

**Aalen model** A \*linear regression model describing the manner in which the \*expected value of a response variable,  $Z(t)$ , depends on explanatory variables  $X_1(t), \dots, X_k(t)$  with time-dependent coefficients  $\beta_0(t), \beta_1(t), \dots, \beta_k(t)$ :

$$E[Z(t)] = \beta_0(t) + \beta_1(t)X_1(t) + \dots + \beta_k(t)X_k(t).$$

**Abbe, Ernst Carl** (1840–1905; b. Eisenach, Germany; d. Jena, Germany) German mathematician and physicist. His father was a book printer and factory worker and his childhood was one of privation. Abbe studied at U Jena and U Göttingen, receiving his PhD in 1861. In 1863 he was appointed to a lectureship at Jena on the basis of a dissertation that, in effect, derived the \*chi-squared distribution. Following an approach from Carl Zeiss, most of his subsequent work was concerned with optics and astronomy. A lunar crater is named after him, also a minor planet, and several schools in Germany.

**abscissa** *See* CARTESIAN COORDINATES.

**absolute difference** The \*absolute value of the difference between two numbers. *See also* MEAN DEVIATION.

**absolute error** The \*absolute value of the difference between (commonly) an \*observation and the value predicted by, or estimated from, some \*model.

**absolute value (modulus)** The value of a number disregarding its sign. Denoted by a pair of ‘|’ signs: thus the modulus of  $-2.5$  is  $|-2.5| = 2.5$ .

**absorbing barrier (absorbing state)** *See* MARKOV PROCESS.

**acceptable quality level** *See* ACCEPTANCE SAMPLING.

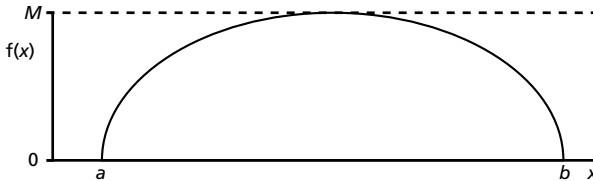
**acceptable risk** In the context of a medical treatment this describes a situation in which the expected benefits outweigh the potential hazards of the treatment.

**acceptance region** The set of values of the \*statistic, in a \*hypothesis test, which lead to acceptance of the \*null hypothesis.

**acceptance–rejection algorithm** A method for generating values of a \*continuous random variable for use in a \*simulation. Suppose that

a

the random variable  $X$ , which takes values in the interval  $(a, b)$ , has \*probability density function  $f$ . Denote the maximum value of  $f(x)$  by  $M$ . Let  $u$  and  $v$  be two random numbers \*uniformly distributed in the interval  $(0, 1)$ . Write  $r = a + (b - a)u$  and  $s = Mv$ , so that  $r$  and  $s$  are uniformly distributed on  $(a, b)$  and  $(0, M)$ , respectively. Calculate  $f(r)$ . If  $f(r) > s$  then  $r$  is accepted as a value of  $X$ . Otherwise, it is rejected and a new pair of values is taken for  $u$  and  $v$ .



**Acceptance–rejection algorithm.** Uniform random numbers are generated in the intervals  $(a, b)$  and  $(0, M)$ . If the point generated lies between the graph of  $f(x)$  and the  $x$ -axis, then the value of  $X$  is accepted.

**acceptance sampling** A method of \*quality control. A random \*sample is taken from a \*batch of output and the decision to accept or reject the batch is based on either the number of \*defectives in the sample (**inspection by attributes**) or on some summary \*statistic such as the \*sample mean (**inspection by variables**).

In the case of inspection by attributes, the \*probability of accepting a batch is a function of the proportion,  $p$ , of defectives in the batch. Any acceptance sampling scheme that does not sample 100% of a batch will lead to the occasional rejection of batches with very low proportions of defectives (the **producer's risk**), and to the occasional acceptance of batches with very high proportions of defectives (the **consumer's risk**).

As  $p$  increases, so the probability that a batch will be rejected increases. The **lot tolerance percent defective (LTPD)** is the value of  $p$  (expressed as a percentage) that the sampling scheme would expect to reject on a given proportion (usually 90%) of occasions.

The maximum proportion of defectives that is regarded as desirable by the consumer is called the **acceptable quality level (AQL)**. The graph relating the probability of acceptance to  $p$  is called the **operating characteristic curve (OC-curve)**. The average quality level of the items in the batches released after inspection is the **average outgoing quality (AOQ)**. The AOQ is usually calculated under the assumption that defective items found during testing will be replaced before the batch is released. The AOQ has minima at  $p = 0$  and  $p = 1$  and its maximum is termed the **average outgoing quality limit (AOQL)**.

Suppose that the proportion of defectives remains constant from batch to batch. Eventually a batch will be rejected. The **average run length (ARL)** is the average number of batches inspected up to and including the one that is rejected. *See* QUALITY CONTROL.

**accessible** *See* MARKOV PROCESS.

**ACF** *See* AUTOCORRELATION.

**acquiescence bias** The tendency of an interviewee to agree with the questioner. For example, if the question is 'Did you vote in the last election?' then the number replying 'Yes' is likely to provide an overestimate of the proportion who voted. On the other hand, if the question had been 'Did you abstain in the last election?', then this would also be likely to provide an overestimate.

**action line** *See* QUALITY CONTROL.

**actuarial statistics** The branch of Statistics concerned with insurance and loss, including reinsurance, \*ruin theory, and \*run-off triangles.

**adaptive sampling** A method of sequential sampling in which the later sampling procedure is affected by the earlier results. Examples include the comparison of rival treatments in \*clinical trials and the use of \*kriging in the exploration for oil.

**addition law for probabilities (total probability law)** Law stating that if two events (*see* SAMPLE SPACE)  $A$  and  $B$  are \*mutually exclusive then

$$P(A \cup B) = P(A) + P(B).$$

For example, the \*probability that, when a normal six-sided die is rolled, it shows a multiple of 3 is

$$\frac{1}{3} = \frac{1}{6} + \frac{1}{6} = P(\text{shows } 3) + P(\text{shows } 6).$$

The generalization to  $n$  mutually exclusive events is the **law of total probability**:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

*See also* UNION.

**additive model** A \*model in which the combined effect of the explanatory variables (and their \*interaction) is equal to the sum of their separate effects.

**ADF test** *See* DICKEY-FULLER TEST.

**adjacent; adjacency matrix** *See* GRAPH.

**adjusted  $R^2$**  *See* ANOVA.

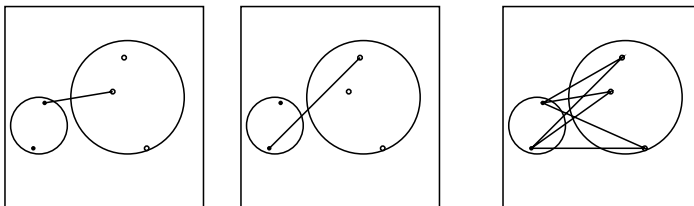
a

**admissibility** A term (introduced by \*Wald in 1939) used in \*statistical inference in several contexts. A procedure is admissible if there is no alternative procedure that performs at least as well under all circumstances and performs better under some circumstances.

**age-specific rate** When a rate, such as a \*birth rate, \*incidence rate, or \*mortality rate is calculated for individuals of a specified age (or age range) then the rate is described as being age-specific.

**agglomerative clustering methods** Methods for grouping \*multivariate data into \*clusters. Suppose there are  $n$  data items. The agglomerative clustering methods start by regarding these as  $n$  separate clusters of size 1. The two clusters judged closest together (on some criterion) are then merged to reduce the number of clusters to  $(n - 1)$ . This procedure could be continued until all the items would be collected into a single cluster.

The three simplest criteria are as follows. In **single linkage clustering** the distance between two clusters is defined as the least distance between an item in one cluster and an item in the other cluster. In **complete linkage clustering**, by contrast, the distance between two clusters is defined as the greatest distance between an item in one cluster and an item in the other cluster. As a compromise, **group-average clustering** uses the average of the distances between every member of one cluster and every member of the other cluster. The process of agglomeration is often represented using a \*dendrogram. *See also* DISTANCE MEASURE; WARD'S METHOD.



Single linkage distance    Complete linkage distance    Group distances to be averaged

**Agglomerative clustering methods.** Examples of the distance definitions used in clustering.

**AH** Abbreviation for \*alternative hypothesis.

**AI** Abbreviation for \*artificial intelligence.

**AIC (Akaike's information criterion)** Criterion, introduced by \*Akaike in 1969, for choosing between competing statistical \*models. For \*categorical data this amounts to choosing the model that minimizes  $G^2 - 2\nu$ , where  $G^2$  is the \*likelihood-ratio goodness-of-fit statistic and  $\nu$  is

the number of \*degrees of freedom associated with the model. An alternative, that usually results in the selection of a simpler model, is the **Bayesian information criterion (BIC)** for which the quantity minimized is  $G^2 - \nu \ln n$ , where  $\ln$  is the \*natural logarithm and  $n$  is the \*sample size. The latter criterion is also called the **Schwarz criterion**. A third alternative is the **Hannan–Quinn criterion** for which the quantity to be minimized is  $G^2 - 2\nu \ln(\ln n)$ . *See also* MALLOWS  $C_p$ ; STEPWISE PROCEDURES.

**AID** Abbreviation for automatic interaction detector.

**Ajne test** *See* CIRCULAR UNIFORM DISTRIBUTION.

**Akaike, Hirotugu** (1927– ; b. Fujinomiya, Japan) Japanese statistician and mathematician. On graduating from U Tokyo in 1952, Akaike joined the staff of the \*Institute of Statistical Mathematics. When he retired from the Institute, in 1994, he was its Director General.

**Akaike's information criterion** *See* AIC.

**aleatory variable** An alternative term for \*random variable.

**algorithm** A procedure consisting of a series of steps, often repetitive, for solving a problem.

**alias** *See* FACTORIAL EXPERIMENT.

**Allen, Sir Roy George Douglas** (1906–83; b. Stoke-on-Trent, England; d. London, England) English economist and statistician. He studied mathematics at Cambridge U, graduating in 1927. In 1928 he joined the faculty of the LSE where he spent his entire career and was appointed Professor of Statistics in 1944. He was President of the \*Econometric Society in 1951 and of the \*RSS in 1969, being awarded the latter's \*Guy Medal in Gold in 1978. He was knighted in 1966.

**Allgemeines Statistisches Archiv** The journal of the \*Deutsche Statistische Gesellschaft. First published in 1890, it was the journal of the Society between 1911 and 2006.

**allocation problem** *See* ASSIGNMENT PROBLEM.

**allometry** The study of the interdependence of size and shape in living organisms.

**Almon model** *See* DISTRIBUTED LAGS MODEL.

**alpha ( $\alpha$ )** The \*probability, in a \*hypothesis test, of rejecting the \*null hypothesis when it is, in fact, true. Usually called the significance level.

**alternating renewal process** A \*renewal process in which the process alternates between states  $A$  and  $B$ . Let the average length of a

a

period in state  $A$  be  $\mu_A$  and the average length of a period in state  $B$  be  $\mu_B$ . The long-run proportion of the time that the system is in state  $A$  is

$$\frac{\mu_A}{\mu_A + \mu_B}.$$

An alternating renewal process is a special case of a \*semi-Markov process. When the states are ‘working’ and ‘under repair’, the \*probability that the system is working at time  $t$  is called the **availability**.

**alternative hypothesis** The hypothesis, in a \*hypothesis test, that will be accepted if the \*null hypothesis is rejected. The term was introduced by \*Neyman and Egon \*Pearson in 1933.

**American Society for Quality (ASQ)** An organization founded in 1946 with the aim of furthering quality in all its aspects. Since 1959 it has published the journal \**Technometrics* in collaboration with the \*American Statistical Association. It awards the \*Shewhart Medal annually.



SEE WEB LINKS

- Society home page.

**American Statistical Association (ASA)** A scientific and educational society founded in 1839 in Boston, Massachusetts. It is the second oldest professional society in the United States. Its current headquarters are in Alexandria, Virginia. The Association has nearly 18 000 members. It publishes nine journals, including the \**Journal of the American Statistical Association* and the \**American Statistician*. Since 1965 the \*Wilks Medal has been presented annually for distinguished contributions to Statistics.



SEE WEB LINKS

- Association home page.

**American Statistician** A quarterly journal published by the \*American Statistical Association, concentrating on statistical methodology. It was first published in 1947.



SEE WEB LINKS

- Journal home page.

**analysis of covariance** See ANOCOVA.

**analysis of variance** See ANOVA.

**ancillary statistic** In the context of \*estimation of an unknown \*parameter, a statistic whose value provides information incidental to the estimation process. The most usual case is where the \*sample size is not fixed. For example, we might wish to know the proportion of a sweet pea mixture that has red flowers. We plant  $N$  seeds, but our estimate will be

based on the ancillary statistic  $n$ , the number of seeds that germinate and produce flowers. The term was introduced by Sir Ronald \*Fisher in 1925.

a

**ANCOVA** See ANOCOVA.

**Anderson, Theodore 'Ted' Wilbur** (1918– ; b. Minneapolis, MN) American mathematical statistician who specialized in the analysis of \*multivariate data. Anderson was a graduate of Northwestern U (1939) and Princeton U (PhD in 1945, supervised by \*Wilks). In 1946 he joined the staff at Columbia U, moving in 1967 to Stanford U. He was Editor of the \**Annals of Mathematical Statistics* 1950–2, and President of the \*IMS in 1962. He was the IMS's \*Wald Lecturer in 1982 and the \*COPSS \*Fisher Lecturer in 1985. He was the \*ASA's \*Wilks Award winner in 1988.

**Anderson–Darling test** A general test, published in 1952, that compares the fit of the observed \*cumulative distribution function with that expected. It was derived by \*Anderson and David A. Darling as a modification of the \*Cramér–von Mises test. The test statistic  $A^2$  is given by

$$A^2 = -\frac{1}{n} \sum_{j=1}^n (2j-1) [\ln\{F(x_{(j)})\} + \ln\{1 - F(x_{(j)})\}] - n,$$

where  $F$  is the hypothesized cumulative distribution function,  $n$  is the \*sample size, and  $x_{(j)}$  is the  $j$ th ordered \*observation ( $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ ). The statistic can also be used to test for \*normal and \*exponential distributions with unknown \*parameters estimated by their sample equivalents. In some cases, as shown in the following table, an adjusted test statistic is required.

		TEST STATISTIC	UPPER TAIL PROBABILITY			
			0.10	0.05	0.025	0.01
Specified distribution	$A^2$	1.933	2.492	3.070	3.857	
Normal, estimated mean ( $n > 20$ )	$A^2$	0.894	1.087	1.285	1.551	
Normal, estimated variance ( $n > 20$ )	$A^2$	1.743	2.308	2.898	3.702	
Normal, estimated mean and variance	$A^2(1 + \frac{3}{4n} + \frac{9}{4n^2})$	0.631	0.752	0.873	1.035	
Exponential, estimated mean	$A^2(1 + \frac{3}{10n})$	1.062	1.321	1.591	1.959	

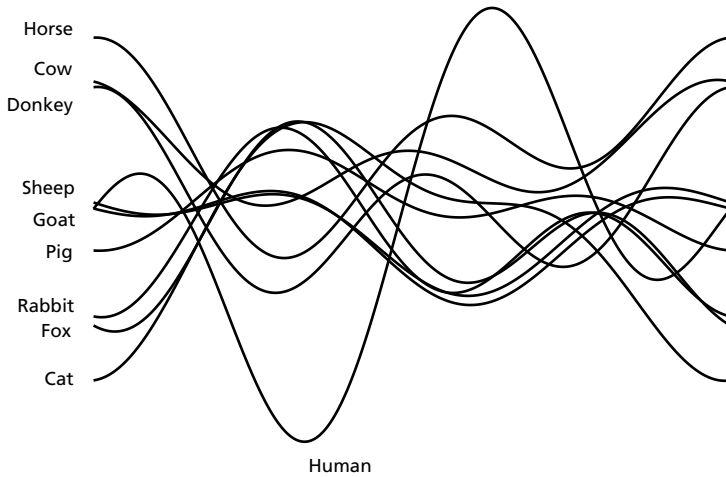
**Andrews plot** A plot suggested in 1972 by David Andrews as an alternative method to \*Chernoff faces for representing \*multivariate data

a

in two dimensions. The plot can help to identify \*outliers and to establish similarities within groups of data items. For an item with values given by the row \*vector  $(a\ b\ c\ d\ e\ \dots)$ , the function  $x(t)$  is defined as

$$x(t) = \frac{a}{\sqrt{2}} + b \sin(t) + c \cos(t) + d \sin(2t) + e \cos(2t) + \dots$$

The resulting graph is drawn for  $(-\pi \leq t \leq \pi)$ , for each data item. The form of the graph is dependent upon the ordering of the characteristics  $a, b, \dots$ ; the usual advice is to order the characteristics in declining order of their (supposed) importance.



**Andrews plot.** This plot compares humans with nine familiar animal species, using five characteristics (body weight, brain weight, hours of sleep, lifespan, and gestation). It appears that humankind is a race apart and that it is difficult to tell the sheep from the goats.

**angular histogram** *See* CIRCULAR HISTOGRAM.

**angular uniform distribution** *See* CIRCULAR UNIFORM DISTRIBUTION.

**anisotropic; anisotropy** *See* ISOTROPY.

***Annals of Applied Probability*** A quarterly publication of the \*Institute of Mathematical Statistics. It was first published in 1973.

 **SEE WEB LINKS**

- Journal home page.

***Annals of Applied Statistics*** A quarterly publication of the \*Institute of Mathematical Statistics. It was first published in 2007.



- Journal home page.

***Annals of Mathematical Statistics*** The single journal of the \*Institute of Mathematical Statistics from 1930 to 1973, before its subdivision into three parts.

***Annals of Probability*** A bi-monthly publication of the \*Institute of Mathematical Statistics. It was first published in 1973.



- Journal home page.

***Annals of Statistics*** A bi-monthly publication of the \*Institute of Mathematical Statistics. It was first published in 1973.



- Journal home page.

***Annals of the Institute of Statistical Mathematics*** A quarterly English language publication of this Japanese institute.



- Journal home page.

**ANOCOVA (ANCOVA; analysis of covariance)** \*ANOVA with a mixture of \*continuous random variables and \*qualitative variables. ANOCOVA \*models can also be thought of as \*multiple regression with some \*dummy variables.

**ANOVA (analysis of variance)** The attribution of variation in a \*variable to variations in one or more explanatory variables. The term was introduced by Sir Ronald \*Fisher in 1918.

A measure of the total variability in a set of \*data is given by the sum of squared differences of the \*observations from their overall \*mean. This is the **total sum of squares** (*TSS*). It is often possible to subdivide this quantity into components that are identified with different causes of variation. The full subdivision is usually set out in an **analysis of variance** table, as suggested by Sir Ronald \*Fisher in his 1925 book *Statistical Methods for Research Workers*. Each row of the table is concerned with one or more of the components of the observed variation. The entries on a row usually include the **sum of squares** (*SS*), the corresponding number of \*degrees of freedom ( $\nu$ ), and their ratio, the **mean square** ( $= SS/\nu$ ).

After the contributions of all the specified sources of variation have been determined, the remainder, often called the **residual sum of squares** (*RSS*) or **error sum of squares**, is attributed to \*random variation. The mean square corresponding to *RSS* is often used as the yardstick for assessing the importance of the specified sources of variation. One

a

method involves comparing ratios of mean squares with the \*critical values of an \* $F$ -distribution.

The proportion of variation explained by the model is

$$R^2 = 1 - \frac{RSS}{TSS},$$

which is sometimes called the \*coefficient of determination.

In an ANOVA analysis each explanatory variable takes one of a small number of values. If, instead, some explanatory variables are \*continuous in nature, then the resulting models are called \*ANOCOVA models. ANOVA can also be thought of as \*multiple regression using only \*dummy variables.

As an example, suppose that four varieties of tomatoes are grown in three grow-bags giving the yields (in g) shown below. The explanatory variables are the grow-bags and the varieties.

T <sub>1</sub>	T <sub>3</sub>	T <sub>4</sub>	T <sub>2</sub>	T <sub>2</sub>	T <sub>1</sub>	T <sub>3</sub>	T <sub>4</sub>	T <sub>3</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>4</sub>
1890	1740	1620	1970	1850	1760	1800	1890	1810	1910	1100	1710
Bag 1				Bag 2				Bag 3			

The following ANOVA table results:

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE
Differences between grow-bags	2	12 950	6 475
Differences between varieties	3	93 425	31 142
Residual	6	72 250	12 042
Total	11	178 625	

Since the mean square for varieties is much greater than that for grow-bags we can conclude that differences between varieties are more important. However, the residual sum of squares amounts to nearly half the total sum of squares, indicating that there are major unexplained sources of variation.

The proportion of the total sum of squares that is attributed to any particular source of variation is referred to as **eta-squared** ( $\eta^2$ ). Thus, the difference between grow-bags has  $\eta^2 = \frac{12\,950}{178\,625} = 0.072$ . A related statistic is **partial eta-squared**, which is the ratio of the sum of squares of interest to the total of that sum of squares and the residual sum of squares. Thus, for grow-bags, the partial eta-squared is  $\frac{12\,950}{12\,950+72\,250} = 0.152$ .