L

Large Deviations and Applications

Francis Comets
Professor
Université Paris Diderot, Paris, France

Large deviations is concerned with the study of rare events and of small probabilities. Let $X_i, 1 \le i \le n$, be independent identically distributed (i.i.d.) real random variables with expectation m, and $\bar{X}_n = (X_1 + \ldots + X_n)/n$ their empirical mean. The law of large numbers shows that, for any Borel set $A \subset \mathbb{R}$ not containing m in its closure, $P(\bar{X}_n \in A) \to 0$ as $n \to \infty$, but does not tell us how fast the probability vanishes. Large deviations theory gives us the rate of decay, which is exponential in n. Cramér's theorem states that,

$$P(\bar{X}_n \in A) = \exp\big(-n\big(\inf\{I(x); x \in A\} + o(1)\big)\big)$$

as $n \to \infty$, for all interval A. The rate function I can be computed as the Legendre conjugate of the logarithmic moment generating function of X,

$$I(x) = \sup\{\lambda x - \ln E \exp(\lambda X_1); \lambda \in \mathbb{R}\},$$

and is called the Cramér transform of the common law of the X_i's. The natural assumption is the finiteness of the ►moment generating function in a neighborhood of the origin, i.e., the property of exponential tails. The function $I : \mathbb{R} \to [0, +\infty]$ is convex with $I(m) = 0$.

- In the Gaussian case $X_i \sim \mathcal{N}(m, \sigma^2)$, we find $I(x) = (x-m)^2/(2\sigma^2)$.
- In the Bernoulli case $P(X_i = 1) = p = 1 - P(X_i = 0)$, we find the entropy function $I(x) = x \ln(x/p) + (1-x) \ln(1-x)/(1-p)$ for $x \in [0,1]$, and $I(x) = +\infty$ otherwise.

To emphasize the importance of rare events, let us mention a consequence, the Erdös–Rényi law: consider an infinite sequence $X_i, i \ge 1$, of Bernoulli i.i.d. variables with parameter p, and let R_n denote the length of the longest consecutive run, contained within the first n tosses, in which the fraction of 1s is at least a ($a > p$). Erdös and Rényi proved that, almost surely as $n \to \infty$,

$$R_n / \ln n \longrightarrow I(a)^{-1},$$

with the function I from the Bernoulli case above. Though it may look paradoxical, large deviations are at the core of this event of full probability. This result is the basis of ►bioinformatics applications like sequence matching, and of statistical tests for sequence randomness.

The theory does not only apply to independent variables, but allows for many variations, including weakly dependent variables in a general state space, Markov or ►Gaussian processes, large deviations from ►ergodic theorems, non-asymptotic bounds, asymptotic expansions (Edgeworth expansions), etc.

Here is the formal definition. Given a Polish space (i.e., a separable complete metric space) $\mathcal{X}$, let $\{\mathbb{P}_n\}$ be a sequence of Borel probability measures on $\mathcal{X}$, let a_n be a positive sequence tending to infinity, and finally let $I : \mathcal{X} \to [0, +\infty]$ be a lower semicontinuous functional on X. We say that the sequence $\{\mathbb{P}_n\}$ satisfies a large deviation principle with speed a_n and rate I, if for each measurable set $E \subset X$

$$-\inf_{x \in E^\circ} I(x) \le \underline{\lim}_n a_n^{-1} \ln \mathbb{P}_n(E) \le \overline{\lim}_n a_n^{-1} \ln \mathbb{P}_n(E) \le -\inf_{x \in \bar{E}} I(x)$$

where $\bar{E}$ and E° denote respectively the closure and interior of E. The rate function can be obtained as

$$I(x) = -\lim_{\delta \searrow 0} \lim_{n \to \infty} a_n^{-1} \ln \mathbb{P}_n(B(x, \delta)),$$

with $B(x, \delta)$ the ball of center x and radius δ.

Sanov's theorem and sampling with replacement: let μ be a probability measure on a set Σ that we assume finite for simplicity, with $\mu(y) > 0$ for all $y \in \Sigma$. Let $Y_i, i \ge 1$, an i.i.d. sequence with law μ, and N_n the score vector of the n-sample,

$$N_n(y) = \sum_{i=1}^{n} \mathbf{1}_y(Y_i).$$

By the law of large numbers, $N_n/n \to \mu$ almost surely. From the ►multinomial distribution, one can check that, for all ν such that $n\nu$ is a possible score vector for the n-sample,

$$(n+1)^{-|\Sigma|} e^{-nH(\nu|\mu)} \le P(n^{-1}N_n = \nu) \le e^{-nH(\nu|\mu)},$$

where $H(\nu|\mu) = \sum_{y \in \Sigma} \nu(y) \ln \frac{\nu(y)}{\mu(y)}$ is the relative entropy of ν with respect to μ. The large deviations theorem holds

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

for the empirical distribution of a general n-sample, with speed n and rate $I(\nu) = H(\nu|\mu)$ given by the natural generalization of the above formula. This result, due to Sanov, has many consequences in information theory and statistical mechanics (Dembo and Zeitouni 1998; den Hollander 2000), and for exponential families in statistics. Applications in statistics also include point estimation (by giving the exponential rate of convergence of M-estimators) and for hypothesis testing (Bahadur efficiency) (Kester 1985), and concentration inequalities (Dembo and Zeitouni 1998).

The *Freidlin–Wentzell theory* deals with diffusion processes with small noise,

$$dX_t^\epsilon = b\left(X_t^\epsilon\right) dt + \sqrt{\epsilon}\, \sigma\left(X_t^\epsilon\right) dB_t\,, \qquad X_0^\epsilon = y.$$

The coefficients b, σ are uniformly lipshitz functions, and B is a standard Brownian motion (see ►Brownian Motion and Diffusions). The sequence X^ϵ can be viewed as $\epsilon \searrow 0$ as a small random perturbation of the ordinary differential equation

$$dx_t = b(x_t)dt\,, \qquad x_0 = y.$$

Indeed, $X^\epsilon \to x$ in the supremum norm on bounded time-intervals. Freidlin and Wentzell have shown that, on a finite time interval $[0, T]$, the sequence X^ϵ with values in the path space obeys the LDP with speed ϵ^{-1} and rate function

$$I(\phi) = \frac{1}{2}\int_0^T \sigma(\phi(t))^{-2}\left(\dot{\phi}(t) - b(\phi(t))\right)^2 dt$$

if ϕ is absolutely continuous with square-integrable derivative and $\phi(0) = y$; $I(\phi) = \infty$ otherwise. (To fit in the above formal definition, take a sequence $\epsilon = \epsilon_n \searrow 0$, and for $\mathbb{P}_n$ the law of X^{ϵ_n}.)

The Freidlin–Wentzell theory has applications in physics (metastability phenomena) and engineering (tracking loops, statistical analysis of signals, stabilization of systems, and algorithms) (Freidlin and Wentzell 1998; Dembo and Zeitouni 1998; Olivieri and Vares 2005).

About the Author

Francis Comets is Professor of Applied Mathematics at University of Paris - Diderot (Paris 7), France. He is the Head of the team "Stochastic Models" in the CNRS laboratory "Probabilité et modèles aléatoires" since 1999. He is the Deputy Director of the Foundation Sciences Mathématiques de Paris since its creation in 2006. He has coauthored 50 research papers, with a focus on random medium, and one book ("Calcul stochastique et modèles de diffusions" in French, Dunod 2006, with Thierry Meyre). He has supervised more than 10 Ph.D thesis, and served as Associate Editor for *Stochastic Processes and their Applications* (1997–2002).

Cross References

►Asymptotic Relative Efficiency in Testing
►Chernoff Bound
►Entropy and Cross Entropy as Diversity and Distance Measures
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Moderate Deviations
►Robust Statistics

References and Further Reading

Dembo A, Zeitouni O (1998) Large deviations techniques and applications. Springer, New York
den Hollander F (2000) Large deviations. Am Math Soc, Providence, RI
Freidlin MI, Wentzell AD (1998) Random perturbations of dynamical systems. Springer, New York
Kester A (1985) Some large deviation results in statistics. CWI Tract, 18. Centrum voor Wiskunde en Informatica, Amsterdam
Olivieri E, Vares ME (2005) Large deviations and metastability. Cambridge University Press, Cambridge

Laws of Large Numbers

Andrew Rosalsky
Professor
University of Florida, Gainesville, FL, USA

The *laws of large numbers* (LLNs) provide bounds on the fluctuation behavior of sums of random variables and, as we will discuss herein, lie at the very foundation of statistical science. They have a history going back over 300 years. The literature on the LLNs is of epic proportions, as this concept is indispensable in probability and statistical theory and their application.

Probability theory, like some other areas of mathematics such as geometry for example, is a subject arising from an attempt to provide a rigorous mathematical model for real world phenomena. In the case of probability theory, the real world phenomena are chance behavior of biological processes or physical systems such as gambling games and their associated monetary gains or losses.

The probability of an event is the abstract counterpart to the notion of the long-run relative frequency of the

occurence of the event through infinitely many replications of the experiment. For example, if a quality control engineer asserts that the probability is 0.98 that a widget produced by her production team meets specifications, then she is asserting that in the long-run, 98% of those widgets meet specifications. The phrase "in the long-run" requires the notion of *limit* as the sample size approaches infinity. The long-run relative frequency approach for describing the probability of an event is natural and intuitive but, nevertheless, it raises serious mathematical questions. Does the limiting relative frequency always exist as the sample size approaches infinity and is the limit the same irrespective of the sequence of experimental outcomes? It is easy to see that the answers are negative. Indeed, in the above example, depending on the sequence of experimental outcomes, the proportion of widgets meeting specifications could fluctuate repeatedly from near 0 to near 1 as the number of widgets sampled approaches infinity. So in what sense can it be asserted that the limit exists and is 0.98? To provide an answer to this question, one needs to apply a LLN.

The LLNs are of two types, viz., *weak* LLNs (WLLNs) and *strong* LLNs (SLLNs). Each type involves a different mode of convergence. In general, a WLLN (resp., a SLLN) involves convergence in probability (resp., convergence almost surely (a.s.)). The definitions of these two modes of convergence will now be reviewed.

Let $\{U_n, n \geq 1\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let $c \in \mathbb{R}$. We say that U_n *converges in probability* to c (denoted $U_n \xrightarrow{P} c$) if

$$\lim_{n\to\infty} P(|U_n - c| > \varepsilon) = 0 \text{ for all } \varepsilon > 0.$$

We say that U_n *converges a.s.* to c (denoted $U_n \to c$ a.s.) if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} U_n(\omega) = c\}) = 1.$$

If $U_n \to c$ a.s., then $U_n \xrightarrow{P} c$; the converse is not true in general.

The celebrated Kolmogorov SLLN (see, e.g., Chow and Teicher [1997], p. 125) is the following result. Let $\{X_n, n \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables and let $c \in \mathbb{R}$. Then

$$\frac{\sum_{i=1}^n X_i}{n} \to c \text{ a.s. if and only if } EX_1 = c. \tag{1}$$

Using statistical terminology, the sufficiency half of (1) asserts that the *sample mean* converges a.s. to the *population mean* as the *sample size n* approaches infinity provided the population mean exists and is finite. This result is of fundamental importance in statistical science. It follows from (1) that

$$\text{if } EX_1 = c \in \mathbb{R}, \text{ then } \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{P} c; \tag{2}$$

this result is the Khintchine WLLN (see, e.g., Petrov [1995], p. 134).

Next, suppose $\{A_n, n \geq 1\}$ is a sequence of independent events all with the same probability p. A special case of the Kolmogorov SLLN is the limit result

$$\hat{p}_n \to p \text{ a.s.} \tag{3}$$

where $\hat{p}_n = \sum_{i=1}^n I_{A_i}/n$ is the proportion of $\{A_1, \ldots, A_n\}$ to occur, $n \geq 1$. (Here I_{A_i} is the indicator function of $A_i, i \geq 1$.) This result is the first SLLN ever proved and was discovered by Emile Borel in 1909. Hence, with probability 1, the *sample proportion* $\hat{p}_n$ approaches the *population proportion* p as the sample size $n \to \infty$. It is this SLLN which thus provides the theoretical justification for the long-run relative frequency approach to interpreting probabilities. Note, however, that the convergence in (3) is not pointwise on Ω but, rather, is pointwise on some subset of Ω *having probability 1.* Consequently, any interpretation of $p = P(A_1)$ via (3) necessitates that one has *a priori* an intuitive understanding of the notion of an event having probability 1.

The SLLN (3) is a key component in the proof of the Glivenko–Cantelli theorem (see ►Glivenko-Cantelli Theorems) which, roughly speaking, asserts that with probability 1, a population distribution function can be uniformly approximated by a sample (or empirical) distribution function as the sample size approaches infinity. This result is referred to by Rényi (1970, p. 400) as the *fundamental theorem of mathematical statistics* and by Loève (1977, p. 20) as the *central statistical theorem*.

In 1689, Jacob Bernoulli (1654–1705) proved the first WLLN

$$\hat{p}_n \xrightarrow{P} p. \tag{4}$$

Bernoulli's renowned book *Ars Conjectandi* (*The Art of Conjecturing*) was published posthumously in 1713, and it is here where the proof of his WLLN was first published. It is interesting to note that there is over a 200 year gap between the WLLN (4) of Bernoulli and the corresponding SLLN (3) of Borel.

An interesting example is the following modification of one of Stout (1974, p. 9). Suppose that the quality control engineer referred to above would like to estimate the proportion p of widgets produced by her production team

that meet specifications. She estimates p by using the proportion $\hat{p}_n$ of the first n widgets produced that meet specifications and she is interested in knowing if there will ever be a point in the sequence of examined widgets such that with probability (at least) a specified large value, $\hat{p}_n$ will be within ε of p *and stay* within ε of p as the sampling continues (where $\varepsilon > 0$ is a prescribed tolerance). The answer is affirmative since (3) is equivalent to the assertion that for a given $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $N_{\varepsilon,\delta}$ such that

$$P\left(\cap_{n=N_{\varepsilon,\delta}}^{\infty}[|\hat{p}_n - p| \le \varepsilon]\right) \ge 1 - \delta.$$

That is, the probability is arbitrarily close to 1 that $\hat{p}_n$ will be arbitrarily close to p simultaneously for all n beyond some point. If one applied instead the WLLN (4), then it could only be asserted that for a given $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $N_{\varepsilon,\delta}$ such that

$$P(|\hat{p}_n - p| \le \varepsilon) \ge 1 - \delta \text{ for all } n \ge N_{\varepsilon,\delta}.$$

There are numerous other versions of the LLNs and we will discuss only a few of them. Note that the expressions in (1) and (2) can be rewritten, respectively, as

$$\frac{\sum_{i=1}^n X_i - nc}{n} \to 0 \text{ a.s. and } \frac{\sum_{i=1}^n X_i - nc}{n} \xrightarrow{P} 0$$

thereby suggesting the following definitions. A sequence of random variables $\{X_n, n \ge 1\}$ is said to obey a general *SLLN* (resp., *WLLN*) with *centering sequence* $\{a_n, n \ge 1\}$ and *norming sequence* $\{b_n, n \ge 1\}$ (where $0 < b_n \to \infty$) if

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \to 0 \text{ a.s. } \left(\text{resp., } \frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{P} 0\right).$$

A famous result of Marcinkiewicz and Zygmund (see, e.g., Chow and Teicher (1997), p. 125) extended the Kolmogorov SLLN as follows. Let $\{X_n, n \ge 1\}$ be a sequence of i.i.d. random variables and let $0 < p < 2$. Then

$$\frac{\sum_{i=1}^n X_i - nc}{n^{1/p}} \to 0 \text{ a.s. for some } c \in \mathbb{R} \text{ if and only if } E|X_1|^p < \infty.$$

In such a case, necessarily $c = EX_1$ if $p \ge 1$ whereas c is arbitrary if $p < 1$.

Feller (1946) extended the Marcinkiewicz–Zygmund SLLN to the case of a more general norming sequence $\{b_n, n \ge 1\}$ satisfying suitable growth conditions.

The following WLLN is ascribed to Feller by Chow and Teicher (1997, p. 128). If $\{X_n, n \ge 1\}$ is a sequence of i.i.d. random variables, then there exist real numbers $a_n, n \ge 1$ such that

$$\frac{\sum_{i=1}^n X_i - a_n}{n} \xrightarrow{P} 0 \tag{5}$$

if and only if

$$nP(|X_1| > n) \to 0 \text{ as } n \to \infty. \tag{6}$$

In such a case, $a_n - nE(X_1 I_{[|X_1| \le n]}) \to 0$ as $n \to \infty$.

The condition (6) is weaker than $E|X_1| < \infty$. If $\{X_n, n \ge 1\}$ is a sequence of i.i.d. random variables where X_1 has probability density function

$$f(x) = \begin{cases} \frac{c}{x^2 \log|x|}, & |x| \ge e \\ 0, & |x| < e \end{cases}$$

where c is a constant, then $E|X_1| = \infty$ and the SLLN $\sum_{i=1}^n X_i/n \to c$ a.s. fails for every $c \in \mathbb{R}$ but (6) and hence the WLLN (5) hold with $a_n = 0, n \ge 1$.

Klass and Teicher (1977) extended the Feller WLLN to the case of a more general norming sequence $\{b_n, n \ge 1\}$ thereby obtaining a WLLN analog of Feller's (1946) extension of the Marcinkiewicz–Zygmund SLLN.

Good references for studying the LLNs are the books by Révész (1968), Stout (1974), Loève (1977), Chow and Teicher (1997), and Petrov (1995). While the LLNs have been studied extensively in the case of independent summands, some of the LLNs presented in these books involve summands obeying a dependence structure other than that of independence.

A large literature of investigation on the LLNs for sequences of Banach space valued random elements has emerged beginning with the pioneering work of Mourier (1953). See the monograph by Taylor (1978) for background material and results up to 1978. Excellent references are the books by Vakhania, Tarieladze, and Chobanyan (1987) and Ledoux and Talagrand (1991). More recent results are provided by Adler et al. (1991), Cantrell and Rosalsky (2004), and the references in these two articles.

About the Author

Professor Rosalsky is Associate Editor, *Journal of Applied Mathematics and Stochastic Analysis* (1989–present), and Associate Editor, *International Journal of Mathematics and Mathematical Sciences* (1994–present). He has collaborated with research workers from 15 different countries. At the University of Florida, he twice received a Teaching Improvement Program Award and on five occasions was named an Anderson Scholar Faculty Honoree, an honor reserved for faculty members designated by undergraduate honor students as being the most effective and inspiring.

Cross References

►Almost Sure Convergence of Random Variables
►Borel–Cantelli Lemma and Its Generalizations

►Chebyshev's Inequality
►Convergence of Random Variables
►Ergodic Theorem
►Estimation: An Overview
►Expected Value
►Foundations of Probability
►Glivenko-Cantelli Theorems
►Probability Theory: An Outline
►Random Field
►Statistics, History of
►Strong Approximations in Probability and Statistics

References and Further Reading

Adler A, Rosalsky A, Taylor RL (1991) A weak law for normed weighted sums of random elements in Rademacher type *p* Banach spaces. J Multivariate Anal 37:259–268

Cantrell A, Rosalsky A (2004) A strong law for compactly uniformly integrable sequences of independent random elements in Banach spaces. Bull Inst Math Acad Sinica 32:15–33

Chow YS, Teicher H (1997) Probability theory: independence, interchangeability, martingales, 3rd edn. Springer, New York

Feller W (1946) A limit theorem for random variables with infinite moments. Am J Math 68:257–262

Klass M, Teicher H (1977) Iterated logarithm laws for asymmetric random variables barely with or without finite mean. Ann Probab 5:861–874

Ledoux M, Talagrand M (1991) Probability in Banach spaces: isoperimetry and processes. Springer, Berlin

Loève M (1977) Probability theory, vol I, 4th edn. Springer, New York

Mourier E (1953) Eléments aléatoires dans un espace de Banach. Annales de l' Institut Henri Poincaré 13:161–244

Petrov VV (1995) Limit theorems of probability theory: sequences of independent random variables. Clarendon, Oxford

Rényi A (1970) Probability theory. North-Holland, Amsterdam

Révész P (1968) The laws of large numbers. Academic, New York

Stout WF (1974) Almost sure convergence. Academic, New York

Taylor RL (1978) Stochastic convergence of weighted sums of random elements in linear spaces. Lecture notes in mathematics, vol 672. Springer, Berlin

Vakhania NN, Tarieladze VI, Chobanyan SA (1987) Probability distributions on Banach spaces. D. Reidel, Dordrecht, Holland

Learning Statistics in a Foreign Language

Khidir M. Abdelbasit
Sultan Qaboos University, Muscat, Sultanate of Oman

Background

The Sultanate of Oman is an Arabic-speaking country, where the medium of instruction in pre-university education is Arabic. In Sultan Qaboos University (SQU) all sciences (including Statistics) are taught in English. The reason is that most of the scientific literature is in English and teaching in the native language may leave graduates at a disadvantage. Since only few instructors speak Arabic, the university adopts a policy of no communication in Arabic in classes and office hours. Students are required to achieve a minimum level in English (about 4.0 IELTS score) before they start their study program. Very few students achieve that level on entry and the majority spends about two semesters doing English only.

Language and Cultural Problems

It is to be expected that students from a non-English-speaking background will face serious difficulties when learning in English especially in the first year or two. Most of the literature discusses problems faced by foreign students pursuing study programs in an English-speaking country, or a minority in a multi-cultural society (see for example Coutis P. and Wood L., Hubbard R, Koh E). Such students live (at least while studying) in an English-speaking community with which they have to interact on a daily basis. These difficulties are more serious for our students who are studying in their own country where English is not the official language. They hardly use English outside classrooms and avoid talking in class as much as they can.

My SQU Experience

Statistical concepts and methods are most effectively taught through real-life examples that the students appreciate and understand. We use the most popular textbooks in the USA for our courses. These textbooks use this approach with US students in mind. Our main problems are:

- Most of the examples and exercises used are completely alien to our students. The discussions meant to maintain the students' interest only serve to put ours off. With limited English they have serious difficulties understanding what is explained and hence tend not to listen to what the instructor is saying. They do not read the textbooks because they contain pages and pages of lengthy explanations and discussions they cannot follow. A direct effect is that students may find the subject boring and quickly lose interest. Their attention then turns to the art of passing tests instead of acquiring the intended knowledge and skills. To pass their tests they use both their class and study times looking through examples, concentrating on what formula to use and where to plug the numbers they have to get the answer. This way they manage to do the mechanics fairly well, but the concepts are almost entirely missed.

- The problem is worse with introductory probability courses where games of chance are extensively used as illustrative examples in textbooks. Most of our students have never seen a deck of playing cards and some may even be offended by discussing card games in a classroom.

The burden of finding strategies to overcome these difficulties falls on the instructor. Statistical terms and concepts such as parameter/statistic, sampling distribution, unbiasedness, consistency, sufficiency, and ideas underlying hypothesis testing are not easy to get across even in the students' own language. To do that in a foreign language is a real challenge. For the Statistics program to be successful, all (or at least most of the) instructors involved should be up to this challenge. This is a time-consuming task with little reward, other than self satisfaction. In SQU the problem is compounded further by the fact that most of the instructors are expatriates on short-term contracts who are more likely to use their time for personal career advancement, rather than time-consuming community service jobs.

What Can Be Done?

For our first Statistics course we produced a manual that contains very brief notes and many samples of previous quizzes, tests, and examinations. It contains a good collection of problems from local culture to motivate the students. The manual was well received by the students, to the extent that students prefer to practice with examples from the manual rather than the textbook.

Textbooks written in English that are brief and to the point are needed. These should include examples and exercises from the students' own culture. A student trying to understand a specific point gets distracted by lengthy explanations and discouraged by thick textbooks to begin with. In a classroom where students' faces clearly indicate that you have got nothing across, it is natural to try explaining more using more examples. In the end of semester evaluation of a course I taught, a student once wrote "The instructor explains things more than needed. He makes simple points difficult." This indicates that, when teaching in a foreign language, lengthy oral or written explanations are not helpful. A better strategy will be to explain concepts and techniques briefly and provide plenty of examples and exercises that will help the students absorb the material by osmosis. The basic statistical concepts can only be effectively communicated to students in their own language. For this reason textbooks should contain a good glossary where technical terms and concepts are explained using the local language.

I expect such textbooks to go a long way to enhance students' understanding of Statistics. An international project can be initiated to produce an introductory statistics textbook, with different versions intended for different geographical areas. The English material will be the same; the examples vary, to some extent, from area to area and glossaries in local languages. Universities in the developing world, naturally, look at western universities as models, and international (western) involvement in such a project is needed for it to succeed. The project will be a major contribution to the promotion of understanding Statistics and excellence in statistical education in developing countries. The international statistical institute takes pride in supporting statistical progress in the developing world. This project can lay the foundation for this progress and hence is worth serious consideration by the institute.

Cross References

►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistical Literacy, Reasoning, and Thinking
►Statistics Education

References and Further Reading

Coutis P, Wood L (2002) Teaching statistics and academic language in culturally diverse classrooms. http://www.math.uoc.gr/~ictm2/Proceedings/pap172.pdf

Hubbard R (1990) Teaching statistics to students who are learning in a foreign language. ICOTS 3

Koh E (1994) Teaching statistics to students with limited language skills using MINITAB http://archives.math.utk.edu/ICTCM/VOL07/C012/paper.pdf

Least Absolute Residuals Procedure

Richard William Farebrother
Honorary Reader in Econometrics
Victoria University of Manchester, Manchester, UK

For $i = 1, 2, \ldots, n$, let $\{x_{i1}, x_{i2}, \ldots, x_{iq}, y_i\}$ represent the ith observation on a set of $q + 1$ variables and suppose that we wish to fit a linear model of the form

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q + \epsilon_i$$

to these n observations. Then, for $p > 0$, the L_p-norm fitting procedure chooses values for $b_1, b_2, \ldots, b_q$ to minimise the L_p-norm of the residuals $[\sum_{i=1}^{n} |e_i|^p]^{1/p}$ where, for $i = 1, 2, \ldots, n$, the ith residual is defined by

$$e_i = y_i - x_{i1}b_1 - x_{i2}b_2 - \ldots - x_{iq}b_q.$$

The most familiar L_p-norm fitting procedure, known as the ►least squares procedure, sets $p = 2$ and chooses values for $b_1, b_2, \ldots, b_q$ to minimize the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$.

A second choice, to be discussed in the present article, sets $p = 1$ and chooses $b_1, b_2, \ldots, b_q$ to minimize the sum of the absolute residuals $\sum_{i=1}^{n} |e_i|$.

A third choice sets $p = \infty$ and chooses $b_1, b_2, \ldots, b_q$ to minimize the largest absolute residual $max_{i=1}^{n}|e_i|$.

Setting $u_i = e_i$ and $v_i = 0$ if $e_i \geq 0$ and $u_i = 0$ and $v_i = -e_i$ if $e_i < 0$, we find that $e_i = u_i - v_i$ so that the least absolute residuals (*LAR*) fitting problem chooses $b_1, b_2, \ldots, b_q$ to minimize the sum of the absolute residuals

$$\sum_{i=1}^{n}(u_i + v_i)$$

subject to

$$x_{i1}b_1 + x_{i2}b_2 + \cdots + x_{iq}b_q + U_i - v_i = y_i \quad \text{for } i = 1, 2, \ldots, n$$

$$\text{and} \quad U_i \geq 0, v_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n.$$

The *LAR* fitting problem thus takes the form of a linear programming problem and is often solved by means of a variant of the dual simplex procedure.

Gauss has noted (when $q = 2$) that solutions of this problem are characterized by the presence of a set of q zero residuals. Such solutions are robust to the presence of outlying observations. Indeed, they remain constant under variations in the other $n - q$ observations provided that these variations do not cause any of the residuals to change their signs.

The *LAR* fitting procedure corresponds to the maximum likelihood estimator when the ϵ-disturbances follow a double exponential (Laplacian) distribution. This estimator is more robust to the presence of outlying observations than is the standard least squares estimator which maximizes the likelihood function when the ϵ-disturbances are normal (Gaussian). Nevertheless, the *LAR* estimator has an asymptotic normal distribution as it is a member of Huber's class of M-estimators.

There are many variants of the basic *LAR* procedure but the one of greatest historical interest is that proposed in 1760 by the Croatian Jesuit scientist Rugjer (or Rudjer) Josip Bošković (1711–1787) (Latin: Rogerius Josephus Boscovich; Italian: Ruggiero Giuseppe Boscovich). In his variant of the standard *LAR* procedure, there are two explanatory variables of which the first is constant $x_{i1} = 1$ and the values of b_1 and b_2 are constrained to satisfy the adding-up condition $\sum_{i=1}^{n}(y_i - b_1 - x_{i2}b_2) = 0$ usually associated with the least squares procedure developed by Gauss in 1795 and published by Legendre in 1805. Computer algorithms implementing this variant of the *LAR* procedure with $q \geq 2$ variables are still to be found in the literature.

For an account of recent developments in this area, see the series of volumes edited by Dodge (1987, 1992, 1997, 2002). For a detailed history of the *LAR* procedure, analyzing the contributions of Bošković, Laplace, Gauss, Edgeworth, Turner, Bowley and Rhodes, see Farebrother (1999). And, for a discussion of the geometrical and mechanical representation of the least squares and *LAR* fitting procedures, see Farebrother (2002).

About the Author

Richard William Farebrother was a member of the teaching staff of the Department of Econometrics and Social Statistics in the (Victoria) University of Manchester from 1970 until 1993 when he took early retirement (he has been blind since 1993). From 1993 until 2001 he was an Honorary Reader in Econometrics in the Department of Economic Studies of the same University. He has published three books: *Linear Least Squares Computations* in 1988, *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900* in 1999, and *Visualizing statistical Models and Concepts* in 2002. He has also published more than 140 research papers in a wide range of subject areas including econometric theory, computer algorithms, statistical distributions, statistical inference, and the history of statistics. Dr. Farebrother was also visiting Associate Professor (1982) at Monash University, Australia, and Visiting Professor (1990) at the Institute for Advanced Studies in Vienna, Austria.

Cross References

►Least Squares
►Residuals

References and Further Reading

Dodge Y (ed) (1987) Statistical data analysis based on the L_1-norm and related methods. North-Holland, Amsterdam
Dodge Y (ed) (1992) L_1-statistical analysis and related methods. North-Holland, Amsterdam
Dodge Y (ed) (1997) L_1-statistical procedures and related topics. Institute of Mathematical Statistics, Hayward
Dodge Y (ed) (2002) Statistical data analysis based on the L_1-norm and related methods. Birkhäuser, Basel

Farebrother RW (1999) Fitting linear relationships: a history of the calculus of observations 1750–1900. Springer, New York
Farebrother RW (2002) Visualizing statistical models and concepts. Marcel Dekker, New York

Least Squares

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

Least Squares (LS) problem involves some algebraic and numerical techniques used in "solving" overdetermined systems $F(x) \approx b$ of equations, where $b \in R^n$ while $F(x)$ is a column of the form

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \dots \\ f_{m-1}(x) \\ f_m(x) \end{bmatrix}$$

with entries $f_i = f_i(x)$, $i = 1, \dots n$, where $x = (x_1, \dots, x_p)^T$. The LS problem is *linear* when each f_i is a linear function, and *nonlinear* – if not.

Linear LS problem refers to a system $Ax = b$ of linear equations. Such a system is overdetermined if $n > p$. If $b \notin range(A)$ the system has no proper solution and will be denoted by $Ax \approx b$. In this situation we are seeking for a solution of some optimization problem. The name "Least Squares" is justified by the l_2-norm commonly used as a measure of imprecision.

The LS problem has a clear statistical interpretation in regression terms. Consider the usual regression model

$$y_i = f_i(x_{i1}, \dots, x_{ip}; \beta_1, \dots, \beta_p) + e_i \text{ for } i = 1, \dots, n \qquad (1)$$

where x_{ij}, $i = 1, \dots n, j = 1, \dots p$, are some constants given by experimental design, f_i, $i = 1, \dots n$, are given functions depending on unknown parameters β_j, $j = 1, \dots, p$, while y_i, $i = 1, \dots, n$, are values of these functions, observed with some random errors e_i. We want to estimate the unknown parameters β_i on the basis of the data set $\{x_{ij}, y_i\}$.

In linear regression each f_i is a linear function of type $f_i = \sum_{j=1}^{p} c_{ij}(x_1, \dots, x_n)\beta_j$ and the model (1) may be presented in vector-matrix notation as

$$y = X\beta + e,$$

where $y = (y_1, \dots, y_n)^T$, $e = (e_1, \dots, e_n)^T$ and $\beta = (\beta_1, \dots, \beta_p)^T$, while X is a $n \times p$ matrix with entries x_{ij}. If $e_1, \dots, e_n$ are not correlated with mean zero and a common (perhaps unknown) variance then the problem of Best Linear Unbiased Estimation (BLUE) of β reduces to finding a vector $\widehat{\beta}$ that minimizes the norm $\left\| y - X\widehat{\beta} \right\|_2 = (y - X\widehat{\beta})^T (y - X\widehat{\beta})$

Such a vector is said to be the *ordinary* LS solution of the overparametrized system $X\beta \approx y$. On the other hand the last one reduces to solving the consistent system

$$X^T X\beta = X^T y$$

of linear equations called normal equations. In particular, if $rank(X) = p$ then the system has a unique solution of the form

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

For linear regression $y_i = \alpha + \beta x_i + e_i$ with one regressor x the BLU estimators of the parameters α and β may be presented in the convenient form as

$$\widehat{\beta} = \frac{ns_{xy}}{ns^2 x} \quad \text{and} \quad \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x},$$

where

$$ns_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n},$$

$$ns_x^2 = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}, \quad \bar{x} = \frac{\sum_i x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_i y_i}{n}$$

For its computation we only need to use a simple pocket calculator.

Example The following table presents the number of residents in thousands (x) and the unemployment rate in % (y) for some cities of Poland. Estimate the parameters β and α.

x_i	131	87	56	312	185	252	157
y_i	8.7	10.2	9.9	6.3	6.1	5.2	11.0

In this case $\sum_i x_i = 1,180$, $\sum_i y_i = 57.4$, $\sum_i x_i^2 = 247,588$ and $\sum_i x_i y_i = 8,713$. Therefore $ns_x^2 = 48,673.71$ and $ns_{xy} = -963$. Thus $\widehat{\beta} = -0.02$ and $\widehat{\alpha} = 11.77$ and hence $f(x) = -0.02x + 11.77$.

If the variance–covariance matrix of the error vector e coincides (except a multiplicative scalar σ^2) with a positive definite matrix V then the Best Linear Unbiased estimation reduces to the minimization of $(y - X\widehat{\beta})^T V(y - X\widehat{\beta})$, called the *weighed* LS problem. Moreover, if $rank(X) = p$ then its solution is given in the form

$$\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

It is worth to add that a nonlinear LS problem is more complicated and its explicit solution is usually not known. Instead of this some algorithms are suggested.

Total least squares problem. The problem has been posed in recent years in numerical analysis as an alternative for the LS problem in the case when all data are affected by errors.

Consider an overdetermined system of n linear equations $Ax \approx b$ with k unknown x. The TLS problem consists in minimizing the Frobenius norm

$$\left\| [A, b] - \left[\widehat{A}, \widehat{b}\right] \right\|_F$$

for all $\widehat{A} \in R^{n \times k}$ and $\widehat{b} \in range(\widehat{A})$, where the Frobenius norm is defined by $\|(a_{ij})\|_F = \sum_{i,j} a_{ij}$. Once a minimizing $\left[\widehat{A}, \widehat{b}\right]$ is found, then any x satisfying $\widehat{A}x = \widehat{b}$ is called a *TLS solution* of the initial system $Ax \approx b$.

The trouble is that the minimization problem may not be solvable, or its solution may not be unique. As an example one can set

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } b = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

It is known that the TLS solution (if exists) is always better than the ordinary LS in the sense that the correction $b - A\widehat{x}$ has smaller l_2-norm. The main tool in solving the TLS problems is the following Singular Value Decomposition:

For any matrix A of $n \times k$ with real entries there exist orthonormal matrices $P = \left[p_1, \ldots, p_n\right]$ and $Q = \left[q_1, \ldots, q_k\right]$ such that

$$P^T AQ = diag(\sigma_1, \ldots, \sigma_m), \text{ where } \sigma_1 \geq \cdots \geq \sigma_m \text{ and } m = \min\{n, k\}.$$

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Absolute Penalty Estimation
►Adaptive Linear Regression
►Analysis of Areal and Spatial Interaction Data
►Autocorrelation in Regression
►Best Linear Unbiased Estimation in Linear Models
►Estimation
►Gauss-Markov Theorem
►General Linear Models
►Least Absolute Residuals Procedure
►Linear Regression Models
►Nonlinear Models
►Optimum Experimental Design
►Partial Least Squares Regression Versus Other Methods
►Simple Linear Regression
►Statistics, History of
►Two-Stage Least Squares

References and Further Reading

Björk A (1996) Numerical methods for least squares problems. SIAM, Philadelphia
Kariya T (2004) Generalized least squares. Wiley, New York
Rao CR, Toutenberg H (1995) Linear models, least squares and alternatives. Springer, New York
Van Huffel S, Vandewalle J (1991) The total least squares problem: computational aspects and analysis. SIAM, Philadelphia
Wolberg J (2005) Data analysis using the method of least squares: extracting the most information from experiments. Springer, New York

Lévy Processes

Mohamed Abdel-Hameed
Professor
United Arab Emirates University, Al Ain, United Arab Emirates

Introduction

Lévy processes have become increasingly popular in engineering (reliability, dams, telecommunication) and mathematical finance. Their applications in reliability stems from the fact that they provide a realistic model for the degradation of devices, while their applications in the mathematical theory of dams as they provide a basis for describing the water input of dams. Their popularity in finance is because they describe the financial markets in a more accurate way than the celebrated Black–Scholes model. The latter model assumes that the rate of returns on assets are normally distributed, thus the process describing the asset price over time is continuous process. In reality, the asset prices have jumps or spikes, and the asset returns exhibit fat tails and ►skewness, which negates the normality assumption inherited in the Black–Scholes model.

Because of the deficiencies in the Black–Scholes model researchers in mathematical finance have been trying to find more suitable models for asset prices. Certain types of Lévy processes have been found to provide a good model for creep of concrete, fatigue crack growth, corroded steel gates, and chloride ingress into concrete. Furthermore, certain types of Lévy processes have been used to model the water input in dams.

In this entry, we will review Lévy processes and give important examples of such processes and state some references to their applications.

Lévy Processes

A stochastic process $X = \{X_t, t \geq 0\}$ that has right continuous sample paths with left limits is said to be a Lévy process if the following hold:

1. X has *stationary increments*, i.e., for every $s, t \geq 0$, the distribution of $X_{t+s} - X_t$ is independent of t.
2. X has independent increments, i.e., for every $t, s \geq 0$, $X_{t+s} - X_t$ is independent of $(X_u, u \leq t)$.
3. X is stochastically continuous, i.e., for every $t \geq 0$ and $\epsilon > 0$:

$$lim_{s \to t} P(|X_t - X_s| > \epsilon) = 0.$$

That is to say a Lévy process is a stochastically continuous process with stationary and independent increments whose sample paths are right continuous with left hand limits.

If $\Phi(z)$ is the characteristic function of a Lévy process, then its characteristic component $\varphi(z) \stackrel{def}{=} \frac{\ln \Phi(z)}{t}$ is of the form

$$\left\{ iza - \frac{z^2 b}{2} + \int_R [\exp(izx) - 1 - izxI_{\{|x|<1\}}] v(dx) \right\}$$

where $a \in R$, $b \in R_+$ and v is a measure on R satisfying $v(\{0\}) = 0$, $\int_R (1 \wedge x^2) v(dx) < \infty$.

The measure v characterizes the size and frequency of the jumps. If the measure is infinite, then the process has infinitely many jumps of very small sizes in any small interval. The constant a defined above is called the drift term of the process, and b is the variance (volatility) term.

The Lévy–Itô decomposition identify any Lévy process as the sum of three independent processes, it is stated as follows:

Given any $a \in R$, $b \in R_+$ and measure v on R satisfying $v(\{0\}) = 0$, $\int_R (1 \wedge x^2) v(dx) < \infty$, there exists a probability space (Ω, F, P) on which a *Lévy* process X is defined. The process X is the sum of three independent processes $\overset{(1)}{X}$, $\overset{(2)}{X}$, and $\overset{(3)}{X}$, where $\overset{(1)}{X}$ is a Brownian motion with drift a and volatility b (in the sense defined below), $\overset{(2)}{X}$ is a compound Poisson process, and $\overset{(3)}{X}$ is a square integrable martingale. The characteristic components of $\overset{(1)}{X}$, $\overset{(2)}{X}$, and $\overset{(3)}{X}$ (denoted by $\overset{(1)}{\varphi}(z)$, $\overset{(2)}{\varphi}(z)$ and $\overset{(3)}{\varphi}(z)$, respectively) are as follows:

$$\overset{(1)}{\varphi}(z) = iza - \tfrac{z^2 b}{2},$$
$$\overset{(2)}{\varphi}(z) = \int_{\{|x| \geq 1\}} (\exp(izx) - 1) v(dx),$$
$$\overset{(3)}{\varphi}(z) = \int_{\{|x|<1\}} (\exp(izx) - 1 - izx) v(dx).$$

Examples of the Lévy Processes

The Brownian Motion

A Lévy process is said to be a Brownian motion (see ►Brownian Motion and Diffusions) with drift μ, and volatility rate σ^2, if $\mu = a$, $b = \sigma^2$, and $v(R) = 0$. Brownian motion is the only nondeterministic Lévy processes with continuous sample paths.

The Inverse Brownian Process

Let X be a Brownian motion with $\mu > 0$ and volatility rate σ^2. For any $x > 0$, let $T_x = \inf\{t : X_t > x\}$. Then T_x is an increasing Lévy process (called inverse Brownian motion), its Lévy measure is given by

$$v(dx) = \frac{1}{\sqrt{2\pi\sigma^2 x^3}} \exp\left(\frac{-x\mu^2}{2\sigma^2}\right).$$

The Compound Poisson Process

The compound Poisson process (see ►Poisson Processes) is a Lévy process where $b = 0$ and v is a finite measure.

The Gamma Process

The gamma process is a nonnegative increasing Lévy process X, where $b = 0$, $a - \int_0^1 xv(dx) = 0$ and its Lévy measure is given by

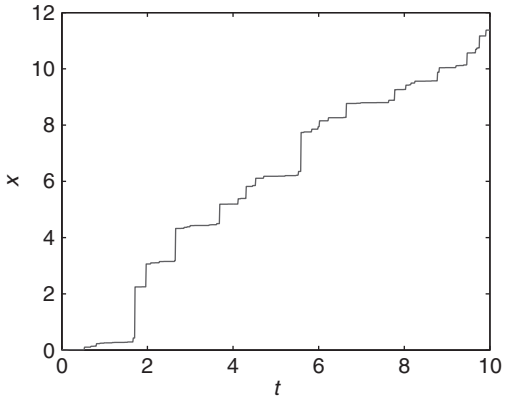
$$v(dx) = \frac{\alpha}{x} exp(-x/\beta) dx, \qquad x > 0$$

where $\alpha, \beta > 0$. It follows that the mean term $(E(X_1))$ and the variance term $(V(X_1))$ for the process are equal to $\alpha\beta$ and $\alpha\beta^2$, respectively.

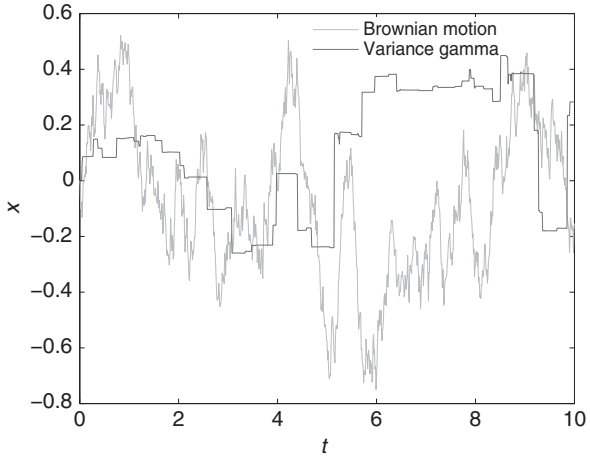
The following is a simulated sample path of a gamma process, where $\alpha = 2$ and $\beta = 0.5$ (Fig. 1).

The Variance Gamma Process

The variance gamma process is a Lévy process that can be represented as either the difference between two independent gamma processes or as a Brownian process subordinated by a gamma process. The latter is accomplished by a random time change, replacing the time of the Brownian process by a gamma process, with a mean term equal to 1. The variance gamma process has three parameters: μ – the



Lévy Processes. Fig. 1 Gamma process



Lévy Processes. Fig. 2 Brownian motion and variance gamma sample paths

Brownian process drift term, σ – the volatility of the Brownian process, and v – the variance term of the the gamma process.

The following are two simulated sample paths, one for a Brownian motion with a drift term $\mu = 0.2$ and volatility term $\sigma = 0.5$ and the other is for a variance gamma process with the same values for the drift term and the volatility terms and $v = 1$ (Fig. 2).

About the Author

Professor Mohamed Abdel-Hameed was the chairman of the Department of Statistics and Operations Research, Kuwait University (1988–1989). He was on the editorial board of *Applied Stochastic Models and Data Analysis* (1983–1990). He was the Editor (jointly with E. Cinlar and J. Quinn) of the text *Survival Models, Maintenance, Replacement Policies and Accelerated Life Testing* (Academic Press 1984). He is an elected member of the International Statistical Institute. He has published numerous papers in Statistics and Operations research.

Cross References

►Non-Uniform Random Variate Generations
►Poisson Processes
►Statistical Modeling of Financial Markets
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Abdel-Hameed MS (1984) Life distribution of devices subject to a Lévy wear process. Math Oper Res 9:606–614

Abdel-Hameed M (2000) Optimal control of a dam using $P^M_{\lambda,\tau}$ policies and penalty cost when the input process is a compound Poisson process with a positive drift. J Appl Prob 37: 408–416

Black F, Scholes MS (1973) The pricing of options and corporate liabilities. J Pol Econ 81:637–654

Cont R, Tankov P (2003) Financial modeling with jump processes. Chapman & Hall/CRC Press, Boca Raton

Madan DB, Carr PP, Chang EC (1998) The variance gamma process and option pricing. Eur Finance Rev 2:79–105

Patel A, Kosko B (2008) Stochastic resonance in continuous and spiking Neutron models with Lévy noise. IEEE Trans Neural Netw 19:1993–2008

van Noortwijk JM (2009) A survey of the applications of gamma processes in maintenance. Reliab Eng Syst Safety 94:2–21

Life Expectancy

Maja Biljan-August
Full Professor
University of Rijeka, Rijeka, Croatia

Life expectancy is defined as the average number of years a person is expected to live from age x, as determined by statistics.

Statistics on life expectancy are derived from a mathematical model known as the ►life table. In order to calculate this indicator, the mortality rate at each age is assumed to be constant. Life expectancy (e_x) can be evaluated at any age and, in a hypothetical stationary population, can be written in discrete form as:

$$e_x = \frac{T_x}{l_x}$$

where x is age; T_x is the number of person-years lived aged x and over; and l_x is the number of survivors at age x according to the life table.

Life expectancy can be calculated for combined sexes or separately for males and females. There can be significant differences between sexes.

Life expectancy at birth (e_0) is the average number of years a newborn child can expect to live if current mortality trends remain constant:

$$e_0 = \frac{T_0}{l_0}$$

where T_0 is the total size of the population and l_0 is the number of births (the original number of persons in the birth cohort).

Life expectancy declines with age. Life expectancy at birth is highly influenced by infant mortality rate. *The paradox of the life table* refers to a situation where life expectancy at birth increases for several years after birth ($e_0 < e_1 < \ldots e_5$ and even beyond). The paradox reflects the higher rates of infant and child mortality in populations in pre-transition and middle stages of the demographic transition.

Life expectancy at birth is a summary measure of mortality in a population. It is a frequently used indicator of health standards and socio-economic living standards. Life expectancy is also one of the most commonly used indicators of social development. This indicator is easily comparable through time and between areas, including countries. Inequalities in life expectancy usually indicate inequalities in health and socio-economic development.

Life expectancy rose throughout human history. In ancient Greece and Rome, the average life expectancy was below 30 years; between the years 1800 and 2000, life expectancy at birth rose from about 30 years to a global average of 67 years, and to more than 75 years in the richest countries (Riley 2001). Furthermore, in most industrialized countries, in the early twenty-first century, life expectancy averaged at about 78 years (WHO). These changes, called the "health transition," are essentially the result of improvements in public health, medicine, and nutrition.

Life expectancy varies significantly across regions and continents: from life expectancies of about 40 years in some central African populations to life expectancies of 80 years and above in many European countries. The more developed regions have an average life expectancy of 76 years, while the population of less developed regions is at birth expected to live an average 12 years less. The two continents that display the most extreme differences in life expectancies are North America (77.6 years) and Africa (49.1 years) where, as of recently, the gap between life expectancies amounts to 29 years (UN, 2000–2005 data).

Countries with the highest life expectancies in the world (82 years) are Australia, Iceland, Italy, Switzerland, and Japan (83 years); Japanese men and women live an average of 79 and 86 years, respectively (WHO 2009).

In countries with a high rate of HIV infection, principally in Sub-Saharan Africa, the average life expectancy is 45 years and below. Some of the world's lowest life expectancies are in Sierra Leone (41 years), Afghanistan (42 years), Lesotho (45 years), and Zimbabwe (45 years).

In nearly all countries, women live longer than men. The world's average life expectancy at birth is 65 years for males and 70 years for females; the gap is about five years. The female-to-male gap is expected to narrow in the more developed regions and widen in the less developed regions. The Russian Federation has the greatest difference in life expectancies between the sexes (13 years less for men), whereas in Tonga, life expectancy for males exceeds that for females by 2 years (WHO 2009).

Life expectancy is assumed to rise continuously. According to estimation by the UN, global life expectancy at birth is likely to rise to an average 74 years by 2045–2050. By 2100, life expectancy is expected to vary across countries from 66 to 97 years. Long-range United Nations population projections predict that by 2300, on average, people can expect to live more than 95 years, from 87 (Liberia) up to 106 years (Japan).

For more details on the calculation of life expectancy, including continuous notation, see Keyfitz (1968, 2005) and Preston et al. (2003).

Cross References

►Biopharmaceutical Research, Statistics in
►Biostatistics
►Demography
►Life Table
►Mean Residual Life
►Measurement of Economic Progress

References and Further Reading

Keyfitz N (1968) Introduction to the Mathematics of Population. Addison-Wesly, Massachusetts
Keyfitz N (2005) Applied mathematical demography. Springer, New York
Preston SH, Heuveline P, Guillot M (2003) Demography: measuring and modelling potion processes. Blackwell, Oxford
Riley JC (2001) Rising life expectancy: a global history. Cambridge University Press, Cambridge
Rowland DT (2003) Demographic methods and concepts. Oxford University Press, New York
Siegel JS, Swanson DA (2004) The methods and materials of demography, 2nd edn. Elsevier Academic Press, Amsterdam
World Health Organization (2009) World health statistics 2009. Geneva

World Population Prospects: The 2004 revision. Highlights. United Nations, New York
World Population to 2300 (2004) United Nations, New York

Life Table

Jan M. Hoem
Professor, Director Emeritus
Max Planck Institute for Demographic Research, Rostock, Germany

The life table is a classical tabular representation of central features of the distribution function F of a positive variable, say X, which normally is taken to represent the lifetime of a newborn individual. The life table was introduced well before modern conventions concerning statistical distributions were developed, and it comes with some special terminology and notation, as follows. Suppose that F has a density $f(x) = \frac{d}{dx}F(x)$ and define *the force of mortality* or *death intensity* at age x as the function

$$\mu(x) = -\frac{d}{dx}\ln\{1 - F(x)\} = \frac{f(x)}{\{1 - F(x)\}}.$$

Heuristically, it is interpreted by the relation $\mu(x)\mathrm{d}x = P\{x < X < x + \mathrm{d}x | X > x\}$. Conversely $F(x) = 1 - \exp\left\{-\int_0^x \mu(s)ds\right\}$. The *survivor function* is defined as $\ell(x) = \ell(0)\{1 - F(x)\}$, normally with $\ell(0) = 100,000$. In mortality applications $\ell(x)$ is the expected number of survivors to exact age x out of an original cohort of 100,000 newborn babies. The *survival probability* is

$$\begin{aligned} {}_tp_x &= P\{X > x + t | X > x\} = \ell(x + t)/\ell(x) \\ &= \exp\left\{-\int_0^t \mu(x + s)ds\right\}, \end{aligned}$$

and the non-survival probability is (the converse) ${}_tq_x = 1 - {}_tp_x$. For $t = 1$ one writes $q_x = {}_1q_x$ and $p_x = {}_1p_x$. In particular, we get $\ell(x + 1) = \ell(x)p_x$. This is a practical recursion formula that permits us to compute all values of $\ell(x)$ once we know the values of p_x for all relevant x.

The life expectancy is $e_0^o = EX = \int_0^\infty \ell(x)dx/\ell(0)$ (The subscript 0 in e_0^o indicates that the expected value is computed at age 0 (i.e., for newborn individuals) and the superscript o indicates that the computation is made in the continuous mode.). The remaining life expectancy at age x is:

$$e_x^o = E(X - x | X > x) = \int_0^\infty \ell(x + t)dt/\ell(x),$$

i.e., it is the expected lifetime remaining to someone who has attained age x.

To turn to the statistical estimation of these various quantities, suppose that the function $\mu(x)$ is piecewise constant, which means that we take it to equal some constant, say μ_j, over each interval (x_j, x_{j+1}) for some partition $\{x_j\}$ of the age axis. For a collection $\{X_i\}$ of independent observations of X, let D_j be the number of X_i that fall in the interval (x_j, x_{j+1}). In mortality applications, this is the number of deaths observed in the given interval. For the cohort of the initially newborn, D_j is the number of individuals who die in the interval (called the *occurrences* in the interval). If individual i dies in the interval, he or she will of course have lived for $X_i - x_j$ time units *during* the interval. Individuals who survive the interval, will have lived for $x_{j+1} - x_j$ time units in the interval, and individuals who do not survive to age x_j, will not have lived during this interval at all. When we aggregate the time units lived in (x_j, x_{j+1}) over all individuals, we get a total R_j which is called the *exposures* for the interval, the idea being that individuals are *exposed to the risk* of death for as long as they live in the interval. In the simple case where there are no relations between the individual parameters μ_j, the collection $\{D_j, R_j\}$ constitutes a statistically sufficient set of observations with a likelihood Λ that satisfies the relation $\ln\Lambda = \sum_j\{-\mu_j R_j + D_j \ln\mu_j\}$ which is easily seen to be maximized by $\hat{\mu}_j = D_j/R_j$. The latter fraction is therefore the maximum-likelihood estimator for μ_j (In some connections an age schedule of mortality will be specified, such as the classical Gompertz–Makeham function $\mu_x = a + bc^x$, which does represent a relationship between the intensity values at the different ages x, normally for single-year age groups. Maximum likelihood estimators can then be found by plugging this functional specification of the intensities into the likelihood function, finding the values $\hat{a}, \hat{b}$, and $\hat{c}$ that maximize Λ, and using $\hat{a} + \hat{b}\hat{c}^x$ for the intensity in the rest of the life table computations. Methods that do not amount to maximum likelihood estimation will sometimes be used because they involve simpler computations. With some luck they provide starting values for the iterative process that must usually be applied to produce the maximum likelihood estimators. For an example, see Forsén (1979)). This whole schema can be extended trivially to cover censoring (*withdrawals*) provided the censoring mechanism is unrelated to the mortality process.

If the force of mortality is constant over a single-year age interval $(x, x+1)$, say, and is estimated by $\hat{\mu}_x$ in this interval, then $\hat{p}_x = e^{-\hat{\mu}_x}$ is an estimator of the single-year survival probability p_x. This allows us to estimate the survival function recursively for all corresponding ages, using $\hat{\ell}(x+1) = \hat{\ell}(x)\hat{p}_x$ for $x = 0, 1, \ldots$, and the rest of the life table computations follow suit. Life table construction consists in the estimation of the parameters and the tabulation of functions like those above from empirical data. The data can be for age at death for individuals, as in the example indicated above, but they can also be observations of duration until recovery from an illness, of intervals between births, of time until breakdown of some piece of machinery, or of any other positive duration variable.

So far we have argued as if the life table is computed for a group of mutually independent individuals who have all been observed in parallel, essentially a cohort that is followed from a significant common starting point (namely from birth in our mortality example) and which is diminished over time due to *decrements* (*attrition*) caused by the risk in question and also subject to reduction due to censoring (withdrawals). The corresponding table is then called a *cohort life table*. It is more common, however, to estimate a $\{p_x\}$ schedule from data collected for the members of a population during a limited time period and to use the mechanics of life-table construction to produce a *period life table* from the p_x values.

Life table techniques are described in detail in most introductory textbooks in actuarial statistics, ►biostatistics, ►demography, and epidemiology. See, e.g., Chiang (1984), Elandt-Johnson and Johnson (1980), Manton and Stallard (1984), Preston et al. (2001). For the history of the topic, consult Seal (1977), Smith and Keyfitz (1977), and Dupâquier (1996).

About the Author

For biography *see* the entry ►Demography.

Cross References

►Demographic Analysis: A Stochastic Approach
►Demography
►Event History Analysis
►Kaplan-Meier Estimator
►Population Projections
►Statistics: An Overview

References and Further Reading

Chiang CL (1984) The life table and its applications. Krieger, Malabar
Dupâquier J (1996) L'invention de la table de mortalité. Presses universitaires de France, Paris
Elandt-Johnson RC, Johnson NL (1980) Survival models and data analysis. Wiley, New York
Forsén L (1979) The efficiency of selected moment methods in Gompertz–Makeham graduation of mortality. Scand Actuarial J 167–178
Manton KG, Stallard E (1984) Recent trends in mortality analysis. Academic Press, Orlando
Preston SH, Heuveline P, Guillot M (2001) Demography. Measuring and modeling populations. Blackwell, Oxford
Seal H (1977) Studies in history of probability and statistics, 35: multiple decrements of competing risks. Biometrica 63(3):429–439
Smith D, Keyfitz N (1977) Mathematical demography: selected papers. Springer, Heidelberg

Likelihood

Nancy Reid
Professor
University of Toronto, Toronto, ON, Canada

Introduction

The likelihood function in a statistical model is proportional to the density function for the random variable to be observed in the model. Most often in applications of likelihood we have a parametric model $f(y;\theta)$, where the parameter θ is assumed to take values in a subset of $\mathbb{R}^k$, and the variable y is assumed to take values in a subset of $\mathbb{R}^n$: the likelihood function is defined by

$$L(\theta) = L(\theta; y) = cf(y;\theta), \tag{1}$$

where c can depend on y but not on θ. In more general settings where the model is semi-parametric or non-parametric the explicit definition is more difficult, because the density needs to be defined relative to a dominating measure, which may not exist: see Van der Vaart (1996) and Murphy and Van der Vaart (1997). This article will consider only finite-dimensional parametric models.

Within the context of the given parametric model, the likelihood function measures the relative plausibility of various values of θ, for a given observed data point y. Values of the likelihood function are only meaningful relative to each other, and for this reason are sometimes standardized by the maximum value of the likelihood function, although other reference points might be of interest depending on the context.

If our model is $f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, y = 0, 1, \ldots, n$; $\theta \in [0,1]$, then the likelihood function is (any function proportional to)

$$L(\theta; y) = \theta^y(1-\theta)^{n-y}$$

and can be plotted as a function of θ for any fixed value of y. The likelihood function is maximized at $\theta = y/n$. This model might be appropriate for a sampling scheme which recorded the number of successes among n independent trials that result in success or failure, each trial having the same probability of success, θ. Another example is the likelihood function for the mean and variance parameters when sampling from a normal distribution with mean μ and variance σ^2:

$$L(\theta; y) = \exp\{-n \log \sigma - (1/2\sigma^2)\Sigma(y_i - \mu)^2\},$$

where $\theta = (\mu, \sigma^2)$. This could also be plotted as a function of μ and σ^2 for a given sample $y_1, \ldots, y_n$, and it is not difficult to show that this likelihood function only depends on the sample through the sample mean $\bar{y} = n^{-1}\Sigma y_i$ and sample variance $s^2 = (n-1)^{-1}\Sigma(y_i - \bar{y})^2$, or equivalently through Σy_i and Σy_i^2. It is a general property of likelihood functions that they depend on the data only through the minimal sufficient statistic.

Inference

The likelihood function was defined in a seminal paper of Fisher (1922), and has since become the basis for most methods of statistical inference. One version of likelihood inference, suggested by Fisher, is to use some rule such as $L(\hat{\theta})/L(\theta) > k$ to define a range of "likely" or "plausible" values of θ. Many authors, including Royall (1997) and Edwards (1960), have promoted the use of plots of the likelihood function, along with interval estimates of plausible values. This approach is somewhat limited, however, as it requires that θ have dimension 1 or possibly 2, or that a likelihood function can be constructed that only depends on a component of θ that is of interest; see section "►Nuisance Parameters" below.

In general, we would wish to calibrate our inference for θ by referring to the probabilistic properties of the inferential method. One way to do this is to introduce a probability measure on the unknown parameter θ, typically called a prior distribution, and use Bayes' rule for conditional probabilities to conclude

$$\pi(\theta \mid y) = L(\theta; y)\pi(\theta) / \int_\theta L(\theta; y)\pi(\theta)d\theta,$$

where $\pi(\theta)$ is the density for the prior measure, and $\pi(\theta \mid y)$ provides a probabilistic assessment of θ after observing $Y = y$ in the model $f(y; \theta)$. We could then make conclusions of the form, "having observed 5 successes in 20 trials, and assuming $\pi(\theta) = 1$, the posterior probability that $\theta > 0.5$ is 0.013," and so on.

This is a very brief description of Bayesian inference, in which probability statements refer to that generated from the prior through the likelihood to the posterior. A major difficulty with this approach is the choice of prior probability function. In some applications there may be an accumulation of previous data that can be incorporated into a probability distribution, but in general there is not, and some rather *ad hoc* choices are often made. Another difficulty is the meaning to be attached to probability statements about the parameter.

Inference based on the likelihood function can also be calibrated with reference to the probability model $f(y; \theta)$, by examining the distribution of $L(\theta; Y)$ as a random function, or more usually, by examining the distribution of various derived quantities. This is the basis for likelihood inference from a frequentist point of view. In particular, it can be shown that $2\log\{L(\hat{\theta}; Y)/L(\theta; Y)\}$, where $\hat{\theta} = \hat{\theta}(Y)$ is the value of θ at which $L(\theta; Y)$ is maximized, is approximately distributed as a χ^2_k random variable, where k is the dimension of θ. To make this precise requires an asymptotic theory for likelihood, which is based on a central limit theorem (see ►Central Limit Theorems) for the *score function*

$$U(\theta; Y) = \frac{\partial}{\partial\theta} \log L(\theta; Y).$$

If $Y = (Y_1, \ldots, Y_n)$ has independent components, then $U(\theta)$ is a sum of n independent components, which under mild regularity conditions will be asymptotically normal. To obtain the χ^2 result quoted above it is also necessary to investigate the convergence of $\hat{\theta}$ to the true value governing the model $f(y; \theta)$. Showing this convergence, usually in probability, but sometimes almost surely, can be difficult: see Scholz (2006) for a summary of some of the issues that arise.

Assuming that $\hat{\theta}$ is consistent for θ, and that $L(\theta; Y)$ has sufficient regularity, the follow asymptotic results can be established:

$$(\hat{\theta} - \theta)^T i(\theta)(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_k, \tag{2}$$

$$U(\theta)^T i^{-1}(\theta) U(\theta) \xrightarrow{d} \chi^2_k, \tag{3}$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_k, \tag{4}$$

where $i(\theta) = E\{-\ell''(\theta; Y); \theta\}$ is the expected Fisher information function, $\ell(\theta) = \log L(\theta)$ is the log-likelihood function, and χ^2_k is the ►chi-square distribution with k degrees of freedom.

These results are all versions of a more general result that the log-likelihood function converges to the quadratic form corresponding to a multivariate normal distribution (see ►Multivariate Normal Distributions), under suitably stated limiting conditions. There is a similar asymptotic result showing that the posterior density is asymptotically

normal, and in fact asymptotically free of the prior distribution, although this result requires that the prior distribution be a proper probability density, i.e., has integral over the parameter space equal to 1.

Nuisance Parameters

In models where the dimension of θ is large, plotting the likelihood function is not possible, and inference based on the multivariate normal distribution for $\hat{\theta}$ or the χ^2_k distribution of the log-likelihood ratio doesn't lead easily to interval estimates for components of θ. However it is possible to use the likelihood function to construct inference for parameters of interest, using various methods that have been proposed to eliminate nuisance parameters.

Suppose in the model $f(y;\theta)$ that $\theta = (\psi, \lambda)$, where ψ is a k_1-dimensional parameter of interest (which will often be 1). The *profile log-likelihood* function of ψ is

$$\ell_P(\psi) = \ell(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the *constrained* maximum likelihood estimate: it maximizes the likelihood function $L(\psi, \lambda)$ when ψ is held fixed. The profile log-likelihood function is also called the concentrated log-likelihood function, especially in econometrics. If the parameter of interest is not expressed explicitly as a subvector of θ, then the constrained maximum likelihood estimate is found using Lagrange multipliers.

It can be verified under suitable smoothness conditions that results similar to those at (2 – 4) hold as well for the profile log-likelihood function: in particular

$$2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\} = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} \xrightarrow{d} \chi^2_{k_1},$$

This method of eliminating nuisance parameters is not completely satisfactory, especially when there are many nuisance parameters: in particular it doesn't allow for errors in estimation of λ. For example the profile likelihood approach to estimation of σ^2 in the linear regression model (see ▶Linear Regression Models) $y \sim N(X\beta, \sigma^2)$ will lead to the estimator $\hat{\sigma}^2 = \Sigma(y_i - \hat{y}_i)^2/n$, whereas the estimator usually preferred has divisor $n-p$, where p is the dimension of β.

Thus a large literature has developed on improvements to the profile log-likelihood. For Bayesian inference such improvements are "automatically" included in the formulation of the marginal posterior density for ψ:

$$\pi_M(\psi \mid y) \propto \int \pi(\psi, \lambda \mid y) d\lambda,$$

but it is typically quite difficult to specify priors for possibly high-dimensional nuisance parameters. For non-Bayesian inference most modifications to the profile log-likelihood are derived by considering conditional or marginal inference in models that admit factorizations, at least approximately, like the following:

$$f(y;\theta) = f_1(y_1;\psi) f_2(y_2 \mid y_1;\lambda), \quad \text{or}$$
$$f(y;\theta) = f_1(y_1 \mid y_2;\psi) f_2(y_2;\lambda).$$

A discussion of conditional inference and density factorisations is given in Reid (1995). This literature is closely tied to that on higher order asymptotic theory for likelihood. The latter theory builds on saddlepoint and Laplace expansions to derive more accurate versions of (2–4): see, for example, Severini (2000) and Brazzale et al. (2007). The direct likelihood approach of Royall (1997) and others does not generalize very well to the nuisance parameter setting, although Royall and Tsou (2003) present some results in this direction.

Extensions to Likelihood

The likelihood function is such an important aspect of inference based on models that it has been extended to "likelihood-like" functions for more complex data settings. Examples include nonparametric and semi-parametric likelihoods: the most famous semi-parametric likelihood is the proportional hazards model of Cox (1972). But many other extensions have been suggested: to empirical likelihood (Owen 1988), which is a type of nonparametric likelihood supported on the observed sample; to quasi-likelihood (McCullagh 1983) which starts from the score function $U(\theta)$ and works backwards to an inference function; to bootstrap likelihood (Davison et al. 1992); and many modifications of profile likelihood (Barndorff-Nielsen and Cox 1994; Fraser 2003). There is recent interest for multi-dimensional responses Y_i in composite likelihoods, which are products of lower dimensional conditional or marginal distributions (Varin 2008). Reid (2000) concluded a review article on likelihood by stating:

> From either a Bayesian or frequentist perspective, the likelihood function plays an essential role in inference. The maximum likelihood estimator, once regarded on an equal footing among competing point estimators, is now typically the estimator of choice, although some refinement is needed in problems with large numbers of nuisance parameters. The likelihood ratio statistic is the basis for most tests of hypotheses and interval estimates. The emergence of the centrality of the likelihood function for inference, partly due to the large increase in computing power, is one of the central developments in the theory of statistics during the latter half of the twentieth century.

Further Reading

The book by Cox and Hinkley (1974) gives a detailed account of likelihood inference and principles of statistical inference; see also Cox (2006). There are several book-length treatments of likelihood inference, including Edwards (1960), Azzalini (1998), Pawitan (2000), and Severini (2000): this last discusses higher order asymptotic theory in detail, as does Barndorff-Nielsen and Cox (1994), and Brazzale, Davison and Reid (2007). A short review paper is Reid (2000). An excellent overview of consistency results for maximum likelihood estimators is Scholz (2006); see also Lehmann and Casella (1998). Foundational issues surrounding likelihood inference are discussed in Berger and Wolpert (1980).

About the Author

Professor Reid is a Past President of the Statistical Society of Canada (2004–2005). During (1996–1997) she served as the President of the Institute of Mathematical Statistics. Among many awards, she received the Emanuel and Carol Parzen Prize for Statistical Innovation (2008) "for leadership in statistical science, for outstanding research in theoretical statistics and highly accurate inference from the likelihood function, and for influential contributions to statistical methods in biology, environmental science, high energy physics, and complex social surveys." She was awarded the Gold Medal, Statistical Society of Canada (2009) and Florence Nightingale David Award, Committee of Presidents of Statistical Societies (2009). She is Associate Editor of *Statistical Science*, (2008–), *Bernoulli* (2007–) and *Metrika* (2008–).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Chi-Square Test: Analysis of Contingency Tables
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Estimation
►Fiducial Inference
►General Linear Models
►Generalized Linear Models
►Generalized Quasi-Likelihood (GQL) Inferences
►Mixture Models
►Philosophical Foundations of Statistics
►Risk Analysis
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview
►Statistics: Nelder's view
►Testing Variance Components in Mixed Linear Models
►Uniform Distribution in Statistics

References and Further Reading

Azzalini A (1998) Statistical inference. Chapman and Hall, London
Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman and Hall, London
Berger JO, Wolpert R (1980) The likelihood principle. Institute of Mathematical Statistics, Hayward
Birnbaum A (1962) On the foundations of statistical inference. Am Stat Assoc 57:269–306
Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics. Cambridge University Press, Cambridge
Cox DR (1972) Regression models and life tables. J R Stat Soc B 34:187–220 (with discussion)
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London
Davison AC, Hinkley DV, Worton B (1992) Bootstrap likelihoods. Biometrika 79:133–130
Edwards AF (1960) Likelihood. Oxford University Press, Oxford
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans R Soc A 222:309–368
Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York
McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
Murphy SA, Van der Vaart A (1997) Semiparametric likelihood ratio inference. Ann Stat 25:1471–1509
Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Pawitan Y (2000) In all likelihood. Oxford University Press, Oxford
Reid N (1995) The roles of conditioning in inference. Stat Sci 10:138–157
Reid N (2000) Likelihood. J Am Stat Assoc 95:1335–1340
Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London
Royall RM, Tsou TS (2003) Interpreting statistical evidence using imperfect models: robust adjusted likelihoodfunctions. J R Stat Soc B 65:391404
Scholz F (2006) Maximum likelihood estimation. In: Encyclopedia of statistical sciences. Wiley, New York, doi: 10.1002/0471667196.ess1571.pub2. Accessed 23 Aug 2009
Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford
Van der Vaart AW (1996) Infinite-dimensional likelihood methods in statistics. http://www.stieltjes.org/archief/biennial9596/frame/node17.html. Accessed 18 Aug 2009
Varin C (2008) On composite marginal likelihood. Adv Stat Anal 92:1–28

Limit Theorems of Probability Theory

Alexandr Alekseevich Borovkov
Professor, Head of Probability and Statistics Chair at the Novosibirsk University
Novosibirsk University, Novosibirsk, Russia

Limit Theorems of Probability Theory is a broad name referring to the most essential and extensive research area in Probability Theory which, at the same time, has the greatest impact on the numerous applications of the latter.

By its very nature, Probability Theory is concerned with asymptotic (limiting) laws that emerge in a long series of observations on random events. Because of this, in the early twentieth century even the very definition of probability of an event was given by a group of specialists (R. von Mises and some others) as the limit of the relative frequency of the occurrence of this event in a long row of independent random experiments. The "stability" of this frequency (i.e., that such a limit always exists) was postulated. After the 1930s, Kolmogorov's axiomatic construction of probability theory has prevailed. One of the main assertions in this axiomatic theory is the *Law of Large Numbers* (LLN) on the convergence of the averages of large numbers of random variables to their expectation. This law implies the aforementioned stability of the relative frequencies and their convergence to the probability of the corresponding event.

The LLN is the simplest limit theorem (LT) of probability theory, elucidating the physical meaning of probability. The LLN is stated as follows: if $X, X_1, X_2, \ldots$ is a sequence of i.i.d. random variables,

$$S_n := \sum_{j=1}^{n} X_j,$$

and the expectation $a := \mathbf{E}X$ exists then $n^{-1}S_n \xrightarrow{a.s.} a$ (almost surely, i.e., with probability 1). Thus the value na can be called the *first order approximation* for the sums S_n. The *Central Limit Theorem* (CLT) gives one a more precise approximation for S_n. It says that, if $\sigma^2 := \mathbf{E}(X-a)^2 < \infty$, then the distribution of the standardized sum $\zeta_n := (S_n - na)/\sigma\sqrt{n}$ converges, as $n \to \infty$, to the standard normal (Gaussian) law. That is, for all x,

$$\mathbf{P}(\zeta_n < x) \to \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

The quantity $n\mathbf{E}\xi + \zeta\sigma\sqrt{n}$, where ζ is a standard normal random variable (so that $\mathbf{P}(\zeta < x) = \Phi(x)$), can be called *the second order approximation* for S_n.

The first LLN (for the Bernoulli scheme) was proved by Jacob Bernoulli in the late 1690s (published posthumously in 1713). The first CLT (also for the Bernoulli scheme) was established by A. de Moivre (first published in 1738 and referred nowadays to as the de Moivre–Laplace theorem). In the beginning of the nineteenth century, P.S. Laplace and C.F. Gauss contributed to the generalization of these assertions and appreciation of their enormous applied importance (in particular, for the *theory of errors of observations*), while later in that century further breakthroughs in both methodology and applicability range of the CLT were achieved by P.L. Chebyshev (1887) and A.M. Lyapunov (1900).

The main directions in which the two aforementioned main LTs have been extended and refined since then are:

1. Relaxing the assumption $\mathbf{E}X^2 < \infty$. When the second moment is infinite, one needs to assume that the "tail" $P(x) := \mathbf{P}(X > x) + \mathbf{P}(X < -x)$ is a function regularly varying at infinity such that the limit $\lim_{x\to\infty} \mathbf{P}(X > x)/P(x)$ exists. Then the distribution of the normalized sum $S_n/\sigma(n)$, where $\sigma(n) := P^{-1}(n^{-1})$, P^{-1} being the generalized inverse of the function P, and we assume that $\mathbf{E}\xi = 0$ when the expectation is finite, converges to one of the so-called stable laws as $n \to \infty$. The ▶characteristic functions of these laws have simple closed-form representations.
2. Relaxing the assumption that the X_j's are identically distributed and proceeding to study the so-called *triangular array scheme*, where the distributions of the summands $X_j = X_{j,n}$ forming the sum S_n depend not only on j but on n as well. In this case, the class of all limit laws for the distribution of S_n (under suitable normalization) is substantially wider: it coincides with the class of the so-called infinitely divisible distributions. An important special case here is the Poisson limit theorem on convergence in distribution of the number of occurrences of rare events to a Poisson law.
3. Relaxing the assumption of independence of the X_j's. Several types of "weak dependence" assumptions on X_j under which the LLN and CLT still hold true have been suggested and investigated. One should also mention here the so-called ergodic theorems (see ▶Ergodic Theorem) for a wide class of random sequences and processes.
4. Refinement of the main LTs and derivation of asymptotic expansions. For instance, in the CLT, bounds of the rate of convergence $\mathbf{P}(\zeta_n < x) - \Phi(x) \to 0$ and

asymptotic expansions for this difference (in the powers of $n^{-1/2}$ in the case of i.i.d. summands) have been obtained under broad assumptions.

5. Studying large deviation probabilities for the sums S_n (theorems on rare events). If $x \to \infty$ together with n then the CLT can only assert that $\mathbf{P}(\zeta_n > x) \to 0$. Theorems on large deviation probabilities aim to find a function $P(x, n)$ such that

$$\frac{\mathbf{P}(\zeta_n > x)}{P(x,n)} \to 1 \quad \text{as} \quad n \to \infty, \quad x \to \infty.$$

The nature of the function $P(x, n)$ essentially depends on the rate of decay of $\mathbf{P}(X > x)$ as $x \to \infty$ and on the "deviation zone," i.e., on the asymptotic behavior of the ratio x/n as $n \to \infty$.

6. Considering observations $X_1, \ldots, X_n$ of a more complex nature – first of all, multivariate random vectors. If $X_j \in \mathbb{R}^d$ then the role of the limit law in the CLT will be played by a d-dimensional normal (Gaussian) distribution with the covariance matrix $\mathbf{E}(X - \mathbf{E}X)(X - \mathbf{E}X)^T$.

The variety of application areas of the LLN and CLT is enormous. Thus, *Mathematical Statistics* is based on these LTs. Let $X_n^* := (X_1, \ldots, X_n)$ be a sample from a distribution F and $F_n^*(u)$ the corresponding empirical distribution function. The fundamental Glivenko–Cantelli theorem (see ▶Glivenko-Cantelli Theorems) stating that $\sup_u |F_n^*(u) - F(u)| \xrightarrow{a.s.} 0$ as $n \to \infty$ is of the same nature as the LLN and basically means that the unknown distribution F can be estimated arbitrary well from the random sample X_n^* of a large enough size n.

The existence of consistent estimators for the unknown parameters $a = \mathbf{E}\xi$ and $\sigma^2 = \mathbf{E}(X - a)^2$ also follows from the LLN since, as $n \to \infty$,

$$a^* := \frac{1}{n}\sum_{j=1}^n X_j \xrightarrow{a.s.} a, \quad (\sigma^2)^* := \frac{1}{n}\sum_{j=1}^n (X_j - a^*)^2$$
$$= \frac{1}{n}\sum_{j=1}^n X_j^2 - (a^*)^2 \xrightarrow{a.s.} \sigma^2.$$

Under additional moment assumptions on the distribution F, one can also construct asymptotic confidence intervals for the parameters a and σ^2, as the distributions of the quantities $\sqrt{n}(a^* - a)$ and $\sqrt{n}\big((\sigma^2)^* - \sigma^2\big)$ converge, as $n \to \infty$, to the normal ones. The same can also be said about other parameters that are "smooth" enough functionals of the unknown distribution F.

The theorem on the ▶asymptotic normality and asymptotic efficiency of maximum likelihood estimators is another classical example of LTs' applications in mathematical statistics (see e.g., Borovkov 1998). Furthermore, in estimation theory and hypotheses testing, one also needs theorems on large deviation probabilities for the respective statistics, as it is statistical procedures with small error probabilities that are often required in applications.

It is worth noting that summation of random variables is by no means the only situation in which LTs appear in Probability Theory.

Generally speaking, the main objective of Probability Theory in applications is finding appropriate stochastic models for objects under study and then determining the distributions or parameters one is interested in. As a rule, the explicit form of these distributions and parameters is not known. LTs can be used to find suitable approximations to the characteristics in question.

At least two possible approaches to this problem should be noted here.

1. Suppose that the unknown distribution F_θ depends on a parameter θ such that, as θ approaches some "critical" value θ_0, the distributions F_θ become "degenerate" in one sense or another. Then, in a number of cases, one can find an approximation for F_θ which is valid for the values of θ that are close to θ_0. For instance, in actuarial studies, ▶queueing theory and some other applications one of the main problems is concerned with the distribution of $\overline{S} := \sup_{k \geq 1}(S_k - \theta k)$, under the assumption that $\mathbf{E}X = 0$. If $\theta > 0$ then $\overline{S}$ is a proper random variable. If, however, $\theta \to 0$ then $\overline{S} \xrightarrow{a.s.} \infty$. Here we deal with the so-called "transient phenomena." It turns out that if $\sigma^2 := \mathrm{Var}\,(X) < \infty$ then there exists the limit

$$\lim_{\theta \downarrow 0} \mathbf{P}(\theta \overline{S} > x) = e^{-2x/\sigma^2}, \quad x > 0.$$

This (Kingman–Prokhorov) LT enables one to find approximations for the distribution of $\overline{S}$ in situations where θ is small.

2. Sometimes one can estimate the "tails" of the unknown distributions, i.e., their asymptotic behavior at infinity. This is of importance in those applications where one needs to evaluate the probabilities of rare events. If the equation $\mathbf{E}e^{\mu(X-\theta)} = 1$ has a solution $\mu_0 > 0$ then, in the above example, one has

$$\mathbf{P}(\overline{S} > x) \sim ce^{-\mu_0 x}, \quad x \to \infty,$$

where c is a known constant. If the distribution F of X is *subexponential* (in this case, $\mathbf{E}e^{\mu(X-\theta)} = \infty$ for any $\mu > 0$) then

$$\mathbf{P}(\overline{S} > x) \sim \frac{1}{\theta}\int_x^\infty (1 - F(t))\,dt, \quad x \to \infty.$$

This LT enables one to find approximations for $\mathbf{P}(\overline{S} > x)$ for large x.

For both approaches, the obtained approximations can be refined.

An important part of Probability Theory is concerned with LTs for *random processes*. Their main objective is to find conditions under which random processes converge, in some sense, to some limit process. An extension of the CLT to that context is the so-called *Functional CLT* (a.k.a. the Donsker–Prokhorov invariance principle) which states that, as $n \to \infty$, the processes $\{\zeta_n(t) := (S_{\lfloor nt \rfloor} - ant)/\sigma\sqrt{n}\}_{t\in[0,1]}$ converge in distribution to the standard Wiener process $\{w(t)\}_{t\in[0,1]}$. The LTs (including large deviation theorems) for a broad class of functionals of the sequence (►random walk) $\{S_1, \ldots, S_n\}$ can also be classified as LTs for ►stochastic processes. The same can be said about Law of iterated logarithm which states that, for an arbitrary $\varepsilon > 0$, the random walk $\{S_k\}_{k=1}^{\infty}$ crosses the boundary $(1-\varepsilon)\sigma\sqrt{2k \ln \ln k}$ infinitely many times but crosses the boundary $(1+\varepsilon)\sigma\sqrt{2k \ln \ln k}$ finitely many times with probability 1. Similar results hold true for trajectories of Wiener processes $\{w(t)\}_{t\in[0,1]}$ and $\{w(t)\}_{t\in[1,\infty)}$.

In mathematical statistics a closely related to functional CLT result says that the so-called "empirical process" $\{\sqrt{n}(F_n^*(u) - F(u))\}_{u\in(-\infty,\infty)}$ converges in distribution to $\{w^0(F(u))\}_{u\in(-\infty,\infty)}$, where $w^0(t) := w(t) - tw(1)$ is the Brownian bridge process. This LT implies ►asymptotic normality of a great many estimators that can be represented as smooth functionals of the empirical distribution function $F_n^*(u)$.

There are many other areas in Probability Theory and its applications where various LTs appear and are extensively used. For instance, convergence theorems for ►martingales, asymptotics of extinction probability of a branching processes and conditional (under non-extinction condition) LTs on a number of particles etc.

About the Author

Professor Borovkov is Head of Probability and Statistics Department of the Sobolev Institute of Mathematics, Novosibirsk (Russian Academy of Sciences), since 1962. He is Head of Probability and Statistics Chair at the Novosibirsk University since 1966. He is Concelour of Russian Academy of Sciences and full member of Russian Academy of Sciences (Academician) (1990). He was awarded the State Prize of the USSR (1979), the Markov Prize of the Russian Academy of Sciences (2003) and Government Prize in Education (2003). Professor Borovkov is Editor-in-Chief of the journal "*Siberian Advances in Mathematics*" and Associated Editor of journals "*Theory of Probability and its Applications*," "*Siberian Mathematical Journal*," "*Mathematical Methods of Statistics*," "*Electronic Journal of Probability*."

Cross References

►Almost Sure Convergence of Random Variables
►Approximations to Distributions
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Central Limit Theorems
►Empirical Processes
►Ergodic Theorem
►Glivenko-Cantelli Theorems
►Large Deviations and Applications
►Laws of Large Numbers
►Martingale Central Limit Theorem
►Probability Theory: An Outline
►Random Matrix Theory
►Strong Approximations in Probability and Statistics
►Weak Convergence of Probability Measures

References and Further Reading

Billingsley P (1968) Convergence of probability measures. Wiley, New York
Borovkov AA (1998) Mathematical statistics. Gordon & Breach, Amsterdam
Gnedenko BV, Kolmogorov AN (1954) Limit distributions for sums of independent random variables. Addison-Wesley, Cambridge
Lévy P (1937) Théorie de l'Addition des Variables Aléatories. Gauthier-Villars, Paris
Loève M (1977–1978) Probability theory, vols I and II, 4th edn. Springer, New York
Petrov VV (1995) Limit theorems of probability theory: sequences of independent random variables. Clarendon/Oxford University Press, New York

Linear Mixed Models

Geert Molenberghs
Professor
Universiteit Hasselt & Katholieke Universiteit Leuven, Leuven, Belgium

In observational studies, repeated measurements may be taken at almost arbitrary time points, resulting in an extremely large number of time points at which only one

or only a few measurements have been taken. Many of the parametric covariance models described so far may then contain too many parameters to make them useful in practice, while other, more parsimonious, models may be based on assumptions which are too simplistic to be realistic. A general, and very flexible, class of parametric models for continuous longitudinal data is formulated as follows:

$$\boldsymbol{y_i}|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i), \quad (1)$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D), \quad (2)$$

where X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters, called the fixed effects, D is a general $(q \times q)$ covariance matrix, and Σ_i is a $(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, i.e., the set of unknown parameters in Σ_i will not depend upon i. Finally, $\boldsymbol{b_i}$ is a vector of subject-specific or random effects.

The above model can be interpreted as a linear regression model (see ▶Linear Regression Models) for the vector $\boldsymbol{y_i}$ of repeated measurements for each unit separately, where some of the regression parameters are specific (random effects, $\boldsymbol{b_i}$), while others are not (fixed effects, $\boldsymbol{\beta}$). The distributional assumptions in (2) with respect to the random effects can be motivated as follows. First, $\mathrm{E}(\boldsymbol{b_i}) = \mathbf{0}$ implies that the mean of $\boldsymbol{y_i}$ still equals $X_i\boldsymbol{\beta}$, such that the fixed effects in the random-effects model (1) can also be interpreted marginally. Not only do they reflect the effect of changing covariates within specific units, they also measure the marginal effect in the population of changing the same covariates. Second, the normality assumption immediately implies that, marginally, $\boldsymbol{y_i}$ also follows a normal distribution with mean vector $X_i\boldsymbol{\beta}$ and with covariance matrix $V_i = Z_iDZ_i' + \Sigma_i$.

Note that the random effects in (1) implicitly imply the marginal covariance matrix V_i of $\boldsymbol{y_i}$ to be of the very specific form $V_i = Z_iDZ_i' + \Sigma_i$. Let us consider two examples under the assumption of conditional independence, i.e., assuming $\Sigma_i = \sigma^2 I_{n_i}$. First, consider the case where the random effects are univariate and represent unit-specific intercepts. This corresponds to covariates Z_i which are n_i-dimensional vectors containing only ones.

The marginal model implied by expressions (1) and (2) is

$$\boldsymbol{y_i} \sim N(X_i\boldsymbol{\beta}, V_i), \quad V_i = Z_iDZ_i' + \Sigma_i$$

which can be viewed as a multivariate linear regression model, with a very particular parameterization of the covariance matrix V_i.

With respect to the estimation of unit-specific parameters $\boldsymbol{b_i}$, the posterior distribution of $\boldsymbol{b_i}$ given the observed data $\boldsymbol{y_i}$ can be shown to be (multivariate) normal with mean vector equal to $DZ_i'V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y_i} - X_i\boldsymbol{\beta})$. Replacing $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by their maximum likelihood estimates, we obtain the so-called empirical Bayes estimates $\widehat{\boldsymbol{b_i}}$ for the $\boldsymbol{b_i}$. A key property of these EB estimates is shrinkage, which is best illustrated by considering the prediction $\widehat{\boldsymbol{y_i}} \equiv X_i\widehat{\boldsymbol{\beta}} + Z_i\widehat{\boldsymbol{b_i}}$ of the ith profile. It can easily be shown that

$$\widehat{\boldsymbol{y_i}} = \Sigma_i V_i^{-1} X_i\widehat{\boldsymbol{\beta}} + \left(I_{n_i} - \Sigma_i V_i^{-1}\right)\boldsymbol{y_i},$$

which can be interpreted as a weighted average of the population-averaged profile $X_i\widehat{\boldsymbol{\beta}}$ and the observed data $\boldsymbol{y_i}$, with weights $\Sigma_i V_i^{-1}$ and $I_{n_i} - \Sigma_i V_i^{-1}$, respectively. Note that the "numerator" of $\Sigma_i V_i^{-1}$ represents within-unit variability and the "denominator" is the overall covariance matrix V_i. Hence, much weight will be given to the overall average profile if the within-unit variability is large in comparison to the between-unit variability (modeled by the random effects), whereas much weight will be given to the observed data if the opposite is true. This phenomenon is referred to as shrinkage toward the average profile $X_i\widehat{\boldsymbol{\beta}}$. An immediate consequence of shrinkage is that the EB estimates show less variability than actually present in the random-effects distribution, i.e., for any linear combination $\boldsymbol{\lambda}$ of the random effects,

$$\mathrm{var}(\lambda'\widehat{\boldsymbol{b_i}}) \leq \mathrm{var}(\lambda'\boldsymbol{b_i}) = \lambda' D \lambda.$$

About the Author

Geert Molenberghs is Professor of Biostatistics at Universiteit Hasselt and Katholieke Universiteit Leuven in Belgium. He was Joint Editor of *Applied Statistics* (2001–2004), and Co-Editor of *Biometrics* (2007–2009). Currently, he is Co-Editor of Biostatistics (2010–2012). He was President of the International Biometric Society (2004–2005), received the Guy Medal in Bronze from the Royal Statistical Society and the Myrto Lefkopoulou award from the Harvard School of Public Health. Geert Molenberghs is founding director of the Center for Statistics. He is also the director of the Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat). Jointly with Geert Verbeke, Mike Kenward, Tomasz Burzykowski, Marc Buyse, and Marc Aerts, he authored books on longitudinal and incomplete data, and on surrogate marker evaluation. Geert Molenberghs received several Excellence in Continuing Education Awards of the American Statistical Association, for courses at Joint Statistical Meetings.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►General Linear Models
►Testing Variance Components in Mixed Linear Models
►Trend Estimation

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, 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

Linear Regression Models

Raoul LePage
Professor
Michigan State University, East Lansing, MI, USA

► I did not want proof, because the theoretical exigencies of the problem would afford that. What I wanted was to be started in the right direction.

(F. Galton)

The *linear regression model* of statistics is any functional relationship $y = f(x, \beta, \varepsilon)$ involving a *dependent* real-valued variable y, *independent* variables x, *model parameters* β and *random variables* ε, such that a measure of central tendency for y in relation to x termed the *regression function* is linear in β. Possible regression functions include the conditional mean $E(y|x, \beta)$ (as when β is itself random as in Bayesian approaches), conditional medians, quantiles or other forms. Perhaps y is corn yield from a given plot of earth and variables x include levels of water, sunlight, fertilization, discrete variables identifying the genetic variety of seed, and combinations of these intended to model interactive effects they may have on y. The form of this linkage is specified by a function f known to the experimenter, one that depends upon parameters β whose values are not known, and also upon unseen random errors ε about which statistical assumptions are made. These models prove surprisingly flexible, as when localized linear regression models are knit together to estimate a regression function nonlinear in β. Draper and Smith (1981) is a plainly written elementary introduction to linear regression models, Rao (1965) is one of many established general references at the calculus level.

Aspects of Data, Model and Notation

Suppose a time varying sound signal is the superposition of sine waves of unknown amplitudes at two fixed known frequencies embedded in white noise background $y[t] = \beta_1 + \beta_2 \sin[.2t] + \beta_3 \sin[.35t] + \varepsilon$. We write $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ for $\beta = (\beta_1, \beta_2, \beta_3), x = (x_1, x_2, x_3), x_1 \equiv 1, x_2(t) = \sin[.2t], x_3(t) = \sin[.35t], t \geq 0$. A natural choice of regression function is $m(x, \beta) = E(y|x, \beta) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ provided $E\varepsilon \equiv 0$. In the *classical linear regression model* one assumes for different instances "i" of observation that random errors satisfy $E\varepsilon_i \equiv 0$, $E\varepsilon_i \varepsilon_k \equiv \sigma^2 > 0$, $i = k \leq n$, $E\varepsilon_i \varepsilon_k \equiv 0$ otherwise. Errors in linear regression models typically depend upon instances i at which we select observations and may in some formulations depend also on the values of x associated with instance i (perhaps the

errors are correlated and that correlation depends upon the x values). What we observe are y_i and associated values of the independent variables x. That is, we observe $(y_i, 1, \sin[.2t_i], \sin[.35t_i]), i \le n$. The linear model on data may be expressed $y = x\beta + \varepsilon$ with y = column vector $\{y_i,\ i \le n\}$, likewise for ε, and matrix x (the *design matrix*) whose $3n$ entries are $x_{ik} = x_k[t_i]$.

Terminology

Independent variables, as employed in this context, is misleading. It derives from language used in connection with mathematical equations and does not refer to statistically independent random variables. Independent variables may be of any dimension, in some applications functions or surfaces. If y is not scalar-valued the model is instead a *multivariate linear regression*. In some versions either x, β or both may also be random and subject to statistical models. Do not confuse multivariate linear regression with *multiple linear regression* which refers to a model having more than one non-constant independent variable.

General Remarks on Fitting Linear Regression Models to Data

Early work (the classical linear model) emphasized independent identically distributed (i.i.d.) additive *normal* errors in linear regression where ▶*least squares* has particular advantages (connections with ▶multivariate normal distributions are discussed below). In that setup least squares would arguably be a principle method of fitting linear regression models to data, perhaps with modifications such as Lasso or other constrained optimizations that achieve reduced sampling variations of coefficient estimators while introducing bias (Efron et al. 2004). Absent a breakthrough enlarging the applicability of the classical linear model other methods gain traction such as Bayesian methods (▶Markov Chain Monte Carlo having enabled their calculation); Non-parametric methods (good performance relative to more relaxed assumptions about errors); Iteratively Reweighted least squares (having under some conditions behavior like maximum likelihood estimators without knowing the precise form of the likelihood). The Dantzig selector is good news for dealing with far fewer observations than independent variables when a relatively small fraction of them matter (Candès and Tao 2007).

Background

C.F. Gauss may have used least squares as early as 1795. In 1801 he was able to predict the apparent position at which asteroid Ceres would re-appear from behind the sun after it had been lost to view following discovery only 40 days before. Gauss' prediction was well removed from all others and he soon followed up with numerous other high-caliber successes, each achieved by fitting relatively simple models motivated by Keppler's Laws, work at which he was very adept and quick. These were fits to imperfect, sometimes limited, yet fairly precise data. Legendre (1805) published a substantive account of least squares following which the method became widely adopted in astronomy and other fields. See Stigler (1986).

By contrast Galton (1877), working with what might today be described as "low correlation" data, discovered deep truths not already known by fitting a straight line. No theoretical model previously available had prepared Galton for these discoveries which were made in a study of his own data w = standard score of weight of parental sweet pea seed(s), y = standard score of seed weights(s) of their immediate issue. Each sweet pea seed has but one parent and the distributions of x and y the same. Working at a time when correlation and its role in regression were yet unknown, Galton found to his astonishment a nearly perfect straight line tracking points (parental seed weight w, median filial seed weight $m(w)$). Since for this data $s_y \sim s_w$ this was the least squares line (also the regression line since the data was bivariate normal) and its slope was $rs_y/s_w = r$ (the correlation). Medians $m(w)$ being essentially equal to means of y for each w greatly facilitated calculations owing to his choice to select equal numbers of parent seeds at weights $0, \pm1, \pm2, \pm3$ standard deviations from the mean of w. Galton gave the name co-relation (later correlation) to the slope $\sim$ 0.33 of this line and for a brief time thought it might be a universal constant. Although the correlation was small, this slope nonetheless gave measure to the previously vague principle of reversion (later regression, as when larger parental examples beget offspring typically not quite so large). Galton deduced the general principle that if $0 < r < 1$ then for a value $w > Ew$ the relation $Ew < m(w) < w$ follows. Having sensibly selected equal numbers of parental seeds at intervals may have helped him observe that points (w, y) departed on each vertical from the regression line by statistically independent $N(0, \theta^2)$ random residuals whose variance $\theta^2 > 0$ was the same for all w. Of course this likewise amazed him and by 1886 he had identified all these properties as a consequence of bivariate normal observations (w, y), (Galton 1886).

Echoes of those long ago events reverberate today in our many statistical models "driven," as we now proclaim, by random errors subject to ever broadening statistical modeling. In the beginning it was very close to truth.

Aspects of Linear Regression Models

Data of the real world seldom conform exactly to any deterministic mathematical model $y = f(x, \beta)$ and through the device of incorporating random errors we have now an established paradigm for fitting models to data (x, y) by statistically estimating model parameters. In consequence we obtain methods for such purposes as predicting what will be the average response y to particular given inputs x; providing *margins of error* (and *prediction error*) for various quantities being estimated or predicted, tests of hypotheses and the like. It is important to note in all this that more than one statistical model may apply to a given problem, the function f and the other model components differing among them. Two statistical models may disagree substantially in structure and yet neither, either or both may produce useful results. In this respect statistical modeling is more a matter of how much we gain from using a statistical model and whether we trust and can agree upon the assumptions placed on the model, at least as a practical expedient. In some cases the regression function conforms precisely to underlying mathematical relationships but that does not reflect the majority of statistical practice. It may be that a given statistical model, although far from being an underlying truth, confers advantage by capturing some portion of the variation of y vis-a-vis x. The method *principle components,* which seeks to find relatively small numbers of linear combinations of the independent variables that together account for most of the variation of y, illustrates this point well. In one application electromagnetic theory was used to generate by computer an elaborate database of theoretical responses of an induced electromagnetic field near a metal surface to various combinations of flaws in the metal. The role of principle components and linear modeling was to establish a simple model reflecting those findings so that a portable device could be devised to make detections in real time based on the model.

If there is any weakness to the statistical approach it lies in the fact that margins of error, statistical tests and the like can be seriously incorrect even if the predictions afforded by a model have apparent value. Refer to Hinkelmann and Kempthorne (2008), Berk (2004), Freedman (2005), Freedman (1991).

Classical Linear Regression Model and Least Squares

The classical linear regression model may be expressed $y = x\beta + \varepsilon$, an abbreviated matrix formulation of the system of equations in which random errors ε are assumed to satisfy $E\varepsilon_I \equiv 0, E\varepsilon_i^2 \equiv \sigma^2 > 0, i \le n$:

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i,\ i \le n. \tag{1}$$

The interpretation is that response y_i is observed for the ith sample in conjunction with numerical values $(x_{i1}, \ldots, x_{ip})$ of the independent variables. If these errors $\{\varepsilon_I\}$ are jointly normally distributed (and therefore statistically independent having been assumed to be uncorrelated) and if the matrix $x^{\mathrm{tr}}x$ is non-singular then the maximum likelihood (ML) estimates of the model coefficients $\{\beta_k, k \le p\}$ are produced by ordinary least squares (LS) as follows:

$$\beta_{ML} = \beta_{LS} = (x^{\mathrm{tr}}x)^{-1}x^{\mathrm{tr}}y = \beta + M\varepsilon \tag{2}$$

for $M = (x^{\mathrm{tr}}x)^{-1}x^{\mathrm{tr}}$ with x^{tr} denoting matrix transpose of x and $(x^{\mathrm{tr}}x)^{-1}$ the matrix inverse. These coefficient estimates β_{LS} are linear functions in y and satisfy the Gauss–Markov properties (3)(4):

$$E(\beta_{LS})_k = \beta_k, k \le p. \tag{3}$$

and, among all unbiased estimators β_k^* (of β_k) that are linear in y,

$$E((\beta_{LS})_k - \beta_k)^2 \le E\left(\beta_k^* - \beta_k\right)^2, \text{ for every } k \le p. \tag{4}$$

Least squares estimator (2) is frequently employed without the assumption of normality owing to the fact that properties (3)(4) must hold in that case as well. Many statistical distributions F, t, *chi-square* have important roles in connection with model (1) either as exact distributions for quantities of interest (normality assumed) or more generally as limit distributions when data are suitably enriched.

Algebraic Properties of Least Squares

Setting all randomness assumptions aside we may examine the algebraic properties of least squares. If $y = x\beta + \varepsilon$ then $\beta_{LS} = My = M(x\beta + \varepsilon) = \beta + M\varepsilon$ as in (2). That is, the least squares estimate of model coefficients acts on $x\beta + \varepsilon$ returning β plus the result of applying least squares to ε. This has nothing to do with the model being correct or ε being error but is purely algebraic. If ε itself has the form $xb + e$ then $M\varepsilon = b + Me$. Another useful observation is that if x has first column identically one, as would typically be the case for a model with constant term, then each row $M_k, k > 1$, of M satisfies $1.M_k = 0$ (i.e., M_k is a *contrast*) so $(\beta_{LS})_k = \beta_k + M_k.\varepsilon$ and $M_k.(\varepsilon + c) = M_k.\varepsilon$ so ε may as well be assumed to be centered for $k > 1$. There are many of these interesting algebraic properties such as $s^2(y - x\beta_{LS}) = (1 - R^2)s^2(y)$ where $s(.)$ denotes the sample standard deviation and R is the *multiple correlation* defined as the correlation between y and the *fitted values* $x\beta_{LS}$. Yet another algebraic identity, this one involving

an interplay of permutations with projections, is exploited to help establish for *exchangeable* errors ε, and contrasts v, a permutation bootstrap of least squares residuals that consistently estimates the *conditional sampling distribution* of $v.(\beta_{LS} - \beta)$ conditional on the ▶order statistics of ε. (See LePage and Podgorski 1996). Freedman and Lane in 1983 advocated tests based on permutation bootstrap of residuals as a descriptive method.

Generalized Least Squares

If errors ε are $N(0, \Sigma)$ distributed for a covariance matrix Σ known up to a constant multiple then the maximum likelihood estimates of coefficients β are produced by a *generalized* least squares solution retaining properties (3)(4) (any positive multiple of Σ will produce the same result) given by:

$$\beta_{ML} = (x^{\mathrm{tr}}\Sigma^{-1}x)^{-1}x^{\mathrm{tr}}\Sigma^{-1}y = \beta + (x^{\mathrm{tr}}\Sigma^{-1}x)^{-1}x^{\mathrm{tr}}\Sigma^{-1}\varepsilon. \quad (5)$$

Generalized least squares solution (5) retains properties (3)(4) even if normality is not assumed. It must not be confused with ▶*generalized linear models* which refers to models equating moments of y to nonlinear functions of $x\beta$.

A very large body of work has been devoted to linear regression models and the closely related subject areas of experimental design, ▶analysis of variance, principle component analysis and their consequent distribution theory.

Reproducing Kernel Generalized Linear Regression Model

Parzen (1961, Sect. 6) developed the reproducing kernel framework extending generalized least squares to spaces of arbitrary finite or infinite dimension when the random error function $\varepsilon = \{\varepsilon(t), t \in T\}$ has zero means $E\varepsilon(t) \equiv 0$ and a covariance function $K(s,t) = E\varepsilon(s)\varepsilon(t)$ that is assumed known up to some positive constant multiple. In this formulation:

$$y(t) = m(t) + \varepsilon(t),\ t \in T,$$
$$m(t) = Ey(t) = \Sigma_i \beta_i w_i(t),\ t \in T,$$

where $w_i(.)$ are *known* linearly independent functions in *the reproducing kernel Hilbert* (*RKHS*) *space* $H(K)$ of the kernel K. For reasons having to do with singularity of Gaussian measures it is assumed that the series defining m is convergent in H(K). Parzen extends to that context and solves the problem of best linear unbiased estimation of the model coefficients β and more generally of *estimable* linear functions of them, developing confidence regions, prediction intervals, exact or approximate distributions, tests and other matters of interest, and establishing the Gauss–Markov properties (3)(4). The *RKHS* setup has been examined from an on-line learning perspective (Vovk 2008).

Joint Normal Distributed (*x*, *y*) as Motivation for the Linear Regression Model and Least Squares

For the moment, think of (x, y) as following a multivariate normal distribution, as might be the case under process control or in natural systems. The (regular) conditional expectation of y relative to x is then, for some β:

$$E(y|x) = Ey + E((y - Ey)|x) = Ey + x.\beta \text{ for every } x$$

and the discrepancies $y - E(y|x)$ are for each x distributed $N(0, \sigma^2)$ for a fixed σ^2, independent for different x.

Comparing Two Basic Linear Regression Models

Freedman (1981) compares the analysis of two superficially similar but differing models:

Errors model: Model (1) above.

Sampling model: Data $(x_{i1}, \ldots, x_{ip}, y_i), i \le n$ represent a *random sample* from a finite population (e.g., an actual physical population).

In the sampling model, $\{\varepsilon_i\}$ are simply the *residual discrepancies* $y\text{-}x\beta_{LS}$ of a least squares fit of linear model $x\beta$ *to the population*. Galton's seed study is an example of this if we regard his data (w, y) as resulting from equal probability without-replacement random sampling of a population of pairs (w, y) with w restricted to be at $0, \pm1, \pm2, \pm3$ standard deviations from the mean. Both with and without-replacement equal-probability sampling are considered by Freedman. Unlike the errors model there is no assumption in the sampling model that the population linear regression model is in any way correct, although least squares may not be recommended if the population residuals depart significantly from i.i.d. normal. Our only purpose is to estimate the coefficients of the population *LS* fit of the model using *LS* fit of the model to our sample, give estimates of the likely proximity of our sample least squares fit to the population fit and estimate the quality of the population fit (e.g., multiple correlation).

Freedman (1981) established the applicability of Efron's Bootstrap to each of the two models above but under different assumptions. His results for the sampling model are a textbook application of Bootstrap since a description of the sampling theory of least squares estimates for the sampling model has complexities largely, as had been said, out of the way when the Bootstrap approach is used. It would be an interesting exercise to examine data, such as Galton's

seed data, analyzing it by the two different models, obtaining confidence intervals for the estimated coefficients of a straight line fit in each case to see how closely they agree.

Balancing Good Fit Against Reproducibility

A balance in the linear regression model is necessary. Including too many independent variables in order to assure a close fit of the model to the data is called overfitting. Models over-fit in this manner tend not to work with fresh data, for example to predict y from a fresh choice of the values of the independent variables. Galton's regression line, although it did not afford very accurate predictions of y from w, owing to the modest correlation ~ 0.33, was arguably best for his bi-variate normal data (w, y). Tossing in another independent variable such as w^2 for a parabolic fit would have over-fit the data, possibly spoiling discovery of the principle of regression to mediocrity.

A model might well be used even when it is understood that incorporating additional independent variables will yield a better fit to data and a model closer to truth. How could this be? If the more parsimonious choice of x accounts for enough of the variation of y in relation to the variables of interest to be useful and if its fewer coefficients β are estimated more reliably perhaps. Intentional use of a simpler model might do a reasonably good job of giving us the estimates we need but at the same time violate assumptions about the errors thereby invalidating confidence intervals and tests. Gauss needed to come close to identifying the location at which Ceres would appear. Going for too much accuracy by complicating the model risked overfitting owing to the limited number of observations available.

One possible resolution to this tradeoff between reliable estimation of a few model coefficients, versus the risk that by doing so too much model related material is left in the error term, is to include all of several hierarchically ordered layers of independent variables, more than may be needed, then remove those that the data suggests are not required to explain the greater share of the variation of y (Raudenbush and Bryk 2001). New results on data compression (Candès and Tao 2007) may offer fresh ideas for reliably removing, in some cases, less relevant independent variables without first arranging them in a hierarchy.

Regression to Mediocrity Versus Reversion to Mediocrity or Beyond

Regression (when applicable) is often used to prove that a high performing group on some scoring, i.e., $X > c > EX$, will not average so highly on another scoring Y, as they do on X, i.e., $E(Y|X > c) < E(X|X > c)$. Termed reversion to mediocrity or beyond by Samuels (1991) this property is easily come by when X, Y have the same distribution. The following result and brief proof are Samuels' except for clarifications made here (***italics***). These comments are addressed only to the *formal mathematical proof* of the paper.

Proposition Let random variables X, Y be *identically distributed* with finite mean EX and fix any $c > \max(0, EX)$. If $P(X > c \text{ and } Y > c) < P(Y > c)$ then there is reversion to mediocrity or beyond for that c.

Proof For any given $c > \max(0, EX)$ define the difference J of indicator random variables $J = (X > c) - (Y > c)$. J is zero unless one indicator is 1 and the other 0. YJ is less *or equal* $cJ = c$ on $J = 1$ (i.e., on $X > c, Y \le c$) and YJ is *strictly less* than $cJ = -c$ on $J = -1$ (i.e., on $X \le c, Y > c$). *Since the event* $J = -1$ *has positive probability by assumption*, the previous implies $E\,YJ < c\,EJ$ and so

$$EY(X > c) = E(Y(Y > c) + YJ) = EX(X > c) + E\,YJ < EX(X > c) + cEJ = EX(X > c),$$

yielding $E(Y|X > c) < E(X|X > c)$. □

Cross References

►Adaptive Linear Regression
►Analysis of Areal and Spatial Interaction Data
►Business Forecasting Methods
►Gauss-Markov Theorem
►General Linear Models
►Heteroscedasticity
►Least Squares
►Optimum Experimental Design
►Partial Least Squares Regression Versus Other Methods
►Regression Diagnostics
►Regression Models with Increasing Numbers of Unknown Parameters
►Ridge and Surrogate Ridge Regressions
►Simple Linear Regression
►Two-Stage Least Squares

References and Further Reading

Berk RA (2004) Regression analysis: a constructive critique. Sage, Newbury park

Candès E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35(6):2313–2351

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26

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

Freedman D (1981) Bootstrapping regression models. Ann Stat 9:1218–1228

Freedman D (1991) Statistical models and shoe leather. Sociol Methodol 21:291–313
Freedman D (2005) Statistical models: theory and practice. Cambridge University Press, New York
Freedman D, Lane D (1983) A non-stochastic interpretation of reported significance levels. J Bus Econ Stat 1:292–298
Galton F (1877) Typical laws of heredity. Nature 15:493–495, 512–514, 532–533
Galton F (1886) Regression towards mediocrity in hereditary stature. J Anthropolocal Inst Great Britain and Ireland 15:246–263
Hinkelmann K, Kempthorne O (2008) Design and analysis of experiments, vol 1, 2, 2nd edn. Wiley, Hoboken
Kendall M, Stuart A (1979) The advanced theory of statistics, volume 2: Inference and relationship. Charles Griffin, London
LePage R, Podgorski K (1996) Resampling permutations in regression without second moments. J Multivariate Anal 57(11):119–141, Elsevier
Parzen E (1961) An approach to time series analysis. Ann Math Stat 32(4):951–989
Rao CR (1965) Linear statistical inference and its applications. Wiley, New York
Raudenbush S, Bryk A (2001) Hierarchical linear models applications and data analysis methods, 2nd edn. Sage, Thousand Oaks
Samuels ML (1991) Statistical reversion toward the mean: more universal than regression toward the mean. Am Stat 45:344–346
Stigler S (1986) The history of statistics: the measurement of uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge
Vovk V (2006) On-line regression competitive with reproducing kernel Hilbert spaces. Lecture notes in computer science, vol 3959. Springer, Berlin, pp 452–463

Local Asymptotic Mixed Normal Family

Ishwar V. Basawa
Professor
University of Georgia, Athens, GA, USA

Suppose $x(n) = (x_1, \ldots, x_n)$ is a sample from a stochastic process $x = \{x_1, x_2, \ldots\}$. Let $p_n(x(n);\theta)$ denote the joint density function of $x(n)$, where $\theta \epsilon \Omega \subset \mathcal{R}^k$ is a parameter. Define the log-likelihood ratio $\Lambda_n = \left[\frac{p_n(x(n);\theta_n)}{p_n(x(n);\theta)}\right]$, where $\theta_n = \theta + n^{-\frac{1}{2}}h$, and h is a $(k \times 1)$ vector. The joint density $p_n(x(n);\theta)$ belongs to a local asymptotic normal (LAN) family if Λ_n satisfies

$$\Lambda_n = n^{-\frac{1}{2}}h^t S_n(\theta) - n^{-1}\left(\frac{1}{2}h^t J_n(\theta)h\right) + o_p(1) \qquad (1)$$

where $S_n(\theta) = \frac{dlnp_n(x(n);\theta)}{d\theta}$, $J_n(\theta) = -\frac{d^2 lnp_n(x(n);\theta)}{d\theta d\theta^t}$, and

$$(i)n^{-\frac{1}{2}}S_n(\theta) \xrightarrow{d} N_k(0, F(\theta)), (ii)n^{-1}J_n(\theta) \xrightarrow{p} F(\theta), \qquad (2)$$

$F(\theta)$ being the limiting Fisher information matrix. Here, $F(\theta)$ ia assumed to be non-random. See LaCam and Yang (1990) for a review of the LAN family.

For the LAN family defined by (1) and (2), it is well known that, under some regularity conditions, the maximum likelihood (ML) estimator $\widehat{\theta}_n$ is consistent asymptotically normal and efficient estimator of θ with

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N_k(0, F^{-1}(\theta)). \qquad (3)$$

A large class of models involving the classical *i.i.d.* (independent and identically distributed) observations are covered by the LAN framework. Many time series models and ►Markov processes also are included in the LAN family.

If the limiting Fisher information matrix $F(\theta)$ is non-degenerate random, we obtain a generalization of the LAN family for which the limit distribution of the ML estimator in (3) will be a mixture of normals (and hence non-normal). If Λ_n satisfies (1) and (2) with $F(\theta)$ random, the density $p_n(x(n);\theta)$ belongs to a local asymptotic mixed normal (LAMN) family. See Basawa and Scott (1983) for a discussion of the LAMN family and related asymptotic inference questions for this family.

For the LAMN family, one can replace the norm $\sqrt{n}$ by a random norm $J_n^{\frac{1}{2}}(\theta)$ to get the limiting normal distribution, viz.,

$$J_n^{\frac{1}{2}}(\theta)(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, I), \qquad (4)$$

where I is the identity matrix.

Two examples belonging to the LAMN family are given below:

Example 1 Variance mixture of normals

Suppose, conditionally on $\mathcal{V} = v$, $(x_1, x_2, \ldots, x_n)$ are *i.i.d.* $N(\theta, v^{-1})$ random variables, and $\mathcal{V}$ is an exponential random variable with mean 1. The marginal joint density of $x(n)$ is then given by $p_n(x(n);\theta) \propto \left[1 + \frac{1}{2}\sum_1^n (x_i - \theta)^2\right]^{-\left(\frac{n}{2}+1\right)}$. It can be verified that $F(\theta)$ is an exponential random variable with mean 1. The ML estimator $\widehat{\theta}_n = \bar{x}$ and $\sqrt{n}(\bar{x} - \theta) \xrightarrow{d} t(2)$. It is interesting to note that the variance of the limit distribution of $\bar{x}$ is ∞!

Example 2 Autoregressive process

Consider a first-order autoregressive process $\{x_t\}$ defined by $x_t = \theta x_{t-1} + e_t, t = 1, 2, \ldots,$ with $x_0 = 0$, where $\{e_t\}$ are assumed to be *i.i.d.* $N(0,1)$ random variables.

We then have $p_n(x(n);\theta) \propto exp\left[-\frac{1}{2}\sum_1^n (x_t - \theta x_{t-1})^2\right]$. For the stationary case, $|\theta| < 1$, this model belongs to the LAN family. However, for $|\theta| > 1$, the model belongs to the LAMN family. For any θ, the ML estimator $\widehat{\theta}_n = \left(\sum_1^n x_i x_{i-1}\right)\left(\sum_1^n x_{i-1}^2\right)^{-1}$. One can verify that $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0,(1-\theta^2)^{-1})$, for $|\theta| < 1$, and $(\theta^2-1)^{-1}\theta^n(\widehat{\theta}_n - \theta) \xrightarrow{d}$ Cauchy, for $|\theta| > 1$.

About the Author

Dr. Ishwar Basawa is a Professor of Statistics at the University of Georgia, USA. He has served as interim head of the department (2000–2003), Executive Editor of the *Journal of Statistical Planning and Inference* (1995–1997), on the editorial board of *Communications in Statistics*, and currently the online *Journal of Probability and Statistics*. Professor Basawa is a Fellow of the Institute of Mathematical Statistics and he was an Elected member of the International Statistical Institute. He has co-authored two books and co-edited eight Proceedings/Monographs/Special Issues of journals. He has authored more than 125 publications. His areas of research include inference for stochastic processes, time series, and asymptotic statistics.

Cross References

►Asymptotic Normality
►Optimal Statistical Inference in Financial Engineering
►Sampling Problems for Stochastic Processes

References and Further Reading

Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York
LeCam L, Yang GL (1990) Asymptotics in statistics. Springer, New York

Location-Scale Distributions

Horst Rinne
Professor Emeritus for Statistics and Econometrics
Justus–Liebig–University Giessen, Giessen, Germany

A random variable X with realization x belongs to the location-scale family when its cumulative distribution is a function only of $(x-a)/b$:

$$F_X(x\,|\,a,b) = \Pr(X \le x|a,b) = F\left(\frac{x-a}{b}\right);\ a \in \mathbb{R},\ b > 0;$$

where $F(\cdot)$ is a distribution having no other parameters. Different $F(\cdot)$'s correspond to different members of the family. (a,b) is called the location–scale parameter, a being the location parameter and b being the scale parameter. For fixed $b = 1$ we have a subfamily which is a location family with parameter a, and for fixed $a = 0$ we have a scale family with parameter b. The variable

$$Y = \frac{X-a}{b}$$

is called the reduced or standardized variable. It has $a = 0$ and $b = 1$. If the distribution of X is absolutely continuous with density function

$$f_X(x\,|\,a,b) = \frac{\mathrm{d}F_X(x\,|\,a,b)}{\mathrm{d}\,x}$$

then (a,b) is a location scale-parameter for the distribution of X if (and only if)

$$f_X(x\,|\,a,b) = \frac{1}{b} f\left(\frac{x-a}{b}\right)$$

for some density $f(\cdot)$, called the reduced density. All distributions in a given family have the same shape, i.e., the same skewness and the same kurtosis. When Y has mean μ_Y and standard deviation σ_Y then, the mean of X is $\mathrm{E}(X) = a + b\,\mu_Y$ and the standard deviation of X is $\sqrt{\mathrm{Var}(X)} = b\,\sigma_Y$.

The location parameter a, $a \in \mathbb{R}$ is responsible for the distribution's position on the abscissa. An enlargement (reduction) of a causes a movement of the distribution to the right (left). The location parameter is either a measure of central tendency e.g., the mean, median and mode or it is an upper or lower threshold parameter. The scale parameter b, $b > 0$, is responsible for the dispersion or variation of the variate X. Increasing (decreasing) b results in an enlargement (reduction) of the spread and a corresponding reduction (enlargement) of the density. b may be the standard deviation, the full or half length of the support, or the length of a central $(1-\alpha)$–interval.

The location-scale family has a great number of members:

- Arc-sine distribution
- Special cases of the beta distribution like the rectangular, the asymmetric triangular, the U–shaped or the power–function distributions
- Cauchy and half–Cauchy distributions
- Special cases of the χ–distribution like the half–normal, the Rayleigh and the Maxwell–Boltzmann distributions
- Ordinary and raised cosine distributions
- Exponential and reflected exponential distributions
- Extreme value distribution of the maximum and the minimum, each of type I
- Hyperbolic secant distribution

- Laplace distribution
- Logistic and half–logistic distributions
- Normal and half–normal distributions
- Parabolic distributions
- Rectangular or uniform distribution
- Semi–elliptical distribution
- Symmetric triangular distribution
- Teissier distribution with reduced density $f(y) = \left[\exp(y) - 1\right] \exp\left[1 + y - \exp(y)\right], y \geq 0$
- V–shaped distribution

For each of the above mentioned distributions we can design a special probability paper. Conventionally, the abscissa is for the realization of the variate and the ordinate, called the probability axis, displays the values of the cumulative distribution function, but its underlying scaling is according to the percentile function. The ordinate value belonging to a given sample data on the abscissa is called plotting position; for its choice see Barnett (1975, 1976), Blom (1958), Kimball (1960). When the sample comes from the probability paper's distribution the plotted data will randomly scatter around a straight line, thus, we have a graphical goodness-fit-test. When we fit the straight line by eye we may read off estimates for a and b as the abscissa or difference on the abscissa for certain percentiles. A more objective method is to fit a least-squares line to the data, and the estimates of a and b will be the the parameters of this line.

The latter approach takes the order statistics $X_{i:n}$, $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ as regressand and the mean of the reduced order statistics $\alpha_{i:n} := \mathrm{E}\left(Y_{i:n}\right)$ as regressor, which under these circumstances acts as plotting position. The regression model reads:

$$X_{i:n} = a + b\,\alpha_{i:n} + \varepsilon_i,$$

where ε_i is a random variable expressing the difference between $X_{i:n}$ and its mean $\mathrm{E}\left(X_{i:n}\right) = a + b\,\alpha_{i:n}$. As the order statistics $X_{i:n}$ and – as a consequence – the disturbance terms ε_i are neither homoscedastic nor uncorrelated we have to use – according to Lloyd (1952) – the general-least-squares method to find best linear unbiased estimators of a and b. Introducing the following vectors and matrices:

$$\boldsymbol{x} := \begin{pmatrix} X_{1:n} \\ X_{2:n} \\ \vdots \\ X_{n:n} \end{pmatrix}, \mathbf{1} := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{\alpha} := \begin{pmatrix} \alpha_{1:n} \\ \alpha_{2:n} \\ \vdots \\ \alpha_{n:n} \end{pmatrix}, \boldsymbol{\varepsilon} := \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \boldsymbol{\theta} := \begin{pmatrix} a \\ b \end{pmatrix},$$

$$\boldsymbol{A} := \begin{pmatrix} \mathbf{1} & \boldsymbol{\alpha} \end{pmatrix}$$

the regression model now reads

$$\boldsymbol{x} = \boldsymbol{A}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

with variance–covariance matrix

$$\mathrm{Var}(\boldsymbol{x}) = b^2\,\boldsymbol{B}.$$

The GLS estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{x}$$

and its variance–covariance matrix reads

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) = b^2\left(\boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{A}\right)^{-1}.$$

The vector $\boldsymbol{\alpha}$ and the matrix $\boldsymbol{B}$ are not always easy to find. For only a few location–scale distributions like the exponential, the reflected exponential, the extreme value, the logistic and the rectangular distributions we have closed-form expressions, in all other cases we have to evaluate the integrals defining $\mathrm{E}\left(Y_{i:n}\right)$ and $\mathrm{E}\left(Y_{i:n}\,Y_{j:n}\right)$. For more details on linear estimation and probability plotting for location-scale distributions and for distributions which can be transformed to location–scale type see Rinne (2010). Maximum likelihood estimation for location–scale distributions is treated by Mi (2006).

About the Author

Dr. Horst Rinne is Professor Emeritus (since 2005) of statistics and econometrics. In 1971 he was awarded the Venia legendi in statistics and econometrics by the Faculty of economics of Berlin Technical University. From 1965 to 1972 he worked as a part-time lecturer of statistics at Berlin Polytechnic and at the Technical University of Hannover. In 1969 he joined Volkswagen AG to do operations research. In 1972 he was appointed full professor of statistics and econometrics at the Justus-Liebig University in Giessen, where later on he was Dean of the faculty of economics and management science for three years. He got further appointments to the universities of Tübingen and Cologne and to Berlin Polytechnic. He was a member of the board of the German Statistical Society and in charge of editing its journal *Allgemeines Statistisches Archiv*, now *AStA – Advances in Statistical Analysis*, from 1981 to 1997. He is Co-founder of the Econometric Board of the German Society of Economics. Since 1985 he is Elected member of the International Statistical Institute. He is Associate editor for *Quality Technology and Quantitative Management*. His scientific interests are wide-spread, ranging from financial mathematics, business and economics statistics, technical statistics to econometrics. He has written numerous papers in these fields and is author of several textbooks and monographs in statistics, econometrics, time series analysis, multivariate statistics and statistical quality control including process capability. He is the author of the text *The Weibull Distribution: A Handbook* (Chapman and Hall/CRC, 2008).

Cross References

►Statistical Distributions: An Overview

References and Further Reading

Barnett V (1975) Probability plotting methods and order statistics. Appl Stat 24:95–108

Barnett V (1976) Convenient probability plotting positions for the normal distribution. Appl Stat 25:47–50

Blom G (1958) Statistical estimates and transformed beta variables. Almqvist and Wiksell, Stockholm

Kimball BF (1960) On the choice of plotting positions on probability paper. J Am Stat Assoc 55:546–560

Lloyd EH (1952) Least-squares estimation of location and scale parameters using order statistics. Biometrika 39:88–95

Mi J (2006) MLE of parameters of location–scale distributions for complete and partially grouped data. J Stat Planning Inference 136:3565–3582

Rinne H (2010) Location-scale distributions – linear estimation and probability plotting; http:geb.uni-giessen/geb/volltexte/2010/7607/

Logistic Normal Distribution

John Hinde
Professor of Statistics
National University of Ireland, Galway, Ireland

The logistic-normal distribution arises by assuming that the *logit* (or logistic transformation) of a proportion has a normal distribution, with an obvious extension to a vector of proportions through taking a logistic transformation of a multivariate normal distribution, see Aitchison and Shen (1980). In the univariate case, this provides a family of distributions on $(0,1)$ that is distinct from the ►*beta distribution*, while the multivariate version is an alternative to the *Dirichlet distribution*. Note that in the multivariate case there is no unique way to define the set of logits for the multinomial proportions (just as in multinomial logit models, see Agresti 2002) and different formulations may be appropriate in particular applications (Aitchison 1982). The univariate distribution has been used, often implicitly, in random effects models for binary data and the multivariate version was pioneered by Aitchison for statistical diagnosis/discrimination (Aitchison and Begg 1976), the Bayesian analysis of contingency tables and the analysis of compositional data (Aitchison 1982, 1986).

The use of the logistic-normal distribution is most easily seen in the analysis of binary data where the logit model (based on a logistic tolerance distribution) is extended to the *logit-normal* model. For grouped binary data with responses r_i out of m_i trials ($i = 1,\ldots,n$), the response probabilities, P_i, are assumed to have a logistic-normal distribution with $\text{logit}(P_i) = \log(P_i/(1-P_i)) \sim N(\mu_i, \sigma^2)$, where μ_i is modelled as a linear function of explanatory variables, $x_1,\ldots,x_p$. The resulting model can be summarized as

$$\begin{aligned} R_i|P_i &\sim \text{Binomial}(m_i, P_i) \\ \text{logit}(P_i)|Z = \eta_i + \sigma Z &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma Z \\ Z &\sim N(0,1) \end{aligned}$$

This is a simple extension of the basic logit model with the inclusion of a single normally distributed random effect in the linear predictor, an example of a *generalized linear mixed model*, see McCulloch and Searle (2001). Maximum likelihood estimation for this model is complicated by the fact that the likelihood has no closed form and involves integration over the normal density, which requires numerical methods using Gaussian quadrature; routines now exist as part of generalized linear mixed model fitting in all major software packages, such as SAS, R, Stata and Genstat. Approximate moment-based estimation methods make use of the fact that if σ^2 is small then, as derived in Williams (1982),

$$\begin{aligned} E[R_i] &= m_i\pi_i \quad \text{and} \\ Var(R_i) &= m_i\pi_i(1-\pi_1)[1+\sigma^2(m_i-1)\pi_i(1-\pi_i)] \end{aligned}$$

where $\text{logit}(\pi_i) = \eta_i$. The form of the variance function shows that this model is *overdispersed* compared to the binomial, that is it exhibits greater variability; the random effect Z allows for unexplained variation across the grouped observations. However, note that for binary data ($m_i = 1$) it is not possible to have overdispersion arising in this way.

About the Author

For biography *see* the entry ►Logistic Distribution.

Cross References

►Logistic Distribution
►Mixed Membership Models
►Multivariate Normal Distributions
►Normal Distribution, Univariate

References and Further Reading

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

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J R Stat Soc Ser B 44:139–177

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London

Aitchison J, Begg CB (1976) Statistical diagnosis when basic cases are not classified with certainty. Biometrika 63:1–12

Aitchison J, Shen SM (1980) Logistic-normal distributions: some properties and uses. Biometrika 67:261–272
McCulloch CE, Searle SR (2001) Generalized, linear and mixed models. Wiley, New York
Williams D (1982) Extra-binomial variation in logistic linear models. Appl Stat 31:144–148

Logistic Regression

Joseph M. Hilbe
Emeritus Professor
University of Hawaii, Honolulu, HI, USA
Adjunct Professor of Statistics
Arizona State University, Tempe, AZ, USA
Solar System Ambassador
California Institute of Technology, Pasadena, CA, USA

Logistic regression is the most common method used to model binary response data. When the response is binary, it typically takes the form of 1/0, with 1 generally indicating a success and 0 a failure. However, the actual values that 1 and 0 can take vary widely, depending on the purpose of the study. For example, for a study of the odds of failure in a school setting, 1 may have the value of *fail*, and 0 of *not-fail*, or pass. The important point is that 1 indicates the foremost subject of interest for which a binary response study is designed. Modeling a binary response variable using normal linear regression introduces substantial bias into the parameter estimates. The standard linear model assumes that the response and error terms are normally or Gaussian distributed, that the variance, σ^2, is constant across observations, and that observations in the model are independent. When a binary variable is modeled using this method, the first two of the above assumptions are violated. Analogical to the normal regression model being based on the Gaussian probability distribution function (*pdf*), a binary response model is derived from a Bernoulli distribution, which is a subset of the binomial *pdf* with the binomial denominator taking the value of 1. The Bernoulli *pdf* may be expressed as:

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \quad (1)$$

Binary logistic regression derives from the canonical form of the Bernoulli distribution. The Bernoulli *pdf* is a member of the exponential family of probability distributions, which has properties allowing for a much easier estimation of its parameters than traditional Newton–Raphson-based maximum likelihood estimation (*MLE*) methods.

In 1972 Nelder and Wedderbrun discovered that it was possible to construct a single algorithm for estimating models based on the exponential family of distributions. The algorithm was termed ▶Generalized linear models (*GLM*), and became a standard method to estimate binary response models such as logistic, probit, and complimentary-loglog regression, count response models such as Poisson and negative binomial regression, and continuous response models such as gamma and inverse Gaussian regression. The standard normal model, or Gaussian regression, is also a generalized linear model, and may be estimated under its algorithm. The form of the exponential distribution appropriate for generalized linear models may be expressed as

$$f(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/\alpha(\phi) + c(y_i;\phi)\}, \quad (2)$$

with θ representing the link function, $\alpha(\phi)$ the scale parameter, $b(\theta)$ the cumulant, and $c(y;\phi)$ the normalization term, which guarantees that the probability function sums to 1. The link, a monotonically increasing function, linearizes the relationship of the expected mean and explanatory predictors. The scale, for binary and count models, is constrained to a value of 1, and the cumulant is used to calculate the model mean and variance functions. The mean is given as the first derivative of the cumulant with respect to θ, $b'(\theta)$; the variance is given as the second derivative, $b''(\theta)$. Taken together, the above four terms define a specific *GLM* model.

We may structure the Bernoulli distribution (3) into exponential family form (2) as:

$$f(y_i;\pi_i) = \exp\{y_i \ln(\pi_i/(1-\pi_i)) + \ln(1-\pi_i)\}. \quad (3)$$

The link function is therefore $\ln(\pi/(1-\pi))$, and cumulant $-\ln(1-\pi)$ or $\ln(1/(1-\pi))$. For the Bernoulli, π is defined as the probability of success. The first derivative of the cumulant is π, the second derivative, $\pi(1-\pi)$. These two values are, respectively, the mean and variance functions of the Bernoulli *pdf*. Recalling that the logistic model is the canonical form of the distribution, meaning that it is the form that is directly derived from the *pdf*, the values expressed in (3), and the values we gave for the mean and variance, are the values for the logistic model.

Estimation of statistical models using the *GLM* algorithm, as well as *MLE*, are both based on the log-likelihood function. The likelihood is simply a re-parameterization of the *pdf* which seeks to estimate π, for example, rather than y. The log-likelihood is formed from the likelihood by taking the natural log of the function, allowing summation

across observations during the estimation process rather than multiplication.

The traditional *GLM* symbol for the mean, μ, is typically substituted for π, when *GLM* is used to estimate a logistic model. In that form, the log-likelihood function for the binary-logistic model is given as:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i/(1-\mu_i)) + \ln(1-\mu_i)\}, \quad (4)$$

or

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) + (1-y_i)\ln(1-\mu_i)\}. \quad (5)$$

The Bernoulli-logistic log-likelihood function is essential to logistic regression. When *GLM* is used to estimate logistic models, many software algorithms use the deviance rather than the log-likelihood function as the basis of convergence. The deviance, which can be used as a goodness-of-fit statistic, is defined as twice the difference of the saturated log-likelihood and model log-likelihood. For logistic model, the deviance is expressed as

$$D = 2\sum_{i=1}^{n} \{y_i \ln(y_i/\mu_i) + (1-y_i)\ln((1-y_i)/(1-\mu_i))\}. \quad (6)$$

Whether estimated using maximum likelihood techniques or as *GLM*, the value of μ for each observation in the model is calculated on the basis of the linear predictor, $x'\beta$. For the normal model, the predicted fit, $\hat{y}$, is identical to $x'\beta$, the right side of (7). However, for logistic models, the response is expressed in terms of the link function, $\ln(\mu_i/(1-\mu_i))$. We have, therefore,

$$x_i'\beta = \ln(\mu_i/(1-\mu_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \quad (7)$$

The value of μ_i, for each observation in the logistic model, is calculated as

$$\mu_i = 1/\left(1+\exp\left(-x_i'\beta\right)\right) = \exp\left(x_i'\beta\right)/\left(1+\exp\left(x_i'\beta\right)\right). \quad (8)$$

The functions to the right of μ are commonly used ways of expressing the logistic inverse link function, which converts the linear predictor to the fitted value. For the logistic model, μ is a probability.

When logistic regression is estimated using a Newton-Raphson type of *MLE* algorithm, the log-likelihood function as parameterized to $x'\beta$ rather than μ. The estimated fit is then determined by taking the first derivative of the log-likelihood function with respect to β, setting it to zero, and solving. The first derivative of the log-likelihood function is commonly referred to as the gradient, or score function. The second derivative of the log-likelihood with respect to β produces the Hessian matrix, from which the standard errors of the predictor parameter estimates are derived. The logistic gradient and hessian functions are given as

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) x_i \quad (9)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \left\{x_i x_i' \mu_i (1-\mu_i)\right\} \quad (10)$$

One of the primary values of using the logistic regression model is the ability to interpret the exponentiated parameter estimates as odds ratios. Note that the link function is the log of the odds of μ, $\ln(\mu/(1-\mu))$, where the odds are understood as the success of μ over its failure, $1-\mu$. The log-odds is commonly referred to as the *logit* function. An example will help clarify the relationship, as well as the interpretation of the odds ratio.

We use data from the 1912 Titanic accident, comparing the odds of survival for adult passengers to children. A tabulation of the data is given as:

Survived	Age (Child vs Adult)		Total
	child	adults	
no	52	765	817
yes	57	442	499
Total	109	1,207	1,316

The odds of survival for adult passengers is 442/765, or 0.578. The odds of survival for children is 57/52, or 1.096. The ratio of the odds of survival for adults to the odds of survival for children is (442/765)/(57/52), or 0.52709552. This value is referred to as the *odds ratio*, or ratio of the two component odds relationships. Using a logistic regression procedure to estimate the odds ratio of age produces the following results

survived	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.5270955	.1058718	−3.19	0.001	.3555642	.7813771

With 1 = *adult* and 0 = *child*, the estimated odds ratio may be interpreted as:

- *The odds of an adult surviving were about half the odds of a child surviving.*

By inverting the estimated odds ratio above, we may conclude that children had [1/.527 ~ 1.9] some 90% – or

nearly two times – greater odds of surviving than did adults.

For continuous predictors, a one-unit increase in a predictor value indicates the change in odds expressed by the displayed odds ratio. For example, if age was recorded as a continuous predictor in the Titanic data, and the odds ratio was calculated as 1.015, we would interpret the relationship as:

- *The odds of surviving is one and a half percent greater for each increasing year of age.*

Non-exponentiated logistic regression parameter estimates are interpreted as log-odds relationships, which carry little meaning in ordinary discourse. Logistic models are typically interpreted in terms of odds ratios, unless a researcher is interested in estimating predicted probabilities for given patterns of model covariates; i.e., in estimating μ.

Logistic regression may also be used for grouped or proportional data. For these models the response consists of a numerator, indicating the number of successes (1s) for a specific covariate pattern, and the denominator (m), the number of observations having the specific covariate pattern. The response y/m is binomially distributed as:

$$f(y_i;\pi_i,m_i) = \binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}, \qquad (11)$$

with a corresponding log-likelihood function expressed as

$$L(\mu_i;y_i,m_i) = \sum_{i=1}^{n}\left\{y_i\ln(\mu_i/(1-\mu_i)) + m_i\ln(1-\mu_i) + \binom{m_i}{y_i}\right\}. \qquad (12)$$

Taking derivatives of the cumulant, $-m_i\ln(1-\mu_i)$, as we did for the binary response model, produces a mean of $\mu_i = m_i\pi_i$ and variance, $\mu_i(1-\mu_i/m_i)$.

Consider the data below:

y	cases	x_1	x_2	x_3
1	3	1	0	1
1	1	1	1	1
2	2	0	0	1
0	1	0	1	1
2	2	1	0	0
0	1	0	1	0

y indicates the number of times a specific pattern of covariates is successful. *Cases* is the number of observations having the specific covariate pattern. The first observation in the table informs us that there are three cases having predictor values of $x_1 = 1, x_2 = 0$, and $x_3 = 1$. Of those three cases, one has a value of y equal to 1, the other two have values of 0. All current commercial software applications estimate this type of logistic model using *GLM* methodology.

y	Odds ratio	OIM std. err.	z	$P>\|z\|$	[95% conf.	interval]
x_1	1.186947	1.769584	0.11	0.908	0.0638853	22.05271
x_2	0.2024631	0.3241584	−1.00	0.318	0.0087803	4.668551
x_3	0.5770337	0.9126937	−0.35	0.728	0.025993	12.8099

The data in the above table may be restructured so that it is in individual observation format, rather than grouped. The new table would have ten observations, having the same logic as described. Modeling would result in identical parameter estimates. It is not uncommon to find an individual-based data set of, for example, 10,000 observations, being grouped into 10–15 rows or observations as above described. Data in tables is nearly always expressed in grouped format.

Logistic models are subject to a variety of fit tests. Some of the more popular tests include the Hosmer-Lemeshow goodness-of-fit test, *ROC* analysis, various information criteria tests, link tests, and residual analysis. The Hosmer–Lemeshow test, once well used, is now only used with caution. The test is heavily influenced by the manner in which tied data is classified. Comparing observed with expected probabilities across levels, it is now preferred to construct tables of risk having different numbers of levels. If there is consistency in results across tables, then the statistic is more trustworthy.

Information criteria tests, e.g., Akaike information Criteria (see ▶Akaike's Information Criterion and ▶Akaike's Information Criterion: Background, Derivation, Properties, and Refinements) (*AIC*) and Bayesian Information Criteria (*BIC*) are the most used of this type of test. Information tests are comparative, with lower values indicating the preferred model. Recent research indicates that *AIC* and *BIC* both are biased when data is correlated to any degree. Statisticians have attempted to develop enhancements of these two tests, but have not been entirely successful. The best advice is to use several different types of tests, aiming for consistency of results.

Several types of residual analyses are typically recommended for logistic models. The references below provide extensive discussion of these methods, together with

appropriate caveats. However, it appears well established that m-asymptotic residual analyses is most appropriate for logistic models having no continuous predictors. m-asymptotics is based on grouping observations with the same covariate pattern, in a similar manner to the grouped or binomial logistic regression discussed earlier. The Hilbe (2009) and Hosmer and Lemeshow (2000) references below provide guidance on how best to construct and interpret this type of residual.

Logistic models have been expanded to include categorical responses, e.g., proportional odds models and multinomial logistic regression. They have also been enhanced to include the modeling of panel and correlated data, e.g., generalized estimating equations, fixed and random effects, and mixed effects logistic models.

Finally, exact logistic regression models have recently been developed to allow the modeling of perfectly predicted data, as well as small and unbalanced datasets. In these cases, logistic models which are estimated using GLM or full maximum likelihood will not converge. Exact models employ entirely different methods of estimation, based on large numbers of permutations.

About the Author

Joseph M. Hilbe is an emeritus professor, University of Hawaii and adjunct professor of statistics, Arizona State University. He is also a Solar System Ambassador with NASA/Jet Propulsion Laboratory, at California Institute of Technology. Hilbe is a Fellow of the American Statistical Association and Elected Member of the International Statistical institute, for which he is founder and chair of the ISI astrostatistics committee and Network, the first global association of astrostatisticians. He is also chair of the ISI sports statistics committee, and was on the founding executive committee of the Health Policy Statistics Section of the American Statistical Association (1994–1996). Hilbe is author of *Negative Binomial Regression* (2007, Cambridge University Press), and *Logistic Regression Models* (2009, Chapman & Hall), two of the leading texts in their respective areas of statistics. He is also co-author (with James Hardin) of *Generalized Estimating Equations* (2002, Chapman & Hall/CRC) and two editions of *Generalized Linear Models and Extensions* (2001, 2007, Stata Press), and with Robert Muenchen is coauthor of the *R for Stata Users* (2010, Springer). Hilbe has also been influential in the production and review of statistical software, serving as Software Reviews Editor for *The American Statistician* for 12 years from 1997–2009. He was founding editor of the *Stata Technical Bulletin* (1991), and was the first to add the negative binomial family into commercial generalized linear models software. Professor Hilbe was presented the Distinguished Alumnus award at California State University, Chico in 2009, two years following his induction into the University's Athletic Hall of Fame (he was two-time US champion track & field athlete). He is the only graduate of the university to be recognized with both honors.

Cross References

►Case-Control Studies
►Categorical Data Analysis
►Generalized Linear Models
►Multivariate Data Analysis: An Overview
►Probit Analysis
►Recursive Partitioning
►Regression Models with Symmetrical Errors
►Robust Regression Estimation in Generalized Linear Models
►Statistics: An Overview
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Collett D (2003) Modeling binary regression, 2nd edn. Chapman & Hall/CRC Cox, London
Cox DR, Snell EJ (1989) Analysis of binary data, 2nd edn. Chapman & Hall, London
Hardin JW, Hilbe JM (2007) Generalized linear models and extensions, 2nd edn. Stata Press, College Station
Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC Press, Boca Raton
Hosmer D, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
Kleinbaum DG (1994) Logistic regression; a self-teaching guide. Springer, New York
Long JS (1997) Regression models for categorical and limited dependent variables. Sage, Thousand Oaks
McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Logistic Distribution

John Hinde
Professor of Statistics
National University of Ireland, Galway, Ireland

The logistic distribution is a location-scale family distribution with a very similar shape to the normal (Gaussian) distribution but with somewhat heavier tails. The distribution has applications in reliability and survival analysis. The cumulative distribution function has been used for

modelling growth functions and as a tolerance distribution in the analysis of binary data, leading to the widely used *logit* model. For a detailed discussion of the properties of the logistic and related distributions, see Johnson et al. (1995).

The probability density function is

$$f(x) = \frac{1}{\tau}\frac{\exp\{-(x-\mu)/\tau\}}{[1+\exp\{-(x-\mu)/\tau\}]^2}, \quad -\infty < x < \infty \quad (1)$$

and the cumulative distribution function is

$$F(x) = \frac{1}{[1+\exp\{-(x-\mu)/\tau\}]}, \quad -\infty < x < \infty .$$

The distribution is symmetric about the mean μ and has variance $\tau^2\pi^2/3$, so that when comparing the standard logistic distribution ($\mu = 0$, $\tau = 1$) with the standard normal distribution, $N(0,1)$, it is important to allow for the different variances. The suitably scaled logistic distribution has a very similar shape to the normal, although the kurtosis is 4.2 which is somewhat larger than the value of 3 for the normal, indicating the heavier tails of the logistic distribution.

In survival analysis, one advantage of the logistic distribution, over the normal, is that both *right-* and *left-censoring* can be easily handled. The *survivor* and *hazard* functions are given by

$$S(x) = \frac{1}{[1+\exp\{(x-\mu)/\tau\}]}, \quad -\infty < x < \infty$$

$$h(x) = \frac{1}{\tau}\frac{1}{[1+\exp\{-(x-\mu)/\tau\}]}.$$

The hazard function has the same logistic form and is monotonically increasing, so the model is only appropriate for ageing systems with an increasing failure rate over time. In modelling the dependence of failure times on explanatory variables, if we use a linear regression model for μ, then the fitted model has an *accelerated failure time* interpretation for the effect of the variables. Fitting of this model to right- and left-censored data is described in Aitkin et al. (2009).

One obvious extension for modelling failure times, T, is to assume a logistic model for $\log T$, giving a *log-logistic* model for T analagous to the *lognormal model*. The resulting hazard function based on the logistic distribution in (1) is

$$h(t) = \frac{\alpha}{\theta}\frac{(t/\theta)^{\alpha-1}}{1+(t/\theta)^{\alpha}}, \quad t, \theta, \alpha > 0$$

where $\theta = e^{\mu}$ and $\alpha = 1/\tau$. For $\alpha \leq 1$ the hazard is monotone decreasing, and for $\alpha > 1$ it has a single maximum as for the lognormal distribution; hazards of this form may be appropriate in the analysis of data such as heart transplant survival – there may be an initial period of increasing hazard associated with rejection, followed by decreasing hazard as the patient survives the procedure and the transplanted organ is accepted.

For the standard logistic distribution ($\mu = 0$, $\tau = 1$), the probability density and the cumulative distribution functions are related through the very simple identity

$$f(x) = F(x)\,[1 - F(x)]$$

which in turn, by elementary calculus, implies that

$$\text{logit}(F(x)) := \log_e\left[\frac{F(x)}{1-F(x)}\right] = x \quad (2)$$

and uniquely characterizes the standard logistic distribution. Equation (2) provides a very simple way for simulating from the standard logistic distribution by setting $X = \log_e[U/(1-U)]$ where $U \sim U(0,1)$; for the general logistic distribution in (1) we take $\tau X + \mu$.

The logit transformation is now very familiar in modelling probabilities for binary responses. Its use goes back to Berkson (1944), who suggested the use of the logistic distribution to replace the normal distribution as the underlying tolerance distribution in quantal bio-assays, where various dose levels are given to groups of subjects (animals) and a simple binary response (e.g., cure, death, etc.) is recorded for each individual (giving r-out-of-n type response data for the groups). The use of the normal distribution in this context had been pioneered by Finney through his work on ▶probit analysis and the same methods mapped across to the logit analysis, see Finney (1978) for a historical treatment of this area. The probability of response, $P(d)$, at a particular dose level d is modelled by a linear logit model

$$\text{logit}(P(d)) = \log_e\left[\frac{P(d)}{1-P(d)}\right] = \beta_0 + \beta_1 d$$

which, by the identity (2), implies a logistic tolerance distribution with parameters $\mu = -\beta_0/\beta_1$ and $\tau = 1/|\beta_1|$. The logit transformation is computationally convenient and has the nice interpretation of modelling the *log-odds* of a response. This goes across to general logistic regression models for binary data where parameter effects are on the log-odds scale and for a two-level factor the fitted effect corresponds to a log-odds-ratio. Approximate methods for parameter estimation involve using the empirical logits of the observed proportions. However, maximum likelihood estimates are easily obtained with standard generalized linear model fitting software, using a binomial response distribution with a logit link function for the response probability; this uses the iteratively reweighted least-squares Fisher-scoring algorithm of

Nelder and Wedderburn (1972), although Newton-based algorithms for maximum likelihood estimation of the logit model appeared well before the unifying treatment of ►generalized linear models. A comprehensive treatment of ►logistic regression including models and applications is given in Agresti (2002) and Hilbe (2009).

About the Author

John Hinde is Professor of Statistics, School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland. He is past President of the Irish Statistical Association (2006–2008), the Statistical Modelling Society (2004–2006) and the European Regional Section of the International Association of Statistical Computing (2000–2002). He has been an active member of the International Biometric Society and is the incoming President of the British and Irish Region (2011). He is an Elected member of the International Statistical Institute (2001). He has authored or coauthored over 50 papers mainly in the area of statistical modelling and statistical computing and is coauthor of several books, including most recently *Statistical Modelling in R* (with M. Aitkin, B. Francis, and R. Darnell, Oxford University Press, 2009). He is currently an Associate Editor of *Statistics and Computing* and was one of the joint Founding Editors of *Statistical Modelling*.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Bivariate Distributions
►Location-Scale Distributions
►Logistic Regression
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Aitkin M, Francis B, Hinde J, Darnell R (2009) Statistical modelling in R. Oxford University Press, Oxford
Berkson J (1944) Application of the logistic function to bio-assay. J Am Stat Assoc 39:357–365
Finney DJ (1978) Statistical method in biological assay, 3rd edn. Griffin, London
Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC Press, Boca Raton
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 2, 2nd edn. Wiley, New York
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Lorenz Curve

Johan Fellman
Professor Emeritus
Folkhälsan Institute of Genetics, Helsinki, Finland

Definition and Properties

It is a general rule that income distributions are skewed. Although various distribution models, such as the Lognormal and the Pareto have been proposed, they are usually applied in specific situations. For general studies, more wide-ranging tools have to be applied, the first and most common tool of which is the Lorenz curve. Lorenz (1905) developed it in order to analyze the distribution of income and wealth within populations, describing it in the following way:

- *Plot along one axis accumulated percentages of the population from poorest to richest, and along the other, wealth held by these percentages of the population.*

The Lorenz curve $L(p)$ is defined as a function of the proportion p of the population. $L(p)$ is a curve starting from the origin and ending at point $(1,1)$ with the following additional properties (I) $L(p)$ is monotone increasing, (II) $L(p) \le p$, (III) $L(p)$ convex, (IV) $L(0) = 0$ and $L(1) = 1$. The Lorenz curve is convex because the income share of the poor is less than their proportion of the population (Fig. 1).

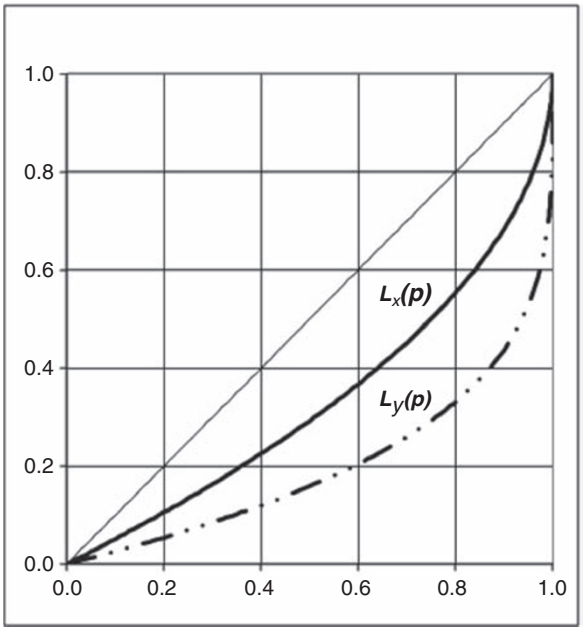
The Lorenz curve satisfies the general rules:

- *A unique Lorenz curve corresponds to every distribution. The contrary does not hold, but every Lorenz $L(p)$ is a common curve for a whole class of distributions $F(\theta x)$ where θ is an arbitrary constant.*

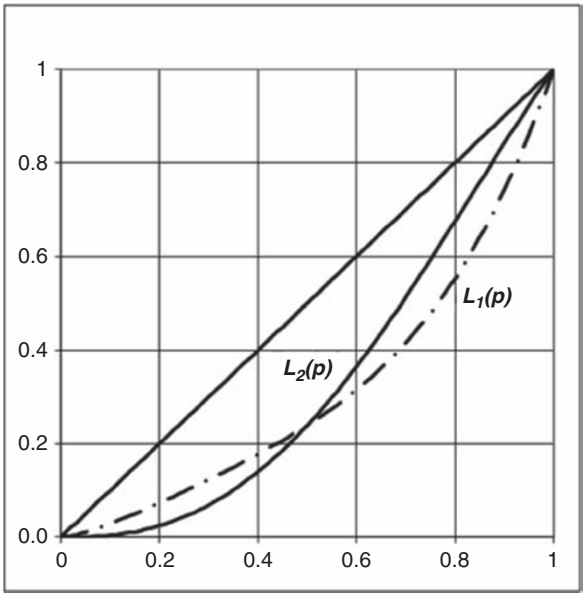
The higher the curve, the less inequality in the income distribution. If all individuals receive the same income, then the Lorenz curve coincides with the diagonal from $(0,0)$ to $(1,1)$. Increasing inequality lowers the Lorenz curve, which can converge towards the lower right corner of the square.

Consider two Lorenz curves $L_X(p)$ and $L_Y(p)$. If $L_X(p) \ge L_Y(p)$ for all p, then the distribution corresponding to $L_X(p)$ has lower inequality than the distribution corresponding to $L_Y(p)$ and is said to Lorenz dominate the other. Figure 1 shows an example of Lorenz curves.

The inequality can be of a different type, the corresponding Lorenz curves may intersect, and for these no Lorenz ordering holds. This case is seen in Fig. 2. Under such circumstances, alternative inequality measures have to be defined, the most frequently used being the Gini index, G, introduced by Gini (1912). This index is the ratio



Lorenz Curve. Fig. 1 Lorenz curves with Lorenz ordering; that is, $L_X(p) \geq L_Y(p)$



Lorenz Curve. Fig. 2 Two intersecting Lorenz curves. Using the Gini index $L_1(p)$ has greater inequality ($G_1 = 0.37$) than $L_2(p)$ ($G_2 = 0.33$)

between the area between the diagonal and the Lorenz curve and the whole area under the diagonal. This definition yields Gini indices satisfying the inequality $0 \leq G \leq 1$. The higher the G value, the greater the inequality in the income distribution.

Income Redistributions

It is a well-known fact that progressive taxation reduces inequality. Similar effects can be obtained by appropriate transfer policies, findings based on the following general theorem (Fellman 1976; Jakobsson 1976; Kakwani 1977):

Theorem *Let $u(x)$ be a continuous monotone increasing function and assume that $\mu_Y = E(u(X))$ exists. Then the Lorenz curve $L_Y(p)$ for $Y = u(X)$ exists and*

(I) $L_Y(p) \geq L_X(p)$ *if* $\frac{u(x)}{x}$ *is monotone decreasing*

(II) $L_Y(p) = L_X(p)$ *if* $\frac{u(x)}{x}$ *is constant*

(III) $L_Y(p) \leq L_X(p)$ *if* $\frac{u(x)}{x}$ *is monotone increasing.*

For progressive taxation rules, $\frac{u(x)}{x}$ measures the proportion of post-tax income to initial income and is a monotone-decreasing function satisfying condition (I), and the Gini index is reduced. Hemming and Keen (1993) gave an alternative condition for the Lorenz dominance, which is that $\frac{u(x)}{x}$ crosses the $\frac{\mu_Y}{\mu_X}$ level once from above. If the taxation rule is a flat tax, then (II) holds and the Lorenz curve and the Gini index remain. The third case in Theorem 1 indicates that the ratio $\frac{u(x)}{x}$ is increasing and the Gini index increases, but this case has only minor practical importance.

A crucial study concerning income distributions and redistributions is the monograph by Lambert (2001).

About the Author

Dr Johan Fellman is Professor Emeritus in statistics. He obtained Ph.D. in mathematics (University of Helsinki) in 1974 with the thesis On the Allocation of Linear Observations and in addition, he has published about 180 printed articles. He served as professor in statistics at Hanken School of Economics, (1977–1994) and has been a scientist at Folkhälsan Institute of Genetics since 1963. He is member of the Editorial Board of *Twin Research and Human Genetics* and of the Advisory Board of *Mathematica Slovaca* and Editor of *InterStat*. He has been awarded the Knight, First Class, of the Order of the White Rose of Finland and the Medal in Silver of Hanken School of Economics. He is Elected member of the Finnish Society of Sciences and Letters and of the International Statistical Institute.

Cross References

►Econometrics
►Economic Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Fellman J (1976) The effect of transformations on Lorenz curves. Econometrica 44:823–824
Gini C (1912) Variabilità e mutabilità. Bologna, Italy
Hemming R, Keen MJ (1993) Single crossing conditions in comparisons of tax progressivity. J Publ Econ 20:373–390
Jakobsson U (1976) On the measurement of the degree of progression. J Publ Econ 5:161–168
Kakwani NC (1977) Applications of Lorenz curves in economic analysis. Econometrica 45:719–727
Lambert PJ (2001) The distribution and redistribution of income: a mathematical analysis, 3rd edn. Manchester university press, Manchester
Lorenz MO (1905) Methods for measuring concentration of wealth. J Am Stat Assoc New Ser 70:209–219

Loss Function

Wolfgang Bischoff
Professor and Dean of the Faculty of Mathematics and Geography
Catholic University Eichstätt–Ingolstadt, Eichstätt, Germany

Loss functions occur at several places in statistics. Here we attach importance to decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview) and regression. For both fields the same loss functions can be used. But the interpretation is different.

Decision theory gives a general framework to define and understand statistics as a mathematical discipline. The loss function is the essential component in decision theory. The loss function judges a decision with respect to the truth by a real value greater or equal to zero. In case the decision coincides with the truth then there is no loss. Therefore the value of the loss function is zero then, otherwise the value gives the loss which is suffered by the decision unequal the truth. The larger the value the larger the loss which is suffered.

To describe this more exactly let Θ be the known set of all outcomes for the problem under consideration on which we have information by data. We assume that one of the values $\theta \in \Theta$ is the true value. Each $d \in \Theta$ is a possible decision. The decision d is chosen according to a rule, more exactly according to a function with values in Θ and defined on the set of all possible data. Since the true value θ is unknown the loss function L has to be defined on $\Theta \times \Theta$, i.e.,

$$L : \Theta \times \Theta \to [0, \infty).$$

The first variable describes the true value, say, and the second one the decision. Thus $L(\theta, a)$ is the loss which is suffered if θ is the true value and a is the decision. Therefore, each (up to technical conditions) function $L : \Theta \times \Theta \to [0, \infty)$ with the property

$$L(\theta, \theta) = 0 \text{ for all } \theta \in \Theta$$

is a possible loss function. The loss function has to be chosen by the statistician according to the problem under consideration.

Next, we describe examples for loss functions. First let us consider a test problem. Then Θ is divided in two disjoint subsets Θ_0 and Θ_1 describing the null hypothesis and the alternative set, $\Theta = \Theta_0 + \Theta_1$. Then the usual loss function is given by

$$L(\theta, \vartheta) = \begin{cases} 0 & \text{if } \theta, \vartheta \in \Theta_0 \quad \text{or} \quad \theta, \vartheta \in \Theta_1 \\ 1 & \text{if } \theta \in \Theta_0, \vartheta \in \Theta_1 \quad \text{or} \quad \theta \in \Theta_1, \vartheta \in \Theta_0 \end{cases}.$$

For point estimation problems we assume that Θ is a normed linear space and let $|\cdot|$ be its norm. Such a space is typical for estimating a location parameter. Then the loss $L(\theta, \vartheta) = |\theta - \vartheta|$, $\theta, \vartheta \in \Theta$, can be used. Next, let us consider the specific case $\Theta = \mathrm{R}$. Then $L(\theta, \vartheta) = \ell(\theta - \vartheta)$ is a typical form for loss functions, where $\ell : IR \to [0, \infty)$ is nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$. ℓ is also called loss function. An important class of such functions is given by choosing $\ell(t) = |t|^p$, where $p > 0$ is a fixed constant. There are two prominent cases, for $p = 2$ we get the classical square loss and for $p = 1$ the robust L_1-loss. Another class of robust losses are the famous Huber losses

$$\ell(t) = t^2/2, \text{ if } |t| \le \gamma, \text{ and } \ell(t) = \gamma|t| - \gamma^2/2, \text{ if } |t| > \gamma,$$

where $\gamma > 0$ is a fixed constant. Up to now we have shown symmetrical losses, i.e., $L(\theta, \vartheta) = L(\vartheta, \theta)$. There are many problems in which underestimating of the true value θ has to be differently judged than overestimating. For such problems Varian (1975) introduced LinEx losses

$$\ell(t) = b(exp(at) - at - 1),$$

where $a, b > 0$ can be chosen suitably. Here underestimating is judged exponentially and overestimating linearly.

For other estimation problems corresponding losses are used. For instance, let us consider the estimation of a

scale parameter and let $\Theta = (0, \infty)$. Then it is usual to consider losses of the form $L(\theta, \vartheta) = \ell(\vartheta/\theta)$, where ℓ must be chosen suitably. It is, however, more convenient to write $\ell(\ln \vartheta - \ln \theta)$. Then ℓ can be chosen as above.

In theoretical works the assumed properties for loss functions can be quite different. Classically it was assumed that the loss is convex (see ►Rao–Blackwell Theorem). If the space Θ is not bounded, then it seems to be more convenient in practice to assume that the loss is bounded which is also assumed in some branches of statistics. In case the loss is not continuous then it must be carefully defined to get no counter intuitive results in practice, see Bischoff (1999).

In case a density of the underlying distribution of the data is known up to an unknown parameter the class of divergence losses can be defined. Specific cases of these losses are the Hellinger and the Kulback-Leibler loss.

In regression, however, the loss is used in a different way. Here it is assumed that the unknown location parameter is an element of a known class $\mathcal{F}$ of real valued functions. Given n observations (data) $y_1, \ldots, y_n$ observed at design points $t_1, \ldots, t_n$ of the experimental region a loss function is used to determine an estimation for the unknown regression function by the 'best approximation', i.e., the function in $\mathcal{F}$ that minimizes $\sum_{i=1}^{n} \ell\left(r_i^f\right), f \in \mathcal{F}$, where $r_i^f = y_i - f(t_i)$ is the residual in the ith design point. Here ℓ is also called loss function and can be chosen as described above. For instance, the least squares estimation is obtained if $\ell(t) = t^2$.

Cross References

►Advantages of Bayesian Structuring: Estimating Ranks and Histograms
►Bayesian Statistics
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Entropy and Cross Entropy as Diversity and Distance Measures
►Sequential Sampling
►Statistical Inference for Quantum Systems

References and Further Reading

Bischoff W (1999) Best ϕ-approximants for bounded weak loss functions. Stat Decis 17:49–61

Varian HR (1975) A Bayesian approach to real estate assessment. In: Fienberg SE, Zellner A (eds) Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage, North Holland, Amsterdam, pp 195–208