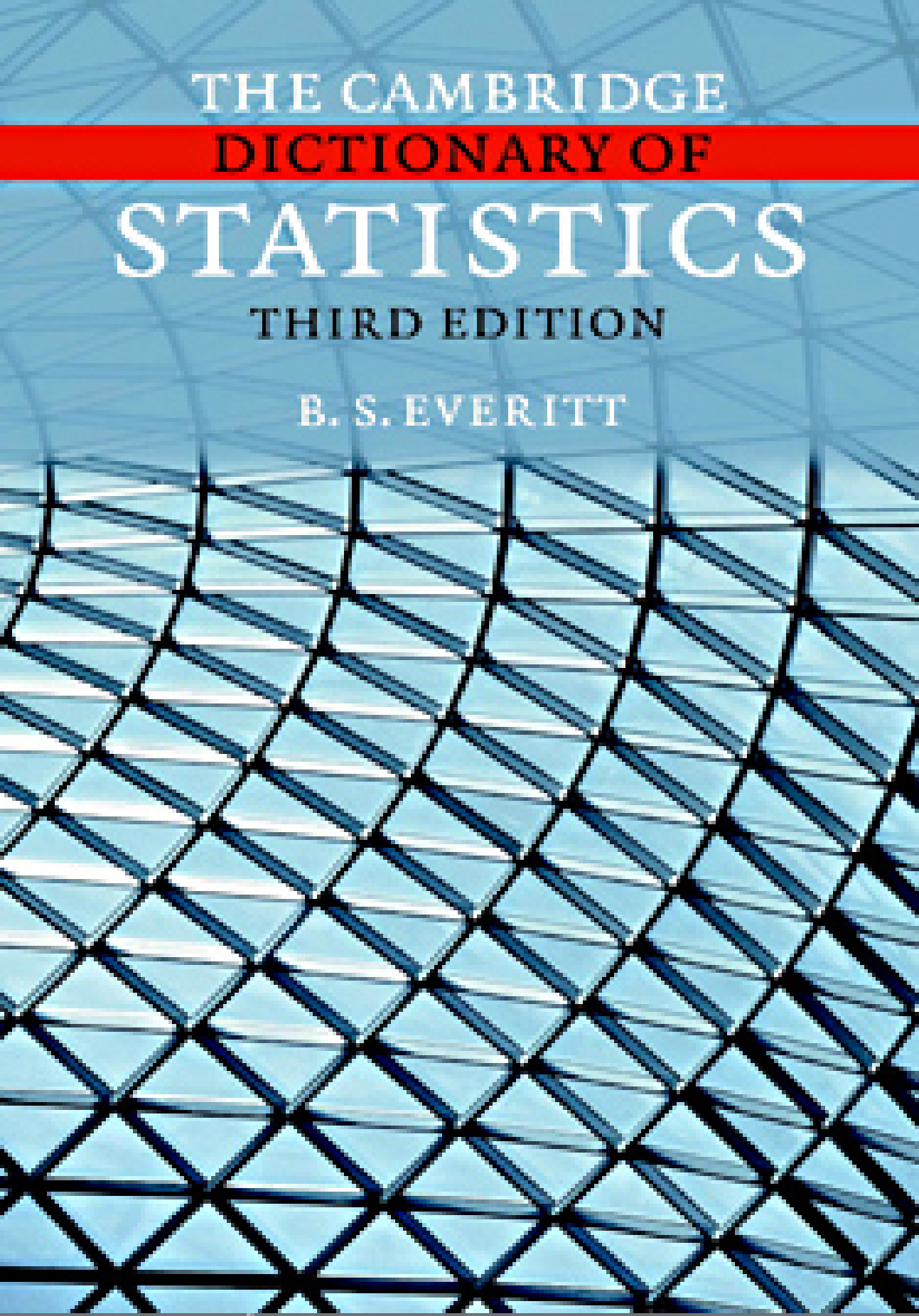
THE CAMBRIDGE DICTIONARY OF STATISTICS

THIRD EDITION

B. S. EVERITT



CAMBRIDGE www.cambridge.org/9780521860390

This page intentionally left blank

THE CAMBRIDGE DICTIONARY OF STATISTICS
THIRD EDITION

If you use statistics and need easy access to simple, reliable definitions and explanations of statistical and statistics-related concepts, then look no further than this dictionary. Over 3600 terms are defined, covering medical, survey, theoretical, and applied statistics, including computational statistics. Entries are provided for standard and specialized statistical software. In addition, short biographies of over 100 important statisticians are given. Definitions provide enough mathematical detail to clarify concepts and give standard formulae when these are helpful. The majority of definitions then give a reference to a book or article where the user can seek further or more specialized information, and many are accompanied by graphical material to aid understanding.

B. S. Everitt is Professor Emeritus of the Institute of Psychiatry, King's College London. He is the author of over 50 books on statistics and computing, including *Medical Statistics from A to Z*, also from Cambridge University Press.

THE CAMBRIDGE DICTIONARY OF Statistics

Third Edition

B. S. EVERITT
Institute of Psychiatry, King's College London

CAMBRIDGE UNIVERSITY PRESS

cambridge university press
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge cb2 2ru, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521860390

First published in print format 2006

isbn-13 978-0-511-24323-3 eBook (Adobe Reader)
isbn-10 0-511-24323-5 eBook (Adobe Reader)

isbn-13 978-0-521-86039-0 hardback
isbn-10 0-521-86039-3 hardback

isbn-13 978-0-521-69027-(paperback
isbn-10 0-521-69027-7 paperback

To the memory of my dear sister Iris

Preface to first edition

The Cambridge Dictionary of Statistics aims to provide students of statistics, working statisticians and researchers in many disciplines who are users of statistics with relatively concise definitions of statistical terms. All areas of statistics are covered, theoretical, applied, medical, etc., although, as in any dictionary, the choice of which terms to include and which to exclude is likely to reflect some aspects of the compiler's main areas of interest, and I have no illusions that this dictionary is any different. My hope is that the dictionary will provide a useful source of reference for both specialists and non-specialists alike. Many definitions necessarily contain some mathematical formulae and/or nomenclature, others contain none. But the difference in mathematical content and level among the definitions will, with luck, largely reflect the type of reader likely to turn to a particular definition. The non-specialist looking up, for example, **Student's t-tests** will hopefully find the simple formulae and associated written material more than adequate to satisfy their curiosity, while the specialist seeking a quick reminder about **spline functions** will find the more extensive technical material just what they need.

The dictionary contains approximately 3000 headwords and short biographies of more than 100 important statisticians (fellow statisticians who regard themselves as 'important' but who are *not* included here should note the single common characteristic of those who are). Several forms of cross-referencing are used. Terms in *slanted roman* in an entry appear as separate headwords, although headwords defining relatively commonly occurring terms such as **random variable**, **probability**, **distribution**, **population**, **sample**, etc., are *not* referred to in this way. Some entries simply refer readers to another entry. This may indicate that the terms are synonyms or, alternatively, that the term is more conveniently discussed under another entry. In the latter case the term is printed in *italics* in the main entry.

Entries are in alphabetical order using the letter-by-letter rather than the word-by-word convention. In terms containing numbers or Greek letters, the numbers or corresponding English word are spelt out and alphabetized accordingly. So, for example, 2×2 table is found under **two-by-two table**, and α-trimmed mean, under **alpha-trimmed mean**. Only headings corresponding to names are inverted, so the entry for William Gosset is found under **Gosset, William** but there is an entry under **Box–Müller transformation** *not* under **Transformation, Box–Müller**.

For those readers seeking more detailed information about a topic, many entries contain either a reference to one or other of the texts listed later, or a more specific reference to a relevant book or journal article. (Entries for software contain the appropriate address.) Additional material is also available in many cases in either the *Encyclopedia of Statistical Sciences*, edited by Kotz and Johnson, or the *Encyclopedia of Biostatistics*, edited by Armitage and Colton, both published by Wiley. Extended biographies of many of the people included in this dictionary can also be found in these two encyclopedias and also in *Leading Personalities in Statistical Sciences* by Johnson and Kotz published in 1997 again by Wiley.

Lastly and paraphrasing Oscar Wilde 'writing one dictionary is suspect, writing two borders on the pathological'. But before readers jump to an obvious conclusion I would like to make it very clear that an anorak has never featured in my wardrobe.

B. S. Everitt, 1998

Preface to third edition

In this third edition of the *Cambridge Dictionary of Statistics* I have added many new entries and taken the opportunity to correct and clarify a number of the previous entries. I have also added biographies of important statisticians whom I overlooked in the first and second editions and, sadly, I have had to include a number of new biographies of statisticians who have died since the publication of the second edition in 2002.

B. S. Everitt, 2005

Acknowledgements

Firstly I would like to thank the many authors who have, unwittingly, provided the basis of a large number of the definitions included in this dictionary through their books and papers. Next thanks are due to many members of the 'allstat' mailing list who helped with references to particular terms. I am also extremely grateful to my colleagues, Dr Sophia Rabe-Hesketh and Dr Sabine Landau, for their careful reading of the text and their numerous helpful suggestions. Lastly I have to thank my secretary, Mrs Harriet Meteyard, for maintaining and typing the many files that contained the material for the dictionary and for her constant reassurance that nothing was lost!

Notation

The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}'$.

Sources

Altman, D.G. (1991) *Practical Statistics for Medical Research*, Chapman and Hall, London. (SMR)

Chatfield, C. (2003) *The Analysis of Time Series: An Introduction*, 6th edition, Chapman and Hall, London. (TMS)

Evans, M., Hastings, N. and Peacock, B. (2000) *Statistical Distributions*, 3rd edition, Wiley, New York. (STD)

Krzanowski, W.J. and Marriot, F.H.C. (1994) *Multivariate Analysis, Part 1*, Edward Arnold, London. (MV1)

Krzanowski, W.J. and Marriot, F.H.C. (1995) *Multivariate Analysis, Part 2*, Edward Arnold, London. (MV2)

McCullagh, P.M. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edition, Chapman and Hall, London. (GLM)

Rawlings, J.O., Sastry, G.P. and Dickey, D.A. (2001) *Applied Regression Analysis: A Research Tool*, Springer-Verlag, New York. (ARA)

Stuart, A. and Ord, K. (1994) *Kendall's Advanced Theory of Statistics, Volume 1*, 6th edition, Edward Arnold, London. (KA1)

Stuart, A. and Ord, K. (1991) *Kendall's Advanced Theory of Statistics, Volume 2*, 5th edition, Edward Arnold, London. (KA2)

A

Aalen–Johansen estimator: An estimator of the *survival function* for a set of *survival times*, when there are competing causes of death. Related to the *Nelson–Aalen estimator*. [*Scandinavian Journal of Statistics*, 1978, **5**, 141–50.]

Aalen's linear regression model: A model for the *hazard function* of a set of survival times given by

$$\alpha(t; \mathbf{z}(t)) = \alpha_0(t) + \alpha_1(t)z_1(t) + \cdots + \alpha_p(t)z_p(t)$$

where $\alpha(t)$ is the hazard function at time t for an individual with covariates $\mathbf{z}(t)' = [z_1(t), \ldots, z_p(t)]$. The 'parameters' in the model are functions of time with $\alpha_0(t)$ the baseline hazard corresponding to $\mathbf{z}(t) = \mathbf{0}$ for all t, and $\alpha_q(t)$, the excess rate at time t per unit increase in $z_q(t)$. See also **Cox's proportional hazards model**. [*Statistics in Medicine*, 1989, **8**, 907–25.]

Abbot's formula: A formula for the proportion of animals (usually insects) dying in a toxicity trial that recognizes that some insects may die during the experiment even when they have not been exposed to the toxin, and among those who have been so exposed, some may die of natural causes. Explicitly the formula is

$$p_i^* = \pi + (1 - \pi)p_i$$

where p_i^* is the observable response proportion, p_i is the expected proportion dying at a given dose and π is the proportion of insects who respond naturally. [*Modelling Binary Data*, 2nd edition, 1993, D. Collett, Chapman and Hall/CRC Press, London.]

ABC method: Abbreviation for **approximate bootstrap confidence method**.

Ability parameter: See **Rasch model**.

Absolute deviation: Synonym for **average deviation**.

Absolute risk: Synonym for **incidence**.

Absorbing barrier: See **random walk**.

Absorption distributions: Probability distributions that represent the number of 'individuals' (e.g. particles) that fail to cross a specified region containing hazards of various kinds. For example, the region may simply be a straight line containing a number of 'absorption' points. When a particle travelling along the line meets such a point, there is a probability p that it will be absorbed. If it is absorbed it fails to make any further progress, but also the point is incapable of absorbing any more particles. When there are M active absorption points, the probability of a particle being absorbed is $[1 - (1 - p^M]$. [*Naval Research Logistics Quarterly*, 1966, **13**, 35–48.]

Abundance matrices: Matrices that occur in ecological applications. They are essentially two-dimensional tables in which the classifications correspond to site and species.

The value in the ijth cell gives the number of species j found at site i. [*Biologica, Bratislava*, 2000, **55**, 357–62.]

Accelerated failure time model: A general model for data consisting of *survival times*, in which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. Consequently the model can be interpreted in terms of the speed of progression of a disease. In the simplest case of comparing two groups of patients, for example, those receiving treatment A and those receiving treatment B, this model assumes that the survival time of an individual on one treatment is a multiple of the survival time on the other treatment; as a result the probability that an individual on treatment A survives beyond time t is the probability that an individual on treatment B survives beyond time ϕt, where ϕ is an unknown positive constant. When the endpoint of interest is the death of a patient, values of ϕ less than one correspond to an acceleration in the time of death of an individual assigned to treatment A, and values of ϕ greater than one indicate the reverse. The parameter ϕ is known as the *acceleration factor*. [*Modelling Survival Data in Medical Research*, 2nd edition, 2003, D. Collett, Chapman and Hall/CRC Press, London.]

Acceleration factor: See **accelerated failure time model**.

Acceptable quality level: See **quality control procedures**.

Acceptable risk: The risk for which the benefits of a particular medical procedure are considered to outweigh the potential hazards. [*Acceptable Risk*, 1984, B. Fischoff, Cambridge University Press, Cambridge.]

Acceptance region: A term associated with statistical significance tests, which gives the set of values of a *test statistic* for which the null hypothesis is to be accepted. Suppose, for example, a *z-test* is being used to test the null hypothesis that the mean blood pressure of men and women is equal against the alternative hypothesis that the two means are not equal. If the chosen significance of the test is 0.05 then the acceptance region consists of values of the test statistic z between –1.96 and 1.96. [*Estimation and Inference in Economics*, 1993, R. Davidson and R. Mackinnon, Oxford University Press, Oxford.]

Acceptance–rejection algorithm: An algorithm for generating random numbers from some probability distribution, $f(x)$, by first generating a random number from some other distribution, $g(x)$, where f and g are related by

$$f(x) \leq kg(x) \text{ for all } x$$

with k a constant. The algorithm works as follows:

- let r be a random number from $g(x)$;
- let s be a random number from a *uniform distribution* on the interval (0,1);
- calculate $c = ksg(r)$;
- if $c > f(r)$ reject r and return to the first step; if $c \leq f(r)$ accept r as a random number from f. [*Statistics in Civil Engineering*, 1997, A.V. Metcalfe, Edward Arnold, London.]

Acceptance sampling: A type of *quality control procedure* in which a sample is taken from a collection or batch of items, and the decision to accept the batch as satisfactory, or reject them as unsatisfactory, is based on the proportion of defective items in the sample. [*Quality Control and Industrial Statistics*, 4th edition, 1974, A.J. Duncan, R.D. Irwin, Homewood, Illinois.]

Accident proneness: A personal psychological factor that affects an individual's probability of suffering an accident. The concept has been studied statistically under a number of different assumptions for accidents:

- pure chance, leading to the *Poisson distribution*;
- true contagion, i.e. the hypothesis that all individuals initially have the same probability of having an accident, but that this probability changes each time an accident happens;
- apparent contagion, i.e. the hypothesis that individuals have constant but unequal probabilities of having an accident.

The study of accident proneness has been valuable in the development of particular statistical methodologies, although in the past two decades the concept has, in general, been out of favour; attention now appears to have moved more towards risk evaluation and analysis. [*Accident Proneness*, 1971, L. Shaw and H.S. Sichel, Pergamon Press, Oxford.]

Accidentally empty cells: Synonym for **sampling zeros**.

Accrual rate: The rate at which eligible patients are entered into a *clinical trial*, measured as persons per unit of time. Often disappointingly low for reasons that may be both physician and patient related. [*Journal of Clinical Oncology*, 2001, **19**, 3554–61.]

Accuracy: The degree of conformity to some recognized standard value. See also **bias**.

ACE: Abbreviation for **alternating conditional expectation**.

ACE model: A genetic epidemiological model that postulates additive genetic factors, common environmental factors, and specific environmental factors in a phenotype. The model is used to quantify the contributions of genetic and environmental influences to variation. [*Encyclopedia of Behavioral Statistics, Volume 1*, 2005, ed. B.S. Everitt and D.C. Howell, Wiley, Chichester.]

ACES: Abbreviation for **active control equivalence studies**.

ACF: Abbreviation for **autocorrelation function**.

ACORN: An acronym for 'A Classification of Residential Neighbourhoods'. It is a system for classifying households according to the demographic, employment and housing characteristics of their immediate neighbourhood. Derived by applying *cluster analysis* to 40 variables describing each neighbourhood including age, class, tenure, dwelling type and car ownership. [*Statistics in Society*, 1999, D. Dorling and S. Simpson eds., Arnold, London.]

Acquiescence bias: The bias produced by respondents in a survey who have the tendency to give positive responses, such as 'true', 'like', 'often' or 'yes' to a question. At its most extreme, the person responds in this way irrespective of the content of the item. Thus a person may respond 'true' to two items like 'I always take my medication on time' and 'I often forget to take my pills'. See also **end-aversion bias**. [*Journal of Intellectual Disability Research*, 1995, **39**, 331–40.]

Action lines: See **quality control procedures.**

Active control equivalence studies (ACES): *Clinical trials* in which the object is simply to show that the new treatment is at least as good as the existing treatment. Such studies are becoming more widespread due to current therapies that reflect previous

successes in the development of new treatments. Such studies rely on an implicit historical control assumption, since to conclude that a new drug is efficacious on the basis of this type of study requires a fundamental assumption that the active control drug would have performed better than a placebo, had a placebo been used in the trial. [*Statistical Issues in Drug Development*, 1997, S. Senn, Wiley, Chichester.]

Active control trials: *Clinical trials* in which the trial drug is compared with some other active compound rather than a placebo. [*Annals of Internal Medicine*, 2000, **135**, 62–4.]

Active life expectancy (ALE): Defined for a given age as the expected remaining years free of disability. A useful index of public health and quality of life for populations. A question of great interest is whether recent trends towards longer *life expectancy* have been accompanied by a comparable increase in ALE. [*New England Journal of Medicine*, 1983, **309**, 1218–24.]

Activ Stats: A commercial computer-aided learning package for statistics. See also **statistics for the terrified**. [Interactive Learning Europe, 124 Cambridge Science Park, Milton Road, Cambridge CB4 4ZS.]

Actuarial estimator: An estimator of the *survival function*, $S(t)$, often used when the data are in grouped form. Given explicitly by

$$S(t) = \prod_{\substack{j \geq 0 \\ t_{(j+1)} \leq t}} \left[1 - \frac{d_j}{N_j - \frac{1}{2} w_j} \right]$$

where the ordered survival times are $0 < t_{(1)} < \cdots < t_{(n)}$, N_i is the number of people at risk at the start of the interval $t_{(i)}$, $t_{(i+1)}$, d_i is the observed number of deaths in the interval and w_i the number of censored observations in the interval. [*Survival Models and Data Analysis*, 1999, R.G. Elandt–Johnson and N.L. Johnson, Wiley, New York.]

Actuarial statistics: The statistics used by actuaries to evaluate risks, calculate liabilities and plan the financial course of insurance, pensions, etc. An example is *life expectancy* for people of various ages, occupations, etc. See also **life table**. [*American Society of Actuaries*, 1961, **13**, 116–20.]

Adaptive cluster sampling: A procedure in which an initial set of subjects is selected by some sampling procedure and, whenever the variable of interest of a selected subject satisfies a given criterion, additional subjects in the neighbourhood of that subject are added to the sample. [*Biometrika*, 1996, **84**, 209–19.]

Adaptive designs: *Clinical trials* that are modified in some way as the data are collected within the trial. For example, the allocation of treatment may be altered as a function of the response to protect patients from ineffective or toxic doses. [*Controlled Clinical Trials*, 1999, **20**, 172–86.]

Adaptive estimator: See **adaptive methods**.

Adaptive methods: Procedures that use various aspects of the sample data to select the most appropriate type of statistical method for analysis. An *adaptive estimator*, T, for the centre of a distribution, for example, might be

$$T = \text{mid-range when } k \leq 2$$
$$= \text{arithmetic mean when } 2 < k < 5$$
$$= \text{median when } k \geq 5$$

where k is the sample *kurtosis*. So if the sample looks as if it arises from a short-tailed distribution, the average of the largest and smallest observations is used; if it looks like a long-tailed situation the median is used, otherwise the mean of the sample is calculated. [*Journal of the American Statistical Association*, 1967, **62**, 1179–86.]

Adaptive sampling design: A *sampling design* in which the procedure for selecting *sampling units* on which to make observations may depend on observed values of the variable of interest. In a survey for estimating the abundance of a natural resource, for example, additional sites (the sampling units in this case) in the vicinity of high observed abundance may be added to the sample during the survey. The main aim in such a design is to achieve gains in precision or efficiency compared to conventional designs of equivalent sample size by taking advantage of observed characteristics of the population. For this type of sampling design the probability of a given sample of units is conditioned on the set of values of the variable of interest in the population. [*Adaptive Sampling*, 1996, S.K. Thompson and G.A.F. Seber, Wiley, New York.]

Added variable plot: A graphical procedure used in all types of regression analysis for identifying whether or not a particular explanatory variable should be included in a model, in the presence of other explanatory variables. The variable that is the candidate for inclusion in the model may be new or it may simply be a higher power of one currently included. If the candidate variable is denoted x_i, then the *residuals* from the regression of the response variable on all the explanatory variables, save x_i, are plotted against the residuals from the regression of x_i on the remaining explanatory variables. A strong linear relationship in the plot indicates the need for x_i in the regression equation (Fig. 1). [*Regression Analysis*, Volume 2, 1993, edited by M.S. Lewis-Beck, Sage Publications, London.]

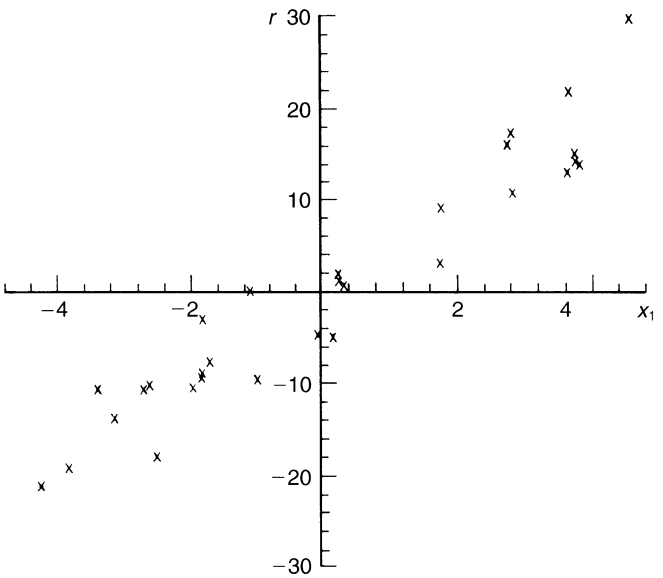


Fig. 1 Added variable plot indicating a variable that could be included in the model.

Addition rule for probabilities: For two events, A and B that are *mutually exclusive*, the probability of either event occurring is the sum of the individual probabilities, i.e.

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

where $\Pr(A)$ denotes the probability of event A etc. For k mutually exclusive events $A_1, A_2, \ldots, A_k$, the more general rule is

$$\Pr(A_1 \text{ or } A_2 \ldots \text{ or } A_k) = \Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_k)$$

See also **multiplication rule for probabilities** and **Boole's inequality**. [KA1 Chapter 8.]

Additive clustering model: A model for *cluster analysis* which attempts to find the structure of a *similarity matrix* with elements s_{ij} by fitting a model of the form

$$s_{ij} = \sum_{k=1}^{K} w_k p_{ik} p_{jk} + \epsilon_{ij}$$

where K is the number of clusters and w_k is a weight representing the salience of the property corresponding to cluster k. If object i has the property of cluster k, then $p_{ik} = 1$, otherwise it is zero. [*Psychological Review*, 1979, **86**, 87–123.]

Additive effect: A term used when the effect of administering two treatments together is the sum of their separate effects. See also **additive model**. [*Journal of Bone Mineral Research*, 1995, **10**, 1303–11.]

Additive genetic variance: The variance of a trait due to the main effects of *genes*. Usually obtained by a factorial *analysis of variance* of trait values on the genes present at one or more loci. [*Statistics in Human Genetics*, 1998, P. Sham, Arnold, London.]

Additive model: A model in which the explanatory variables have an *additive effect* on the response variable. So, for example, if variable A has an effect of size a on some response measure and variable B one of size b on the same response, then in an assumed additive model for A and B their combined effect would be $a + b$.

Additive outlier: A term applied to an observation in a *time series* which is affected by a non-repetitive intervention such as a strike, a war, etc. Only the level of the particular observation is considered affected. In contrast an *innovational outlier* is one which corresponds to an extraordinary shock at some time point T which also influences subsequent observations in the series. [*Journal of the American Statistical Association*, 1996, **91**, 123–31.]

Additive tree: A connected, *undirected graph* where every pair of nodes is connected by a unique path and where the distances between the nodes are such that

$$d_{xy} + d_{uv} \leq \max[d_{xu} + d_{yv}, d_{xv} + d_{yu}] \text{ for all } x, y, u, \text{ and } v$$

An example of such a tree is shown in Fig. 2. See also **ultrametric tree**. [*Tree Models of Similarity and Association*, 1996, J.E. Corter, Sage University Papers 112, Sage Publications, Thousand Oaks.]

Adequate subset: A term used in regression analysis for a subset of the explanatory variables that is thought to contain as much information about the response variable as the complete set. See also **selection methods in regression**.

Adjacency matrix: A matrix with elements, x_{ij}, used to indicate the connections in a *directed graph*. If node i relates to node j, $x_{ij} = 1$, otherwise $x_{ij} = 0$. For a simple graph with no self-loops, the adjacency matrix must have zeros on the diagonal. For an undirected graph the adjacency matrix is symmetric. [*Introductory Graph Theory*, 1985, G. Chartrand, Dover, New York.]

a. **Dissimiliarities**

	A	B	C	D	E
Worker A	--				
Worker B	15	--			
Worker C	20	25	--		
Worker D	18	23	6	--	
Worker E	20	25	20	18	--

b. Additive Tree

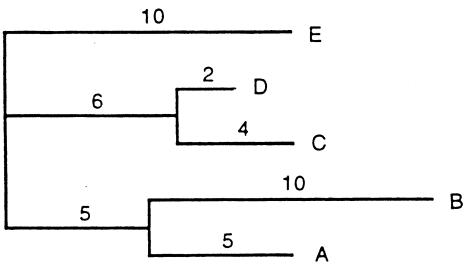


Fig. 2 An example of an additive tree. (Reproduced by permission of Sage Publications from *Tree Models of Similarity and Association*, 1996, J.E. Corter.)

Adjusted correlation matrix: A *correlation matrix* in which the diagonal elements are replaced by *communalities*. The basis of *principal factor analysis*.

Adjusted treatment means: Usually used for estimates of the treatment means in an *analysis of covariance*, after adjusting all treatments to the same mean level for the covariate(s), using the estimated relationship between the covariate(s) and the response variable. [*Biostatistics*, 1993, L.D. Fisher and G. Van Belle, Wiley, New York.]

Adjusting for baseline: The process of allowing for the effect of *baseline characteristics* on the response variable usually in the context of a *longitudinal study*. A number of methods might be used, for example, the analysis of simple *change scores*, the analysis of percentage change, or, in some cases, the analysis of more complicated variables. In general it is preferable to use the adjusted variable that has least dependence on the baseline measure. For a longitudinal study in which the correlations between the repeated measures over time are moderate to large, then using the baseline values as covariates in an *analysis of covariance* is known to be more efficient than analysing change scores. See also **baseline balance.** [*Statistical Issues in Drug Development*, 1997, S. Senn, Wiley, Chichester.]

Administrative databases: *Databases* storing information routinely collected for purposes of managing a health-care system. Used by hospitals and insurers to examine admissions, procedures and lengths of stay. [*Healthcare Management Forum*, 1995, **8**, 5–13.]

Admissibility: A very general concept that is applicable to any procedure of statistical inference. The underlying notion is that a procedure is admissible if and only if there does not exist within that class of procedures another one which performs uniformly at least as well as the procedure in question and performs better than it in at least one case. Here 'uniformly' means for all values of the parameters that determine the probability distribution of the random variables under investigation. [KA2 Chapter 31.]

Admixture in human populations: The inter-breeding between two or more populations that were previously isolated from each other for geographical or cultural reasons. Population admixture can be a source of spurious associations between diseases and *alleles* that are both more common in one ancestral population than the others. However, populations that have been admixed for several generations may be useful for mapping disease *genes*, because spurious associations tend to be dissipated more rapidly than true associations in successive generations of random mating. [*Statistics in Human Genetics*, 1998, P. Sham, Arnold, London.]

Adoption studies: Studies of the rearing of a nonbiological child in a family. Such studies have played an important role in the assessment of genetic variation in human and animal traits. [*Foundations of Behavior Genetics*, 1978, J.L. Fulker and W.R. Thompson, Mosby, St. Louis.]

Aetiological fraction: Synonym for **attributable risk.**

Affine invariance: A term applied to statistical procedures which give identical results after the data has been subjected to an *affine transformation*. An example is *Hotelling's* T^2 *test*. [*Canadian Journal of Statistics*, 2003, **31**, 437–55.]

Affine transformation: The transformation, $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ where $\mathbf{A}$ is a nonsingular matrix and $\mathbf{b}$ is any vector of real numbers. Important in many areas of statistics particularly *multivariate analysis*.

Age-dependent birth and death process: A *birth and death process* where the birth rate and death rate are not constant over time, but change in a manner which is dependent on the age of the individual. [*Stochastic Modelling of Scientific Data*, 1995, P. Guttorp, CRC/Chapman and Hall, London.]

Age heaping: A term applied to the collection of data on ages when these are accurate only to the nearest year, half year or month. Occurs because many people (particularly older people) tend not to give their exact age in a survey. Instead they round their age up or down to the nearest number that ends in 0 or 5. See also **coarse data** and **Whipple index**. [*Geographic Journal*, 1992, **28**, 427–42.]

Age–period–cohort model: A model important in many *observational studies* when it is reasonable to suppose that age, number of years exposed to risk factor, and age when first exposed to risk factor, all contribute to disease risk. Unfortunately all three factors cannot be entered simultaneously into a model since this would result in *collinearity*, because 'age first exposed to risk factor' + 'years exposed to risk factor' is equal to 'age'. Various methods have been suggested for disentangling the dependence of the factors, although most commonly one of the factors is simply not included in the modelling process. See also **Lexis diagram**. [*Statistics in Medicine*, 1984, **3**, 113–30.]

Age-related reference ranges: Ranges of values of a measurement that give the upper and lower limits of normality in a population according to a subject's age. [*Journal of the Royal Statistical Society, Series A*, 1998, **161**, 79–101.]

Age-specific death rates: Death rates calculated within a number of relatively narrow age bands. For example, for 20–30 year olds,

$$\mathrm{DR}_{20,30} = \frac{\text{number of deaths among 20–30 year olds in a year}}{\text{average population size in 20–30 year olds in the year}}$$

Calculating death rates in this way is usually necessary since such rates almost

invariably differ widely with age, a variation not reflected in the *crude death rate*. See also **cause-specific death rates** and **standardized mortality ratio.** [*Biostatistics*, 1993, L.D. Fisher and G. Van Belle, Wiley, New York.]

Age-specific failure rate: A synonym for *hazard function* when the time scale is age. [*Family Planning Perspectives*, 1999, **31**, 241–5.]

Age-specific incidence rate: *Incidence rates* calculated within a number of relatively narrow age bands. See also **age-specific death rates.**

Agglomerative hierarchical clustering methods: Methods of *cluster analysis* that begin with each individual in a separate cluster and then, in a series of steps, combine individuals and later, clusters, into new, larger clusters until a final stage is reached where all individuals are members of a single group. At each stage the individuals or clusters that are 'closest', according to some particular definition of distance are joined. The whole process can be summarized by a *dendrogram*. Solutions corresponding to particular numbers of clusters are found by 'cutting' the dendrogram at the appropriate level. See also **average linkage, complete linkage, single linkage, Ward's method, Mojena's test, K-means cluster analysis** and **divisive methods.** [MV2 Chapter 10.]

Agresti's α: A generalization of the *odds ratio* for *2×2 contingency tables* to larger *contingency tables* arising from data where there are different degrees of severity of a disease and differing amounts of exposure. [*Analysis of Ordinal Categorical Data*, 1984, A. Agresti, Wiley, New York.]

Agronomy trials: A general term for a variety of different types of agricultural field experiments including fertilizer studies, time, rate and density of planting, tillage studies, and pest and weed control studies. Because the response to changes in the level of one factor is often conditioned by the levels of other factors it is almost essential that the treatments in such trials include combinations of multiple levels of two or more production factors. [*An Introduction to Statistical Science in Agriculture*, 4th edition, 1972, D.J. Finney, Blackwell, Oxford.]

AI: Abbreviation for **artificial intelligence.**

AIC: Abbreviation for **Akaike's information criterion**.

Aickin's measure of agreement: A chance-corrected measure of agreement which is similar to the *kappa coefficient* but based on a different definition of agreement by chance. [*Biometrics*, 1990, **46**, 293–302.]

AID: Abbreviation for **automatic interaction detector.**

Aitchison distributions: A broad class of distributions that includes the *Dirichlet distribution* and *logistic normal distributions* as special cases. [*Journal of the Royal Statistical Society, Series B*, 1985, **47**, 136–46.]

Aitken, Alexander Craig (1895–1967): Born in Dunedin, New Zealand, Aitken first studied classical languages at Otago University, but after service during the First World War he was given a scholarship to study mathematics in Edinburgh. After being awarded a D.Sc., Aitken became a member of the Mathematics Department in Edinburgh and in 1946 was given the Chair of Mathematics which he held until his retirement in 1965. The author of many papers on least squares and the fitting of polynomials,

Aitken had a legendary ability at arithmetic and was reputed to be able to dictate rapidly the first 707 digits of π. He was a Fellow of the Royal Society and of the Royal Society of Literature. Aitken died on 3 November 1967 in Edinburgh.

Ajne's test: A *distribution free method* for testing the uniformity of a *circular distribution*. The test statistic A_n is defined as

$$A_n = \int_0^{2\pi} [N(\theta) - n/2]^2 \mathrm{d}\theta$$

where $N(\theta)$ is the number of sample observations that lie in the semicircle, θ to $\theta + \pi$. Values close to zero lead to acceptance of the hypothesis of uniformity. [*Scandinavian Audiology*, 1996, 201–6.]

Akaike's information criterion (AIC): An index used in a number of areas as an aid to choosing between competing models. It is defined as

$$-2L_m + 2m$$

where L_m is the maximized *log-likelihood* and m is the number of parameters in the model. The index takes into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters. Lower values of the index indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data. See also **parsimony principle** and **Schwarz's criterion.** [MV2 Chapter 11.]

ALE: Abbreviation for **active life expectancy**.

Algorithm: A well-defined set of rules which, when routinely applied, lead to a solution of a particular class of mathematical or computational problem. [*Introduction to Algorithms*, 1989, T.H. Cormen, C.E. Leiserson, and R.L. Rivest, McGraw-Hill, New York.]

Alias: See **confounding.**

Allele: The DNA sequence that exists at a genetic location that shows sequence variation in a population. Sequence variation may take the form of insertion, deletion, substitution, or variable repeat length of a regular motif, for example, CACACA. [*Statistics in Human Genetics*, 1998, P. Sham, Arnold, London.]

Allocation ratio: Synonym for **treatment allocation ratio.**

Allocation rule: See **discriminant analysis.**

Allometry: The study of changes in shape as an organism grows. [MV1 Chapter 4.]

All subsets regression: A form of regression analysis in which all possible models are considered and the 'best' selected by comparing the values of some appropriate criterion, for example, *Mallow's C_k statistic*, calculated on each. If there are q explanatory variables, there are a total of $2^q - 1$ models to be examined. The *leaps-and-bounds algorithm* is generally used so that only a small fraction of the possible models have to be examined. See also **selection methods in regression.** [ARA Chapter 7.]

Almon lag technique: A method for estimating the coefficients, $\beta_0, \beta_1, \ldots, \beta_r$, in a model of the form

$$y_t = \beta_0 x_t + \cdots + \beta_r x_{t-r} + \epsilon_t$$

where y_t is the value of the dependent variable at time t, $x_t, \ldots, x_{t-r}$ are the values of the explanatory variable at times $t, t-1, \ldots, t-r$ and ϵ_t is a disturbance term at time t. If r is finite and less than the number of observations, the regression coefficients can be found by *least squares estimation*. However, because of the possible problem of a high degree of *multicollinearity* in the variables $x_t, \ldots, x_{t-r}$ the approach is to estimate the coefficients subject to the restriction that they lie on a polynomial of degree p, i.e. it is assumed that there exist parameters $\lambda_0, \lambda_1, \ldots, \lambda_p$ such that

$$\beta_i = \lambda_0 + \lambda_1 i + \cdots + \lambda_p i^p, \quad i = 0, 1, \ldots, r, \quad p \leq r$$

This reduces the number of parameters from $r+1$ to $p+1$. When $r = p$ the technique is equivalent to least squares. In practice several different values of r and/or p need to be investigated. [*The American Statistician*, 1972, **26**, 32–5.]

Alpha(α): The probability of a type I error. See also **significance level.**

Alpha factoring: A method of *factor analysis* in which the variables are considered samples from a population of variables.

Alpha spending function: An approach to *interim analysis* in a *clinical trial* that allows the control of the type I error rate while giving flexibility in how many interim analyses are to be conducted and at what time. [*Statistics in Medicine*, 1996, **15**, 1739–46.]

Alpha(α)-trimmed mean: A method of estimating the mean of a population that is less affected by the presence of *outliers* than the usual estimator, namely the sample average. Calculating the statistic involves dropping a proportion α (approximately) of the observations from both ends of the sample before calculating the mean of the remainder. If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ represent the ordered sample values then the measure is given by

$$\alpha_{\text{trimmed mean}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

where k is the smallest integer greater than or equal to αn. See also **M-estimators.** [*Biostatistics*, 1993, L.D. Fisher and G. Van Belle, Wiley, New York.]

Alpha(α)-Winsorized mean: A method of estimating the mean of a population that is less affected by the presence of *outliers* than the usual estimator, namely the sample average. Essentially the k smallest and k largest observations, where k is the smallest integer greater than or equal to αn, are respectively increased or reduced in size to the next remaining observation and counted as though they had these values. Specifically given by

$$\alpha_{\text{Winsorized mean}} = \frac{1}{n}\left[(k+1)(x_{(k+1)} + x_{(n-k)}) + \sum_{i=k+2}^{n-k-1} x_{(i)}\right]$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the ordered sample values. See also **M-estimators.** [*Biostatistics*, 1993, L.D. Fisher and G. Van Belle, Wiley, New York.]

Alshuler's estimator: An estimator of the *survival function* given by

$$\prod_{j=1}^{k} \exp(-d_j/n_j)$$

where d_j is the number of deaths at time $t_{(j)}$, n_j the number of individuals alive just before $t_{(j)}$ and $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$ are the ordered survival times. See also **product limit estimator.** [*Modelling Survival Data in Medical Research*, 2nd edition, 2003, D. Collett, Chapman and Hall/CRC Press, London.]

Alternate allocations: A method of allocating patients to treatments in a *clinical trial* in which alternate patients are allocated to treatment A and treatment B. Not to be recommended since it is open to abuse. [SMR Chapter 15.]

Alternating conditional expectation (ACE): A procedure for estimating optimal transformations for regression analysis and correlation. Given explanatory variables $x_1, \ldots, x_q$ and response variable y, the method finds the transformations $g(y)$ and $s_1(x_1), \ldots, s_q(x_q)$ that maximize the correlation between y and its predicted value. The technique allows for arbitrary, smooth transformations of both response and explanatory variables. [*Biometrika*, 1995, **82**, 369–83.]

Alternating least squares: A method most often used in some methods of *multidimensional scaling*, where a goodness-of-fit measure for some configuration of points is minimized in a series of steps, each involving the application of least squares. [MV1 Chapter 8.]

Alternating logistic regression: A method of *logistic regression* used in the analysis of *longitudinal data* when the response variable is binary. Based on *generalized estimating equations*. [*Analysis of Longitudinal Data*, 2nd edition, 2002, P.J. Diggle, K.-Y. Liang and S.L. Zeger, Oxford Science Publications, Oxford.]

Alternative hypothesis: The hypothesis against which the null hypothesis is tested.

Aly's statistic: A statistic used in a *permutation test* for comparing variances, and given by

$$\delta = \sum_{i=1}^{m-1} i(m-i)(X_{(i+1)} - X_{(i)})$$

where $X_{(1)} < X_{(2)} < \cdots < X_{(m)}$ are the *order statistics* of the first sample. [*Statistics and Probability Letters*, 1990, **9**, 323–5.]

Amersham model: A model used for *dose–response curves* in immunoassay and given by

$$y = 100(2(1-\beta_1)\beta_2)/(\beta_3 + \beta_2 + \beta_4 + x + [(\beta_3 - \beta_2 + \beta_4 + x)^2 + 4\beta_3\beta_2]^{\frac{1}{2}}) + \beta_1$$

where y is percentage binding and x is the analyte concentration. Estimates of the four parameters, $\beta_1, \beta_2, \beta_3, \beta_4$, may be obtained in a variety of ways. [*Medical Physics*, 2004 **31**, 2501–8.]

AML: Abbreviation for **asymmetric maximum likelihood.**

Amplitude: A term used in relation to *time series*, for the value of the series at its peak or trough taken from some mean value or trend line.

Amplitude gain: See **linear filters.**

Analysis as-randomized: Synonym for **intention-to-treat analysis.**

Analysis of covariance (ANCOVA): Originally used for an extension of the *analysis of variance* that allows for the possible effects of continuous concomitant variables (covariates) on the response variable, in addition to the effects of the factor or treatment variables. Usually assumed that covariates are unaffected by treatments and that their relationship to the response is linear. If such a relationship exists then inclusion of covariates in this way decreases the *error mean square* and hence increases the sensitivity of the *F-tests* used in assessing treatment differences. The term now appears to also be more generally used for almost any analysis seeking to assess the relationship between a response variable and a number of explanatory variables. See also **parallelism in ANCOVA** and **generalized linear models.** [KA2 Chapter 29.]

Analysis of dispersion: Synonym for **multivariate analysis of variance.**

Analysis of variance (ANOVA): The separation of variance attributable to one cause from the variance attributable to others. By partitioning the total variance of a set of observations into parts due to particular factors, for example, sex, treatment group etc., and comparing variances (mean squares) by way of *F-tests*, differences between means can be assessed. The simplest analysis of this type involves a *one-way design*, in which N subjects are allocated, usually at random, to the k different levels of a single factor. The total variation in the observations is then divided into a part due to differences between level means (the *between groups sum of squares*) and a part due to the differences between subjects in the same group (the *within groups sum of squares*, also known as the *residual sum of squares*). These terms are usually arranged as an *analysis of variance table*.

Source	**df**	**SS**	**MS**	**MSR**
Bet. grps.	$k-1$	SSB	$\text{SSB}/(k-1)$	$\frac{\text{SSB}/(k-1)}{\text{SSW}/(N-k)}$
With. grps.	$N-k$	SSW	$\text{SSW}/(N-k)$	
Total	$N-1$			

SS = sum of squares; MS = mean square; MSR = mean square ratio.

If the means of the populations represented by the factor levels are the same, then within the limits of random variation, the *between groups mean square* and *within groups mean square*, should be the same. Whether this is so can, if certain assumptions are met, be assessed by a suitable F-test on the mean square ratio. The necessary assumptions for the validity of the F-test are that the response variable is normally distributed in each population and that the populations have the same variance. Essentially an example of the *generalized linear model* with an identity *link function* and normally distributed error terms. See also **analysis of covariance, parallel groups design** and **factorial designs**. [SMR Chapter 9.]

Analysis of variance table: See **analysis of variance.**

Analytic epidemiology: A term for epidemiological studies, such as *case-control studies*, that obtain individual-level information on the association between disease status and exposures of interest. [*Journal of the National Cancer Institute*, 1996, **88**, 1738–47.]

Ancillary statistic: A term applied to the statistic C in situations where the *minimal sufficient statistic*, S, for a parameter θ, can be written as $S = (T, C)$ and C has a *marginal distribution* not depending on θ. For example, let N be a random variable with a known distribution $p_n = \Pr(N = n)(n = 1, 2, \ldots)$, and let $Y_1, Y_2, \ldots, Y_N$ be independently and identically distributed random variables from the *exponential family distribution* with parameter, θ. The *likelihood* of the data $(n, y_1, y_2, \ldots, y_n)$ is

$$p_n \exp\left\{a(\theta)\sum_{j=1}^{n} b(y_j) + nc(\theta) + \sum_{j=1}^{n} d(y_j)\right\}$$

so that $S = [\sum_{j=1}^{N} b(Y_j), N]$ is sufficient for θ and N is an ancillary statistic. Important in the application of *conditional likelihood* for estimation. [KA2 Chapter 31.]

ANCOVA: Acronym for **analysis of covariance.**

Anderson–Darling test: A test that a given sample of observations arises from some specified theoretical probability distribution. For testing the normality of the data, for example, the test statistic is

$$A_n^2 = -\frac{1}{n}\left[\sum_{i=1}^{n}(2i-1)\{\log z_i + \log(1 - z_{n+1-i})\}\right] - n$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the ordered observations, s^2 is the sample variance, and

$$z_i = \Phi\left(\frac{x_{(i)} - \bar{x}}{s}\right)$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

The null hypothesis of normality is rejected for 'large' values of A_n^2. Critical values of the test statistic are available. See also **Shapiro–Wilk test.** [*Journal of the American Statistical Society*, 1954, **49**, 765–9.]

Anderson–Gill model: A model for analysing *multiple time response data* in which each subject is treated as a multi-event *counting process* with essentially independent increments. [*Annals of Statistics*, 1982, **10**, 1100–20.]

Anderson, John Anthony (1939–1983): Anderson studied mathematics at Oxford, obtaining a first degree in 1963, and in 1968 he was awarded a D.Phil. for work on statistical methods in medical diagnosis. After working in the Department of Biomathematics in Oxford for some years, Anderson eventually moved to Newcastle University, becoming professor in 1982. Contributed to *multivariate analysis*, particularly *discriminant analysis* based on *logistic regression*. He died on 7 February 1983, in Newcastle.

Anderson, Oskar Nikolayevick (1887–1960): Born in Minsk, Byelorussia, Anderson studied mathematics at the University of Kazan. Later he took a law degree in St Petersburg and travelled to Turkestan to make a survey of agricultural production under irrigation in the Syr Darya River area. Anderson trained in statistics at the Commercial Institute in Kiev and from the mid-1920s he was a member of the Supreme Statistical Council of the Bulgarian government during which time he successfully advocated the use of sampling techniques. In 1942 Anderson accepted an appointment at the University of Kiel, Germany and from 1947 until his death he was Professor of Statistics in the Economics Department at the University of Munich. Anderson was a pioneer of applied sample-survey techniques.

Andrews' plots: A graphical display of multivariate data in which an observation, $\mathbf{x}' = [x_1, x_2, \ldots, x_q]$ is represented by a function of the form

$$f_{\mathbf{x}}(t) = x_1/\sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots$$

plotted over the range of values $-\pi \leq t \leq \pi$. A set of multivariate observations is displayed as a collection of these plots and it can be shown that those functions that remain close together for all values of t correspond to observations that are close to one another in terms of their *Euclidean distance*. This property means that such plots can often be used to both detect groups of similar observations and identify *outliers* in multivariate data. The example shown at Fig. 3 consists of plots for a sample of 30 observations each having five variable values. The plot indicates the presence of three groups in the data. Such plots can cope only with a moderate number of observations before becoming very difficult to unravel. See also **Chernoff faces** and **glyphs.** [MV1 Chapter 3.]

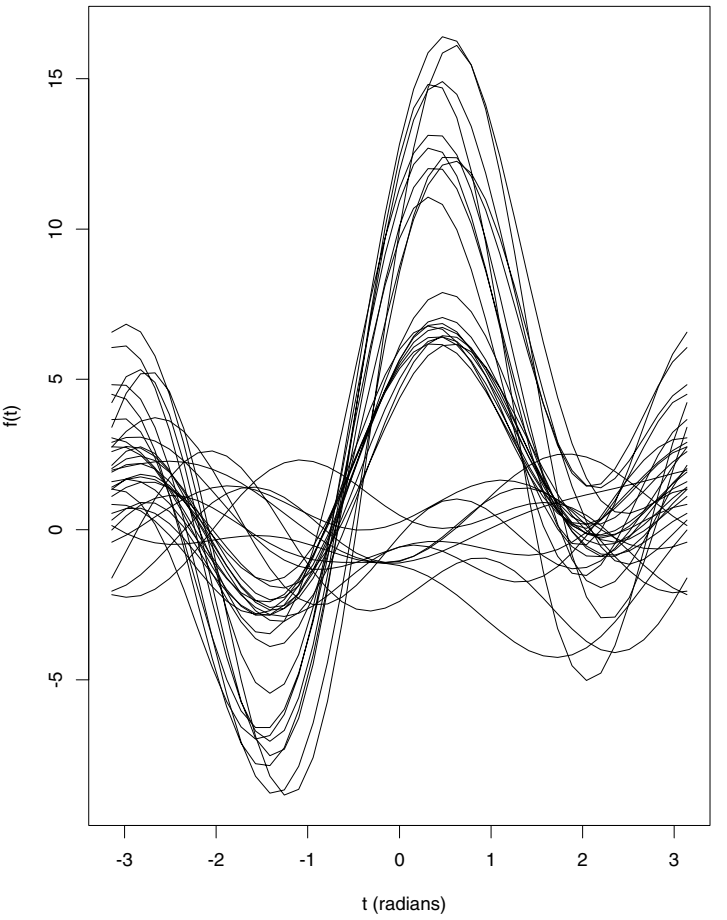


Fig. 3 Andrews' plot for 30, five-dimensional observations constructed to contain three relatively distinct groups.

Angle count method: A method for estimating the proportion of the area of a forest that is actually covered by the bases of trees. An observer goes to each of a number of points in the forest, chosen either randomly or systematically, and counts the number of trees that subtend, at that point, an angle greater than or equal to some predetermined fixed angle 2α. [*Spatial Data Analysis by Example*, Volume 1, 1985, G. Upton and B. Fingleton, Wiley, New York.]

Angler survey: A survey used by sport fishery managers to estimate the total catch, fishing effort and catch rate for a given body of water. For example, the total effort might be estimated in angler-hours and the catch rate in fish per angler-hour. The total catch is then estimated as the product of the estimates of total effort and average catch rate. [*Fisheries Techniques*, 1983, L.A. Nielson and D.C. Johnson, eds., American Fisheries Society, Bethesda, Maryland.]

Angular histogram: A method for displaying *circular data*, which involves wrapping the usual histogram around a circle. Each bar in the histogram is centred at the midpoint of the group interval with the length of the bar proportional to the frequency in the group. Figure 4 shows such a display for arrival times on a 24 hour clock of 254

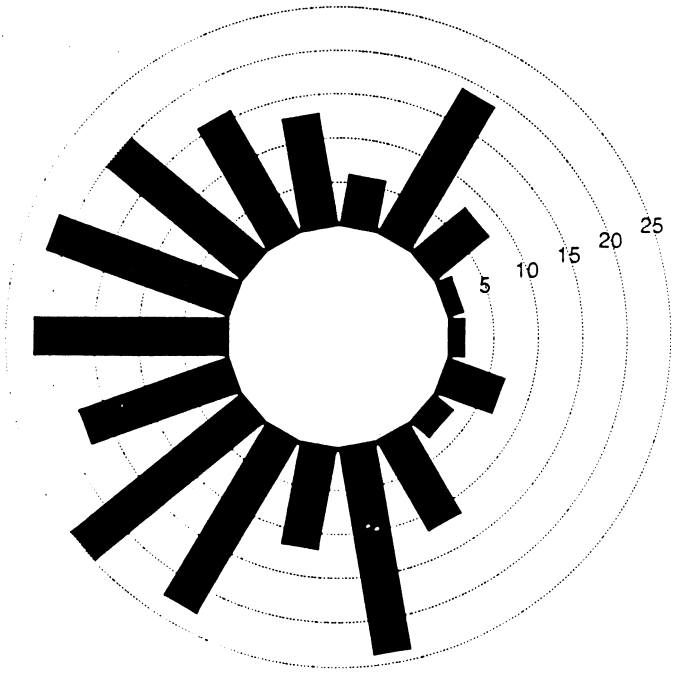


Fig. 4 Angular histogram for arrival times at an intensive care unit. (Reproduced by permission of Cambridge University Press from *Statistical Analysis of Circular Data*, 1993, N.I. Fisher.)

patients at an intensive care unit, over a period of 12 months. See also **rose diagram.** [*Statistical Analysis of Circular Data*, 1993, N.I. Fisher, Cambridge University Press, Cambridge.]

Angular transformation: Synonym for **arc sine transformation.**

Angular uniform distribution: A probability distribution for a *circular random variable*, θ, given by

$$f(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta \leq 2\pi$$

[*Statistical Analysis of Circular Data*, 1993, N.I. Fisher, Cambridge University Press, Cambridge.]

Annealing algorithm: Synonym for **simulated annealing.**

ANOVA: Acronym for **analysis of variance.**

Ansari–Bradley test: A test for the equality of variances of two populations having the same median. The test has rather poor efficiency relative to the *F-test* when the populations are normal. See also **Conover test** and **Klotz test.** [*Annals of Mathematical Statistics*, 1960, **31**, 1174–89.]

Anscombe residual: A *residual* based on the difference between some function of the observed value of a response and the same function of the fitted value under some assumed model. The function is chosen to make the residuals as normal as possible and for *generalized linear models* is obtained from

$$\int \frac{\mathrm{d}x}{[V(x)]^{\frac{1}{3}}}$$

where $V(x)$ is the function specifying the relationship between the mean and variance of the response variable of interest. For a variable with a *Poisson distribution*, for example, $V(x) = x$ and so residuals might be based on $y^{\frac{2}{3}} - \hat{y}^{\frac{2}{3}}$. [*Modelling Binary Data*, 2nd edition, 2002, D. Collett, Chapman and Hall/CRC Press, London.]

Antagonism: See **synergism.**

Antidependence models: A family of structures for the *variance-covariance matrix* of a set of *longitudinal data*, with the model of order r requiring that the sequence of random variables, $Y_1, Y_2, \ldots, Y_T$ is such that for every $t > r$

$$Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-r}$$

is conditionally independent of $Y_{t-r-1}, \ldots, Y_1$. In other words once account has been taken of the r observations preceding Y_t, the remaining preceding observations carry no additional information about Y_t. The model imposes no constraints on the constancy of variance or covariance with respect to time so that in terms of second-order moments, it is not *stationary*. This is a very useful property in practice since the data from many longitudinal studies often have increasing variance with time. [MV2 Chapter 13.]

Anthropometry: A term used for studies involving measuring the human body. Direct measures such as height and weight or indirect measures such as surface area may be of interest. See also **body mass index**. [*Human Growth and Development*, 1970, R. McCammon, Wiley, New York.]

Anti-ranks: For a random sample $X_1, \ldots, X_n$, the random variables $D_1, \ldots, D_n$ such that

$$Z_1 = |X_{D_1}| \leq \cdots \leq Z_n = |X_{D_n}|$$

If, for example, $D_1 = 2$ then X_2 is the smallest absolute value and Z_1 has rank 1. [*Robust Nonparametric Statistical Methods*, 1998, T.P. Hettmansperger and J.W. McKean, Arnold, London.]

Antithetic variable: A term that arises in some approaches to simulation in which successive simulation runs are undertaken to obtain identically distributed unbiased run estimators that rather than being independent are negatively correlated. The value of this approach is that it results in an unbiased estimator (the average of the estimates from all runs) that has a smaller variance than would the average of identically distributed run estimates that are independent. For example, if r is a random variable between 0 and 1 then so is $s = 1 - r$. Here the two simulation runs would involve $r_1, r_2, \ldots, r_m$ and $1 - r_1, 1 - r_2, \ldots, 1 - r_m$, which are clearly not independent. [*Proceedings of the Cambridge Philosophical Society*, 1956, **52**, 449–75.]

A-optimal design: See **criteria of optimality.**

A posteriori comparisons: Synonym for **post-hoc comparisons.**

Apparent error rate: Synonym for **resubstitution error rate.**

Approximate bootstrap confidence (ABC) method: A method for approximating confidence intervals obtained by using the *bootstrap* approach, that do not use any Monte Carlo replications. [*An Introduction to the Bootstrap*, 1994, B. Efron and R.J. Tibshirani, CRC/Chapman and Hall.]

Approximation: A result that is not exact but is sufficiently close for required purposes to be of practical use.

A priori comparisons: Synonym for **planned comparisons.**

Aranda–Ordaz transformations: A family of transformations for a proportion, p, given by

$$y = \ln\left[\frac{(1-p)^{-\alpha} - 1}{\alpha}\right]$$

When $\alpha = 1$, the formula reduces to the *logistic transformation* of p. As $\alpha \to 0$ the result is the *complementary log-log transformation*. [*Modelling Binary Data*, 2nd edition, 2002, D. Collett, Chapman and Hall/CRC Press, London.]

Arbuthnot, John (1667–1735): Born in Inverbervie, Grampian, Arbuthnot was physician to Queen Anne from 1709 until her death in 1714. A friend of Jonathan Swift who is best known to posterity as the author of satirical pamphlets against the Duke of Marlborough and creator of the prototypical Englishman, John Bull. His statistical claim to fame is based on a short note published in the *Philosophical Transactions of the Royal Society* in 1710, entitled 'An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes.' In this note he claimed to demonstrate that divine providence, not chance governed the sex ratio at birth, and presented data on christenings in London for the eighty-two-year period 1629–1710 to support his claim. Part of his reasoning is recognizable as what would now be known as a *sign test*. Arbuthnot was elected a Fellow of the Royal Society in 1704. He died on 27 February 1735 in London.

Archetypal analysis: An approach to the analysis of multivariate data which seeks to represent each individual in the data as a mixture of individuals of pure type or archetypes. The archetypes themselves are restricted to being mixtures of individuals in the data set. Explicitly the problem is to find a set of $q \times 1$ vectors $\mathbf{z}_1, \ldots, \mathbf{z}_p$ that characterize the archetypal patterns in the multivariate data, $\mathbf{X}$. For fixed $\mathbf{z}_1, \ldots, \mathbf{z}_p$ where

$$\mathbf{z}_k = \sum_{j=1}^{n} \beta_{kj}\mathbf{x}_j \qquad k = 1, \ldots, p$$

and $\beta_{ki} \geq 0$, $\sum_i \beta_{ki} = 1$, define $\{\alpha_{ik}\}$, $k = 1, \ldots, p$ as the minimizers of

$$\left\| \mathbf{x}_i - \sum_{k=1}^{p} \alpha_{ik}\mathbf{z}_k \right\|^2$$

under the constraints, $\alpha_{ik} \geq 0$, $\sum \alpha_{ik} = 1$. Then define the archetypal patterns or archetypes as the mixtures $\mathbf{z}_1, \ldots, \mathbf{z}_p$ that minimize

$$\sum_i \left\| \mathbf{x}_i - \sum_{k=1}^{p} \alpha_{ik}\mathbf{z}_k \right\|^2$$

For $p > 1$ the archetypes fall on the *convex hull* of the data; they are extreme data values such that all the data can be represented as convex mixtures of the archetypes. However, the archetypes themselves are not wholly mythological because each is constrained to be a mixture of points in the data. [*Technometrics*, 1994, **36**, 338–47.]

Arc sine distribution: A *beta distribution* with $\alpha = \beta = 0.5$.

Arc sine law: An approximation applicable to a simple *random walk* taking values 1 and -1 with probabilities $\frac{1}{2}$ which allows easy computation of the probability of the fraction of time that the accumulated score is either positive or negative. The

approximation can be stated thus; for fixed α ($0 < \alpha < 1$) and $n \to \infty$ the probability that the fraction k/n of time that the accumulated score is positive is less than α tends to

$$2\pi^{-1}\arcsin(\alpha^{\frac{1}{2}})$$

For example, if an unbiased coin is tossed once per second for a total of 365 days, there is a probability of 0.05 that the more fortunate player will be in the lead for more than 364 days and 10 hours. Few people will believe that a perfect coin will produce sequences in which no change of lead occurs for millions of trials in succession and yet this is what such a coin will do rather regularly. Intuitively most people feel that values of k/n close to $\frac{1}{2}$ are most likely. The opposite is in fact true. The possible values close to $\frac{1}{2}$ are least probable and the extreme values $k/n = 1$ and $k/n = 0$ are most probable. Figure 5 shows the results of an experiment simulating 5000 tosses of a fair coin ($\Pr(\text{Heads}) = \Pr(\text{Tails}) = \frac{1}{2}$) in which a head is given a score of 1 and a tail -1. Note the length of the waves between successive crossings of $y = 0$, i.e., successive changes of lead. [*An Introduction to Probability Theory and its Applications*, Volume 1, 3rd edition, 1968, W. Feller, Wiley, New York.]

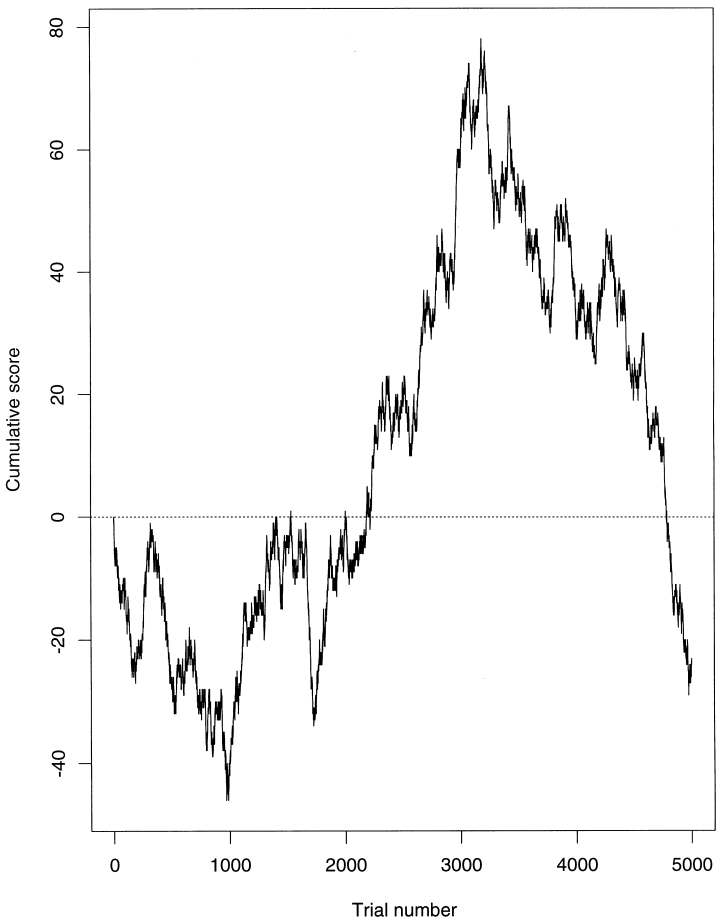


Fig. 5 Result of 5000 tosses of a fair coin scoring 1 for heads and -1 for tails.

Arc sine transformation: A transformation for a proportion, p, designed to stabilize its variance and produce values more suitable for techniques such as *analysis of variance* and regression analysis. The transformation is given by

$$y = \sin^{-1}\sqrt{p}$$

[*Modelling Binary Data*, 2nd edition, 2002, D. Collett, Chapman and Hall/CRC Press, London.]

ARE: Abbreviation for **asymptotic relative efficiency.**

Area sampling: A method of sampling where a geographical region is subdivided into smaller areas (counties, villages, city blocks, etc.), some of which are selected at random, and the chosen areas are then subsampled or completely surveyed. See also **cluster sampling.** [*Handbook of Area Sampling*, 1959, J. Monroe and A.L. Fisher, Chilton, New York.]

Area under curve (AUC): Often a useful way of summarizing the information from a series of measurements made on an individual over time, for example, those collected in a *longitudinal study* or for a *dose–response curve*. Usually calculated by adding the areas under the curve between each pair of consecutive observations, using, for example, the *trapezium rule*. Often a predictor of biological effects such as toxicity or efficacy. See also $\mathbf{C_{max}}$, **response feature analysis** and $\mathbf{T_{max}}$. [SMR Chapter 14.]

Arfwedson distribution: The probability distribution of the number of zero values (M_0) among k random variables having a *multinomial distribution* with $p_1 = p_2 = \cdots = p_k$. If the sum of the k random variables is n then the distribution is given by

$$\Pr(M_0 = m) = \binom{k}{m}\sum_{i=0}^{m}(-1)^i\binom{m}{i}\left(\frac{m-i}{k}\right)^n \qquad m = 0, 1, \ldots, k-1$$

[*Skandinavisk Aktuarletidskrift*, 1951, **34**, 121–32.]

ARIMA: Abbreviation for **autoregressive integrated moving-average model.**

Arithmetic mean: See **mean.**

Arjas plot: A procedure for checking the fit of *Cox's proportional hazards model* by comparing the observed and expected number of events, as a function of time, for various subgroups of covariate values. [*Journal of the American Statistical Association*, 1988, **83**, 204–12.]

ARMA: Abbreviation for **autoregressive moving-average model.**

Armitage–Doll model: A model of carcinogenesis in which the central idea is that the important variable determining the change in risk is not age, but time. The model proposes that cancer of a particular tissue develops according to the following process:

- a normal cell develops into a cancer cell by means of a small number of transitions through a series of intermediate steps;
- initially, the number of normal cells at risk is very large, and for each cell a transition is a rare event;
- the transitions are independent of one another.

[*Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 1961, L.M. Le Cam and J. Neyman (eds.) University of California Press, Berkeley.]

Armitage–Hill test: A test for *carry-over effect* in a *two-by-two crossover design* where the response is a *binary variable*. [*Diabetic Medicine*, 2004, **21**, 769–74.]

Artificial intelligence: A discipline that attempts to understand intelligent behaviour in the broadest sense, by getting computers to reproduce it, and to produce machines that behave intelligently, no matter what their underlying mechanism. (Intelligent behaviour is taken to include reasoning, thinking and learning.) See also **artificial neural network hand pattern recognition.** [*Artificial Intelligence Frontiers in Statistics*, 1993, D.J. Hand, Chapman and Hall/CRC Press, London.]

Artificial neural network: A mathematical structure modelled on the human neural network and designed to attack many statistical problems, particularly in the areas of *pattern recognition*, *multivariate analysis*, learning and memory. The essential feature of such a structure is a network of simple processing elements (*artificial neurons*) coupled together (either in the hardware or software), so that they can cooperate. From a set of 'inputs' and an associated set of parameters, the artificial neurons produce an 'output' that provides a possible solution to the problem under investigation. In many neural networks the relationship between the input received by a neuron and its output is determined by a *generalized linear model*. The most common form is the *feed-forward network* which is essentially an extension of the idea of the *perceptron*. In such a network the vertices can be numbered so that all connections go from a vertex to one with a higher number; the vertices are arranged in layers, with connections only to higher layers. This is illustrated in Fig. 6. Each neuron sums its inputs to form a total input x_j and applies a function f_j to x_j to give output y_j. The links have weights w_{ij} which multiply the signals travelling along them by that factor. Many ideas and activities familiar to statisticians can be expressed in a neural-network notation, including regression analysis, *generalized additive models*, and *discriminant analysis*. In any practical problem the statistical equivalent of specifying the architecture of a suitable network is specifying a suitable model, and training the network to perform well with reference to a training set is equivalent to estimating the parameters of the model given a set of data. [*Pattern Recognition and Neural Networks*, 1996, B.D. Ripley, Cambridge University Press, Cambridge.]

Artificial neuron: See **artificial neural network.**

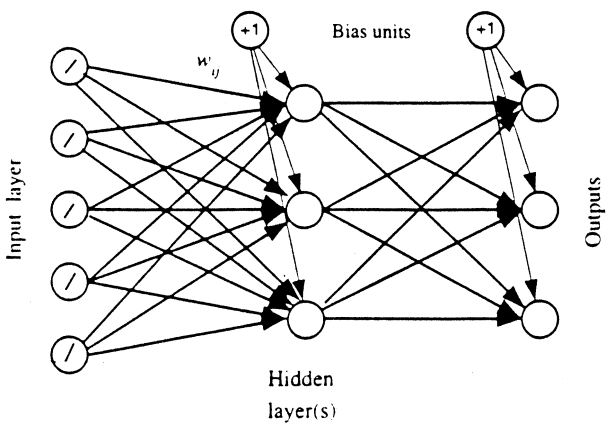


Fig. 6 A diagram illustrating a feed-forward network.

Artificial pairing: See **paired samples.**

Ascertainment bias: A possible form of bias, particularly in *retrospective studies*, that arises from a relationship between the exposure to a risk factor and the probability of detecting an event of interest. In a study comparing women with cervical cancer and a control group, for example, an excess of oral contraceptive use among the cases might possibly be due to more frequent screening for the disease among women known to be taking the pill. [SMR Chapter 5.]

ASN: Abbreviation for **average sample number.**

As-randomized analysis: Synonym for **intention-to-treat analysis.**

Assignment method: Synonym for **discriminant analysis.**

Association: A general term used to describe the relationship between two variables. Essentially synonymous with correlation. Most often applied in the context of binary variables forming a *two-by-two contingency table*. See also **phi-coefficient** and **Goodman–Kruskal measures of association.** [SMR Chapter 11.]

Assortative mating: A form of non-random mating where the probability of mating between two individuals is influenced by their *phenotypes* (*phenotypic assortment*), *genotypes* (*genotypic assortment*) or environments (*cultural assortment*). [*Statistics in Human Genetics*, 1998, P. Sham, Arnold, London.]

Assumptions: The conditions under which statistical techniques give valid results. For example, *analysis of variance* generally assumes normality, homogeneity of variance and independence of the observations.

Asymmetrical distribution: A probability distribution or frequency distribution which is not symmetrical about some central value. Examples include the *exponential distribution* and *J-shaped distribution*. [KA1 Chapter 1.]

Asymmetric maximum likelihood (AML): A variant of *maximum likelihood estimation* that is useful for estimating and describing *overdispersion* in a *generalized linear model*. [*IEEE Proceedings Part F – Communications, Radar and Signal Processing*, 1982, **129**, 331–40.]

Asymmetric proximity matrices: *Proximity matrices* in which the off-diagonal elements, in the ith row and jth column and the jth row and ith column, are not necessarily equal. Examples are provided by the number of marriages between men of one nationality and women of another, immigration/emigration statistics and the number of citations of one journal by another. *Multidimensional scaling* methods for such matrices generally rely on their canonical decomposition into the sum of a symmetric matrix and a skew symmetric matrix. [MV1 Chapter 5.]

Asymptotically unbiased estimator: An estimator of a parameter which tends to being unbiased as the sample size, n, increases. For example,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is not an unbiased estimator of the population variance σ^2 since its expected value is

$$\frac{n-1}{n}\sigma^2$$

but it is asymptotically unbiased. [*Normal Approximation and Asymptotic Expansions*, 1976, R.N. Bhattacharya and R. Rao, Wiley, New York.]

Asymptotic distribution: The limiting probability distribution of a random variable calculated in some way from n other random variables, as $n \to \infty$. For example, the mean of n random variables from a *uniform distribution* has a normal distribution for large n. [KA2 Chapter 25.]

Asymptotic efficiency: A term applied when the estimate of a parameter has a normal distribution around the true value as mean and with a variance achieving the *Cramér–Rao lower bound*. See also **superefficient.** [KA2 Chapter 25.]

Asymptotic method: Synonym for **large sample method.**

Asymptotic relative efficiency: The *relative efficiency* of two estimators of a parameter in the limit as the sample size increases. [KA2 Chapter 25.]

Atlas mapping: A biogeographical method used to investigate species-specific distributional status, in which observations are recorded in a grid of cells. Such maps are examples of *geographical information systems*. [*Biometrics*, 1995, **51**, 393–404.]

Attack rate: A term often used for the *incidence* of a disease or condition in a particular group, or during a limited period of time, or under special circumstances such as an epidemic. A specific example would be one involving outbreaks of food poisoning, where the attack rates would be calculated for those people who have eaten a particular item and for those who have not. [*Epidemiology Principles and Methods*, 1970, B. MacMahon and T.F. Pugh, Little, Brown and Company, Boston.]

Attenuation: A term applied to the correlation between two variables when both are subject to measurement error, to indicate that the value of the correlation between the 'true values' is likely to be underestimated. See also **regression dilution.** [*Biostatistics*, 1993, L.D. Fisher and G. Van Belle, Wiley, New York.]

Attitude scaling: The process of estimating the positions of individuals on scales purporting to measure attitudes, for example a *liberal–conservative scale*, or a *risk-willingness scale*. Scaling is achieved by developing or selecting a number of stimuli, or items which measure varying levels of the attitude being studied. See also **Likert scale** and **multidimensional scaling**. [*Sociological Methodology*, 1999, **29**, 113–46.]

Attributable response function: A function $N(x, x_0)$ which can be used to summarize the effect of a numerical covariate x on a binary response probability. Assuming that in a finite population there are $m(x)$ individuals with covariate level x who respond with probability $\pi(x)$, then $N(x, x_0)$ is defined as

$$N(x, x_0) = m(x)\{\pi(x) - \pi(x_0)\}$$

The function represents the response attributable to the covariate having value x rather than x_0. When plotted against $x \geq x_0$ this function summarizes the importance of different covariate values in the total response. [*Biometrika*, 1996, **83**, 563–73.]

Attributable risk: A measure of the association between exposure to a particular factor and the risk of a particular outcome, calculated as

$$\frac{\text{incidence rate among exposed} - \text{incidence rate among nonexposed}}{\text{incidence rate among exposed}}$$

Measures the amount of the *incidence* that can be attributed to one particular factor.

See also **relative risk** and **prevented fraction.** [*An Introduction to Epidemiology*, 1983, M. Alderson, Macmillan, London.]

Attrition: A term used to describe the loss of subjects over the period of a *longitudinal study*. May occur for a variety of reasons, for example, subjects moving out of the area, subjects dropping out because they feel the treatment is producing adverse side effects, etc. Such a phenomenon may cause problems in the analysis of data from such studies. See also **missing values** and **Diggle–Kenward model for dropouts.**

AUC: Abbreviation for **area under curve.**

Audit in clinical trials: The process of ensuring that data collected in complex *clinical trials* are of high quality. [*Controlled Clinical Trials*, 1995, **16**, 104–36.]

Audit trail: A computer program that keeps a record of changes made to a *database*.

Autocorrelation: The internal correlation of the observations in a *time series*, usually expressed as a function of the time lag between observations. Also used for the correlations between points different distances apart in a set of *spatial data* (*spatial autocorrelation*). The autocorrelation at lag k, $\gamma(k)$, is defined mathematically as

$$\gamma(k) = \frac{E(X_t - \mu)(X_{t+k} - \mu)}{E(X_t - \mu)^2}$$

where $X_t, t = 0, \pm 1, \pm 2, \ldots$ represent the values of the series and μ is the mean of the series. E denotes expected value. The corresponding sample statistic is calculated as

$$\hat{\gamma}(k) = \frac{\sum_{i=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{i=1}^{n}(x_t - \bar{x})^2}$$

where $\bar{x}$ is the mean of the series of observed values, $x_1, x_2, \ldots, x_n$. A plot of the sample values of the autocorrelation against the lag is known as the *autocorrelation function* or *correlogram* and is a basic tool in the analysis of time series particularly for indicating possibly suitable models for the series. An example is shown in Fig. 7. The term in the numerator of $\gamma(k)$ is the *autocovariance*. A plot of the autocovariance against lag is called the *autocovariance function*. [TMS Chapter 2.]

Autocorrelation function: See **autocorrelation.**

Autocovariance: See **autocorrelation.**

Autocovariance function: See **autocorrelation.**

Automatic interaction detector (AID): A method that uses a set of categorical explanatory variables to divide data into groups that are relatively homogeneous with respect to the value of some continuous response variable of interest. At each stage, the division of a group into two parts is defined by one of the explanatory variables, a subset of its categories defining one of the parts and the remaining categories the other part. Of the possible splits, the one chosen is that which maximizes the between groups sum of squares of the response variable. The groups eventually formed may often be useful in predicting the value of the response variable for some future observation. See also **classification and regression tree technique** and **chi-squared automated interaction detector.** [*Journal of the American Statistical Society*, 1963, **58**, 415–34.]

Autoregressive integrated moving-average models: See **autoregressive moving-average model.**

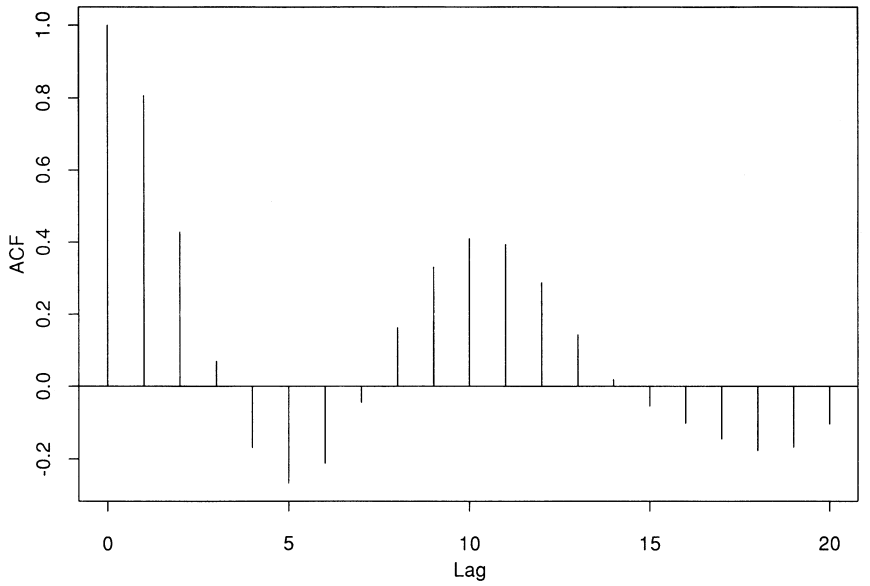


Fig. 7 An example of an autocorrelation function.

Autoregressive model: A model used primarily in the analysis of *time series* in which the observation, x_t, at time t, is postulated to be a linear function of previous values of the series. So, for example, a *first-order autoregressive model* is of the form

$$x_t = \phi x_{t-1} + a_t$$

where a_t is a random disturbance and ϕ is a parameter of the model. The corresponding model of order p is

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t$$

which includes the p parameters, $\phi_1, \phi_2, \ldots, \phi_p$. [TMS Chapter 4.]

Autoregressive moving-average model: A model for a *time series* that combines both an *autoregressive model* and a *moving-average model*. The general model of order p, q (usually denoted ARMA(p, q)) is

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model and $a_t, a_{t-1}, \ldots$ are a *white noise sequence*. In some cases such models are applied to the time series observations after *differencing* to achieve *stationarity*, in which case they are known as *autoregressive integrated moving-average models*. [TMS Chapter 4.]

Auxiliary variable techniques: Techniques for improving the performance of *Gibbs sampling* in the context of *Bayesian inference* for *hierarchical models*. [*Journal of the Royal Statistical Society, Series B*, 1993, **55**, 25–37.]

Available case analysis: An approach to handling *missing values* in a set of multivariate data, in which means, variances, covariances, etc., are calculated from all available subjects with non-missing values for the variable or pair of variables involved. Although this approach makes use of as much of the data as possible it has disadvantages. One is that summary statistics will be based on different numbers of observations. More problematic however is that this method can lead to *variance–*

covariance matrices and *correlation matrices* with properties that make them unsuitable for many methods of multivariate analysis such as *principal components analysis* and *factor analysis*. [*Analysis of Incomplete Multivariate Data*, 1997, J.L. Schafer, Chapman and Hall/CRC Press, London.]

Average: Most often used for the arithmetic mean of a sample of observations, but can also be used for other measures of location such as the median.

Average age at death: A flawed statistic summarizing *life expectancy* and other aspects of mortality. For example, a study comparing average age at death for male symphony orchestra conductors and for the entire US male population showed that, on average, the conductors lived about four years longer. The difference is, however, illusory, because as age at entry was birth, those in the US male population who died in infancy and childhood were included in the calculation of the average life span, whereas only men who survived to become conductors could enter the conductor cohort. The apparent difference in longevity disappeared after accounting for infant and perinatal mortality. [*Methodological Errors in Medical Research*, 1990, B. Andersen, Blackwell Scientific, Oxford.]

Average deviation: A little-used measure of the spread of a sample of observations. It is defined as

$$\text{Average deviation} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the sample values, and $\bar{x}$ their mean.

Average linkage: An *agglomerative hierarchical clustering method* that uses the average distance from members of one cluster to members of another cluster as the measure of inter-group distance. This distance is illustrated in Fig. 8. [MV2 Chapter 10.]

Average man: See **Quetelet, Adolphe (1796–1874)**.

Average sample number (ASN): A quantity used to describe the performance of a *sequential analysis* given by the expected value of the sample size required to reach a decision to accept the null hypothesis or the alternative hypothesis and therefore to discontinue sampling. [KA2 Chapter 24.]

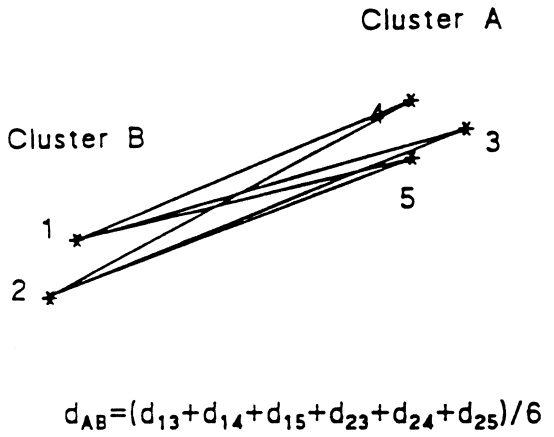


Fig. 8 Average linkage distance for two clusters.

B

b632 method: A procedure of *error rate estimation* in *discriminant analysis* based on the *bootstrap*, which consists of the following steps:

(1) Randomly sample (with replacement) a *bootstrap sample* from the original data.
(2) Classify the observations omitted from the bootstrap sample using the classification rule calculated from the bootstrap sample.
(3) Repeat (1) and (2) many times and calculate the mean bootstrap *classification matrix*, C_b.
(4) Calculate the resubstitution classification matrix, C_r, based on the original data.
(5) The b632 estimator of the classification matrix is $0.368C_r + 0.632C_b$, from which the required error rate estimate can be obtained. [*Technometrics*, 1996, **38**, 289–99.]

B_k method: A form of *cluster analysis* which produces overlapping clusters. A maximum of $k - 1$ objects may belong to the overlap between any pair of clusters. When $k = 1$ the procedure becomes *single linkage clustering*. [*Classification*, 2nd edition, 1999, A.D. Gordon, CRC/Chapman and Hall, London.]

Babbage, Charles (1792–1871): Born near Teignmouth in Devon, Babbage read mathematics at Trinity College, Cambridge, graduating in 1814. His early work was in the theory of functions and modern algebra. Babbage was elected a Fellow of the Royal Society in 1816. Between 1828 and 1839 he held the Lucasian Chair of Mathematics at Trinity College. In the 1820s Babbage developed a 'Difference Engine' to form and print mathematical tables for navigation and spent much time and money developing and perfecting his calculating machines. His ideas were too ambitious to be realized by the mechanical devices available at the time, but can now be seen to contain the essential germ of today's electronic computer. Babbage is rightly seen as the pioneer of modern computers.

Back-calculation: Synonym for **back-projection.**

Back-projection: A term most often applied to a procedure for reconstructing plausible HIV incidence curves from AIDS incidence data. The method assumes that the probability distribution of the *incubation period* of AIDS has been estimated precisely from separate *cohort studies* and uses this distribution to project the AIDS incidence data backwards to reconstruct an HIV *epidemic curve* that could plausibly have led to the observed AIDS incidence data. [*Statistics in Medicine*, 1994, **13**, 1865–80.]

Back-to-back stem-and-leaf plots: A method for comparing two distributions by 'hanging' the two sets of leaves in the *stem-and-leaf plots* of the two sets of data, off either side of the same stem. An example appears in Fig. 9.

Backward elimination procedure: See **selection methods in regression.**

Before						After		
			12	9				
			13	1	6			
	8	6	14	5	7	7		
		4	15	1	7			
9	7	0	16	5	6	8	8	
6	4	3	17	9				
	7	5	18	0				
		8	19					
		0	20	1				

Fig. 9 Back-to-back stem-and-leaf plot of systolic blood pressure of fifteen subjects before and two hours after taking the drug captoril.

Backward-looking study: An alternative term for *retrospective study*.

Backward shift operator: A mathematical operator denoted by B, met in the analysis of *time series*. When applied to such a series the operator moves the observations back one time unit, so that if x_t represents the values of the series then, for example,

$$Bx_t = x_{t-1}$$
$$B(Bx_t) = B(x_{t-1}) = x_{t-2}$$

Bagging: A term used for producing replicates of the *training set* in a classification problem and producing an *allocation rule* on each replicate. The basis of *bagging predictors* which involve multiple versions of a predictor that are used to get an aggregated predictor. [*Statistical Pattern Recognition*, 1999, A. Webb, Arnold, London.]

Bagging predictors: See **bagging**.

Bagplot: An approach to detecting *outliers* in *bivariate data*. The plot visualizes location, spread, correlation, *skewness* and the tails of the data without making assumptions about the data being symmetrically distributed. [*American Statistician*, 1999, **53**, 382–7.]

Balaam's design: A design for testing differences between two treatments A and B in which patients are randomly allocated to one of four sequences, AA, AB, BA, or BB. See also **crossover design.** [*Statistics in Medicine*, 1988, **7**, 471–82.]

Balanced design: A term usually applied to any experimental design in which the same number of observations is taken for each combination of the experimental factors.

Balanced incomplete block design: A design in which not all treatments are used in all *blocks*. Such designs have the following properties:

- each block contains the same number of units;
- each treatment occurs the same number of times in all blocks;
- each pair of treatment combinations occurs together in a block the same number of times as any other pair of treatments.

In medicine this type of design might be employed to avoid asking subjects to attend for treatment an unrealistic number of times, and thus possibly preventing problems with missing values. For example, in a study with five treatments, it might be thought that subjects could realistically only be asked to make three visits. A possible balanced incomplete design in this case would be the following:

Patient	Visit 1	Visit 2	Visit 3
1	T_4	T_5	T_1
2	T_4	T_2	T_5
3	T_2	T_4	T_1
4	T_5	T_3	T_1
5	T_3	T_4	T_5
6	T_2	T_3	T_1
7	T_3	T_1	T_4
8	T_3	T_5	T_2
9	T_2	T_3	T_4
10	T_5	T_1	T_2

[*Experimental Designs*, 2nd edition, 1992, W. Cochran and G. Cox, Wiley, New York.]

Balanced incomplete repeated measures design (BIRMD): An arrangement of N randomly selected experimental units and k treatments in which every unit receives k_1 treatments $1 \leq k_1 < k$, each treatment is administered to r experimental units and each pair of treatments occurs together λ times. See also **balanced incomplete blocks.**

Balanced longitudinal data: See **longitudinal data.**

Balanced repeated replication (BRR): A popular method for variance estimation in surveys which works by creating a set of 'balanced' pseudoreplicated datasets from the original dataset. For an estimator, $\hat{\theta}$, of a parameter, θ, the estimated variance is obtained as the average of the squared deviations, $\hat{\theta}^{(r)} - \hat{\theta}$, where $\hat{\theta}^{(r)}$ is the estimate based on the rth replicated data set. See also **jackknife**. [*Journal of the American Statistical Association*, 1970, **65**, 1071–94.]

Balancing score: Synonymous with **propensity score**.

Ballot theorem: Let $X_1, X_2, \ldots, X_n$ be independent random variables each with a *Bernoulli distribution* with $\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}$. Define S_k as the sum of the first k of the observed values of these variables, i.e. $S_k = X_1 + X_2 + \cdots + X_k$ and let a and b be nonnegative integers such that $a - b > 0$ and $a + b = n$, then

$$\Pr(S_1 > 0, S_2 > 0, \ldots, S_n > 0 | S_n = a - b) = \frac{a - b}{a + b}$$

If $+1$ is interpreted as a vote for candidate A and -1 as a vote for candidate B, then S_k is the difference in numbers of votes cast for A and B at the time when k votes have been recorded; the probability given is that A is always ahead of B given that A receives a votes in all and B receives b votes. [*An Introduction to Probability Theory and its Applications*, Volume 1, 3rd edition, 1968, W. Feller, Wiley, New York.]

BAN: Abbreviation for **best asymptotically normal estimator**.

Banach's match-box problem: A person carries two boxes of matches, one in their left and one in their right pocket. Initially they contain N matches each. When the person wants a match, a pocket is selected at random, the successive choices thus constituting *Bernoulli trials* with $p = \frac{1}{2}$. On the first occasion that the person finds that a box is empty the other box may contain $0, 1, 2, \ldots, N$ matches. The probability distribution of the number of matches, R, left in the other box is given by:

$$\Pr(R = r) = \binom{2N - r}{N} \frac{1}{2}^{(2N-r)}$$

So, for example, for $N = 50$ the probability of there being not more than 10 matches in the second box is 0.754. [*An Introduction to Probability Theory and its Applications*, Volume 1, 3rd edition, 1968, W. Feller, Wiley, New York.]

Bancroft, Theodore Alfonso (1907–1986): Born in Columbus, Mississippi, Bancroft received a first degree in mathematics from the University of Florida. In 1943 he completed his doctorate in mathematical statistics with a dissertation entitled 'Tests of Significance Considered as an Aid in Statistical Methodology'. In 1950 he became Head of the Department of Statistics of the Iowa Agriculture and Home Economics Experiment Station. His principal area of research was incompletely specified models. Bancroft served as President of the American Statistical Association in 1970. He died on 26 July 1986 in Ames, Iowa.

Bandwidth: See **kernel estimation.**

Bar chart: A form of graphical representation for displaying data classified into a number of (usually unordered) categories. Equal-width rectangular bars are constructed over each category with height equal to the observed frequency of the category as shown in Fig. 10. See also **histogram** and **component bar chart.**

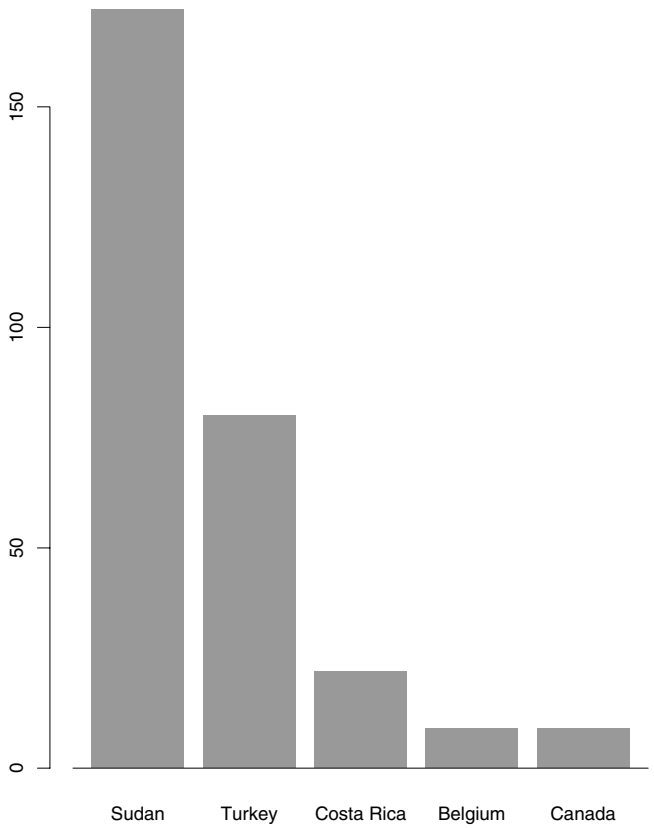


Fig. 10 Bar chart of mortality rates per 1000 live births for children under five years of age in five different countries.

Barnard, George Alfred (1915–2002): Born in Walthamstow in the east end of London, Barnard gained a scholarship to St. John's College, Cambridge, where he graduated in mathematics in 1936. For the next three years he studied mathematical logic at Princeton, New Jersey, and then in 1940 joined the engineering firm, Plessey. After three years acting as a mathematical consultant for engineers, Barnard joined the Ministry of Supply and it was here that his interest in statistics developed. In 1945 he went to Imperial College London, and then in 1966 he moved to a chair in the newly created University of Essex, where he stayed until his retirement in 1975. Barnard made major and important contributions to several fundamental areas of inference, including likelihood and 2×2 tables. He was made President of the Royal Statistical Society in 1971–2 and also received the Society's Guy medal in gold. He died in Brightlingsea, Essex, on 30 July 2002.

Barrett and Marshall model for conception: A biologically plausible model for the probability of conception in a particular menstrual cycle, which assumes that batches of sperm introduced on different days behave independently. The model is

$$P(\text{conception in cycle } k|\{X_{ik}\}) = 1 - \prod_i (1 - p_i)^{X_{ik}}$$

where the X_{ik} are 0,1 variables corresponding to whether there was intercourse or not on a particular day relative to the estimated day of ovulation (day 0). The parameter p_i is interpreted as the probability that conception would occur following intercourse on day i only. See also **EU model**. [*Biometrics*, 2001, **57**, 1067–73.]

Bartholomew's likelihood function: The joint probability of obtaining the observed known-complete survival times as well as the so-far survived measurements of individuals who are still alive at the date of completion of the study or other endpoint of the period of observation. [*Journal of the American Statistical Association*, 1957, **52**, 350–5.]

Bartlett, Maurice Stevenson (1910–2002): Born in Chiswick, London, Bartlett won a scholarship to Latymer Upper School, where his interest in probability was awakened by a chapter on the topic in Hall and Knight's *Algebra*. In 1929 he went to Queen's College, Cambridge to read mathematics, and in his final undergraduate year in 1932 published his first paper (jointly with *John Wishart*), on second-order moments in a normal system. On leaving Cambridge in 1933 Bartlett became Assistant Lecturer in the new Statistics Department at University College London, where his colleagues included *Egon Pearson*, *Fisher* and *Neyman*. In 1934 he joined Imperial Chemical Industries (ICI) as a statistician. During four very creative years Bartlett published some two-dozen papers on topics as varied as the theory of inbreeding and the effect of non-normality on the t-distribution. From ICI he moved to a lectureship at the University of Cambridge, and then during World War II he was placed in the Ministry of Supply. After the war he returned to Cambridge and began his studies of *time series* and diffusion processes. In 1947 Bartlett was given the Chair of Mathematical Statistics at the University of Manchester where he spent the next 13 years, publishing two important books, *An Introduction to Stochastic Processes* (in 1955) and *Stochastic Population Models in Ecology and Epidemiology* (in 1960) as well as a stream of papers on *stochastic processes*, etc. It was in 1960 that Bartlett returned to University College taking the Chair in Statistics, his work now taking in stochastic path integrals, spatial patterns and *multivariate analysis*. His final post was at Oxford where he held the Chair of Biomathematics from 1967 until his retirement eight years later. Bartlett received many honours and awards in his long and productive career,

including being made a Fellow of the Royal Society in 1961 and being President of the Royal Statistical Society for 1966–7. He died on 8 January 2002, in Exmouth, Devon.

Bartlett's adjustment factor: A correction term for the *likelihood ratio* that makes the *chi-squared distribution* a more accurate approximation to its probability distribution. [*Multivariate Analysis*, 1979, K.V. Mardia, J.T. Kent, and J.M. Bibby, Academic Press, London.]

Bartlett's identity: A matrix identity useful in several areas of multivariate analysis and given by

$$(\mathbf{A} + c\mathbf{b}\mathbf{b}')^{-1} = \mathbf{A}^{-1} - \frac{c}{1 + c\mathbf{b}'\mathbf{A}^{-1}\mathbf{b}}\mathbf{A}^{-1}\mathbf{b}\mathbf{b}'\mathbf{A}^{-1}$$

where $\mathbf{A}$ is $q \times q$ and nonsingular, $\mathbf{b}$ is a $q \times 1$ vector and c is a scalar.

Bartlett's test for eigenvalues: A large-sample test for the null hypothesis that the last $(q - k)$ eigenvalues, $\lambda_{k+1}, \ldots, \lambda_q$, of a *variance–covariance matrix* are zero. The test statistic is

$$X^2 = -\nu \sum_{j=k+1}^{q} \ln(\lambda_j) + \nu(q-k)\ln\left[\frac{\sum_{j=k+1}^{q}\lambda_j}{q-k}\right]$$

Under the null hypothesis, X^2 has a *chi-squared distribution* with $(1/2)(q-k-1)(q-k+2)$ degrees of freedom, where ν is the degrees of freedom associated with the covariance matrix. Used mainly in *principal components analysis*. [MV1 Chapter 4.]

Bartlett's test for variances: A test for the equality of the variances of a number (k) of populations. The test statistic is given by

$$B = \left[\nu \ln s^2 + \sum_{i=1}^{k} \nu_i \ln s_i^2\right]/C$$

where s_i^2 is an estimate of the variance of population i based on ν_i degrees of freedom, and ν and s^2 are given by

$$\nu = \sum_{i=1}^{k} \nu_i$$

$$s^2 = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{\nu}$$

and

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{\nu}\right]$$

Under the hypothesis that the populations all have the same variance, B has a *chi-squared distribution* with $k - 1$ degrees of freedom. Sometimes used prior to applying *analysis of variance* techniques to assess the assumption of homogeneity of variance. Of limited practical value because of its known sensitivity to non-normality, so that a significant result might be due to departures from normality rather than to different variances. See also **Box's test** and **Hartley's test.** [SMR Chapter 9.]

Baseline balance: A term used to describe, in some sense, the equality of the observed *baseline characteristics* among the groups in, say, a *clinical trial*. Conventional practice dictates that before proceeding to assess the treatment effects from the clinical outcomes, the groups must be shown to be comparable in terms of these baseline

measurements and observations, usually by carrying out appropriate significant tests. Such tests are frequently criticized by statisticians who usually prefer important prognostic variables to be identified prior to the trial and then used in an *analysis of covariance*. [SMR Chapter 15.]

Baseline characteristics: Observations and measurements collected on subjects or patients at the time of entry into a study before undergoing any treatment. The term can be applied to demographic characteristics of the subject such as sex, measurements taken prior to treatment of the same variable which is to be used as a measure of outcome, and measurements taken prior to treatment on variables thought likely to be correlated with the response variable. At first sight, these three types of baseline seem to be quite different, but from the point-of-view of many powerful approaches to analysing data, for example, *analysis of covariance*, there is no essential distinction between them. [SMR Chapter 1.]

Baseline hazard function: See **Cox's proportional hazards model.**

BASIC: Acronym for Beginners All-Purpose Symbolic Instruction Code, a programming language once widely used for writing microcomputer programs.

Basic reproduction number: A term used in the theory of infectious diseases for the number of secondary cases which one case would produce in a completely susceptible population. The number depends on the duration of the *infectious period*, the probability of infecting a susceptible individual during one contact, and the number of new susceptible individuals contacted per unit time, with the consequence that it may vary considerably for different infectious diseases and also for the same disease in different populations. [*Applied Statistics*, 2001, **50**, 251–92.]

Basu's theorem: This theorem states that if T is a complete *sufficient statistic* for a family of probability measures and V is an *ancillary statistic*, then T and V are independent. The theorem shows the connection between sufficiency, ancillarity and independence, and has led to a deeper understanding of the interrelationship between the three concepts. [*Sankhyā*, 1955, **15**, 377–80.]

Bathtub curve: The shape taken by the *hazard function* for the event of death in human beings; it is relatively high during the first year of life, decreases fairly soon to a minimum and begins to climb again sometime around 45–50. See Fig. 11. [*Technometrics*, 1980, **22**, 195–9.]

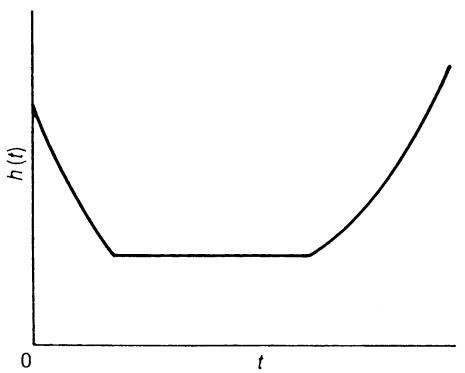


Fig. 11 Bathtub curve shown by hazard function for death in human beings.