



MAD– Acronym for *mean, median or mode absolute deviation*.

Mahalanobis D^2 – A measure of distance involving **multivariate data** useful in discriminating between two **populations**. It was proposed by P. C. Mahalanobis to assess the divergence between two populations based on **observations** on p characters or **variates**. The square of the distance (D^2) can be expressed as the **Euclidean distance** squared. It is related to **Fisher's discriminant function** and **Hotelling's T^2** . It has found extensive applications in many fields including **cluster analysis, profile analysis, and discriminant analysis**.

Mahalanobis generalized distance– Same as *Mahalanobis D^2* .

main effect– In an **analysis of variance** or **regression** involving two or more **factors**, where each factor may have a separate **effect**, the main effect is an **estimate** of the effect of an experimental **variable** or **treatment** on the **dependent variable** that is separable from the other factors' effect and from the **interaction effect**. In a **factorial experiment**, the main effect of a factor is the **average** change in response produced by changing the **levels** of the factor.

mainframe– A high-speed digital computer with very large capacity. Originally, the term was employed to refer to the main framework of a central processing unit (CPU) on which the arithmetic unit and associated logic circuits were mounted.

Mallow's C_p statistic– A diagnostic index used in **regression analysis** in the selection of the "best" set of **predictor variables**. The index is defined as

$$C_p = \sum_{i=1}^n (y_i - \hat{y}_{i(p)})^2 / s_e^2 - n + 2p$$

where y_i is the i th observed value of the **dependent variable**, $\hat{y}_{i(p)}$ is the predicted value based on a particular subset of p **explanatory variables**, s_p^2 is the full regression **residual mean square**, and n is the number of **observations**. A **model** with the smallest value of the C_p statistic is considered to provide the best fit.

Malthusian theory– The theory that the population tends to increase faster than the natural resources needed to sustain it. More specifically, the theory states that the population grows at a **geometric progression** while the food supply increases only in **arithmetic progression**.

manifest variable– A term used to describe an **observed variable** that can be measured in contrast to a **latent variable**, which cannot be measured directly. For example, intelligence is a latent variable that cannot be measured directly. But it can be measured in terms of a manifest variable such as an IQ test score. A manifest variable is also called an indicator variable.

Mann–Whitney *U* test– A **nonparametric test** for detecting differences between two **location parameters** based on the analysis of two **independent samples**. The **test statistic** is formed by counting all the bivariate pairs from the two **samples** in which one sample value is smaller than the other. It is equivalent to the **Wilcoxon rank-sum test**. The procedure is used for comparing two independent samples of **scores** that cannot be compared by means of a **two-sample *t* test** either because the scores are ordinal in nature or the **normality** or **homogeneity of variance** assumptions cannot be satisfied. See also **normal scores test**.

Mann–Whitney–Wilcoxon test– Same as *Mann–Whitney *U* test* or *Wilcoxon rank-sum test*.

MANOVA– Acronym for *multivariate analysis of variance*.

Mantel–Haenszel chi-square test– A summary **chi-square test** involving two or more **two-by-two contingency tables**. It is used for stratified **data** involving several 2×2 **tables** with a view to adjust or control for **confounding**. After stratifying the data by the categories of the **confounding variable**, such as age, sex, occupation, etc., the results are pooled together to produce a single summary test based on **chi-square distribution** with one **degree of freedom**.

Mantel–Haenszel estimator– In a **stratified analysis** involving a series of 2×2 **tables**, an **estimator** of the common **odds ratio** that may be derived from matched and unmatched **data sets**. The estimator is a type of **weighted average** of the odds ratio estimators from each individual table where the weights are inversely proportional to the **variances** of the individual **estimates**. Thus, estimates with smaller variance (higher precision) are given more weight, whereas those with larger variance (lower precision) are given less weight. It is calculated by the formula

$$\sum_{i=1}^k a_i d_i / \sum_{i=1}^k b_i c_i$$

where a_i , b_i , c_i , and d_i are the four **cell counts** in the i th table and k is the number of 2×2 tables. It produces an adjusted estimate of the overall odds ratio and provides a method of controlling **confounding** by stratifying a **sample** into a series of strata that are homogeneous with respect to the **confounding variable**. Two common applications of the Mantel–Haenszel estimate are the analysis of **case-control studies** and **meta-analysis**. See also *Peto's method*.

Mardia's test– A statistical procedure for testing the **normality** of a **multivariate data set**.

marginal density function– The **probability density function** of one of the (continuous) **random variables** of a set of jointly distributed (continuous) random variables. It is obtained by integrating the **joint density function** with respect to other random variables.

marginal distribution– The **probability distribution** of one of the **random variables** of a set of jointly distributed random variables obtained from a **joint distribution** by summing out, or integrating out, all the other variables.

marginal frequencies (probabilities)– The sum of the **frequencies (probabilities)** in one of the rows or in one of the columns of a two-way table. The marginal frequencies (probabilities) are usually shown at the margins of the table.

marginal frequency (probability) distribution– See *marginal distribution*.

marginally significant– A term used to refer to **statistical significance** of research results that barely reach the **critical value** needed to be **statistically significant**.

marginal probability function– **Probability function** of one of the (discrete) **random variables** of a set of jointly distributed (discrete) random variables. It is obtained by summing the **joint probability function** with respect to other random variables.

marginals– A clipped form for **marginal frequencies** or totals.

marginal totals– Same as *marginal frequencies*.

Markov chain– Same as *Markov process*.

Markov inequality– If a **random variable** X with **mean** μ and finite **variance** can take only positive values, then the Markov inequality states that $P(X \leq x) \leq 1 - \mu/x$.

Markov process– A **discrete stochastic process** in which, in a series of **trials**, the **probability** of an **event** depends upon the results of the event immediately preceding it. Thus, the state of the process is unaffected by the past, except the immediate past.

masking – Same as *blinding*.

matched case-control study– A **case-control study** in which **cases** and **controls** are matched on certain characteristics known to be associated to both disease and the **risk factor**. Some examples of commonly used **matching** variables are age, sex, occupation, and socioeconomic status.

matched groups– Same as *matched samples*.

matched-groups t test– Same as *paired t test*.

matched pairs– See *matched-pair samples*.

matched-pair samples– Two **samples** taken such that each **experimental unit** in one group has been matched with a unit from another group. In matched-pair samples any **sample observation** about a unit in one group automatically yields an associated **observation** about a unit in another group.

matched-pairs t test– Same as *paired t test*.

matched samples– **Samples** where two or more groups of subjects are matched or paired according to one or more relevant **variables** such as age, sex, or sociodemographics. See also *matched-pair samples*.

matched-samples t test– Same as *paired t test*.

matched set– In a **case-control study**, a form of **matching** in which a number of **controls**, known as a matched set, are matched to each **case**. This form of matching is normally used to increase the **sensitivity** of the design, especially when controls are more economical.

matched-subjects designs– These are **experimental designs** that test two or more groups of subjects, matched according to one or more relevant **variables**. Studies involving identical twins are the ideal examples of such designs. The **scores** for each pair or set of subjects are treated as correlated measures. See also *matched-pair samples*, *matched samples*.

matching– The process of making two groups of subjects or **experimental units** homogeneous on possible **confounding factors** by matching them according to relevant factors causing **confounding**. Matching can be individual matching, in which study and comparison subjects are paired on the basis of matching variables, or frequency matching, in which the **frequency distribution** of matched variables is similar in study and comparison groups. It is usually done prior to **randomization in clinical trials**. See also *matched-pair samples*, *matched samples*, *matched-subjects designs*.

maternal death rate– A measure of **risk** of dying from causes associated with child birth. It is obtained as the number of deaths actually observed due to puerperal causes during a calendar year divided by the total number of births (live + still) (expressed per 100 or 1000).

maternal mortality rate– Same as *maternal death rate*.

mathematical expectation– Same as *expected value*.

mathematical model– A mathematical equation used in a **mathematical modeling**.

mathematical modeling– A term used to describe a mathematical formulation that characterizes the behavior of one or more **variables** that may influence some natural phenomenon or causal system.

matrix– A **rectangular array** of numbers (called elements) or mathematical objects arranged into rows and columns. Matrices are denoted by capital Roman letters **A**, **B**, **C**, etc. Two examples of matrices are

$$\mathbf{A} = \begin{pmatrix} 3 & 5 & 9 \\ 4 & 6 & 2 \\ 2 & 8 & 3 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_{11} & \cdot & \cdot & \cdot & b_{1n} \\ b_{21} & \cdot & \cdot & \cdot & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{m1} & \cdot & \cdot & \cdot & b_{mn} \end{pmatrix}$$

matrix algebra– A system of algebra in which basic elements and symbols for unknown quantities including arithmetical operations are presented in terms of matrix notation.

matrix of correlation– Same as *correlation matrix*.

maximax criterion– One of several nonprobabilistic criteria for making an optimal decision under **uncertainty**. According to this criterion, a decision maker determines the maximum benefit associated with each possible action, searches for the maximum among these maxima, and then chooses the action associated with this maximum of maxima.

maximin criterion– One of several nonprobabilistic criteria for making an optimal decision under **uncertainty**. According to this criterion, a decision maker determines the minimum benefit associated with each possible action, searches for the maximum among these minima, and chooses the action associated with this maximum of minima.

maximum *F*-ratio test– Same as *Hartley's test*.

maximum likelihood criterion– One of several probabilistic criteria for making an optimal decision under **uncertainty**. The maximum likelihood criterion is based on the assumption that the most likely factor or factors have generated the most probable **sample**. It attaches the greatest **probability** to the observed **event** and the degree of **reliability** of such values. According to this criterion, a decision maker identifies the event most likely to occur and selects the action that produces the maximum benefit associated with this most likely event.

maximum likelihood estimation– A method of **estimation** of one or more **parameters** of a **population** by maximizing the **likelihood** or **log-likelihood function** of the **sample** with respect to the parameter(s). The maximum likelihood estimators are functions of the **sample observations** that make the **likelihood function** greatest. The procedure consists of computing the **probability** that the particular **sample statistic** would have occurred if it were the true value of the parameter. Then for the **estimate**, we select the particular value for which the probability of the actual observed value is greatest. Maximum likelihood estimates are determined by using methods of calculus for maximization and minimization of a function. These estimates possess many desirable properties such as **consistency**, **asymptotic normality**, and **asymptotic efficiency**.

maximum likelihood estimate/estimator– See *maximum likelihood estimation*.

maximum likelihood method– See *maximum likelihood estimation*.

maximum likelihood principle– See *maximum likelihood estimation*.

maximum likelihood procedure– See *maximum likelihood estimation*.

maximum tolerance dose– The highest level of dose of a drug that a patient can tolerate with an acceptable level of toxicity. This is especially important in cytotoxic therapy of cancer where the treatment generally produces some serious side effects.

McNemar's chi-square test– Same as *McNemar's test*.

McNemar's test– A **nonparametric test** for comparing two correlated **proportions** arising from two **dependent** or **paired groups**. It is calculated by the formula $X^2 = (b - c)^2 / (b + c)$, where b is the number of pairs in which the individual from group A has positive result and the individual from group B does not; and c is the number of pairs for which it is just the reverse. Under the **null hypothesis** that the **probability** of positive response is the same in two groups, X^2 has a **chi-square distribution** with one **degree of freedom**. It is a special case of the **Mantel-Haenszel chi-square test** for a single 2×2 table.

mean– A **measure of location** or the **central tendency** of a **data set**. It is the arithmetic **average** computed by summing all the values in the **data set** and dividing the sum by the number of **data values**. Given x_1, x_2, \dots, x_n , a set of n numbers, it is defined by the formula $\bar{x} = \sum_{i=1}^n x_i / n$. It is the most stable and useful measure of central tendency. For a data set with values 7, 8, 8, 9, 12, 13, the mean is $\bar{x} = (7 + 8 + 8 + 9 + 12 + 13) / 6 = 9.5$. The physical interpretation of the mean is illustrated in the figure below, where it is the value on

the **horizontal axis** that serves as a balance point. When used without any qualification, mean refers to **arithmetic mean**. It is the most widely used and best understood data summary in all **statistics**. Two other means used in statistics are **geometric mean** and **harmonic mean**. The mean is a reliable measure of location if the underlying data set has a **symmetrical distribution**. If the **distribution** in question is skewed, mean does not provide a useful measure, since it is greatly influenced by **extreme observations**. See also *population mean*, *sample mean*.

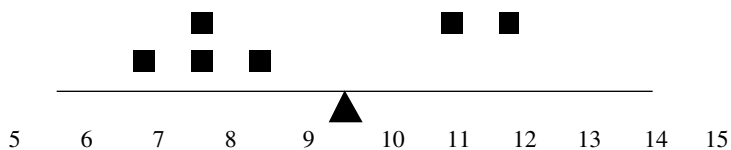


Figure showing the mean as a balance point

mean absolute deviation— See *average absolute deviation*.

mean absolute error— Same as *average absolute deviation*.

mean deviation— See *average absolute deviation*.

mean error— Same as *average absolute deviation*.

mean of squared deviations— Same as *mean square deviation*.

mean square— In an **analysis of variance**, the **sum of squares** divided by its corresponding **degrees of freedom**. This quantity is used in the *F* ratio to determine if there exist significant differences in **population means**.

mean square between— Same as *mean square between groups*.

mean square between (among) groups— In a **one-way analysis of variance** design, the measure of **variation** between **group means** obtained by dividing the **sum of squares between groups** by its **degrees of freedom**.

mean square contingency coefficient— See *phi coefficient*.

mean square deviation— The square of the **deviation** of a value of a **data set** from the **mean**. The concept is used extensively in many statistical applications, including **correlation**, **variance**, and **least squares regression**.

mean square error— A measure of **error** of an **estimator** defined as the **expected value** of the squared difference between the estimator and the true value of the **parameter**. For an **unbiased estimator**, mean square error equals the **variance**; for a **biased estimator**, it is equal to the variance plus **bias** square. The square root of the mean square error is referred to as the root mean square error.

mean square for columns— In a **two-way analysis of variance** design, the measure of the differences between columns **means** obtained by dividing the **sum of squares for columns** by its **degrees of freedom**.

mean square for error— In a **one-, two-, or multiway analysis of variance** design, the measure of the **variance** due to individual differences between subjects, **measurement**

errors, uncontrolled **variations** in experimental procedures, and so on. It is obtained by dividing the **error sum of squares** by the corresponding **degrees of freedom**.

mean square for interaction– In a **two-** or **multiway analysis of variance** design, the measure of the **interaction** between any two **treatment factors** obtained by dividing the **sum of squares for interaction** by its **degrees of freedom**.

mean square for regression– Same as *regression mean square*.

mean square for rows– In a **two-way analysis of variance** design, the measure of the differences between row **means**, obtained by dividing the **sum of squares for rows** by its **degrees of freedom**.

mean square for treatment– In an **analysis of variance**, an estimate of the **population variance**, based on the observed **variation** among the **treatment groups**. It is obtained by dividing the **treatment sum of squares** by the corresponding **degrees of freedom**.

mean square ratio– In an **analysis of variance**, the **ratio** of two **mean squares**. See also *F statistic*.

mean square within– Same as *mean square within groups*.

mean square within groups– In a **one-way analysis of variance** design, the measure of **variation** obtained by dividing the **within group sum of squares** by its **degrees of freedom**. It is a measure of the **deviations** of the individual **observations** from their respective **group means**.

mean variation– Same as *average absolute deviation*.

mean vector– In a **data set** comprising **multivariate observations**, it is the **vector** containing the **mean** value of each **variable**. It is a multivariate analogue of the mean of a **univariate data set**.

measurement– The process of assigning a label, number, or numerical value to characteristics that are being observed, according to a set of rules.

measurement class– Same as *measurement interval*.

measurement errors– **Errors** in reading, calculating, or recording a value caused by flaws in the measuring instruments, such as faulty calibration, or the experimenter making the **observations**, as contrasted with other errors, or unknown **variation**.

measurement interval– A range of values assumed by a **variable** into which **observations** can be grouped.

measurement scale– Same as *scale of measurement*.

measure of association– Any numerical measure that shows the degree of relationship between two **variables**. More precisely, it is a numerical index of the strength of the statistical dependence of two or more **qualitative variables**. A measure of association is usually a **statistic** that shows direction and magnitude of the relationship. Examples of measures of association include **coefficient of correlation**, **lambda**, **gamma**, and **odds ratio**, among others. See also *asymmetric measure of association*, *symmetric measure of association*.

measure of risk– Any of various **measures of association**, such as **risk difference**, **risk ratio**, and **odds ratio**, used to measure **association** between a **risk factor** and the disease or condition of interest.

measures of central tendency– **Summary indices** or **statistics** describing the central or middle point, or the most typical value, of a set of **measurements** around which **observations** tend to cluster. They are also frequently referred to as **average** values. See also *mean, median, mode*.

measures of dispersion– **Summary indices** or **statistics** that describe the **scatter** or **spread** of **observations** about the **central location**. They show the extent to which individual values in a **data set** differ from one another and, hence, differ from their central location. See also *range, standard deviation, variance*.

measures of location– Same as *measures of central tendency*.

measures of shape– Indices or numbers that indicate either the degree of **asymmetry** or the peakedness in a **frequency distribution**. The term is used in contrast to measures of **skewness** and **kurtosis**.

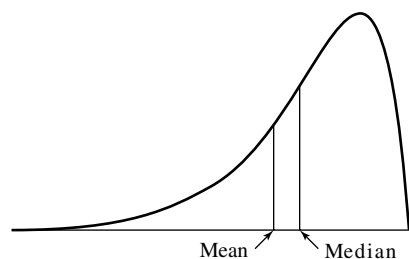
measures of spread– Same as *measures of dispersion*.

measures of variability– Same as *measures of dispersion*.

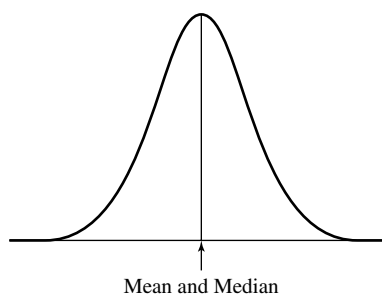
measures of variation– Same as *measures of dispersion*.

median– A **measure of location** or **central tendency** of a **data set**. It is the value that divides the data set into two equal groups; one with values greater than or equal to the median, and the other with values less than or equal to the median. It is an ordinal measure of central tendency. It is the middle value in a data set which divides a **distribution** exactly in half so that 50 percent of its **scores** are higher than it and 50 percent are lower. Thus, the median is also referred as the 50th **percentile**. In a **frequency distribution**, median is calculated by first ascertaining the **class interval** within which it is located, and then finding its value within this class interval by **interpolation**. For a right-skewed distribution, **mean** is larger than the median; for a left-skewed distribution, mean is smaller than the median; and for a **symmetric distribution**, mean and median are equal. The median is one of several types of **averages** currently in use; and its principal advantage is that it is not unduly influenced by **extreme observations**. It is often used in describing the typical income of a group of individuals. The name “median” was first used by Francis Galton in 1883. See also *population median, sample median*.

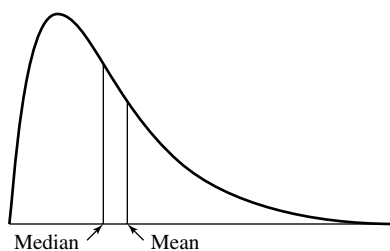
median absolute deviation– See *average absolute deviation*.



The median and mean of a left-skewed distribution



The median and mean of a symmetric distribution



The median and mean of a right-skewed distribution

median class– In a **grouped frequency distribution**, the **class interval** that contains the **median**.

median deviation– The **median** of the **absolute values** of the **deviations** about some **measure of central tendency**. It is also called median error and sometimes, improperly, **probable error**.

median effective concentration– Same as *median lethal dose*.

median effective dose– Same as *median lethal dose*.

median error– Same as *median deviation*.

median lethal concentration– Same as *median lethal dose*.

median lethal dose– In **biological assay** involving a toxic substance, the amount of stimulus or quantity of a dose that will result in a desired response (say mortality) in 50% of the subjects in the **population** under study during a specified period of time. It is denoted by LD50 for lethal dose, ED50 for effective dose, LC50 for lethal concentration, EC50 for effective concentration, and TIm 50 for tolerance limit.

median test– A **nonparametric test** performed to test the **hypothesis** that two **populations** have the same **median**.

median tolerance limit– Same as *median lethal dose*.

median unbiased estimator– An **estimator** is said to be median unbiased if its **median** equals the true value of the **parameter** being estimated. See also *unbiased estimator*.

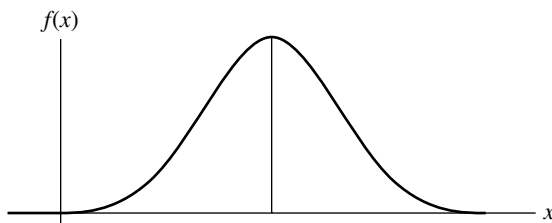
median unbiasedness– The term is used to indicate the property of a **median unbiased estimator**.

medical decision making– The application of **decision analysis** in making diagnostic and/or treatment inferences in clinical medicine. It synthesizes all the accumulated evidence and other relevant information concerning diagnostic and/or treatment alternatives and associated **risks**, consequences of a particular diagnosis or treatment, and **uncertainties** in making decisions about diagnoses or treatments. Its aim is to assist the physician in making the correct diagnosis and choosing the appropriate therapy.

medical record– A file of information containing cumulative narrative history of a patient, the treatment given, final diagnosis, and continuing care following release. The full range of **data** in a medical record includes a variety of other clinical, sociodemographic, economic, administrative and behavioral information.

medical statistics— Statistical methods and techniques applied to the study of medical and health-related problems. In the United States the term is synonymous with **biostatistics**.

mesokurtic— A **frequency distribution** or **curve** is said to be mesokurtic when it exhibits a moderate clustering of **scores** around the **mean** as does the **normal curve**, which by definition, is mesokurtic. See also *leptokurtic*, *platykurtic*.

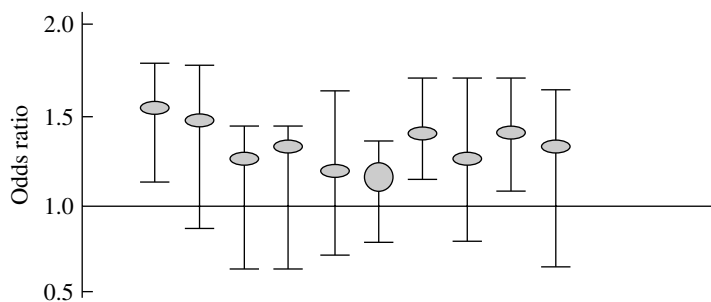


A mesokurtic distribution ($\beta_2 = 3$)

mesokurtic curve— See *mesokurtic*.

mesokurtic distribution— See *mesokurtic*.

meta-analysis— The process of using statistical methods for combining or summarizing the results from several independent studies of the same **outcome** so that an overall **effect size** and **p value** may be determined. Meta-analysis is frequently used in pooling results from several smaller studies, none of them large enough to show **statistically significant** differences, but the pooling increases the **power** of the study. The pooling is usually done by taking a **weighted average** of the individual results according to their study size. It uses methods such as the **Mantel-Haenszel estimator** and **Peto's method** to calculate the combined **estimate**. The technique is particularly popular among researchers interested in summarizing results from **randomized controlled trials** of therapies or interventions. However, it is also being used in many epidemiological studies involving **risk factors** or **diagnostic tests**. Meta-analysis suffers from several **biases** and limitations. Some of the controversies surrounding meta-analysis include **publication bias**, **heterogeneity of effect size**, use of individual or aggregated **data**, and choice of **fixed** or **random effects models**.



Meta-analysis of 9 hypothetical randomized clinical trials: observed odds ratio and 95% confidence limits. The overall odds ratio is shown by a circle.

method of least squares– Same as *least squares estimation*.

method of maximum likelihood– Same as *maximum likelihood estimation*.

method of moment estimation– Same as *method of moments*.

method of moments– A method of **estimation of parameters** by equating the **sample moments** to their respective population values. It is the oldest general method for estimating unknown parameters, and was proposed by Karl Pearson about 1891. It is generally applicable and provides a fairly simple method for obtaining **estimates** in most cases. The method, however, yields **estimators** that, in certain cases, are less efficient than those obtained by the **method of maximum likelihood**. See also *least squares estimation*.

Michael's test– A test of **normality** based on **order statistics** from **sample data**. See also *Anseron–Darling test*, *Cramér–von Mises test*, *D'Agostino test*, *Shapiro–Francia test*, *Shapiro–Wilk W test*.

midpoint– The value located halfway between the **lower** and **upper real limits** of an interval. It is obtained as the **mean** of the lower and upper real limits of an interval. In plotting **grouped data**, the **midpoints** are used to represent the **observations** within each interval.

mid- p value– A modification of the conventional **p value** that is used in some analyses involving a **test statistic** based on a **discrete distribution**. Let T denote a test statistic based on a discrete distribution and t be the observed number of **outcomes**. Then, the mid- p value is defined as

$$\text{mid } p = \frac{1}{2}P(T = t) + P(T \geq t + 1)$$

while the conventional p value is determined as $p = P(T \geq t)$. In other words, mid p averages the exact p value for the observed number of outcomes t and $t + 1$.

midrange– The **mean** of the smallest and largest values in a **data set**. Given a set of values x_1, x_2, \dots, x_n arranged in ascending or descending order of magnitude, the midrange is defined as $(x_1 + x_n)/2$. It provides a crude **estimate** of the center of a **symmetrical distribution**.

midvalue– Same as *midpoint*.

minimax criterion– In **decision** or **game theory**, one of several nonprobabilistic criteria for making an optimal decision under **uncertainty**. According to this criterion, a decision maker determines the maximum cost associated with each possible action, searches the minimum of these maxima, and chooses the action associated with this minimum of maxima.

minimax regret criterion– In **decision** or **game theory**, one of several nonprobabilistic criteria for making an optimal decision under **uncertainty**. According to this criterion, a decision maker finds the maximum regret value associated with each possible action, searches the minimum among these maxima, and chooses the action associated with this minimum of maxima.

minimax strategy– Same as *minimax criterion*.

minimin criterion– In **decision** or **game theory**, one of several nonprobabilistic criteria for making an optimal decision under **uncertainty**. According to this criterion, a decision

maker who seeks to minimize some cost or loss determines the minimum cost associated with each possible action, searches the minimum among these minima, and chooses the action associated with this minimum of minima.

minimum chi-square estimation– A method of **estimation** in which an **estimate** of a **parameter** is determined by minimizing a **chi-square statistic**. The procedure involves determining the values of the parameters so as to minimize X^2 calculated from **observed frequencies** and **expected frequencies** expressed in terms of the parameters. The minimum chi-square estimators are asymptotically equivalent to the **maximum likelihood estimators**.

minimum effective dose– The lowest level of dose of a drug that can produce the desired clinical effect in a patient.

MINITAB– A general-purpose **statistical software package** designed to perform interactive **data analysis**. The package is very easy to use and proved to be very popular with both students and instructors. It includes a wide variety of methods for statistical and graphical analysis. It is based on a two-dimensional spreadsheet concept in which columns are **variables** and rows are cases.

Minkowski distance– A generalized measure of the distance between two points as determined by the location of their coordinates. It includes **Euclidean distance** as a special case.

missing data– Same as *missing values*.

missing values– **Observations** missing from a **data set** for a variety of reasons. For example, information may not be available because a subject may drop out of the study or may fail to answer one of the questions in a **survey**, or certain measuring instrument may break down, or animals and plants may die during the course of the **experiment**. The presence of missing values greatly complicates the methods of analysis. Several approaches for analyzing **data** containing missing values have been developed, but none of them seem to be entirely satisfactory. See also *imputation*.

mixed data– **Data** containing a mixture of **continuous** and **discrete data**.

mixed effects model– An **analysis of variance** model in which at least one **treatment level** is fixed and at least one treatment level is **random**, excluding the **residual term** which is always considered random. It is also called Model III. See also *fixed effects model*, *random effects model*.

mixed model– Same as *mixed effects model*.

mixed time-series model– A **time-series model** that is a mixture of **additive** and **multiplicative time-series models**; for example, $Y = T \times C \times I + S$.

MLE– Acronym for *maximum likelihood estimation*.

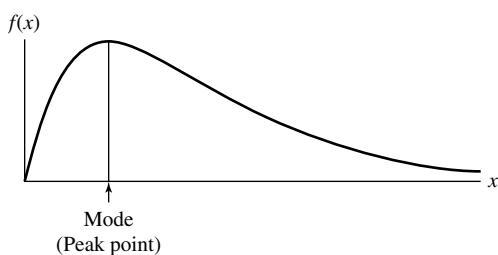
modal class– The **class interval** (generally from a **frequency table** or **histogram**) that contains the highest **frequency** of **observations**. See also *mode*.

modal group– Same as *modal class*.

modal interval– Same as *modal class*.

modal range– Same as *modal class*.

mode– A measure of central tendency or location of a data set. It is defined as the **data value** that occurs most frequently. When **grouped data** are involved, the **class interval** having the highest **frequency** is called the **modal class**. Its **midpoint** is often used to represent the mode. More precisely, it can be calculated by first ascertaining the class interval within which it is located, and then finding its value within this interval by **interpolation**. In a **frequency distribution** involving a **categorical variable**, the name of the category of **scores** that has the highest frequency is referred to as mode. It is the most primitive measure of central tendency. A set of **data** can have more than one mode or no mode when all values are different. Like the **median**, the mode is not influenced by unusually high or low values, but it is used less frequently in statistical analysis than either the median or the **mean**.



Mode of a continuous distribution

mode absolute deviation– See *average absolute deviation*.

model– A construct or formulation that provides a description of the assumed structure of a set of **data**. A model involves a set of **assumptions** about relationships used to describe the data structure in a manner that may aid in understanding the process assumed to have generated the data. See also *deterministic model*, *mathematical model*, *mathematical modeling*, *probability model*, *stochastic model*.

model building– A procedure for finding the simplest **model** that provides an adequate description of the **data**.

model equation– Mathematical equation used in a **model**.

Model I– Same as *fixed effects model*.

Model II– Same as *random effects model*.

Model III– Same as *mixed effects model*.

model misspecification– The use of an incorrect **model** to fit a given set of **data**.

moment generating function– A function of a **variable** t associated with the **probability distribution** of a **random variable** X , and defined by

$$M_x(t) = E(e^{tX}) \quad \text{for } -h < t < h$$

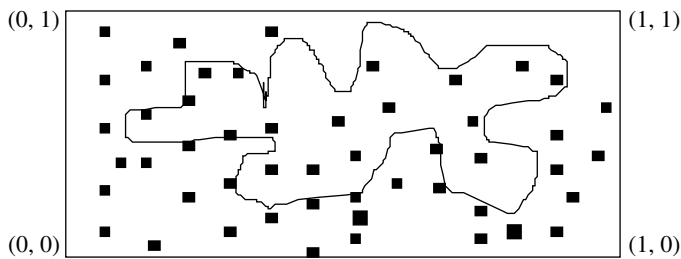
If $M_x(t)$ is expanded as a power series in t , the coefficient of $t^k/k!$ gives the k th **moment** of X about the origin. See also *characteristic function*, *probability generating function*.

moments— Values used to characterize the **probability distribution** of a **random variable** or describe a set of **data**. For a random variable X , its k th moment about the origin is defined as $\mu'_k = E(X^k)$, so that μ'_1 is simply the **mean** of the distribution and is commonly denoted by μ . The k th moment about the mean is defined as $\mu_k = E(X - \mu)^k$, so that μ_2 is the **variance** of the distribution and is commonly denoted by σ^2 . For a set of **sample observations** x_1, x_2, \dots, x_n , the k th moment about the origin is defined as $m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k$, so that m'_1 is simply the **sample mean** and is commonly denoted by \bar{x} . The k th sample moment about the mean is defined as $m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$, so that m_2 is the **sample variance** and is commonly denoted by s^2 . The k th moment about the mean is also known as the k th central moment.

moments about the origin— See *moments*.

monitoring— A term used in a **clinical trial** to describe the **follow-up** and **observation** of the conduct and progress of an ongoing trial according to a set of predefined guidelines contained in the **protocol**.

Monte Carlo method— A term that has most commonly been used in the solution of any mathematical and statistical problem by performing **sampling** experiments involving generation of **random numbers** from a given **probability distribution**. It provides an **empirical** method of finding solutions to many mathematical and statistical problems for which no simple analytical solutions are available. For example, suppose we want to find the area of the closed curve of an irregular shape contained within a unit square as shown below. It is evident that the area in question is rather complicated and there does not seem to be a simple method for determining it. Now, suppose a pair of random numbers (x, y) , such that $0 \leq x \leq 1$, $0 \leq y \leq 1$, is selected and the point (x, y) is plotted within the unit square. The process is continued a large number of times, say N , and let n be the number of points that have fallen within the closed curve. Then by a famous theorem in **probability theory**, called the **law of large numbers**, it follows that the ratio n/N approaches to the true value of the area, provided the points selected are truly **random**.



Finding the area of a closed curve by the Monte Carlo method

Monte Carlo simulation— Same as *Monte Carlo method*.

Mood's test— A **nonparametric procedure** for testing the equality of **variances** of two **populations** having a **symmetric distribution** with a common **median**. The procedure is based on the assignment of **ranks** to the original **observations** in the combined **sample**

arranged in ascending order. The **asymptotic relative efficiency** of the Mood's test compared to the classical **F test** is 0.76, which is slightly higher than the Siegel–Tukey **efficiency** measure of 0.61. See also *Ansari–Bradley test*, *Barton–David test*, *Conover test*, *F test for two population variances*, *Klotz test*, *Rosenbaum test*, *Siegel–Tukey test*.

morbidity– A term used to describe sickness, illness, or any other disorder in a human population.

morbidity rate– The number of subjects in a given population who develop sickness, illness, or any other morbid condition over a given period of time divided by the total number of people at **risk** during that period. The term is indiscriminately used to refer to **incidence** or **prevalence rates** of disease and should preferably be avoided.

more-than-fair gamble– A game of chance in which the expected monetary **payoff** of what is being lost is less than the expected monetary gain of what is being received.

mortality– A term used in **vital statistics** to describe deaths in a human population. Mortality data are usually obtained from the information contained in death certificates.

mortality rate– Same as *death rate*.

most powerful test– A test of a **null hypothesis** that provides the maximum **power** against a given **alternative hypothesis**.

moving averages– In the **time-series analysis**, an artificially constructed series obtained by successively averaging overlapping groups of two or more consecutive values in a set of **time-series data** and substituting the **average** value in each group by the group's average. For instance, one begins by selecting a fixed number of successive items in a series, computing the average, then dropping the first item and adding the next succeeding one, computing the average of this second group, dropping the second item and adding the next succeeding one, computing the average of this third group, and so on. It is a method of **smoothing** the curve representing the **data**. The method is used primarily for the smoothing of **time series** and elimination of **seasonal variation**, in which each **observation** is substituted by a **weighted average** of the observations and its neighboring values.

MSE– Acronym for *mean square error*.

multicenter clinical trial– A **clinical trial** conducted at a number of research centers in which all follow a common set of predefined guidelines with independent **randomization** performed within each center. Such a study allows a larger **sample size** and permits generalization of findings to a much greater and diverse group of patients and treatment settings than would normally be possible if the study were to be performed at a single location.

multicollinearity– The presence of high or near-perfect intercorrelations between or among various **independent variables** in a **multiple regression analysis**. The multicollinearity results in imprecise **estimates** of the **regression coefficients**, and this makes it difficult to determine their separate **effects** on the **dependent variable**. Extreme multicollinearity can also cause problems in estimating regression coefficients. The use of **exploratory analysis** prior to the model fitting can usually clarify any problems arising from the high **correlations** between **predictor variables** and between **predictors** and **outcome variables**.

multidimensionality– A term generally used to refer to a phenomenon having more than one aspect or dimension. The term is employed to describe attitudes requiring a multiphase decision.

multidimensional scaling– A class of **multivariate techniques** involving a **graphical representation** of statistical similarities or differences with a view to trace a map of how individuals' attitudes or characteristics cluster. The procedure consists of plotting pairs of values with highest **correlations** closest together and those with the lowest correlations farther apart.

multilevel modeling– A term used to refer to a class of **statistical models** such as **regression analysis** where observational **data** have a hierarchical or clustered structure. Many kinds of data in social and biological sciences have a natural hierarchy. For example, many animal and human studies deal with hierarchies where offspring are grouped within families. Similarly, studies on school children involve a hierarchy where children are grouped within schools. Many designed **experiments** such as **clinical trials** also have a hierarchy where subjects are grouped into several randomly chosen centers. A hierarchy usually consists of units grouped at different levels. For instance, offspring may be the level 1 units in a two-level structure where the level 2 units are the families; students may be the level 1 units clustered within schools that are the level 2 units. Multilevel models are designed to take into account differences between levels of a hierarchy.

multilevel models– See *multilevel modeling*.

multilevel regression– An extension or generalization of ordinary **multiple regression** to take into account differences between different levels of a hierarchy. In a multilevel regression, when a higher order **interaction** term is included, all the lower order terms are also included. See also *multilevel modeling*.

multimodal distribution– A **frequency** or **probability distribution** in which two or more different values occur with the highest or nearly highest **frequency** indicating **data values** with more than one **mode**. Such a **distribution** probably indicates that several distributions of relatively distinct groups of **observations** are present. See also *bimodal distribution*, *trimodal distribution*, *unimodal distribution*.

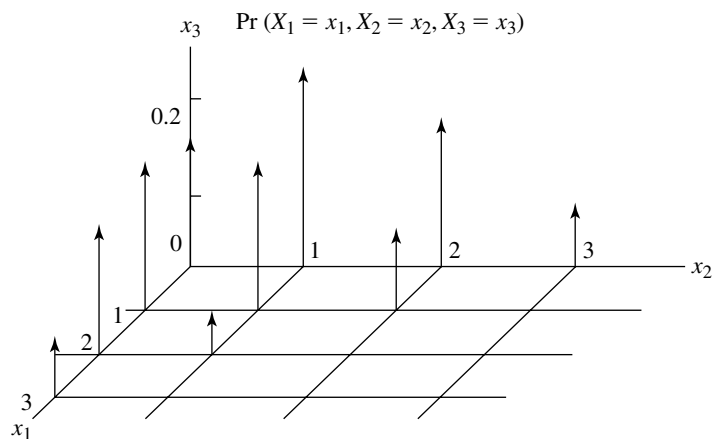
multimodal frequency (probability) distribution– See *multimodal distribution*.

multinomial coefficient– The number of distinct arrangements in which n distinguishable objects with n_1 of the first kind, n_2 of the second kind, . . . , n_k of the k th kind can be distributed into k compartments. It is given by the formula $n!/(n_1!n_2! \dots n_k!)$.

multinomial distribution– A generalization of the **binomial distribution** when there are more than two **outcomes** for each **Bernoulli trial**. The **probability function** of a multinomial distribution is given by the formula

$$P(r_1, r_2, \dots, r_k) = \frac{n!}{r_1!r_2! \dots r_k!} (p_1)^{r_1} (p_2)^{r_2} \dots (p_k)^{r_k}$$

where r_1, r_2, \dots, r_k are the numbers of observations corresponding to k different outcomes with respective **probabilities** of occurrence p_1, p_2, \dots, p_k ($\sum_{i=1}^k r_i = n$, $\sum_{i=1}^k p_i = 1$). It can be shown that the **expected value (mean)** of X_i is np_i , its **variance** is $np_i(1 - p_i)$, and the **covariance** between X_i and X_j is $-np_i p_j$.



Multinomial distribution for $n = 3$, $p_1 = 0.2$, and $p_2 = 0.3$

multinomial experiment— A sequence of n **independent trials** of a **random experiment** where each trial can result in one of k possible **outcomes**. When $k = 2$, the experiment is known as the **binomial experiment**. When $k = 3$, the experiment is known as **trinomial experiment**, and so forth.

multinomial qualitative variable— Same as *multinomial variable*.

multinomial variable— A nominally scaled or **qualitative variable** in which there are more than two categories or classes of **observations**.

multinormal distribution— Same as *multivariate normal distribution*.

multiphase sampling— An extension of two-phase or **double sampling** to three or more phases.

multiple causation— A term used to describe the view that any “effect” is produced by multiplicity of causes.

multiple coefficient of determination— Same as *coefficient of multiple determination*.

multiple comparison— A statistical procedure that, on the basis of the same **data set**, makes a number (more than one) of tests (comparisons) concerning the various **parameters** of interest controlling for the overall **error rate**. If an overall error rate is fixed at 5 percent, then each test must be performed at a **significance level** less than 5 percent. In an **analysis of variance**, multiple comparison is used to test which **mean** (or a set of means) differs from which other (or a set of means). It is used as a follow-up to significant **F tests**. It is also called a **posthoc comparison**. No single test is found to be best in all situations, and a major difference between them lies in the manner in which they control the increase in **type I error** due to multiple testing. Some most commonly used multiple comparison tests are the **Bonferroni procedure**, **Duncan multiple range test**, **Dunnett’s multiple comparison test**, **Newman–Keuls test**, **Scheffe’s test**, and **Tukey’s test**.

multiple comparison test— Same as *multiple comparison*.

multiple correlation— Same as *multiple correlation coefficient*.

multiple correlation analysis– A method of analysis for determining **correlations** among many **variables** simultaneously.

multiple correlation coefficient– The **product moment correlation** between the actual values of the **dependent variable** and the predicted values as determined by the **multiple regression equation**. It is a measure of the degree of **linear association** between more than two **variables** and is equal to the square root of the **coefficient of multiple determination**. The square of the multiple correlation coefficient provides a measure of the **proportion of variation** of the **response variable** that is explained by the **explanatory variables** and is denoted by R^2 .

multiple correspondence analysis– See *homogeneity analysis*.

multiple discriminant analysis– See *discriminant analysis*.

multiple logistic regression– The **logistic regression** involving several **independent variables**. If X_1, X_2, \dots, X_p are p independent variables and Y is a **binary response variable** with **probability** of success equal to p , then the multiple logistic regression model is given by

$$p = \frac{e^{\alpha + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\alpha + \beta_1 x_1 + \dots + \beta_p x_p}} = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \dots + \beta_p x_p)}}$$

where e is the (natural) exponential function. After applying the log odds **transformation**, the **regression model** is written as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Note that the effect of each **explanatory variable** is to multiply the baseline log odds. In epidemiological studies, multiple logistic regression is frequently used for controlling **confounding** or assessing **interactions**. The results from logistic regression are often expressed in terms of an **odds ratio**.

multiple logistic regression model– See *multiple logistic regression*.

multiple R– Same as *multiple correlation coefficient*.

multiple regression– Same as *multiple regression analysis*.

multiple regression analysis– An **analysis of regression** involving two or more **independent variables** as **predictors** to estimate the value of a single **dependent** or **response variable**. The dependent variable is usually continuous, but the independent variables can be continuous or categorical. The **regression model** being fitted is $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$, where Y is the dependent or response variable, X_1, X_2, \dots, X_p are the independent variables, β_0 is the intercept, and $\beta_1, \beta_2, \dots, \beta_p$ are the corresponding **regression coefficients**. The **parameters** $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are generally estimated by the **method of least squares**. Each regression coefficient is interpreted as the change in the magnitude of the dependent variable corresponding to a unit change in the appropriate independent variable while holding the **effects** of other independent variables as **constants**. See also *regression analysis*.

multiple regression coefficient– See *regression coefficient*.

multiple regression equation– In a **multiple regression analysis**, an algebraic equation relating the **independent variables** to the **expected value** of the **dependent variable**.

multiple regression model– See *multiple regression analysis*.

multiple significance testing– See *multiple comparison*.

multiple-stage sampling– **Sampling** by stages, where the **sampling units** at each stage are subsampled from the larger units chosen at the previous stage. Thus, a municipality may be divided into a certain number of zones and a number of those zones are selected randomly. Within each zone drawn in the **sample**, a number of schools are chosen **at random**. Within each school drawn in the sample, a sample of students can be randomly selected. This is an example of a three-stage sampling where students drawn within schools compose the sample to be analyzed. It is often used in combination with **area sampling** and **cluster sampling**.

multiple testing– See *multiple comparison*.

multiple time series– A multivariate analogue of a univariate **time series** comprising a set of ordered observation vectors measured on several quantitative characteristics taken at different points in time.

multiplication rule for probabilities– A **probability rule** used to determine the **probability** of an intersection of two or more **events**. For any two arbitrary events A and B , it is given by the formula $P(A \cap B) = P(A)P(B|A)$ or $P(A \cap B) = P(B)P(A|B)$. For two **independent events**, it reduces to $P(A \cap B) = P(A)P(B)$. In a series of **independent trials**, the probability that each of a specified series of events takes place is the product of the probabilities of the individual events.

multiplicative model– A **model** in which the combined **effect** of a number of **factors** is taken as the product of effects that can be attributed to the individual factors. See also *additive model*.

multiplicative time-series model– A **classical time-series model** that expresses the actual value of a time series as the product of its components; for example, $Y = T \times C \times S \times I$. See also *additive time-series model*, *mixed time-series model*.

multistage sampling– Same as *multiple-stage sampling*.

multivariable analysis– A term sometimes used in contradistinction to **multivariate analysis**. When there are several **independent variables**, but only a single **dependent variable**, the term ‘multiple’ or ‘multivariable’ is preferable to multivariate.

multivariate analysis– A class of statistical methods and techniques involving multiple **independent** or **dependent variables**. Examples of multivariate analysis include **factor analysis**, **discriminant analysis**, **multiple regression** and **correlation analysis**, and many other techniques. Such techniques play an important role in investigating **multivariate data**. See also *bivariate analysis*, *univariate analysis*.

multivariate analysis of variance– An advanced statistical procedure that provides an overall test when there are multiple measures of **dependent variables** and the **independent variables** are nominal. It is a generalization of the univariate **analysis of variance** with multiple outcome measures for the dependent variable. It is used to test group differences on profiles of **measurements**, in contrast to the use of ANOVA to test group differences on measurements of a single variable. It is widely used in business, psychological, and social science research. Unlike the univariate case where **F tests** are used to test **hypotheses** of interest, in the multivariate case there does not exist a single optimal **test**

procedure. Three most commonly used test criteria are: Wilk's lambda (λ), Roy's largest root criterion, and the Hotelling–Lawley trace. If the dependent variables are not correlated, separate ANOVAs for each dependent variable would suffice.

multivariate contingency table– An extension of a **contingency table** for **bivariate data** to **multivariate data**.

multivariate contingency table analysis– Methods and techniques for analyzing relationships among several **categorical variables** forming a **multivariate contingency table**.

multivariate data– Same as *multivariate data set*.

multivariate data set– A **data set** containing information on two or more **variables**. Such data are usually displayed in the form of a **data matrix**.

multivariate density function– A multivariable continuous function $f(x_1, x_2, \dots, x_p)$ defined for all possible p -tuples (x_1, x_2, \dots, x_p) in the range of **continuous random variables** X_1, X_2, \dots, X_p , such that $f(x_1, x_2, \dots, x_p) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_p) dx_1 dx_2 \cdots dx_p = 1$$

See also *bivariate density function, joint density function*.

multivariate distribution– Same as *multivariate probability distribution*.

multivariate methods– See *multivariate analysis*.

multivariate normal distribution– A generalization of a **bivariate normal distribution** to three or more **random variables**. Geometrically, it can be represented as concentric ellipsoids of **constant** density in multidimensional space. The form of its **probability density function**, however, involves the use of complex matrix notations and can be found in any book on **multivariate analysis**. Like its univariate and bivariate counterparts, the distribution has a number of simple properties that make its use as a **probability model** for observed **multivariate data** very popular. See also *normal distribution, trivariate normal distribution*.

multivariate observations– Same as *multivariate data*.

multivariate probability distribution– See *joint probability distribution*.

multivariate probability function– A multivariate discrete function $p(x_1, x_2, \dots, x_p)$ defined for all possible p -tuples (x_1, x_2, \dots, x_p) in the range of **discrete random variables** X_1, X_2, \dots, X_p , such that $p(x_1, x_2, \dots, x_p) \geq 0$ and $\sum_{x_1, x_2, \dots, x_p} p(x_1, x_2, \dots, x_p) = 1$. See also *joint probability function*.

multivariate statistical analysis– Same as *multivariate analysis*.

multivariate statistical methods– See *multivariate analysis*.

multivariate statistical procedures– See *multivariate analysis*.

multivariate statistical techniques– See *multivariate analysis*.

multivariate techniques– See *multivariate analysis*.

multivariate time series— Same as *multiple time series*.

multiway analysis of variance— An **analysis of variance** procedure involving the study of several **factors** simultaneously. It is an extension of the analysis of variance methodology for the case of two factors to three or more factors involving a single experiment. Multi-factor ANOVA designs usually provide more information and often can be even more economical than separate one-way or two-way designs. See also *one-way analysis of variance*, *two-way analysis of variance*, *three-way analysis of variance*.

multiway classification— A classification of a set of **observations** according to three or more characteristics or **factors**. See also *one-way classification*, *two-way classification*.

mutual independence— In **probability theory**, when each subset of a set of n **events** defined on the same **sample space**, e.g., $((A_i, A_j; i < j, = 1, 2, \dots, n)$, $(A_i, A_j, A_k; i < j < k = 1, 2, \dots, n)$, etc., is independent, the set are said to be mutually independent. For example, three events A_1, A_2 , and A_3 defined on the same sample space are mutually independent if

$$P(A_1 \cap A_2) = P(A_1)P(A_2), P(A_1 \cap A_3) = P(A_1)P(A_3), P(A_2 \cap A_3) = P(A_2)P(A_3)$$

and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

See also *pairwise independence*.

mutually exclusive events— In **probability theory**, two or more **events** are said to be mutually exclusive if they cannot occur simultaneously or do not have any simple elements in common. A single toss of a coin for example must result in either a head or a tail. These outcomes are mutually exclusive. Compare *nonmutually exclusive events*.

