispute between Box and Tukey about the proper robustness concept was a question of the chicken and the egg: which comes first, the least favorable model of Nature, or the robust minimax procedure of the Statistician? See Huber and Ronchetti (2009), Chapter 15, in particular p. 325.

## Finite Sample Results and Robust Tests

In his decision theoretic formalization, Huber (1964) had imposed an unpleasant restriction on Nature by allowing only symmetric contaminations. It seems to be little known that this restriction is irrelevant; it can be removed by an approach through finite sample robust tests, Huber (1965, 1968). The extension of robust tests beyond the single-parameter case, however, is difficult; see Huber and Ronchetti (2009), Chapter 13.

## Heuristic Aspects of Robustness

There are no rigorous optimality results available once one leaves the single-parameter case. Admittedly, the perceived need for mathematical rigor and proven optimality properties has faded away after the 1960s. But at least, one should subject one's procedures to a worst case analysis in some neighborhood of the model. Even this is difficult and rigorously feasible only in few cases. A good heuristic alternative is a combination of infinitesimal approaches (influence function or shrinking neighborhoods) with breakdown point aspects. Shrinking neighborhoods were first dealt with by Huber-Carol (1970) in her thesis, and a comprehensive treatment was given by Rieder (1994).

In my opinion the crucial attribute of robust methods is stability under small perturbations of the model. I am tempted to claim that robustness is not a collection of procedures, but rather a state of mind: a statistician should keep in mind that *all* aspects of a data analytic setup (experimental design, data collection, models, procedures)

should be such that minor deviations from the assumptions cannot have large effects on the results (a robustness problem), and that major deviations can be discovered (a diagnostic problem). Compromises are unavoidable. For example, the so-called "optimal" linear regression designs, which evenly distribute the observations on the $d$ corners of a $(d-1)$-dimensional simplex, on one hand lack redundancy to spot deviations of the response surface from linearity, and on the other hand, already subliminal deviations from linearity may impair optimality to such an extent that the "naive" design (which distributes the observations evenly over the entire design space) is superior. Moreover, if there is a problem at a single corner of the simplex, affecting half of the observations there, then this can cause breakdown, leading to a breakdown point no better than $\varepsilon^* \cong 1/(2d)$. See Huber and Ronchetti (2009), Chapter 9, and Chapter 11, p. 285.

## About the Author

Professor Huber is a Fellow of the American Academy of Arts and Sciences. He received a Humboldt Award in 1988. He was a Professor of statistics at ETH Zurich (Switzerland), Harvard University, Massachusetts Institute of Technology, and the University of Bayreuth (Germany). Peter Huber has published four books and over 70 papers on statistics and data analysis, including the fundamental paper on robust statistics "Robust Estimation of a Location Parameter" (Annals of Mathematical Statistics, (1964) Volume 35, Number 1, 73–101), and the text *Robust Statistics* (Wiley, 1981; republished in paperback 2004). In addition to his fundamental results in robust statistics, Peter Huber made important contributions to computational statistics, strategies in data analysis, and applications of statistics in fields such as crystallography, EEGs, and human growth curves.

## Cross References

►Bayesian Statistics
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Imprecise Probability
►Large Deviations and Applications
►Misuse of Statistics
►Multivariate Technique: Robustness
►Optimality and Robustness in Statistical Forecasting
►Robust Inference
►Robust Statistical Methods
►Statistical Fallacies: Misconceptions, and Myths

## References and Further Reading

Anonymous (1821) Dissertation sur la recherche du milieu le plus probable. Ann Math Pures et Appl 12:181–204
Berger JO (1994) An overview of robust Bayesian analysis. Test 3: 5–124
Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335
Davies PL (1993) Aspects of robust linear regression. Ann Stat 21:1843–1899
Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds) A festschrift for Erich L. Lehmann. Wadsworth, Belmont
Hampel FR (1968) Contributions to the theory of robust estimation, Ph.D. Thesis. University of California, Berkeley
Hampel FR (1971) A general qualitative definition of robustness. Ann Math Stat 42:1887–1896
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 62:1179–1186
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence. Wiley, New York
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
Huber PJ (1965) A robust version of the probability ratio test. Ann Math Stat 36:1753–1758
Huber PJ (1968) Robust confidence limits. Z Wahrscheinlichkeitstheorie Verw Gebiete 10:269–278
Huber PJ (1981) Robust statistics. Wiley, New York
Huber PJ (2009) On the non-optimality of optimal procedures. In: Rojo J (ed) Optimality. The third E. L. Lehmann symposium. Institute of Mathematical Statistics, Lecture Notes Vol. 57. Beachwood, Ohio, USA, pp 31–46
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, New York
Huber-Carol C (1970) Etude asymptotique de tests robustes, Ph.D. Thesis, Eidgen. Technische Hochschule, Zürich
Maronna RA (1976) Robust M-estimators of multivariate location and scatter. Ann Stat 4:51–67
Rieder H (1994) Robust asymptotic statistics. Springer, Berlin
Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I (ed) Contributions to probability and statistics, Stanford University Press, Stanford



# ROC Curves

Lino Sant
Professor, Head of Department of Statistics & Operations Research
University of Malta, Msida, Malta

Classification problems, arising in different forms within various contexts, have stimulated a lot of statistical research with a thread of development stretching back to Fisher's discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues

and Problems) and leading right to the core of statistical learning theory. Along this line ROC (Receiver Operating Characteristic) has come to occupy a privileged position. Weaving within its theory a central role for two classification errors types, it manages to give a statistically sound way of evaluating the diagnostic accuracy of classifier variables.

*ROC* saw its birth within signal detection theory (Green and Swets 1966). It was cultivated for a time by researchers in psychophysics and later on much promoted within the biomedical sciences (Pepe 2003; Zhou et al. 2002). Interest in the technique and the theoretical tools it offers has extended to many areas these days. The problem it addresses is fairly simple:

A population $\Pi$ of entities, be they individuals, signal emitters, images, ecosystems, whatever, is made up of two disjoint subpopulations: $\Pi = M \cup N$. An attribute of relevance, say a biomarker like BMI, intensity of an electrical signal, or environmental variable like Air Quality Index, is being measured across both subpopulations. This attribute will be modeled by random variable $X$ with values on a continuous or ordinal scale. $F$ is the probability distribution of $X$ restricted to $M$ with probability density function $f$, $G$ that restricted to $N$ with density $g$. The classification problem is that of determining appurtenance to one subpopulation of an object whose $X$-reading was $x$. Assuming the mean corresponding to $F$ is smaller than that of $G$, it is natural to set up some number $c$ and declare the object to belong to $M$ if $x < c$, or to $N$ if $x \geq c$. With reference to the figure below, the location of $c$ determines a number of classification probabilities.

Corresponding to this rule we have two consequences for each decisions: assigning object with value $x < c$ to $M$ incurs the risk of committing an error whose probability is denoted by false negative fraction (*FNF*) $P[X < c|N]$ or else being correct with probability called true negative fraction (*TNF*) $P[X < c|M]$. The "negative" epithet comes from the medical context where $F$ would correspond to a healthy group who are not afflicted by some disease under study. Conversely, assigning object with value $x \geq c$ to $N$ incurs the risk of committing an error, whose probability is called the false positive fraction (*FPF*) $P[X > c|M]$, or else being correct with probability called true positive fraction (*TPF*) $P[X > c|N]$.

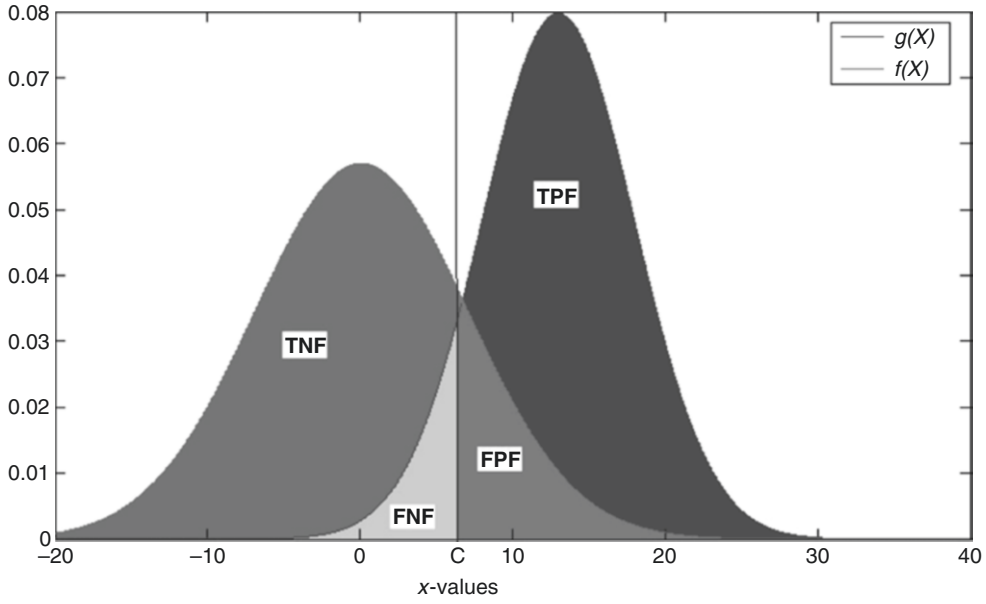
The performance of a classifier variable, in particular its diagnostic accuracy, can be studied in depth by looking at the graph of the *sensitivity* (another name for *TPF*) against the 1-*specificity* (another name for *TNF*) of the classifier for each possible value of $c$. This graph is called the receiver operating characteristic curve, *ROC*. Using distribution functions and hiding $c$ implicitly we have: $ROC(t) = 1 - G(F^{-1}(1-t))$ for $0 \leq t \leq 1$ and $c$ is given by: $c = G^{-1}(1 - ROC(t))$. A typical *ROC* curve is shown in the figure below.

The higher the graph reaches toward the top left corner the better the classifier behaves. One way of gauging this property is through the area under the curve, denoted *AUC*, and defined as: $AUC = \int_0^1 ROC(t)dt$. This quantity corresponds to the probability that a randomly selected pair of objects, one from each subpopulation, is correctly classified by a test using the classifier. This statistic allows comparisons to be made between classifiers. Classifiers with large AUC are to be preferred. The above analysis can be suitably adapted to random variables with discrete distributions (Figs. 1 and 2).

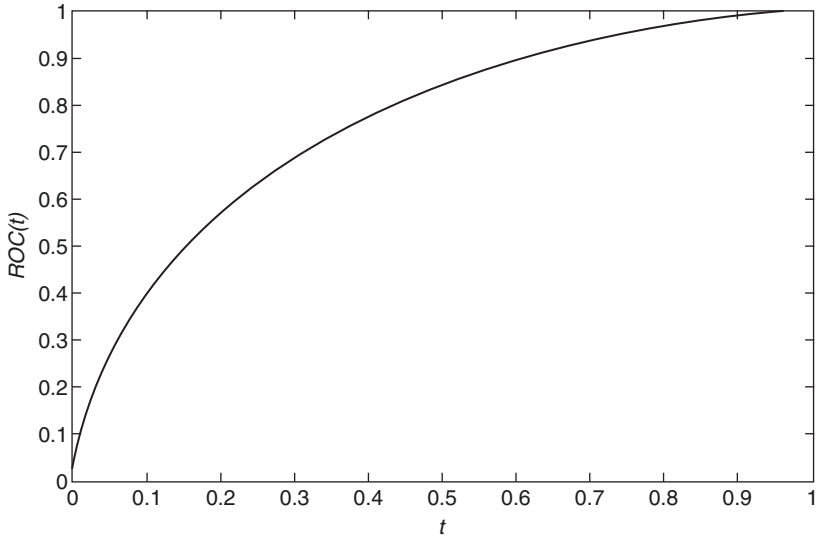
In practice all the population quantities above are not known explicitly. They have to be estimated from actual data. The estimation procedure starts with the procurement of samples of size $m$ (resp. $n$) selected from subpopulation $M$ (resp. $N$). The values obtained are used to obtain optimal values of $c$ as well as to compare different classifiers. *ROC* curves can be estimated from such data using a number of techniques which vary across the whole spectrum of estimation techniques. There are parametric methods using known underlying distribution types. The sample from $M$ (resp. $N$) gives estimates for the corresponding parameters and an *ROC* curve can be derived from the definition above using distribution functions explicitly.

Nonparametric models using empirical distributions are popular in areas where identification of the underlying distributions has not been definitively established. Using results from ▶empirical processes and asymptotic theory a number of very useful statistical results have been obtained for nonparametric models. The most popular method, called the binormal model, is in fact semiparametric. It derives a lot of its sampling distributional results from the Komlós–Major–Tusnády Brownian bridge construction (Hsieh and Turnbull 1996). Though it assumes underlying normal distributions, it can be shown to be valid in cases where the distributions can be transformed to normal distributions. Furthermore the method has shown itself to be robust to departures from normality.

A large number of other estimation techniques have been proposed in the literature like minimum distance and Bayesian estimators. The former are defined relative to some specific metric, or penalty function if you will, on some suitable space of probability distributions. This idea ties up nicely with the hypothesis testing aspect of *ROC* theory. In practice good values of cut-off point $c$




**ROC Curves. Fig. 1** Superimposed graphs for pdf's *f* and *g*




**ROC Curves. Fig. 2** A typical ROC graph for continuous distributions

obtained from reliable *ROC* curve estimators would be needed. "Good" varies from one application to the other, but in general it means values which minimize costs related to consequences following from taking the wrong decision, which are tied up to probabilities *FNF* and *TPF*. So in general we need to take care of some penalty function, say linear function: $\alpha_0 + \alpha_{TP}\boldsymbol{P}[TP] + \alpha_{TN}\boldsymbol{P}[TN] + \alpha_{FP}\boldsymbol{P}[FP] + \alpha_{FN}\boldsymbol{P}[FN]$ where the coefficients $\alpha_{AB}$ give the costs corresponding to eventuality *AB*.

*ROC* was, and still is, extensively used and developed within the biomedical sciences. One important current line of research tries to locate canonical theory within a GLM context. Nevertheless these last thirty years have seen an enormous amount of interest in the technique

amongst computer science researchers interested in disciplines related to statistical classification and machine learning (Krzanowski and Hand 2009).

## Cross References

►Discriminant Analysis: An Overview
►Nonparametric Predictive Inference
►Pattern Recognition, Aspects of
►Statistical Signal Processing

## References and Further Reading

Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York

Hsieh F, Turnbull BW (1996) Nonparametric and semi-parametric estimation of the receiver operating characteristic curve. Ann Stat 24(1):25–40

Krzanowski WJ, Hand J (2009) ROC curves for continuous data. CRC/Chapman and Hall, Boca Raton

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. University Press, Oxford

Zhou KH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley, New York

# Role of Statistics

Ashok Sahai[1] , Miodrag Lovric[2]
[1]Professor
St. Augustine Campus of the University of the West Indies at Trinidad, St. Augustine, Trinidad and Tobago
[2]Professor
University of Kragujevac, Kragujevac, Serbia

► *"Modern statistics, like telescopes, microscopes, X-rays, radar, and medical scans, enables us to see things invisible to the naked eye. Modern statistics enables us to see through the mists and confusion of the world about us, to grasp the underlying reality."*

David Hand

## Introduction

Despite some recent vigorously promulgated criticisms of statistical methods (particularly significance tests), methodological limitations, and misuses of statistics (see Ziliak and McCloskey 2008; Hurlbert and Lombardi 2009; Marcus 2009; especially Siegfried 2010, among others), we are the ones still "living in the golden age of statistics" (Efron 1998).

Statistics play a vital role in collecting, summarizing, analyzing, and interpreting data in almost all branches of science such as agriculture, astronomy, biology, business, chemistry, economics, education, engineering, genetics, government, medicine, pharmacy, physics, psychology, sociology, etc. Statistical concepts and principles are ubiquitous in science: "as researchers, we use them to design experiments, analyze data, report results, and interpret the published findings of others" (Curran-Everett et al. 1998). Statistical analysis has become an indispensable and fundamental component and vehicle of modern research.

Why is there such a dependence on statistical methods? One of the possible reasons is that statistical thinking has a universal value, as a process that recognizes that variation is present in all phenomena and that the study of variation leads to new knowledge and better decisions. According to Suppes (2007), "the new work, the new concepts, the new efforts, always lead initially, and, often for a long time, to uncertain results. It is…only by an understanding of probability and statistics that a philosopher of science can come to appreciate, in any sort of sophisticated way, the nature of uncertainty that is at the heart of contemporary science…Without statistical methods, it is often impossible to convert the natural, seemingly confused, uncertainty of many results in science into highly probable ones." Straf, in his presidential ASA address (2003), points out that statistics is special "not only because it advances discoveries across the breadth of scientific disciplines and advances the development of technologies, but also because it has an important connection to the human side of scientific and technological development." According to him, the role of statistics is "to increase our understanding, to promote human welfare, and to improve our quality of life and well-being by advancing the discovery and effective use of knowledge from data."

## The Importance of Statistics

Since this Encyclopedia contains many entries on the specific role of statistics in different sciences, we will list here only several selected sources underlying the importance of statistics and its versatile usefulness. For obtaining a further appreciation of the role of statistics, the interested reader is referred to those entries, list of references and is urged to "virtually attend" a lecture given by Sir David Cox, by downloading the video file "The Role of Statistics in Science and Technology." Additionally, readers (including all researchers and writers of introductory textbooks on statistics) are advised to read carefully the elucidating paper written by James Brewer (1985) on myths and misconceptions in statistics textbooks.

(a) **Climate research**. Zwiers and Storch (2004) emphasize the importance of statistical methods "for a whole gamut of activities that contribute to the ultimate synthesis of climate knowledge, ranging from the collection of primary data, to the interpretation and analysis of the resulting high-level data sets" (see also ▶Statistics and Climate Change).

(b) **Economics and social studies**. Statistical analysis has proved useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty, etc. "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth, methods adequate to the treatment of economic data, in so far as these exist, have only been developed in the study of biology and the other sciences" (Fisher 1925).

(c) **Engineering**. According to Johnson et al. (2004, p. 7) "there are few areas where the impact of the recent growth of statistics has been felt more strongly than in engineering and industrial management. Indeed, it would be difficult to overestimate the contributions statistics has made to solving production problems, to the effective use of materials and labor, to basic research, and to the development of new products." Statistics in engineering can be effectively used to solve, for example, the following diversified tasks: "calculating the average length of the downtimes of a computer, collecting and presenting data on the numbers of persons attending seminars on solar energy, evaluating the effectiveness of commercial products, predicting the reliability of a rocket, or studying the vibrations of airplane wings" (op. cit., p. 5).

(d) **Genomics**. Ben-Hui Liu (1998) emphasizes that statistics is a tool to solve problems that cannot be solved only through biological observation or qualitative analysis and that this is especially true for the statistics used in genomic mapping.

(e) **Information systems**. Dudewicz and Karian (1999) indicate that the role of statistics in information systems and information technology in general "can be substantial, yielding more nearly optimal performance of problems at the emerging frontiers in all their aspects."

(f) **Kinetic theory of gases**. Von Mises (1930, p. 207) believes that "not even the tiniest little theorem in the kinetic theory of gases follows from classical physics alone, without additional assumptions of a statistical kind."

(g) **Medical research**. Statisticians are at the "forefront of medical research, helping to produce the evidence for new drugs or discovering links between health and disease and the way we lead our lives" (Oxford Brookes University web site (http://tech.brookes.ac.uk/teaching/pg/msc-in-medical-statistics). According to Sprent (2003) the role of statistics in medical research "starts at the planning stage of a clinical trial or laboratory experiment to establish the design and size of an experiment that will ensure a good prospect of detecting effects of clinical or scientific interest. Statistics is again used during the analysis of data (sample data) to make inferences valid in a wider population." Feinstein (2001) points out that the statistical citation of results has become one of the most common, striking phenomena of modern medical literature (see also ▶Medical Statistics and ▶Medical Research, Statistics in).

(h) **Ophthalmology**. Coleman (2009) believes that statistics play a vital role in "helping us to make decisions about new diagnostic tools and treatments and the care of our patients in the face of uncertainty because, when dealing with patients, we are never 100% certain about an outcome."

(i) **Pharmacogenomics**. Kirkwood (2003) argues that statistical theory and probability will play an expanded role in understanding genetic information through the development of new analytical methodology and the novel application of traditional statistical theory. He points out that "the combination of statistical applications and genomic technologies is a key to understanding the genetic differences that identify patients susceptible to disease, stratify patients by clinical outcome, indicate treatment response, or predict adverse event occurrences."

(j) **Policy and world development**. For example, Moore (1998), in his presidential address to the American Statistical Association (ASA), claimed that it is difficult to think of policy questions that have no statistical component, and argued that statistics is a general and fundamental method because data, variation, and chance are omnipresent in modern life. High-quality statistics also "improve the transparency and accountability of policy making, both of which are essential for good governance, by enabling electorates to judge the success of government policies and

to hold their government to account for those policies… Statistics play a vital role in poverty reduction and world development" (Paris21). However, many developing countries still lack the capacity to produce and analyze good-quality data and use the range of appropriate statistical techniques required to support effective development progress (see also the entries ►Promoting, Fostering and Development of Statistics in Developing Countries, The Role of Statistics – Developing Country Perspective and ►Selection of Appropriate Statistical Methods in Developing Countries).

(k) **Psychiatry**. Hand (1985) believes that statistics has a major role in modern psychiatry, and that "awareness and understanding of statistical concepts is of increasing importance to all psychiatrists, but especially those who wish to advance the field by undertaking research themselves" (see also ►Psychiatry, Statistics in).

(l) **Quality management**. The role of statistics includes control and improvement of the quality of industrial products, during and after the production process through statistical quality control (Srivastava and Rego 2008).

(m) **Quantum theory**. Karl Popper (2002, p. 217) argues that the concept that "quantum theory should be interpreted statistically was suggested by various aspects of the problem situation. Its most important task—the deduction of the atomic spectra—had to be regarded as a statistical task ever since Einstein's hypothesis of photons (or light-quanta)" (see also ►Statistical Inference for Quantum Systems).

(n) **Science**. Magder (2007) points out that the role of statistics in science should be to quantify the strength of evidence in a study so other scientists can integrate the new results with other information to make scientific judgments.

(o) **Seismology**. According to Vere-Jones, one of the pioneers of statistical seismology, "the last decade has seen an influx of new concepts, new data, and new procedures, which combine to make the present time as exciting as any for statistical seismology. New concepts include new mathematical structures, such as self-similarity, fractal growth and dimension and self-organizing criticality, for which existing statistical techniques, based as most of them are on assumptions of stationarity and ergodicity, are inappropriate. In this area at least, seismology is once more challenging the statisticians to enlarge and update their tool box" (Vere-Jones 2006).

(p) **Sociology**. Statistical methods have had a successful half-century in sociology, contributing to a greatly improved standard of scientific rigor in the discipline (Raftery 2001). The overall trend has been toward using more complex statistical methods to match the data, starting from cross-tabulation, measures of association, and log-linear models in the late 1940s; LISREL-type causal models and event-history analysis in the 1960s; and social networks, simulation models, etc. in the late 1980s (see also ►Sociology, Statistics in).

## Statistics and Uncertainty

Human life always confronts challenging situations calling for decision-making under uncertainties. While the role of statistics is to minimize the uncertainty associated with the impugned phenomena under investigation, the uncertainty could be measured by the concept of probability. Probability is sine qua non for statistics and statistical modeling, the most important covariate that is omnipresent in all realistic situations challenging the scientists. In fact, the role of statistics encompasses the two fundamentally relevant areas of approximation (any model is an approximation of the real-life phenomenon) and that of optimization (to achieve the minimization of the "gap" between the model and reality).

The role of statistics could, very comprehensibly, be summarized as the "statistical game" being played by the statistician/scientist(s) using statistics against nature as the second player. And this statistical game is quite different from the well-known "zero-sum two person game" in the mathematical setup, inasmuch as the second player is not a conscious player trying to be strategic with the choice of the playing strategies of the first player (statistician/scientist(s) using statistics), and in that the loss incurred by the first player is not a gain for the second player, namely, nature (so that this game is not zero-sum). For example, nature will not cause rains if the statistician/person, guided by weather scientists predicting empirically (statistically), is not carrying an umbrella/raincoat. And vice versa if the person is not carrying the protection, that nature will cause the rain. Nature will cause/not cause the rain if it had to do so for whatever reasons not fully known to us/scientists.

The previous discussion is related to the quantum physics phenomenon. If we go at the microphysics level, as we would attempt to do with the help of a powerful microscope, any matter is not deterministic. Actually, the most decisive conceptual event of twentieth-century physics has been the discovery that the world is not deterministic (Hacking 1990, p. 1). In fact, as we know, at the microscopic level, as to whether or not there will be occupancy/a particle or the absence of it at a particular point in the space

occupied by the relevant matter at any specific point in time, the best physicist in the world, as of today, cannot tell. The best that one could do will be the statement of probability of occupancy reckoned empirically (i.e., based on the experimental data, and that too, only statistically and probabilistically) subject to approximation error.

## Conclusion and Recommendation

We agree with Provost and Norman (1990, p. 43) that the 21st century will place even greater demands on society for statistical thinking throughout industry, government, and education.

*However, if statistics aspire to be an essential element in the description and understanding of the actual phenomena in the world around us, it is an imperative that we, statisticians, conduct a critical evaluation of statistics in the first place.* To achieve that, we need to begin with building a bridge between Bayesians and frequentists, may be with the help of a new Ronald Fisher. Equally importantly, we need to find a way to explain more clearly the usage of statistical methods, along with their advantages and disadvantages, to overcome the generalized confusion in the public and among many researchers over many statistical issues and also to educate statistical practitioners at all levels.

> *"A chisel in a skillful artist's hand can produce a beautiful sculpture and a scalpel in an experienced surgeon's hand can save a person's life. Similarly, statistical techniques used properly by an honest and knowledgeable scientist can be equally impressive at illuminating complex phenomena, thus promoting scientific understanding, and shortening the time between scientific discovery and its impact on societal problems. If misused, they can produce the counterproductive results… Such erroneous results, however, should not be viewed as a failing of Statistics"*
>
> (ASA unedited letter to the editor of the *Science News* in response to the "Odds Are, It's Wrong" paper.)

## Acknowledgment

Professor Sahai dedicates this write-up to the fond memory of his most beloved and inspiring Professor, Late Dr. A. R. Roy (Stanford University).

## Cross References

►Careers in Statistics
►Environmental Monitoring, Statistics Role in
►Frequentist Hypothesis Testing: A Defense
►Medical Research, Statistics in
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Rise of Statistics in the Twenty First Century
►Role of Statistics in Advancing Quantitative Education
►Role of Statistics: Developing Country Perspective
►Statistics: An Overview

## References and Further Reading

Brewer JK (1985) Behavioral statistics textbooks: source of myths and misconceptions? J Edu Stat, 10(3) Available at: http://www.jstor.org/stable/1164796 (Special issue: Teaching statistics)

Coleman AL (2009) The role of statistics in ophthalmology. Am J Ophthalmol 147(3):387–388

Cox D (2008) The role of statistics in science and technology. Video file, available at: http://video.google.com/videoplay?docid=1739298413105326425#

Curran-Everett D, Taylor S, Kafadar K (1998) Fundamental concepts in statistics: elucidation and illustration. J Appl Physiol 85: 775–786

Dudewicz EJ, Karian ZA (1999) The role of statistics in IS/IT: practical gains from mined data. Inform Syst Frontiers 1(3):259–266

Efron B (1998) R. A. Fisher in the 21st century. Stat Sci 13(2):95–122

Feinstein AR (2001) Principles of medical statistics. Chapman and Hall/CRC, London

Fisher R (1925) Statistical methods for research workers, Oliver and Boyd, Edinburgh

Hacking I (1990) The taming of chance. Cambridge University Press, Cambridge

Hand D (1985) The role of statistics in psychiatry. Psychol Med 15:471–476

Hand D (2008) Statistics: a very short introduction. Oxford University Press, Oxford

Hurlbert SH, Lombardi CM (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Ann Zool Fenn 46:311–349

Johnson R, Miller I, Freund J (2004) Miller & Freund's probability and statistics for engineers. 7th edn. Prentice Hall, Englewood Cliffs, NJ

Kirkwood SC (2003) The role of statistics in pharmacogenomics. J Japan Soc Comp Stat 15(2):3–13

Liu BH (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press, Boca Raton, p x

Magder L (2007) Against statistical inference: a commentary on the role of statistics in public health research, The 135th APHA annual meeting & exposition of APHA, Washington DC

Marcus A (2009) Fraud case rocks anesthesiology community: Mass. researcher implicated in falsification of data, other misdeeds. Anesthesiology News, 35, 3

Moore DS (1998) Statistics among the liberal arts. J Am Stat Assoc 93(444):1253–1259

Morrison D, Henkel R (eds) (2006) The significance test controversy: a reader. Aldine transaction, Piscataway, USA (reprint)

Paris21 (The partnership in statistics for development in the 21st Century) Counting down poverty: the role of statistics in world development. Available at http://www.paris21.org/documents/2532.pdf

Popper K (2002) The logic of scientific discovery. (trans: Logik der Forschung, Vienna, 1934). Routledge, London

Provost LP, Norman CL (1990) Variation through the ages. Quality Progress Special Issue on Variation, ASQC
Raftery AE (2001) Statistics in sociology, 1950–2000: a selective review. Sociol Methodol 31(1):1–45
Siegfried T (2010) Odds are, it's wrong: science fails to face the shortcomings of statistics, Science News, 177, 26
Sprent P (2003) Statistics in medical research. Swiss Med Wkly. 133(39–40), 522–529
Srivastava TN, Rego S (2008) Statistics for management, Tata McGraw Hill, New Delhi
Straf ML (2003) Statistics: the next generation. J Am Stat Assoc 98:461 (Presidential address)
Suppes P (2007) Statistical concepts in philosophy of science. Synthese 154:485–496
v Mises R (1930) Über kausale und statistische Gesetzmäßigkeit in der Physik. Die Naturwissenschaften 18(7):145–153
Vere-Jones D (2006) The development of statistical seismology: a personal experience. Tectonophysics 413:5–12
Ziliak ST, McCloskey DN (2008) The cult of statistical significance: how the standard error costs us jobs, justice, and lives. University of Michigan Press
Zwiers FW, Storch HV (2004) On the role of statistics in climate research. Int J Climatol 24:665–680

# Role of Statistics in Advancing Quantitative Education

James J. Cochran
College of Business, Louisiana Tech University, Ruston, LA, USA

## Introduction

Education policy makers and educators from all disciplines generally agree that as data have become more plentiful and more readily available, the importance of statistical literacy has grown (Utts 2003; Garfield and Ben-Zvi 2008). Since trends in availability of data show no signs of abating (indicators actually point to accelerating increases in data availability), the importance of teaching statistical thinking to students at all levels is difficult to overstate or overestimate. It is not an exaggeration to state that statistical/quantitative literacy is almost as critical to the future success of students as is reading literacy (Steen 2002). Thus, it is vitally important that both society and the statistics community understand the vital role of statistics in quantitative education.

## Statistics' Role in Quantitative Education

Several phrases are used somewhat interchangeably in reference to an individual's ability to work with numbers and relationships and understand the implications of her or his results. While the generally accepted definitions of these phrases overlap, there are subtle but important differences. These phrases (and generic versions of their generally acceptable definitions) include:

- Numeracy – this phrase, first used in the 1959 Crowther Report on education in the United Kingdom (Jarman 1960), comprises the aptitude to use reason to solve sophisticated quantitative problems; a fundamental understanding of the scientific method; and the ability to communicate with others about everyday quantitative issues, questions, and concerns. In explaining this phrase, Steen (1990) wrote:
  - ▶ Numeracy is to mathematics as literacy is to language. Each represents a distinctive means of communication that is indispensable to civilized life.
- Quantitative Literacy – this phrase refers to minimal levels of comfort with, competency in, and disposition toward working with numerical data and concepts necessary to function at a reasonable level in society.
- Quantitative Reasoning – this phrase represents the manifestation of basic logic applied by an individual to the construction of rigorous and valid arguments as well as the evaluation of the rigor and validity of arguments made by others. Thus, quantitative reasoning emphasizes the higher-order analytic and critical thinking skills needed to understand, create, and cope with sophisticated arguments (which are frequently supported by quantitative data).

See Madison and Steen (2008) for a brief history of the evolution of these terms.

If one considers these three concepts to be the cornerstones of quantitative education, then their meanings must provide the basis of a broad definition of quantitative education. The definition of quantitative education that results from this perspective is *the effort to imbue students with numeracy, quantitative literacy, and quantitative reasoning.*

While quantitative education certainly subsumes statistics education, statistics education is without question a vital and critical component of quantitative education. The strict association most individuals place on quantitative literacy and mathematics, in combination with the manner in which mathematics is generally taught at lower levels of education, generates an overwhelming and ill-advised emphasis on thinking of quantitative issues deterministically. This emphasis on a deterministic treatment of quantitative problems reinforces the notion that a problem has a single correct answer and cultivates the more damaging conclusion that there is a single correct way to solve

a problem, which ultimately robs the student of the opportunity to fully comprehend and appreciate the nature of quantitative problems and problem solving. This issue can be confronted directly through the integration of statistics (and probability) into quantitative education at all levels. Integration of statistics and probability into quantitative education can also be used to address the common misconceptions that (1) quantitative concepts can only be memorized and cannot be understood by average students; (2) quantitative concepts have little relevance to everyday life; (3) quantitative approaches to problem solving lead to conclusive and consistent conclusions; and (4) quantitative analysis is a solitary activity to be pursued by individuals in isolation.

Florence Nightingale (Cook and Nash 1936) expressed her explicit recognition and appreciation of statistics' role in quantitative education when she stated:

> Statistics is the most important science in the whole world, for upon it depends the practical application of every other science and of every art; the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience.

and

> To understand God's thoughts we must study statistics, for these are the measure of his purpose.

As a critically important component of quantitative education, statistics educators should strive to encourage students to develop numeracy, quantitative literacy, and quantitative reasoning skills. Thus, it is vitally important that statistical education focus on the components of these three objectives.

## Important Objectives of Statistics Education and their Links to Quantitative Education

Statistics education contributes to quantitative education through its strong emphasis on the development of numeracy, quantitative literacy, and quantitative reasoning skills. Statistics education can address the various components of numeracy, quantitative literacy, and quantitative reasoning skills in a myriad of ways. In the following sections the author discusses several ways in which statistics education naturally and organically encourages students to develop each component of numeracy, quantitative literacy, and quantitative reasoning skills (identified in the preceding definitions).

### Aptitude to use Reason to Solve Sophisticated Quantitative Problems

Statistics instructors have abundant opportunities to help students develop their aptitude to use reason to solve sophisticated quantitative problems. To solve problems in statistics, students must appreciate the fundamental difference between certainty and uncertainty as well as the ramifications of uncertainty; development of this appreciation certainly fosters the aptitude to use reason in resolving sophisticated quantitative problems. The student must use sophisticated logic and reason to understand the p-value as a conditional probability (the probability of taking a sample in a prescribed manner and collecting results at least as counter to the null hypothesis *given that the condition that the null hypothesis is true*). Similarly, comprehension of the distinctions between conditional and joint probability, precision and accuracy, and experimentation and observation also require extensive use of sophisticated reasoning and logic.

### Fundamental Understanding of the Scientific Method

The *scientific method* is the logical and rational order of steps by which a scientist analytically and rigorously tests a conjecture and reaches a conclusion while minimizing the biases s/he introduces into a scientific inquiry. *Statistical inference* is unquestionably an overt embodiment of the scientific method. Indeed, the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE 2005) report of the American Statistical Association maintains that statistics is a problem-solving process that comprises four steps

- Formulation of questions – developing hypotheses and selecting appropriate analytic methods and decision making criteria;
- Data collection – making decisions on what data to collect and how to collect the data, as well as executing the actual process of collecting data;
- Data analysis – using descriptive and inferential approaches that lead to an understanding of what the sample data that have been collected can reveal about their population; and
- Interpretation of results – disseminating and explaining the results to the appropriate audience, considering implementation issues, and making suggestions to make similar future efforts more scientifically sound.

These four steps and the steps of the scientific method have a bijective relationship (this is why many consider statistics to be the purest science). By stressing this relationship in statistics courses, statistics educators can help students

understand the relevance of the scientific method to their lives. Furthermore, statistics educators have opportunities to reinforce the appreciation of the scientific method through coverage of the justifications for sampling, as it is the use of sample data in lieu of census data that necessitates the use of the scientific method.

## Ability to Communicate about Everyday Quantitative Issues, Questions, and Concerns

Because statistics students are generally learning about concepts and ideas that are sophisticated and unfamiliar to them, statistics courses provide natural environments for the nurturing of communications skills. As a statistics instructor works with students to enable them to connect with and understand statistical concepts, s/he can also stress the importance of students' attempts to emulate the instructor's efforts to communicate; through these efforts students can develop the ability to explain statistical concepts, methodologies, and results with individuals who are unfamiliar with and intimidated by statistics. Students will naturally embrace this skill once they understand its desirability and marketability. A student who can correctly explain the underlying principle of statistical inference in an understandable manner or clarify the distinction between the concepts of association and causality, replication and repeated measurement, or parametric and nonparametric approaches will have great advantages in her/his academic and career pursuits.

## Minimal Levels of Comfort with, Competency in, and Disposition toward Working with Numerical Data and Concepts Necessary to Function in Society

A sound background in basic statistics provides an individual with an important level of self-sufficiency with respect to numerical data and concepts. For example, statistics provides its users with systematic methods for dealing with variation; the quantitatively literate individual not only appreciates the ubiquity of variation but also understands the need to consider variation when interpreting observed phenomena and making decisions. Such an individual is capable of quantifying and explaining variability; she or he also recognizes that variability can be the result of patterns in the values of the variable of interest, relationships between the variable of interest and other variables, and/or randomness. Understanding of these notions leads directly to an understanding of randomness (and its importance) as well as the distinction between experimentation and observation (and the associated ramifications).

While every concept covered in an introductory statistics course provides statistics students with an occasion to become more comfortable with numerical concepts, perhaps none presents a greater opportunity (and challenge) than the notion of a central limit theorem (see ▶Central Limit Theorems). An instructor will surely fail to communicate with all but the most highly motivated students by explaining that:

- A central limit theorem is any weak-convergence theorem that expresses the tendency for a sum of several independent identically distributed random variables with a positive variance to converge in distribution to a member of a known and predictable family of distributions.

On the other hand, a statistics instructor can open her/his students' eyes to the elegance and beauty of this concept if s/he explains instead that:

- One version of the central limit theorem states that given a sample is taken from a population whose distribution has mean $\mu$ and variance $\sigma^2$, the distribution of the potential values of the resulting sample mean $\bar{X}$ approaches a normal distribution with a mean $\mu_{\bar{X}} = \mu$ and a variance $\sigma_{\bar{X}}{}^2 = {}^{\sigma^2}/_{n}$ as the sample size n increases.

The second explanation allows the statistics instructor to further elaborate on how the probability of collecting a sample with an extreme mean decreases rapidly as the sample size increases because an extreme sample mean can only result from a sample that consists primarily of extreme observations, and the probability of collecting a sample that consists primarily of extreme observations decreases rapidly as the number of observations in the sample increases. This explanation both appeals to and nurtures the student's levels of comfort with, competency in, and disposition toward working with numerical data and concepts.

## Development/Enhancement of the Ability to Use the Higher-Order Analytic and Critical Thinking Skills Needed to Understand and Create Sophisticated Arguments

In properly designing and executing a statistical investigation, one must use higher-order analytic and critical thinking skills to create quantitatively and logically sophisticated arguments. Thus, by teaching the process of statistical investigation, a statistics instructor is implicitly assisting the student in the development and enhancement of these higher-order analytic and critical thinking skills. For example, because students generally think of mathematics deterministically, they tend to adhere to the rhetorical tactic of using examples to support an argument (Lawton 2009; Sotos et al. 2009). Because of the uncertainty

embedded in sample data (and so is embedded in any statistical investigation), one cannot use examples to support a null hypothesis; sample data is evaluated in terms of the strength of the evidence it provides *against the null hypothesis*. This distinction, between logical/rhetorical and empirical arguments, is initially difficult for students to comprehend. However, with clear and cogent explanations the statistics instructor can help the student understand the sophisticated argument behind this distinction; this certainly constitutes the development and enhancement of these higher-order analytic and critical thinking skills.

In another example, consider the introductory statistics student's strong tendency to fall victim to the *cum hoc ergo propter hoc* fallacy. When these students find a strong correlation between two random variables, they often immediately infer that a causal relationship exists between these two variables. Enhancement of their higher-order analytic and critical thinking skills is necessary to facilitate their understanding that correlation is a necessary but not sufficient condition for causality. The students must further refine these skills in order to achieve an understanding of the concepts of spurious correlation, reverse causation, two way causation, and common causal variables (examples of which can be found throughout the popular media). Thus, through the development of the ability to properly interpret the results of a statistical analysis (in this case, a simple correlation), a student is enhancing her/his ability to use the higher-order analytic and critical thinking skills needed to understand and create sophisticated arguments.

## Conclusions

Statistics education has a critical role in each of the primary components of quantitative education. Through statistics education students can become more numerate and strengthen their analytic and critical thinking skills. Statistics instructors can and should increase their students' and the public's appreciation for statistics by closely aligning their course objectives with the broad definition of quantitative education.

## About the Author

Dr. James J. Cochran is Associate Professor (the Bank of Ruston Endowed Research Professor) of Quantitative Analysis and Computational Modeling; Senior Scientist, Center For Information Assurance; and Senior Scientist/Analytic Group Director, Center For Secure Cyberspace at Louisiana Tech University. He has previously been on the faculty at Wright State University, Drexel University, Miami University, and the University of Cincinnati. He has also held the position of Visiting Scholar with Stanford University, the University of South Africa, and the Universidad de Talca. Professor Cochran has published over thirty articles in statistics and operations research journals. He is a Coeditor of the *Anthology of Statistics in Sports*, and is the founding Editor-in-Chief of the *Wiley Encyclopedia of Operations Research and Management Science*. Professor Cochran is also the Editor-in-Chief of *INFORMS Transactions on Education*. Dr Cochran was General Chair of the 2005 INFORMS Conference, and President of INFORM-ED (the INFORMS Education Forum) from 2002–2005 and the founding Chair of the INFORMS Section on OR in SpORts from 2004–2008. He has received the INFORMS Prize for the Teaching of OR/MS Practice and the American Statistical Association's Significant Contributor to Statistics in Sports Award (both in 2008). He established and has organized INFORMS' Teaching Effectiveness Colloquium series and annual Case Competition as well as the annual INFORMS/IFORS International Education Workshop series. He is also a founding co-chair of *Statistics Without Borders*, and was elected to the International Statistics Institute in 2005.

## Cross References

►Careers in Statistics
►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Statistical Literacy, Reasoning, and Thinking
►Statistics Education

## References and Further Reading

Cook ET, Nash RN (1936) A Short Life of Florence Nightingale. Macmillan, New York

GAISE (2005) Guidelines for assessment and instruction in statistics education. Retrieved November 2, 2009 from www.amstat.org/Education/gaise/GAISECollege.htm

Garfield JB, Ben-Zvi D (2008) Developing students' statistical reasoning: connecting research and teaching practice. Springer, New York

Jarman TL (1960) Developments in English Education in 1959: The year of the Crowther Report. Internationale Zeitschrift für Erziehungswissenschaft/Revue Internationale l' Éducation 6(1):231–234

Lawton L (2009) An exercise for illustrating the logic of hypothesis testing. J Stat Educ 17(2). Available at: www.amstat.org/publications/jse/v17n2/lawton.html

Madison BL, Steen LA (2008) Evolution of numeracy and the national numeracy network. Numeracy 1(1). Available at: http://services.bepress.com/numeracy/vol1/iss1/art2

Sotos AEC, Vanhoof S, Van den Noortgate W, Onghena P (2009) How confident are students in their misconceptions about hypothesis tests? J Stat Educ 17(2). Available at: www.amstat.org/publications/jse/v17n2/castrosotos.html

Steen LA (1990) Numeracy. Daedalus 119(2):211–231
Steen LA (2002) Quantitative literacy: why numeracy matters for schools and colleges. Focus 22(2):8–9
Utts J (2003) What educated citizens should know about statistics and probability. Am Stat 57(2):74–79

# Role of Statistics: Developing Country Perspective

Dimitri Sanga
Acting Director
African Centre for Statistics, Addis Ababa, Ethiopia

## Statistics: A Part of the Enabling Environment for Development

The main role of statistical development is to help National Statistical Systems (NSS) efficiently produce good statistics. Good statistics are characterized, *inter alia*, by the quality (reliability, accuracy, accessibility, timeliness, etc.) with which they are produced. They are said to be good only to the extent that they meet users' needs. They need to be available to a broad range of public and private users and be trusted to be objective and reliable. In addition, they must meet all policy needs and inform the public so that the latter can evaluate the effectiveness of government's actions.

Good statistics are needed to assess, identify issues, support the choice of interventions, forecast the future, monitor progress and evaluate the results and impacts of policies and programmes. They provide a basis for good decision-making, support governments in identifying the best courses of action in addressing problems, are essential to manage the effective delivery of basic services, and are indispensable for accountability and transparency. They are also essential for providing a sound basis for the design, management, monitoring, and evaluation of national policy frameworks such as the Poverty Reduction Strategies (PRSs) and for monitoring progress towards national, sub regional, regional, and international development goals including the Millennium Development Goals initiatives (MDGs). Accordingly, good statistics are considered to be part of the enabling environment for development.

## Initiatives Aimed at Supporting Developing Countries in Statistics

In recognition of the importance of statistics in their development process, developing countries have been struggling to provide their users with quality statistical information. However, the last decade of the twentieth century has witnessed an unprecedented increase in the demand for quality and timely statistics following an emergence of initiatives aimed at tackling development issues including those enshrined in the Millennium Declaration. In fact, there is increasing recognition that the successful implementation, evaluation, and monitoring of national, sub regional, regional, and international development agendas rely on the production and use of quality statistics. This has challenged already weak and vulnerable NSSs in developing countries.

In response to this challenge, several initiatives have been launched at the international level to support developing countries to meet their respective users' needs. Among these is the Marrakech Action Plan for Statistics (MAPS) adopted in 2004 during the Second Round Table on Managing for Development Results. It consists of a global agenda aimed at improving the availability and use of quality statistics in support of PRSs according to a well-defined budget covering a specific period of time. The MAPS recommends, *inter alia*, that every low-income country designs and implements a National Strategy for the Development of Statistics (NSDS) aimed at providing the country with strategic orientations and appropriate mechanisms to guide and accelerate the development of its statistical capacity in a sustainable manner.

## Some Issues and Challenges Facing Developing Countries

Key issues and problems confronting statistical development in developing countries include: inadequate political commitment to statistical development especially at the national level; limited coordination, collaboration, networking and information sharing among stakeholders; inadequate resources (financial, human, and technical) for statistical production; inadequate infrastructure (physical and statistical) for statistical production; limited institutional capacity; poor quality data; inadequate data management (archiving, analysis, and dissemination) systems; lack of long-term planning for statistical development; and inappropriate profiles of National Statistical Offices (NSOs) in government hierarchy.

In this context, those delivering NSSs in developing countries face specific challenges including: creating greater awareness among data users and especially planners, policy makers and decision makers about the strategic importance of statistics in their work, particularly in evidence-based macro-economic management, policy formulation and poverty measurement and monitoring; playing an advocacy role to ensure that statistical production and use are given high priority by national governments; building ample capacity to make user needs assessments for data of improved quality and keep abreast of the data needs of policy makers, the private sector and civil society;

building capacity to harness technology and to improve the way data are collected and disseminated to users; building competent user groups (policy makers, researchers, media) to properly understand and interpret available statistical data; building competence in Survey Management in NSOs; and promoting co-ordination and synergy among institutions involved in statistical activities.

## Conclusion

Several efforts are being made at international, regional, sub-regional and national levels to support NSS of developing countries. In spite of these efforts, the majority of developing countries still do not have statistical systems that are capable of providing, in a sustainable manner, good statistical data and information required for evidence-based planning and policy formulation, democratic governance and accountability, political and personal decisions. It is therefore imperative that those supporting statistical development efforts in developing countries address the above-mentioned challenges to help statistics play their role of enablers of development.

## Acknowledgments

The views expressed in this paper are personal to the author and do not necessarily represent those of the United Nations Economic Commission for Africa or its subsidiary organs.

## About the Author

Dr. Dimitri Sanga is currently the Acting Director of the African Centre for Statistics (ACS) at the United Nations Economic Commission for Africa (ECA). Until end of July 2009, he was Chief of the Statistical Development and Data Management Section of the ACS. In this capacity, and formerly, he contributed to the revamping of the statistical function at ECA and most notably the inception of ACS. Before joining the United Nations, he served as Senior Economic Statistician at Statistics Canada, occupying several posts in areas such as price statistics, national accounts and household surveys undertaking and analysis. He was also part time Professor of economics, econometrics, and statistics in a number of Canadian universities namely Laval, Sherbrooke, and Ottawa. He has substantively published in refereed journals and produced a number of textbooks in economics with special interest in index number theory and practices. An elected member of the International Statistical Institute (ISI), he currently serves on the Editorial Board of the *African Statistical Journal* and the *African Statistical Newsletter*. He is also member of several international expert groups including the 2010 United Nations Expert Group on Population and Housing Censuses, the Inter Agency and Expert Group on the Millennium Development Goals Indicators, and the Inter Agency and Expert Group on Gender Statistics. He received the joint Natural Resources of Canada and Groupe de recherche en économie de l'énergie, de l'environnement et des ressources naturelles de l'Université Laval (GREEN) awards for 1993, 1994, 1995, 1996, 1997, and 1998 successively.

## Cross References

►African Population Censuses
►Aggregation Schemes
►Careers in Statistics
►Economic Growth and Well-Being: Statistical Perspective
►Measurement of Economic Progress
►Promoting, Fostering and Development of Statistics in Developing Countries
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics in Advancing Quantitative Education
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistics and Climate Change

# Rubin Causal Model

Donald B. Rubin
John L. Loeb Professor of Statistics
Harvard University, Cambridge, MA, USA

The Rubin Causal Model (RCM) is a formal mathematical framework for causal inference, first given that name by Holland (1986) for a series of previous articles developing the perspective (Rubin 1974, 1975, 1976, 1977, 1978, 1979, 1980). This framework, as formulated in these articles, has two essential parts (the definition of causal effects using the concept of potential outcomes and the definition of a model for the assignment mechanism) and one optional part (which involves the specification of a model for quantities treated as fixed by the first two parts). These three parts are briefly described, emphasizing the implications for practice. A longer encyclopedic entry on the RCM is Imbens and Rubin (2008), chapter length summaries are included in Rubin (2006, 2008) and a full-length text from this perspective is Imbens and Rubin (2010).

The first part is conceptual and implies that we should always start an inquiry into a causal question by carefully defining all causal estimands (quantities to be estimated)

in terms of potential outcomes, which are all values that could be observed in some real or hypothetical experiment comparing the results under a well-defined active treatment versus the results under a well-defined control treatment in a population of units, each of which can be exposed, in principle, to either treatment. That is, causal effects are defined by comparisons of (a) the values that would be observed if the active treatment were applied and (b) the values that would be observed if instead the control treatment were applied to this population of units. This step contrasts with the common practice of defining causal effects in observational studies in terms of parameters in some regression model, where the manipulations defining the active versus control treatments are often left implicit and ill-defined, with the resulting causal inferences correspondingly ambiguous. See, for example, the discussions by Mealli and Rubin (2003) and Angrist et al. (1996). This first step of the RCM can take place before any data are observed or even collected. The set of potential outcomes under the active treatment and the control treatment defines the "science" – the scientific objective of causal inference in all studies, whether randomized [see the entries on experiments by Hinkelman (2010) and Cox (2010)], observational or entirely hypothetical. It appears that the first use of the formal concept of potential outcomes to define causal effects was Neyman (1923) in the context of randomization-based inference in randomized experiments, but this notation was not extended to nonrandomized studies until Rubin (1974); Rubin (2010) provides some historical perspective. The science also includes covariates (background variables) that describe the units in the population.

The second part of the RCM, the formulation of the assignment mechanism, implies that after having defined the science, we should continue by explicating the design of the real or hypothetical study being used to estimate that science. The assignment mechanism mathematically describes why some study units will be (or were) exposed to the active treatment and why other study units will be (or were) exposed to the control treatment. Sometimes the assignment mechanism involves the consideration of background (i.e., pre-treatment) variables for the purpose of creating strata of similar units to be exposed to the active treatment and the control treatment, thereby improving the balance of treatment and control groups with respect to these background variables (i.e., covariates). A true experiment automatically cannot use any outcome (post-assignment) variables to influence design because they are not yet observed. If the observed data were not generated by a true experiment, but rather by an assignment mechanism corresponding to a nonrandomized observational data set, there still should be an explicit design phase. That is, in an observational study, the same guidelines as in an experiment should be followed.

More explicitly, the design step in the analysis of an observational data set for causal inference should structure the data to approximate (or reconstruct or replicate) a true randomized experiment as closely as possible. In this design step, the researcher never uses or even examines any final outcome data, but rather, identifies subsets of units such that the active and control treatments can be thought of as being randomly assigned within the subsets. This assumed randomness of treatment assignment is assessed by examining, within these subsets of units, the similarity of the distributions of the observed covariates in the treatment group and in the control group.

The third part of the RCM is optional; it derives inferences for causal effects from the observed data by conceptualizing the problem as one of imputing the missing potential outcomes. That is, once some outcome data are observed (i.e., observations of the potential outcomes corresponding to the treatments actually received by the various units), then the modeling of the outcome data given the covariates should be structured to derive predictive distributions of those potential outcomes that would have been observed if the treatment assignments had been different. This modeling generates stochastic predictions (i.e., imputations) for all missing potential outcomes in the study, which, when combined with the actually observed potential outcomes, allows the calculation of causal–effect estimands. Because the imputations of the missing potential outcomes are stochastic, repeating the process results in different values for the causal–effect estimands. This variation across the multiple imputations (Rubin 1987, 2004) generates interval estimates and tests for the causal estimands. Typically in practice, this third part is implemented using simulation-based methods, such as Markov chain Monte Carlo computation (see ▶Markov Chain Monte Carlo) applied to calculate posterior distributions under Bayesian models.

The conceptual clarity in the first two parts of the RCM often allows previously difficult causal inference situations to be easily formulated. The optional third part often extends this successful formulating by relying on modern computational power to handle analytically intractable problems. Recent, albeit somewhat idiosyncratic, examples include Hirano et al. (2000), Jin and Rubin (2008), and Rubin and Zell (2010).

## About the Author

Professor Donald B. Rubin is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University. He was Chair of the department during 1985–1994 and 2000–2004. He is an Elected Fellow/Member of: the American

Statistical Association (1977), the Institute of Mathematical Statistics (1977), the International Statistical Institute (1984), the American Association for the Advancement of Science (1984), the American Academy of Arts and Sciences (1993), and the National Academy of Sciences (2010). He has authored/coauthored over 350 publications (including 10 books) and has made important contributions to statistical theory and methodology, particularly in causal inference, design and analysis of experiments and sample surveys, treatment of missing data, and Bayesian data analysis. Among his many awards, Professor Rubin has received the Samuel S. Wilks Medal (American Statistical Association, 1995), the Parzen Prize for Statistical Innovation (1996), the Fisher Lectureship (2004), and the George W. Snedecor Award of the Committee of Presidents of Statistical Societies (2007). He was named Statistician of the Year, American Statistical Association, Boston Chapter (1995); and Statistician of the Year, American Statistical Association, Chicago Chapter (2001). He was Associate Editor for: *Journal of Educational Statistics* (1976–1979), *Theory and Methods, The Journal of American Statistical Association* (1975–1979), Editor Elect, *The Journal of American Statistical Association* (1979), Coordinating Editor and Applications Editor, *The Journal of American Statistical Association* (1980–1982), *Biometrika* (1992–1995), *Survey Methodology* (1988–1993), *Statistica Sinica* (1993-2004). Professor Rubin has been, for many years, one of the mostly cited authors in mathematics in the world (according to ISI Science Watch); in 2002 he was ranked Seventh in the World for the Decade 1991–2000. His biography is included in many places including Who's Who in The World. In 2008 Professor Rubin was elected a Honorary Member of the European Association of Methodology, and in 2009 he was elected a Corresponding (foreign) Fellow of the British Academy.

## Cross References

►Causal Diagrams
►Causation and Causal Inference
►Design of Experiments: A Pattern of Progress
►Experimental Design: An Introduction
►Imputation
►Markov Chain Monte Carlo
►Multiple Imputation
►Randomization

## References and Further Reading

Angrist J, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables (with discussion). *J Am Stat Assoc*, Applications Invited Discussion Article with discussion and rejoinder 91(434):444–472

Cox DR (2010) Design of experiments: a pattern of progress (this volume)

Hinkelman K (2010) Introduction to experimental design. (this volume)

Hirano K, Imbens G, Rubin DB, Zhou X (2000) Estimating the effect of an influenza vaccine in an encouragement design. Biostatistics 1:69–88

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

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–970

Imbens GW, Rubin DB (2010) Causal Inference in Statistics and the Medical and Social Sciences. Cambridge University Press, Cambridge, U.K

Jin H, Rubin DB (2008) Principal stratification for causal inference with extended partial compliance: application to Efron-Feldman data. J Am Stat Assoc 103:101–111

Mealli F, Rubin DB (2003) Commentary: assumptions allowing the estimation of direct causal effects, J Economet 112: 79–87

Neyman J (1923) On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in Statistical Science 5(465–480):1990

Neyman J (1935) Statistical problems in agricultural experimentation. Supplement to J R Stat Soc B 2:107–108 (with discussion). (With cooperation of K. Kwaskiewicz and St. Kolodziejczyk)

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66: 688–701

Rubin DB (1975) Bayesian inference for causality: the importance of randomization. Proceedings of the Social Statistics Section of the American Statistical Association, pp 233–239

Rubin DB (1976) Inference and missing data. Biometrika 63: 581–592

Rubin DB (1977) Assignment of treatment group on the basis of a covariate. J Educ Stat 2:1–26

Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58

Rubin DB (1979) Discussion of "conditional independence in statistical theory," by A.P. Dawid. J R Stat Soc B 41:27–28

Rubin DB (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. J Am Stat Assoc 75:591–593

Rubin DB (1987) (2004) Multiple Imputation for Nonresponse in Surveys. 1st edn. and Wiley, Classic edn. Wiley, New York

Rubin DB (2006) Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In: Rao CR, Sinharay S (eds) Handbook of Statistics, Vol. 26: Psychometrics. Elsevier, The Netherlands, Chapter 24, pp 769–800

Rubin DB (2008) Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In: Rao CR, Miller JP, Rao DC (eds) Handbook of statistics: epidemiology and medical statistics, Chapter 2. Elsevier, The Netherlands, pp 28–63

Rubin DB (2010) Reflections stimulated by the comments of Shadish (2009) and West and Thoemmes. Psychol Methods 15(1): 38–46

Rubin DB and Zell ER (2010) Dealing with noncompliance and missing outcomes in a randomized trial using Bayesian technology: prevention of perinatal sepsis clinical trial, Soweto, South Africa. Stat Methodol 7(3):338–350