



saddle point– In a **zero-sum game**, the pure strategies of two players constitute a saddle point if the corresponding entry of the **payoff matrix** is simultaneously a maximum of row minima and a minimum of column maxima.

Sakoda coefficient– A **measure of association** or relationship between two **categorical** or **qualitative variables** whose **data** are cross-classified in an $r \times c$ **contingency table**. It is calculated by the formula

$$S = \sqrt{\frac{p\chi^2}{(p-1)(n + \chi^2)}}$$

where $p = \min(r, c)$, χ^2 is the usual **chi-square statistic** for testing the **independence**, and n is the **sample size**. See also *contingency coefficient*, *phi coefficient*, *Tschuprov coefficient*.

sample– A sample is a subset or a portion of the entire aggregate of a **population**. A sample is usually selected according to some specified criteria. In many statistical applications, samples are used to draw **inferences** about the population characteristics, that is, to generalize results from sample to population. To be useful, a sample must be representative of the population from which it is drawn; that is, it must have characteristics similar to those of the population. **Random** or **probability samples** often produce a **representative sample**. There are various methods and techniques available for selecting a sample and drawing inferences from it. See also *judgment sample*, *nonprobability sample*.

sample autocorrelation– Same as *autocorrelation*.

sample coefficient of correlation– A standardized measure of the **linear relationship** between two **variables** using **sample data**. See also *correlation coefficient*, *population coefficient of correlation*.

sample coefficient of determination– Same as *sample coefficient of multiple determination*.

sample coefficient of multiple correlation– An **estimate** of the degree of **linear relationship** between more than two **variables** obtained by using **sample data**. See also *coefficient of multiple correlation*, *coefficient of multiple determination*.

sample coefficient of multiple determination– An estimate of the **goodness of fit** of the estimated **regression plane** (or **hyperplane**) obtained by using the **sample data**. See also *coefficient of multiple determination, population coefficient of multiple determination*.

sample coefficient of partial correlation– The square root of the **sample coefficient of partial determination**. See also *coefficient of partial determination*.

sample coefficient of partial determination– An estimate of the **coefficient of partial determination** obtained by using the **sample data**.

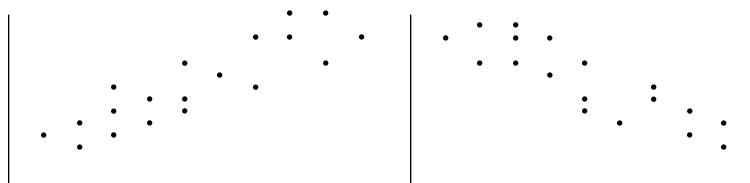
sample correlation– Same as *sample coefficient of correlation*.

sample correlation coefficient– Same as *sample coefficient of correlation*.

sample covariance– An unstandardized measure of **linear relationship** between the two **variables** X and Y using **sample data**. If a **sample** of n **observations** is $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, then the sample covariance, denoted by S_{xy} , is defined as

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

A positive **covariance** will result if the factors $(x_i - \bar{x})$ and $(y_i - \bar{y})$ tend to be either both positive or both negative. This happens if there is a tendency for both variables to increase or decrease at the same time. On the other hand, a negative covariance will result if the both factors $(x_i - \bar{x})$ and $(y_i - \bar{y})$ tend to be of opposite sign. This happens if there is a tendency for y to decrease as x increases.



Examples of positive and negative covariance

sample covariance matrix– A **covariance matrix** where **variances** and **covariances** are **sample estimates** of the corresponding **population parameters**. See also *population covariance matrix*.

sample data– **Data** obtained from a **sample** rather than from the entire **population**.

sampled population– The **population** from which the **sample** is actually selected. See also *parent population, target population*.

sample estimate– Same as *estimate*.

sample estimator– Same as *estimator*.

sample frame– Same as *frame*.

sample mean– The most commonly used **estimate** of the **population mean**. For a **sample** of size n , with **measurement** values x_1, x_2, \dots, x_n , the sample mean is defined as $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$. See also *mean, population mean*.

sample median– The value that divides the **sample data** into two equal groups. For an odd **sample size**, say a sample of $2n + 1$ **observations**, denote the ordered values by $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(2n+1)}$. Then the sample median is $x_{(n+1)}$, the observation that occupies position $(n + 1)$ in the list. For an even sample size, say a sample of $2n$ observations, listed in order as $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(2n)}$, it is customary to make the sample median unique by defining it as the **mean** of $x_{(n)}$ and $x_{(n+1)}$, that is, $(x_{(n)} + x_{(n+1)})/2$. The sample median provides a **measure of central tendency** that is more appropriate for **skewed distributions**. It is also relatively insensitive to presence of **outliers**. See also *population median*.

The data arranged in an increasing order of magnitude



Median

Schematic representation of a sample median

sample moments– See *moments*.

sample observations– Same as *sample data*.

sample point– The individual **outcome** of a **random experiment** is called a sample point.

sample proportion– The **estimate** of a **binomial proportion** based on **sample data**. It is calculated by the formula x/n where x is the number of successes in n **independent trials**.

sample range– The **range** calculated from a **sample data** rather than from the entire **population**.

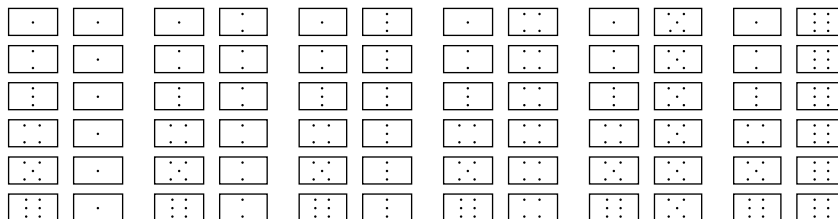
sample regression line– Same as *estimated regression line*.

sample size– Same as *size of a sample*.

sample space– In **probability theory**, the collection of all possible **outcomes** of an **experiment** is called sample space.

Sample space for five tossed coins

HHHHH	HTHHH	THHHH	TTHHH	HHHHT	HTHHT	THHHT	TTHHT
HHHHT	HTHTH	THTHH	TTHTH	HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH	HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH	HHTTT	HTTTT	THHTT	TTTTT



Sample space for two tossed dice

sample standard deviation– The most commonly used **estimate** of the **population standard deviation**. For a **sample** of size n , with **measurement** values x_1, x_2, \dots, x_n , the sample standard deviation is defined as

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

where \bar{x} is the **sample mean**. See also *standard deviation*.

sample statistic– Same as *statistic*.

sample survey– A **survey** in which **observations** are made about one or more characteristics of interest for only a **sample** of human populations, business, industry, or other institutions. The **sample observations** are used to estimate particular population characteristics of interest.

sample unit– Same as *sampling unit*.

sample values– Same as *sample data*.

sample variance– The most commonly used **estimate** of the **population variance**. It is equal to the square of the **sample standard deviation**. See also *variance*.

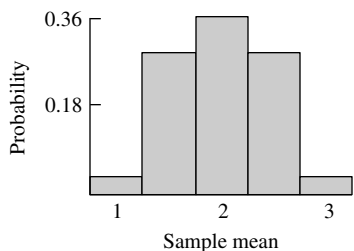
sampling– The process of selecting a **sample** from a **population**, in order to use the sample information to draw conclusions of the source population. The aim of sampling is to provide the required information with minimum investment of time, effort, and money. In some cases samples may provide more accurate results than a **census** or complete enumeration. Two types of sampling procedures commonly used are: (1) **probability sampling**, in which each unit is chosen with a given **chance** of being selected, and (2) **nonprobability sampling**, in which the selection of the sample is based on convenience or judgment. See also *cluster sampling, random sampling, simple random sampling, stratified random sampling, systematic sampling*.

sampling design– A procedure for drawing a **sample** from a given **population**. The term “sampling design” is often understood to mean all the necessary steps and procedures in the selection of a sample and subsequent analysis, including choice of a **sample frame**, recruiting and training of interviewers, data collection procedures, and methods of **estimation** and **hypothesis testing**.

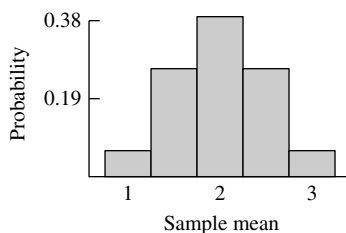
sampling distribution– A theoretical **probability distribution** of any **statistic** that results from drawing all possible **samples** of a given size from a **population** and can be calculated on the basis of a sample of a given size. Some useful examples are the sampling distribution of **means** or **proportions** and the sampling distribution of the difference between two means or proportions. It is described by showing all possible values of a **sample statistic** and its corresponding **probabilities**. Sampling distribution is useful in drawing **inferences** about the population based on the statistic in question.

sampling distribution of mean– See *sampling distribution*.

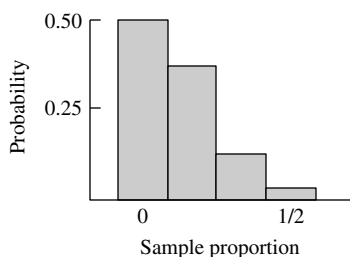
sampling distribution of proportion– See *sampling distribution*.



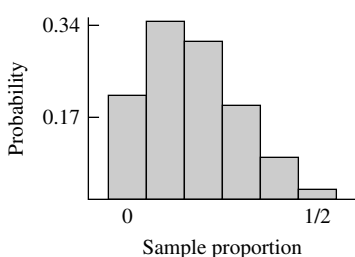
Distribution of the sample mean of samples of size 2 drawn (without replacement) from the population {1, 1, 2, 2, 2, 2, 3, 3}



Distribution of the sample mean of samples of size 2 drawn (with replacement) from the population {1, 1, 2, 2, 2, 2, 3, 3}



Distribution of the sample proportion for samples of size 6 with $p = 0.1$



Distribution of the sample proportion for samples of size 16 with $p = 0.1$

sampling error– The difference between a **point estimate** and the value of the **population parameter** being estimated. It is a measure of inaccuracies in estimating a parameter because a **sample** rather than the entire **population** has been taken. Although the sampling error is usually unknown, with an appropriate **sampling design** it can usually be kept small within a desired level of **precision**.

sampling fraction– The **proportion** of **sampling units** to be drawn from a specified **population** for selection in the **sample**. It is obtained as the **ratio** of **sample size** to **population size**. A 5% sample has a sampling fraction of $\frac{1}{20}$.

sampling frame– Same as *frame*.

sampling procedure– Same as *sampling design*.

sampling scheme– Same as *sampling design*.

sampling unit– The unit of selection in the sampling process, e.g., a person, a household, a district, etc. It is not necessarily the unit of **observation** or study.

sampling variability– Same as *sampling variation*.

sampling variance– The **variance** of the **sampling distribution** of a **statistic** or the square of its **standard error**.

sampling variation– The unaccounted fluctuations (**random error**) in results as exhibited from one **sample** to the other. See also *sampling distribution*.

sampling without replacement– A method of **sampling** such that once a **sampling unit** from a **population** has been selected, it is removed from the population and cannot be selected in a second or subsequent draw.

sampling with replacement– A method of **sampling** such that, as each **sampling unit** from a **population** is selected, it is returned to the population being sampled. It is possible that a previously selected item may be selected again and, therefore, appear in the **sample** more than once.

SAS– A widely used **statistical computing package** for data management, report writing, and statistical analysis. It is an acronym for Statistical Analysis System. SAS is a powerful **statistical software package**, and is currently available on thousands of computing facilities throughout the world. The package includes a great variety of elementary and advanced statistical procedures suitable for myriads of business and scientific applications. It is an extremely flexible package containing a complete range of statistical procedures with powerful graphical capabilities, and all can be accessed with a single run.

Satterthwaite's approximation– Same as *Satterthwaite's procedure*.

Satterthwaite's procedure– A general procedure for approximating the **probability distribution** of a **linear combination** of independent **random variables** where each variable has a scaled **chi-square distribution** with known **degrees of freedom**. The procedure is frequently employed for constructing **confidence intervals** for the **mean** and the **variance components** in a **random** or **mixed effects analysis of variance**.

saturated model– A **model** that contains as many **parameters** as there are **cells** or **means** and consequently results in a perfect fit for a given set of **data**.

Savage's test– A **nonparametric procedure** for testing the difference between two **cumulative distribution functions**. See also *goodness-of-fit test*, *Kolmogorov–Smirnov test*.

scalar– A single number in contrast to a **vector** in a matrix context.

scale– A term used to describe the property of a **distribution** that is related to the scale of the **variable**, e.g., the **standard deviation** of a **normal distribution**.

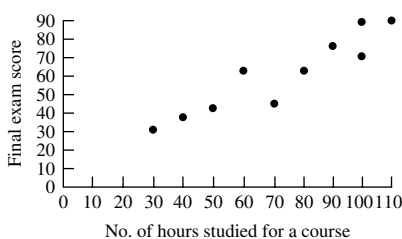
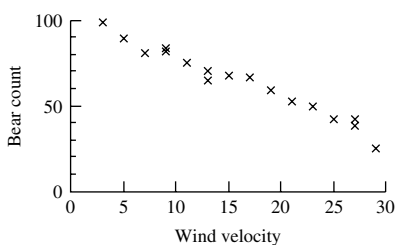
scale of measurement– A term used to describe the degree of **precision** with which an **attribute** or characteristic can be measured. It is generally classified into **nominal**, **ordinal**, **interval**, and **ratio scales**. These four scales are arranged in order of strength, from the lowest to the highest. The **data** obtained at a higher scale of measurement can usually be described with a lower scale of measurement, but the converse is not true.

scale parameter– A term generally used to refer to a **parameter** of a **distribution** that determines its scale.

scatter– The extent to which the **data points** in a **scatter diagram** fail to fall into alignment. The term is often used as a synonym for **variability**.

scatter diagram– A two-dimensional graph displaying the relationship between two characteristics or **variables** of a set of **bivariate data** in which one variable appears on the

horizontal axis and the other appears on the **vertical axis**. It is drawn on a **cartesian plane** where one set of **score** values is displayed on the horizontal **x axis**, called the **abscissa**; the other is displayed on the vertical **y axis**, called the **ordinate**. Each **data point** represents a pair of scores, an x value and a y value. For example, x and y may represent height and weight, and each dot represents the associated height and weight. The **independent variable** usually appears on the horizontal axis and the **dependent variable** appears on the vertical axis. A set of n (x, y) **observations** thus provides n points on the plot and the **scatter** or clustering of the points exhibits the relationship between x and y . In using a **regression model** in order to assess the **association** between the two variables, it is always useful to draw a scatter diagram. The diagram provides an important visual aid in assessing the type of relationship between the two variables. See also *bivariate plot*, *correlation coefficient*, *measure of association*.



Scatter diagrams of two bivariate data sets

scattergram— Same as *scatter diagram*.

scatter plot— Same as *scatter diagram*.

s chart— A **graphical device** used to control the **variance** of a process by inspecting the **standard deviation** of a set of **measurements** taken from various batches or subgroups. The values of standard deviation taken from each subgroup are plotted along the **vertical axis** and can then be used to control within subgroup **spread**. The **center line** of the s chart is the **average** of standard deviations (\bar{s}) from a pilot set (about 20 rational subgroups). The **control lines** are set at $\bar{s} \pm 3(0.389/0.9123)\bar{s}$. In practice, the engineer sets the limits at $B_3\bar{s}$ and $B_4\bar{s}$, where B_3 and B_4 are obtained from some specially prepared tables. See also *c chart*, *control chart*, *p chart*, *run chart*, *x-bar chart*.

Scheffé's test— A **multiple comparison** procedure for comparing **means** following a significant **F test** in an **analysis of variance**. It can be used to make any comparisons among means, not simply pairwise. It is one of the most conservative of all multiple comparison procedures. The method is equally applicable with both equal and unequal **sample sizes**. See also *Bonferroni procedure*, *Duncan multiple range test*, *Dunnnett multiple comparison test*, *Newman-Keuls test*, *Tukey's test*.

scientific sample— Another term for a **probability sample** commonly used in popular media and scientific publications.

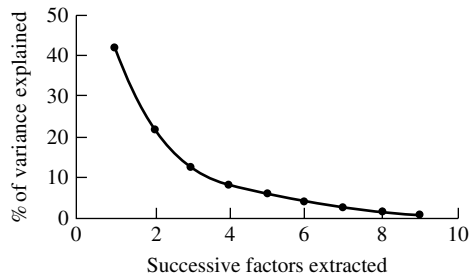
score— A numerical value assigned to a **measurement** or **observation**.

score data— **Numerical data** that have some of the characteristics or properties of **data** measured on an **interval scale** but are inherently ordinal in nature.

score interval– Same as *class interval*.

scoring method– A numerical **algorithm** normally employed for optimization of a mathematical function. It is an **iterative procedure** that is useful for solving nonlinear maximum likelihood equations.

scree diagram– A plot used in the **principal components analysis** to provide a visual aid for determining the number of **factors** that explain most of the **variability** in the **data set**.



Schematic illustration of a scree diagram showing the percentage of total variance accounted for by each of nine successively extracted factors

screening– Screening is an initial attempt to identify the presence or absence of a disease or disorder by means of test procedures that can be administered rather quickly and economically. In general, the term is used to refer to clinical, laboratory, radiological, or any other procedure performed for the purpose of identifying **risk factors** for a disease.

screening test– Same as *screening*.

SD– Acronym for *standard deviation*.

SE– Acronym for *standard error*.

seasonal chart– A graph showing a plot of a **time series** month by month or quarter by quarter for each of a number of years. Seasonal charts are used as a preliminary to estimating regular **seasonal components**.

seasonal component– In **time-series analysis**, narrow up and down swings of the series of interest around the **trend** and **cyclical components**, with the swings generally repeating each other within periods of 1 year or less. The seasonal components repeat each other regularly with more or less the same intensity as a result of **seasonality**. Although the term is used to denote yearly **cycles**, it is sometimes used to indicate other periodic movements. There are various statistical methods currently available for estimating seasonal components. They are used for making seasonal corrections to **data** such as those in calculating business ratios and budgeting. See also *time series*.

seasonal fluctuation– Same as *seasonal variation*.

seasonality– A term sometimes used to refer to **seasonal variation** of a **time-series data**.

seasonally adjusted– In **time-series analysis**, a term used to refer to series from which periodic fluctuations with a period of 1 year have been eliminated.

seasonal variation– A term used to indicate regular **variation** in **seasonal components** such as variations in sales turnover or current costs, due to regularly recurring seasonal factors.

secondary data– This refers to the **data** that are published by an organization different from the one that originally collected and published them. The *Statistical Abstract of the United States*, which compiles data from several primary government sources and is updated annually, is a popular source of secondary data. Compare *primary data*.

second quartile– The 0.50 **fractile** or 50th **percentile point** in a **data set**, below which half of all **observations** lie. See also *first quartile, median, quartiles, third quartile*.

secular trend– Same as *trend*.

selection bias– A systematic tendency to favor the inclusion in a **sample** of certain selected **elementary units** with particular characteristics, while excluding those with other characteristics. A selection bias leads to a systematic difference between the characteristics of a sample and its source **population**. Conclusions drawn from a sample with selection bias are not generalizable to the entire population. Many medical and epidemiological studies are prone to selection bias. For example, in **case-control studies**, **cases** with higher levels of **exposure** are more likely to be diagnosed and therefore to be included in the studies. In **clinical trials**, a selection bias can occur because of methods of allocation, which may lead to imbalance between **treatment groups** with respect to important **prognostic factors**. Thus, a selection bias causes a sample to be unrepresentative of the population from which it is drawn.

self-controlled study– An investigation in which the **study subjects** serve as their own **controls**. This is usually achieved by measuring the **response measure** of interest before and after administering the **treatment**.

semi-interquartile range– A **measure of variability** or **dispersion** obtained by dividing the **interquartile range** by 2. Thus, semi-interquartile range is one half of the distance between two **quartiles** of a **sample** or a **distribution**. It is also called **quartile deviation**.

semilogarithmic chart– See *logarithmic chart*.

semilog paper– See *semilogarithmic chart*.

sensitivity– In a **screening** or **diagnostic test**, the **probability** that the test will yield a positive result when administered to a person who has the disease or condition of interest. It is the measure of the goodness of a diagnostic test in detecting individuals who have the disease or condition in question. In **hypothesis testing**, it refers to the **power** of a **statistical test** to detect **deviations** from some specified hypothetical value. Compare *specificity*. See also *Bayes' theorem, receiver operating characteristic curve*.

sensitivity analysis– A term used to describe a method for determining how the final outcome of an analysis changes as a function of varying one or more input parameters. A sensitivity analysis quantifies how changes in the values of the input parameters affect the values of the **outcome variable**. Sensitivity analysis is frequently carried out to assess the impact of different assumptions or scenarios on the results of a study. For example, it can be used to calculate the **sample size** requirements for different values of **significance level, power**, expected differences between groups, and **variability of measurements**, among others. In **meta-analysis**, sensitivity analysis can be used to assess the impact of removing

some studies, which may be of poorer quality, from the overviews. See also *uncertainty analysis*.

separate variance t test– Same as *unequal variance t test*.

sequential analysis– See *sequential sampling*.

sequential sampling– A method of **sampling** in which the **sample size** is not fixed in advance, but in which a decision is made, after the selection of each unit, as to whether to continue the sampling. In sequential sampling, the **sample units** are drawn one by one or in groups of a given size, and the decision is made after each **observation** whether to continue or terminate the sampling. The sample size is thus not fixed in advance and depends on the actual observations and varies from one **sample** to another. This kind of sampling often results in much fewer observations than would be required if the sample size were fixed in order to provide the same control over **type I** and **type II errors**. This is often used in **quality control** procedures where testing is expensive, i.e., where it involves destruction in testing for estimating length of life.

serial correlation– In a **longitudinal study**, the term is used to describe the **correlation** between pairs of **measurements** on the same subject. The magnitude of such correlation usually depends on the time lag between the measurements; as the time lag increases, the correlation usually becomes weaker. In a **time-series analysis**, the term is used to refer to the correlation between **observations** that either lead or lag by a specified time interval. A **statistical test**, based on the **ratio** of the mean square successive difference to the **variance**, can be used to test the significance of serial correlation.

serial measurements– In a **longitudinal study**, the **observations** made on the same subject at different points in time.

set– In **set theory**, a collection, class, or aggregate of objects or things. The objects of a set are called elements.

set theory– A branch of mathematics that is concerned with the study of the characteristics and relations among sets.

shape– A term used to describe the degree of **asymmetry** or the peakedness in a **frequency distribution**. It is that aspect of the form of a **distribution** that is distinct from the property of **skewness** and **kurtosis**.

shape parameter– A term generally used to refer to a **parameter** of a **distribution** that determines its **shape**, in contrast to **location** and **scale** of the distribution. The term was earlier thought to be associated with **skewness** and **kurtosis**, but the usual measures of skewness and kurtosis are not good representations of shape.

Shapiro–Francia test– A test of **normality** based on **order statistics** from **sample data**. It is a modification of the **Shapiro–Wilk W test**, and the **null distribution** of the **test statistic** can be approximated by the **standard normal distribution**. See also *Anderson–Darling test*, *Cramér–von Mises test*, *D’Agostino’s test*, *Michael’s test*.

Shapiro–Wilk W test– A test of **normality** based on **order statistics** from **sample data**. The **test statistic** is calculated as the **ratio** of the square of a **linear combination** of sample **order statistics** to the usual **sample variance**. A **statistic** commonly reported in addition to the test statistic W is V , which is equal to 1 if the **data** are normally distributed, or greater than 1 if not. It is one of most powerful omnibus tests for normality. The test has

been found to be good against short- or very long-tailed **distributions** even for **samples** as small as 10. See also *Anderson–Darling test*, *Cramér–von Mises test*, *D’Agostino’s test*, *Michael’s test*, *Shapiro–Francia test*.

Sheppard’s corrections– The corrections used in the computation of **moments** due to the approximation introduced by considering the values of a **grouped frequency distribution** as if they were concentrated at the **midpoints** of **class intervals**. For example, if the **distribution** is continuous and tails off smoothly, the second moment about the origin calculated from grouped frequencies should be corrected by subtracting from it $h^2/12$, where h is the length of the interval.

short-term forecast– A business **forecast** which usually extends as much as six quarters ahead of the current period. Forecasts for the short term are usually more popular than those involving medium or long time intervals.

Siegel–Tukey test– A **nonparametric procedure** for testing the equality of **variances** of two **populations** having the common **median**. It is a modification of the **Wilcoxon rank-sum test**; the **test statistics** of both tests have the same **null distribution**. For this test, one assigns a **rank** of 1 to the smallest **observation**, a rank of 2 to the largest, a rank of 3 to the second largest, a rank of 4 to the second smallest, a rank of 5 to the third smallest, a rank of 6 to the third largest, and so on. If the **null hypothesis** of no difference in **spread** is true, then the **means** of rank values in two **samples** should be nearly equal. If the populations also differ in **locations**, the Siegel–Tukey test may not be useful, since the rejection of the null hypothesis of equal variances may result from differences in locations. If it is known that the populations differ in location, the **data** should be adjusted by subtracting the appropriate means or medians from each observation. The test procedure should then be performed on the adjusted data. The **asymptotic relative efficiency** of this test compared to the classical **F test** for **normal populations** is only 0.61. However, for the **double exponential distribution**, the efficiency increases to 0.94. See also *Ansari–Bradley test*, *Barton–David test*, *Conover test*, *F test for two population variances*, *Klotz test*, *Mood test*, *Rosenbaum test*.

signed-rank test– Same as *Wilcoxon signed-rank test*.

significance level– The significance level of a **statistical test** refers to the **probability** level at which the investigator is prepared to reject the **null hypothesis** as being very unlikely and to favor the **alternative hypothesis** instead. It is the **probability** of selecting a value of the **test statistic** that is as extreme or more extreme than the value observed. It is interpreted as the probability level of a difference arising largely by **chance**, below which it is considered sufficiently unlikely for the difference to be **statistically significant**. In many scientific investigations, it is usually set by the researcher and is conventionally taken as 0.05. It is the probability of committing **type I error** and is denoted by the Greek letter α . See also *p value*.

significance probability– Same as *p value*. See also *significance level*.

significance test– Same as *statistical test*.

significant– Same as *statistically significant*.

sign test– A **nonparametric procedure** for detecting differences between the **locations** of two **populations** by the analysis of two **matched** or **paired samples**. It is based on the number of plus or minus signs of pairwise differences, which is then considered a **sample**

from a **binomial population**. The test is also applicable for testing a **hypothesis** about the **median**. The sign test is one of simplest and oldest of all **nonparametric statistical tests** available.

simple correlation– Same as *correlation*.

simple correlation analysis– **Correlation analysis** that measures the **association** or **correlation** between two **variables** only.

simple event– Same as *elementary event*.

simple hypothesis– A **hypothesis** that completely specifies the **distribution** of a **random variable**.

simple linear regression analysis– Same as *simple regression analysis*.

simple random sample– A **sample** selected from a **population** of size N in such a manner that each possible sample of a given size n has the same **probability** of being selected. Thus, in a simple random sample, all the $\binom{N}{n}$ samples have the same probability of being selected. For an infinite population, it is a sample selected such that each item comes from the same population and each item is selected independently of the other. See also *simple random sampling*.

simple random sampling– The method of **sampling** that gives all **sampling units** in a specified **sampling frame** an equal **chance** of being selected for inclusion in the **sample**, and an equal chance for selection for each of all possible samples of the same size. Thus, a simple random sampling of n objects from a **population** of N objects is any procedure that assures that each possible sample of size n has an equal chance or **probability** of being selected. The procedure also assures that each possible sample has the probability $1/\binom{N}{n}$ of being selected. Simple random sampling is not very simple to use in field work, particularly when the population is large and the individuals are not numbered. See also *simple random sample*.

simple randomized design– Another term for the basic **one-way analysis of variance** design, the so-called **completely randomized design**.

simple regression analysis– **Regression analysis** that uses a single **variable** as the **predictor** of a **dependent variable**. It is used in contrast to **multiple regression analysis** in which two or more predictors are used to explain one dependent variable. The **regression model** for a simple linear regression analysis is $E(Y) = \alpha + \beta X$ where Y is the dependent or response variable, X is the **independent variable**, α is the **intercept**, and β is the **regression coefficient**. The **parameters** α and β are generally estimated by the **method of least squares**. The regression coefficient β measures the change in the magnitude of Y corresponding to a unit change in the magnitude of X .

simple regression model– See *simple regression analysis*.

simplex algorithm– An optimization **algorithm** for minimizing and maximizing a function of several variables.

Simpson's paradox– A phenomenon that occurs when either the magnitude or direction of the **association** between two **variables** is influenced by a third variable which may act as a **confounder**. By failing to control for its effect, the value of the observed association may appear to be greater than the reality.

simulation– Same as *Monte Carlo method*.

simultaneous confidence intervals– **Confidence intervals** for several **parameters** being determined simultaneously. In an ordinary confidence interval, we make a **probability** statement about a single parameter while in simultaneous confidence intervals, the probability statement is valid for intervals for more than one parameters simultaneously.

single-blind study– Same as *single-blind trial*.

single-blind trial– A **clinical trial** in which the patient has no knowledge of the **treatment** he is receiving. See also *blind study*, *double-blind trial*, *triple-blind trial*.

single-factor experiment– **Experiment** or design that entails only one **factor**. It is also called **one-way classification**.

single-masked study– Same as *single-blind trial*.

single-masked trial– Same as *single-blind trial*.

single-sample *t* test– Same as *one-sample *t* test*.

size of a sample– The number of cases or **observations** included in a specific **sample**. It is usually denoted by the letter *n*. It is generally determined to estimate a **parameter** with a given bound of error or to detect an effect of a particular size for given values of **type I** and **type II errors**. In complex surveys involving **multistage sampling**, it refers to the number of units at the final stage in the **sampling**.

size of the test– Same as *significance level*.

skewed distribution– An **asymmetrical distribution** of values of a **variable** that is characterized by extreme values at one end of the **distribution** or the other. In a **skewed distribution**, the **scores** accumulate at one end and spread out markedly toward the other. If the skew, or thin end, points to the right, the distribution is positively skewed. If the skew points to the left, the distribution is negatively skewed. See also *symmetrical distribution*.



A positively or right-skewed distribution

A negatively or left-skewed distribution

skewness– The lack of **symmetry** in a **distribution**. It is the property of a distribution that refers to the extent of its **asymmetry**. See also *coefficient of skewness*, *skewed distribution*.

slope of the regression line– The slope of the **regression line** $E(Y) = \alpha + \beta X$ is equal to the **coefficient** β and specifies the amount of increase in the **ordinate** or **y axis** for each unit increase in the **abscissa** or **x axis**. It is analogous to the concept of grade or angle of inclination in surveying or road building.

smoothing– In **time-series analysis**, a statistical technique such as the construction of a **moving averages** series, that reduces or averages out fluctuations in a series.

smoothing constant– In **time-series analysis** and **forecasting**, a **parameter** employed in the **exponential smoothing** formula.

SMR– Acronym for *standardized mortality ratio*.

Snedecor's F distribution– Same as *F distribution*.

snowball sampling– A method of selecting a **sample** from a human population in which individuals selected in the sample are asked to provide information about other potential individuals to be included in the sample.

software– Same as *computer package*.

software package– Same as *computer package*.

Somer's D – An **asymmetric measure of association** in a **contingency table** where row and column **variables** are measured on an **ordinal scale**. The measure is appropriate when one variable is considered dependent and the other independent. See also *measure of association*, *symmetric measure of association*.

Spearman's rank correlation– Same as *Spearman's rho*.

Spearman's rank correlation coefficient– Same as *Spearman's rho*.

Spearman's rho (ρ)– A **correlation coefficient** between two **random variables** whose paired values have been replaced by their **ranks** within their respective **samples** or which are based on **rank order** measured on an **ordinal scale**. It provides a measure of the **linear relationship** between two **variables**. This measure is usually used for correlating variable(s) measured with rank-order **scores**. It is calculated by the formula

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of the i th pair. This correlation is equal to the coefficient of correlation when there are no ties. See also *Kendall's rank correlation*.

specific death rate– **Mortality rate** calculated for a specific subgroup of a population. See also *death rate*.

specificity– In a **screening** or **diagnostic test**, the **probability** that the test will yield a negative result when given to a person who does not have the disease or condition of interest. It is a measure of the goodness of a diagnostic test in detecting individuals who do not have the disease or condition in question. Compare *sensitivity*. See also *Bayes' theorem*, *receiver operating characteristic curve*.

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

specific rate– A **rate** calculated for a special group or segment of the population. Some examples are **age-specific fertility rate** and **cause-specific death rate**.

split-half method– A method of estimating the **reliability** of a test by dividing it into two comparable halves (usually the odd- versus even-numbered items), and then calculating the **correlation** between the **scores** of the two halves. In order to estimate the reliability of a test twice as long as each half, split-half correlations are increased by a factor to correspond to the length of the original test.

split-plot design– An **experimental design** that introduces an additional **factor** into the **experiment** by dividing an **experimental unit** known as a whole plot into smaller units called split plots or subplots. Any one of the experimental designs can be used for this purpose in which each unit can be divided into smaller units. For example, in industrial experimentation, **levels** of one factor may require a rather large bulk of experimental materials, such as types of furnaces for the preparation of alloys, but the levels of the other factor can be compared through use of small materials, such as the molds into which alloy is poured. Such an experiment can be run through the use of a split-plot design where large materials are applied to whole plots and small materials are applied to split plots. A split-plot design provides more precise information about one factor (whose levels are applied to split plots) and the **interaction** between the two, but less precise information about the other factor (whose levels are applied to the whole plots). The design given below shows a split-plot arrangement obtained by using a **randomized block design** in which whole plots within a block are used to allocate three levels of factor A and the split plots are used to allocate four levels of the factor B.

Layout of a split-plot design

Block I			Block II			Block III		
A_2	A_1	A_3	A_1	A_3	A_2	A_3	A_1	A_2
B_3	B_4	B_1	B_3	B_2	B_4	B_2	B_4	B_1
B_2	B_3	B_3	B_2	B_4	B_3	B_1	B_1	B_3
B_1	B_1	B_2	B_1	B_3	B_2	B_3	B_3	B_4
B_4	B_2	B_4	B_4	B_1	B_1	B_4	B_2	B_2

split-split-plot design– In a **split-plot design**, each one of the subplots may be further subdivided into a number of sub-subplots to which a third set of treatments may be applied. Such a design is known as split-split-plot design, where three sets of treatments are assigned to various levels of **experimental units** using three distinct stages of **randomization**. The details of statistical analysis follow the some general pattern as that of the split-plot design.

S-PLUS– A general-purpose, command-driven, highly interactive **software package**. It includes hundreds of functions that operate on **scalars**, **vectors**, **matrices**, and more complex objects. Statistical procedures available in S-PLUS are extremely versatile and offer powerful tools for comprehensive **data analysis**.

spot sample– A small **sample** taken on the spot without regard to its **randomness** or **representativeness**.

spread– Same as *variability*.

SPSS– A popular **statistical computing package** for data management and statistical analysis. This is an integrated system of **computer programs** initially developed for the analysis of social science data. It is an acronym for Statistical Package for the Social Sciences.

spurious correlation– High **positive** or **negative correlation** observed between two **variables** in spite of the original **observations** being made on uncorrelated **variates**. It is

usually caused by a third variable and there is no causal link between the two variables. When the effects of the third variable are removed, the observed correlation usually disappears.

SQC– Acronym for *statistical quality control*.

square matrix– A **matrix** having the same number of rows and columns.

square-root transformation– A **transformation** of the form $y = \sqrt{x}$ often used to stabilize **variance** of the **data** suspected to follow a **Poisson distribution**. If some of the **observations** are very small (particularly zero), the **homogeneity of variance** is more likely to be achieved by the transformations of the form $y = \sqrt{x + 0.5}$ or $y = \sqrt{\left(x + \frac{3}{8}\right)}$. See also *arc-sine transformation*, *logarithmic transformation*, *power transformation*, *reciprocal transformation*, *square transformation*.

square transformation– A **transformation** of the form $y = x^2$ that is often useful to stabilize the **variance** of a **data set** when the **distribution** is skewed to the left. See also *arc-sine transformation*, *logarithmic transformation*, *power transformation*, *reciprocal transformation*, *square-root transformation*.

stable population– A population that has been growing at a constant rate over a number of years.

standard deviation– A **measure of variability** or **dispersion** of a **data set** calculated by taking the positive square root of the **variance**. It can be interpreted as the average distance of the individual **observations** from the **mean**. The standard deviation is expressed in the same units as the **measurements** in question. It is usually employed in conjunction with the mean to summarize a data set. It is the most widely used measure of the dispersion and plays a central role in statistical theory and methods. It is commonly used to express the **spread** of the individual observations around the mean. See also *population standard deviation*, *sample standard deviation*.

standard deviation of the population– Same as *population standard deviation*.

standard error– The **standard deviation** of the **sampling distribution** of a **statistic** or the positive square root of the **sampling variance**. The standard error can be interpreted as a **measure of variation** that might be expected to occur merely by **chance** in the various characteristics of **samples** drawn equally randomly from one and the same **population**. Its magnitude depends on the **sample size** and **variability** of **measurements**. It indicates the degree of **uncertainty** in calculating an **estimate** from a **data set**. The smaller the standard error, the better the **sample statistic** is as an **estimate** of the **population parameter**.

standard error of the difference between sample means– The name given to the **standard deviation** of the **sampling distribution** of the difference between two **sample means**. The estimated standard error of the difference between sample means is used as the denominator in the **t test for independent samples**.

standard error of the mean difference– The standard error of the mean difference is the **standard deviation** of the **sampling distribution** of mean differences based on **paired**

data. The estimated standard error of the mean difference is used as the denominator in the *t* test for correlated samples.

standard error of the sample mean– The **standard deviation** of the **sampling distribution** of the **sample mean** is called the standard error of the mean. It is calculated by the formula σ/\sqrt{n} where σ is the **standard deviation of the population** and n is the **sample size**. Since all possible sample means are usually not available, one rarely works with the actual standard error of the means and generally uses an estimate based on **sample data**.

standard error of the sample proportion– The **standard deviation** of the **sampling distribution** of the **sample proportion** (\bar{p}) is called the standard error of the proportion. It is calculated by the formula $\sqrt{pq/n}$ where p is the **proportion** in the **population** having the characteristic, $q = 1 - p$ and n is the **sample size**.

standardization– The process of adjusting a crude **mortality** or **morbidity rate** in order to remove as far as possible the effects of differences in age, sex, ethnicity/race, or other **confounding variables** when comparing two or more populations. The rationale for standardization is the potential for **confounding** that exists in many **observational studies** and may lead to biased or erroneous results. The usual procedure involves computing **weighted averages** of **rates** applicable to different confounding variables according to specific **distribution** of these variables. There are two commonly used procedures for standardization, known as direct standardization and indirect standardization. In direct standardization, the **specific rates** of the **study population** are averaged by using weights as the distribution of a specified **reference** or **standard population**. In indirect standardization, the specific rates of the reference population are averaged by using weights as the distribution of the study population. This rate shows what the mortality or morbidity would be in the study population if it had the same distribution as the reference population with respect to the **variable** for which the adjustments are being made. For example, to compare cancer mortality rates between two populations, one younger and the other older, **age-specific mortality rates** from each of the two populations would be applied to the age distribution of a reference population to yield mortality rates that could be directly compared. In indirect standardization, the specific rates of the reference population are averaged by using weight as the distribution of the study population. This rate shows what the mortality in the reference population would be if it had the same distribution as the study population. In the example above, age-specific cancer mortality rates in the reference population would be applied separately to the age distribution of the two populations to determine the expected number of deaths in each. These would then be combined with the observed number of deaths in the two populations to determine comparable mortality rates. This method is normally used when the specific rates of the study population are unreliable or unknown. The term is also sometimes used in the context of standardizing a variable by dividing by its standard deviation so that the new variable has unit **variance**.

standardized coefficient– Same as *standardized regression coefficient*.

standardized death rate– A measure of **mortality** of a population that takes into account the age and sex composition of the population involved. Compare *crude death rate*. See also *standardization*.

standardized deviate– The value of a **deviate** that is reduced to standardized form (zero mean and unit **variance**) by subtracting the mean and then dividing it by the **standard deviation**. See also *standard normal deviate*, *standard score*.

standardized event rate– A **mortality** or **morbidity rate** commonly adjusted for age and sex distribution of the population. See also *standardization*.

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

standardized mortality ratio– It is the **ratio** of the observed to the expected number of deaths in the **study population** if it had the same age and sex-specific rate structure as the **standard population** (expressed per 1000).

standardized rate– See *standardization*.

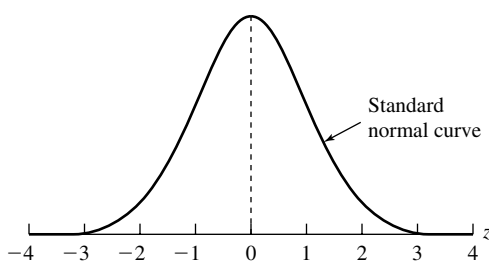
standardized regression coefficient– A **regression coefficient** that removes the effect of the **measurement scale** so that the relative size of the **coefficients** can be compared. A standardized regression coefficient measures the change in the **dependent variable** for an increase of one **standard deviation** in the **independent variable**. It can be compared directly with another and with the beta coefficients of other **regression models**. It is calculated by using **standard scores** for all the variables. It can also be obtained from the corresponding (raw) regression coefficient by multiplying it by the standard deviation of the independent variable.

standardized score– Same as *standard score*.

standard normal curve– It is a curve represented by the **probability density function** of the **normal distribution** with a **mean** of zero and a **standard deviation** of one. Some important properties of the standard normal curve are:

- The total area under the standard normal curve is equal to 1.
- The standard normal curve extends indefinitely in both directions, approaching but never touching the **horizontal axis**.
- The standard normal curve is symmetric about 0. That is, the part of the curve to the left of the vertical line through 0 is identical to the part of the curve to the right of it.
- Almost all the area under the standard normal curve lies between -3 and 3 .

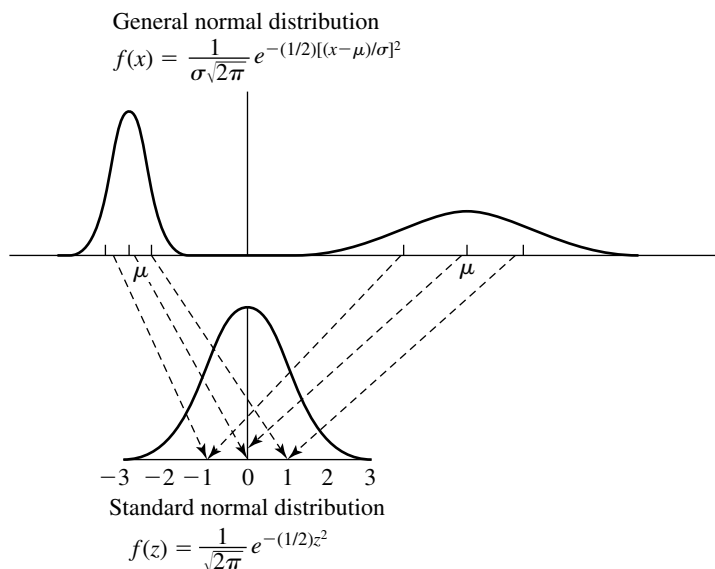
See also *normal curve*, *standard normal distribution*.



The standard normal curve

standard normal deviate– The values of the **deviation** from the **mean** of any normally distributed **random variable** measured in units of **standard deviations**.

standard normal distribution– A **normal distribution** with a **mean** of zero and a **standard deviation** of one is called the standard normal distribution. A general normal distribution with mean μ and standard deviation σ can be converted to the standard normal distribution by the **linear transformation** $z = (x - \mu)/\sigma$.



Schematic diagram illustrating a linear transformation of a general normal distribution to a standard normal distribution

standard normal probability density function– The **probability density function** of a **standard normal distribution**. It is represented by the mathematical equation

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2} \quad -\infty < x < \infty$$

standard normal variable– A **random variable** having the **standard normal distribution**.

standard partial regression coefficient– In a **multiple regression analysis**, the **standardized regression coefficient** of an **independent variable** in the **regression equation** involving all the independent variables under consideration. A standard partial regression coefficient measures the change in the **dependent variable** for an increase of one **standard deviation** in the independent variable when the values of other independent variables are kept constant. See also *estimated partial regression coefficient*, *partial regression coefficient*.

standard population– See *reference population*.

standard score– A standard or *z* score, like a **percentile rank**, is used to express the relative standing of a **score** with respect to the **distribution** to which it belongs. The **mean** of any standard score is always 0 and the **standard deviation** is always 1. The standard score for a particular **raw score** expresses its distance from the mean, expressed in units of standard deviation. It is calculated by subtracting the mean from each score and dividing by the standard deviation.

STATA– A general-purpose, command-oriented interactive statistical and graphical **software package**. It is one of the most complete and comprehensive software packages for routine **data analysis**. It is especially useful for longitudinal and epidemiological data

analysis. Graphical capabilities of STATA include numerous charts, graphs, and plots for **quantitative** and **qualitative data**.

stationary population– A population with no migration and for which the **crude birth rate** is equal to the **crude death rate**.

statistic– A numerical value used as a **summary measure** for a **sample** calculated according to certain rules or procedures. Some examples are the **sample mean** and the **sample standard deviation**. A statistic when derived to estimate some **parameter** is called an **estimator**.

STATISTICA– A general-purpose, menu-driven **statistical software package**. The package contains well-integrated modules for data management, statistical analysis, and high-quality graphics for **numerical** and **qualitative data**. It supports a wide variety of statistical procedures for routine as well as specialized **data analysis**.

statistical algorithms– The term is employed to refer to the **algorithms** having useful applications to problems encountered in **statistics**.

statistical computing package– Same as *computer package*.

statistical data– Same as *data*.

statistical description– A term used to refer to the use of **descriptive statistics** in describing a **data set**.

statistical estimation– Same as *estimation*.

statistical hypothesis– A proposition or statement about one or more **population**. A statistical hypothesis stems from questions, such as, “Does cigarette smoking cause lung cancer?” “Is treatment A better than treatment B in treating a disease?” See also *alternative hypothesis, hypothesis, null hypothesis*.

statistical inference– Same as *inferential statistics*.

statistically nonsignificant– In **hypothesis testing** any **sample** result that does not lead to the rejection of the **null hypothesis**. A nonsignificant result should not be interpreted as the “null hypothesis is true” but rather as “the data have not shown that the null hypothesis is false.” See also *p value, statistically significant, statistical significance*.

statistically significant– In **hypothesis testing**, any **sample** result that leads to the rejection of the **null hypothesis** because it has a low **probability** of occurring when that **hypothesis** is true is called statistically significant. Thus, when a sample result is declared statistically significant, it means that the result deviates from some hypothetical value by more than can be reasonably attributed to the **chance** errors of **sampling**. See also *p value, statistically nonsignificant, statistical significance*.

statistical map– A **graphical representation** of **data** for area units by such devices as differentiated cross-hatching or shading of these units on a geographic map.

statistical measure– Same as *statistical description*.

statistical model– Same as *stochastic model*.

statistical package– Same as *computer package*.

statistical population– Same as *population*.

statistical power– Same as *power*.

statistical process control– Same as *quality control*.

statistical quality control– Same as *quality control*.

statistical significance– Said of a result of **hypothesis testing** if the value of the **test statistic** used to test it is smaller or larger than the value that would be expected to occur by **chance** alone, assuming that the **null hypothesis** is true. It is generally interpreted as a result that would occur by chance less than 1 time in 20, with a **p value** less than or equal to 0.05. It is said to occur when the investigator rejects the null hypothesis. When this happens, conclusions based on a **sample** of **observations** also hold true for the **population** from which the sample is selected. See also *statistically nonsignificant*, *statistically significant*.

statistical software– Same as *computer package*.

statistical software package– Same as *computer package*.

statistical table– A presentation of numerical facts usually arranged in the form of columns and rows. A statistical table either summarizes or displays the results of a statistical analysis.

statistical test– A statistical procedure or any of several tests of **statistical significance** used to test a **null hypothesis**. The test assesses the compatibility of the **experimental data** with the null hypothesis. The procedure rejects the null hypothesis if an observed difference (or a more extreme one) would have a small **probability** if the null hypothesis were true. Some examples of statistical tests are *t*, χ^2 , and *F* tests.

statistical tolerance intervals– A statistical tolerance interval establishes limits that include a specified **proportion** of the response in a **population** or a processes with a prescribed degree of confidence.

statistician– A person trained in statistical methods and **data analysis**. Statisticians are found in a variety of fields, ranging from business and engineering to psychology and medicine.

statistics– A field of study that is concerned with making decision in the face of **uncertainty**. In particular, it is the study of inferential process, especially the planning and analysis of **experiments** and **surveys**. It develops and utilizes techniques for the collection, presentation, analysis, and interpretation of **numerical data** relating to aggregates of individuals. The term is also applied to the numerical data themselves. See also *descriptive statistics*, *inferential statistics*.

STATXACT– A powerful **statistical package** for personal computers that supports exact inference for the analysis of **binary**, **categorical**, and **continuous data**. The programs in STATXACT produce exact **p values** and **confidence intervals** for small **sample data**. It includes over 80 **test procedures** covering all the important problems of interest to a data analyst. A related package is LOGXACT, which provides exact inference for **logistic regression models**, including conditional and unconditional inferences. It produces *p* values and confidence intervals that remain valid for small **samples**.

steepest descent– An optimization **algorithm** for finding the maximum or minimum value of a function of several variables by looking in the direction of positive (negative) gradient of the function with respect to the **parameters**. See also *Newton–Raphson method*, *simplex method*.

stem-and-leaf diagram– Same as *stem-and-leaf plot*.

stem-and-leaf plot– An **exploratory data analysis** technique pioneered by John W. Tukey that simultaneously rank-orders the **data** and provides representation of the shape of the underlying **frequency distribution**. It presents **raw data** in a **histogram**-like display and combines features of both a **frequency table** and a histogram. For example, consider the following data on cholesterol levels of 20 patients in an hypothetical study:

Cholesterol levels of 20 patients (in mg/100 mL) in a hypothetical study

211	210	213	209	218	208	211	204	209	211
211	200	216	222	214	219	203	219	201	215

To construct a stem-and-leaf plot, since these data are three-digit numbers, we use the first two digits as the stems and the third digit as the leaves. A stem-and-leaf plot for the cholesterol levels is then given as follows:

Stem-and-leaf plot for cholesterol level data

20	9	8	4	9	0	3	1					
21	1	0	3	8	1	1	1	6	4	9	9	5
22	2											

The stem-and-leaf plot displayed above is not very useful because there are very few stems. We can construct a better stem-and-leaf plot by using two lines for each stem, with the first line for the leaf digits 0 to 4 and the second line for the leaf digits 5 to 9. This stem-and-leaf plot is shown below.

**Stem-and-leaf plot for cholesterol level data
using two lines per stem**

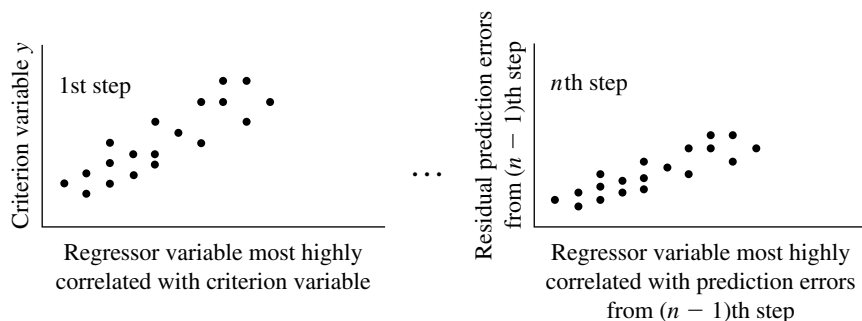
20	4	0	3	1				
20	9	8	9					
21	1	0	3	1	1	1	4	
21	8	6	9	9	5			
22	2							
22								

See also *back-to-back stem-and-leaf plot*.

stepwise procedure– Same as *stepwise regression*.

stepwise regression– In a **multiple regression analysis**, a technique for selecting the “best” set of **independent variables** to be included in the final **regression model** by entering or deleting **regressor variables** sequentially in various combinations and orders. The technique begins by including the **variables** one at a time (forward) or by starting with the entire set of variables and deleting them one at a time (backward). Thus, the stepwise regression combines **forward selection** and **backward elimination procedures**. Variables are selected and eliminated until there are no more that meet the criterion. The criterion for entering or deleting a variable depends on the extent to which it alters the **multiple correlation coefficient** or, equivalently, the **error variance**. The rationale behind the stepwise regression is the need to develop a parsimonious prediction model that excludes highly

correlated and redundant **predictor variables**. A researcher usually collects information on a number of potential **explanatory variables** and wishes to find which of them provides a stable and optimum predictive model.



Schematic illustration of the stepwise regression procedure

stillbirth— A late fetal death; i.e., a fetal death that occurred after 28 completed weeks of gestation.

stillbirth rate— The number of **stillbirths** actually observed during a given calendar year divided by the total births occurring during the calendar year (expressed per 1000). Perinatal mortality rate is based on stillbirths plus deaths in the first year of life.

Stirling's formula— A highly accurate mathematical formula for evaluation of the values of $n!$ (factorial). It is given by $n! = \sqrt{2\pi} n^{n+0.5} e^{-n}$. It gives an asymptotic approximation of $n!$ in the sense that $n!/\sqrt{2\pi} n^{n+0.5} e^{-n} \approx 1$. For $n = 5$, the percentage error in using the formula is roughly 2%. For $n = 10$, it is 0.8%, and for $n = 100$ it is just 0.08%. Stirling's formula is used in calculating **probabilities** of a **binomial distribution** for large values of n .

stochastic independence— In **probability theory**, this term is synonymous with **mutual independence**. See also *independent events*, *independent random variables*, *pairwise independence*.

stochastic model— A **mathematical model** containing **random** or probabilistic elements. It is based on a **stochastic relationship** between two or more **variables** where specific statistical **assumptions** are made to allow for **error**. Compare *deterministic model*.

stochastic process— A physical process that is governed at least in part by some **random** mechanism. In a stochastic process, the **probabilities** of the occurrence of an **event** change over time and one is especially concerned with interdependence and limiting behavior of **empirical probabilities**. A stochastic process can be discrete or continuous in time, and its value at any given time can be a value of a **discrete** or a **continuous variable**. An example of a stochastic process is provided by the growth of populations such as bacterial colonies.

stochastic relationship— A relationship between any two **variables**, X and Y , such that any possible values of Y can be associated with any one value of X .

stochastic variable— Same as *random variable*.

stopping rule— A procedure for performing **interim analysis** at certain specified periods of time.

strata— Levels of a **categorical variable** such as age, sex, or age–sex groups, where each **stratum** corresponds to a single level or combination of levels of one or more **factors**. See also *stratification*.

stratification— The division of a **population** into a number of subpopulations commonly known as **strata**. Stratification is normally used for the purpose of drawing a **stratified sample**. The term is also used to describe the process of performing a statistical procedure separately in groups (strata) in order to reduce the effects of the **stratifying variable**. Thus, separate **estimates** and **significance tests** for each **stratum** of a **confounding variable** are performed in order to produce a single estimate or **test statistic** across all strata. In **clinical trials** or other **experimental studies**, the term is used for the creation of strata for the purpose of implementing a **stratified randomization**.

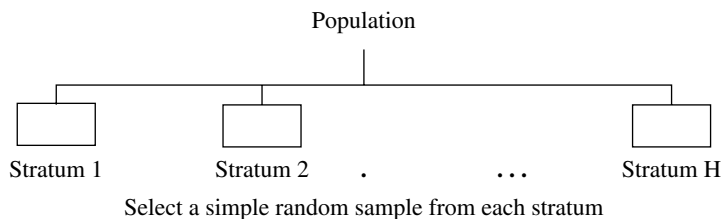
stratified analysis— A term commonly used in epidemiologic **data analysis** to refer to a statistical procedure for evaluating and removing **confounding** by stratifying a **sample** into a series of **strata** that are homogenous with respect to the **confounding variable**. See also *Mantel–Haenszel chi-square test*, *Mantel–Haenszel estimator*.

stratified logrank test— A **nonparametric statistical procedure** for comparing two **survival curves** when the subjects are stratified by age, sex, or some other **prognostic variable**. See also *logrank test*.

stratified randomization— A method of **randomization** in which subjects are classified by sex and age, usually 5- or 10-year age groups. Then the subjects in each sex and age **stratum** are randomly assigned to one of the two **treatment groups** from that stratum. Sometimes, patients are also stratified by severity of disease, inasmuch as severity generally has an effect on the outcome of the disease. Participants are then assigned to a **treatment** or **control group** within each category of severity. The goal of stratified randomization is to achieve approximate balance of important **prognostic factors** while retaining the advantages of randomization. Stratified randomization in conjunction with **block randomization** reduces the **variability** due to the stratifying variables.

stratified (random) sample— A **sample** consisting of **random samples** selected from each **stratum** or subpopulation of a **population**. It is used to ensure that each subpopulation of a large heterogeneous population is appropriately represented in the sample. A stratified random sample usually leads to better **precision** than the **simple random sample**.

stratified (random) sampling— A **sampling** procedure in which the **population** is first divided into parts, known as **strata**, and a **simple random sample** is selected from each one of the strata. The procedure gives every individual in a **stratum** an equal and independent chance of appearing in the **sample**. The strata are formed such that they are internally homogenous, but differ from one another with respect to some characteristics of interest. For example, in the sample used for the **Current Population Survey**, which is conducted monthly by the **U.S. Bureau of the Census**, all the 31,000 counties in the United States are classified into 333 strata and sample counties or groups of counties are chosen from each stratum. In the construction of strata, such characteristics as geographic area, **population size**, income, occupation, and race/ethnicity are taken into account, so that the counties in any given stratum are similar. The goal of a stratified random sampling is to select a sample that is representative of all strata in a given population and to minimize the size of the whole sample for a given level of representativeness. Usually, the same **proportion** of individuals is selected from each stratum, so that the composition in the population is reflected in the sample. See also *stratified random sample*.

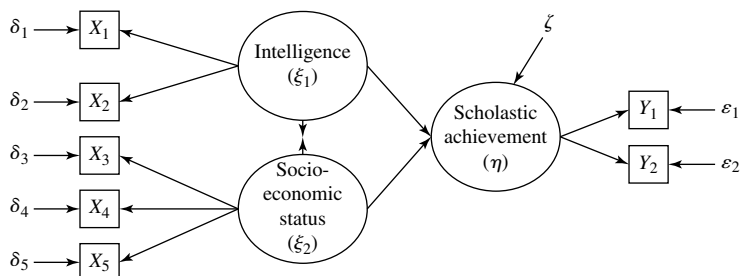


Schematic diagram for stratified random sampling

stratifying variable— A **variable** used to create **strata** for the purpose of drawing a **stratified random sample** or to control for **confounding** in an epidemiological study. Some examples of a stratifying variable are age, sex, income, or geographical boundary.

stratum— A single subpopulation formed by a single **level** or combination of levels of one or more **factors**; a singular form of **strata**.

structural equation model— The structural equation model refers to a method of analyzing relations between the sets of **endogenous** and **exogenous variables**. The procedure consists of the combined application of **multiple regression** and **factor analysis** to investigate the relationships between the **variables**. Equations describing **causal relations** among the variables are formulated and estimated by the **method of maximum likelihood** or **least squares theory**. Most often the endogenous and exogenous variables used in a structural equation model are theoretical constructs or **latent variables**. The purpose of the analysis is to assess the adequacy of the **causal model** proposed by the researcher. See also *LISEREL*, *path analysis*.



The observed variables are the indicators of intelligence, X_1 = Wechsler IQ score and X_2 = Stanford–Binet IQ score; the indicators of socioeconomic status, X_3 = father's education, X_4 = mother's education, and X_5 = parent's total income; and the indicators of scholastic achievement. Y_1 = verbal score and Y_2 = quantitative score on a scholastic achievement test

Schematic illustration of a structural equation model for scholastic achievement as endogenous and intelligence and socioeconomic status as exogenous latent variables

Studentized range— Same as *studentized range statistic*.

Studentized range statistic— A statistic defined by the formula

$$q = \frac{\bar{x}_{\max} - \bar{x}_{\min}}{\sqrt{\text{MSE}/n}}$$

where \bar{x}_{\max} and \bar{x}_{\min} are the maximum and minimum values among a set of **group means** and **MSE** is the **error mean square** from an **analysis of variance** of the groups. It is widely used in **multiple comparison** tests.

Studentized residuals– Same as *jackknife residuals*.

Student's *t* distribution– Same as *t distribution*.

Student's *t* statistic– Same as *t statistic*.

Student's *t* test– Same as *t test*.

study design– A logical plan for selecting a **sample** and collecting **data** necessary to answer a research question by estimating **parameters** or **testing hypotheses**. Some examples of a study design are **clinical trial**, **cohort study**, and **case-control study**.

study group– Same as *study sample*.

study population– A **population** used to select the **study sample**. Sometimes the term is used to refer to the group of people from whom data are collected. See also *sampled population*, *target population*.

study sample– A **sample** of subjects selected to undertake a study.

study subjects– Same as *study sample*.

sturdy statistics– Same as *robust statistics*.

Sturges' rule– A general rule for determining the number of **class intervals** in a **grouped frequency distribution**. It is given by the formula $k = 1 + 3.322 \log_{10} n$ where k denotes the number of class intervals and n is the number of values in the **data set**. For example, with $n = 29$, the rule gives $k = 1 + 3.322 \log_{10} 29 \approx 5$ groups.

subgroup analysis– Statistical analysis performed on a subgroup of cases with certain common characteristics, such as males, females, elderly, and urban/rural. The purpose of a subgroup analysis is to know whether the results of an analysis differ from one group of cases to the other.

subjective probability– The definition of **probability** based on an individual's subjective judgment or belief in the occurrence or nonoccurrence of an **event** or phenomenon. It expresses a purely personal degree of belief in the likelihood of specific occurrence of an event or phenomenon. A subjective probability may differ from one individual to the other who may assign different probabilities to the same event. Subjective probabilities are useful in **bayesian inference** to develop **prior distributions** for the **parameters** of interest. See also *empirical probability*, *objective probability*, *posterior probabilities*, *prior probabilities*.

substantive significance– Same as *practical significance*.

sufficient statistic– A **statistic** that in certain sense contains all the information obtainable from a **sample** of **observations** about a particular **parameter** it is used to estimate.

summary indices– Numerical values summarizing a set of **observations**.

summary measures– **Descriptive statistics**, such as, **mean**, **median**, **proportion**, **standard deviation**, etc.

summary statistics– Same as *summary measures*.

sum of squares– In an **analysis of variance**, the sum of squared deviations around a particular mean, i.e., either the **grand mean** or the individual **group mean**. See also *sum of*

squares between groups, sum of squares for columns, sum of squares for error, sum of squares for interaction, sum of squares for rows, sum of squares for total, sum of squares within groups.

sum of squares between groups– In a **one-way analysis of variance**, the sum of the squared deviations of the **group means** from the **grand mean**. It is calculated by subtracting each group mean from the grand mean, squaring these differences for all items and then summing them.

sum of squares due to regression– Same as *regression sum of squares*.

sum of squares due to residuals– Same as *residual sum of squares*.

sum of squares for columns– In a **two-way analysis of variance**, the **variability** between **treatments**, which are represented in the columns, calculated as the sum of the squared deviations of the column means from the **grand mean**, weighted by the number of cases in the column.

sum of squares for error– In a **two-way analysis of variance**, the **variability** due to individual differences between subjects, **measurement errors**, uncontrolled variations in experimental procedures, and so on, calculated by subtracting the **sum of squares for rows, columns, and interaction** from the **total sum of squares**.

sum of squares for interaction– In a **two-way analysis of variance**, the **variability** due to the **interaction** between the two experimental **factors**.

sum of squares for rows– In a **two-way analysis of variance**, the **variability** between blocks of subjects, which are represented in the rows, calculated as the sum of the squared deviations of the row means from the **grand mean**, weighted by the number of cases in the row.

sum of squares for total– In an **analysis of variance**, the overall sum of squared deviations within all the groups. It is obtained by subtracting each individual **observation** from the **mean** of all observations, squaring, and summing these values.

sum of squares for treatment– In an **analysis of variance**, the sum of the squared deviations between each **treatment mean** and the **grand mean**. It is the component of the **total sum of squares** that can be attributed to possible differences among the **treatments**.

sum of squares within groups– In a **one-way analysis of variance**, the overall sum of squared deviations within all the groups. It is obtained by subtracting each **observation** from its **group mean**, squaring these differences, and then summing them.

suppression of zero– A term used for choice of scales in a misleading graph that does not use a break or a jagged line in the **y axis** to show that part of the scale which has been omitted.

surrogate outcome– A term used in **clinical trial** to refer to an **outcome measure** that can be used as a substitute for a definitive clinical **outcome** or disease. In order to be useful, a surrogate outcome should be highly correlated with the outcome of interest. Some examples include prostatic specific antigen (PSA) as a surrogate for prostate cancer and blood pressure as a surrogate for cardiovascular disease.

survey– A research or study of a **population**, usually human subjects, to collect **data** regarding social, economic, or political issues of the day without any particular control over

other factors that may affect the characteristics of interest being observed. The information collected is usually of quantitative nature or a type that can be summarized in quantitative terms. Surveys are **observational studies** usually conducted by studying a cross section of the **target population**. In order to ensure **reliability** of the results, it is important that surveys are conducted by using a **probability sample**. See also *opinion survey*, *sample survey*

survey data– Data obtained from a **sample survey** rather than from the enumeration of the entire population.

survey design– Same as *sampling design*.

survey research– Same as *survey*.

survey sampling design– Same as *sampling design*.

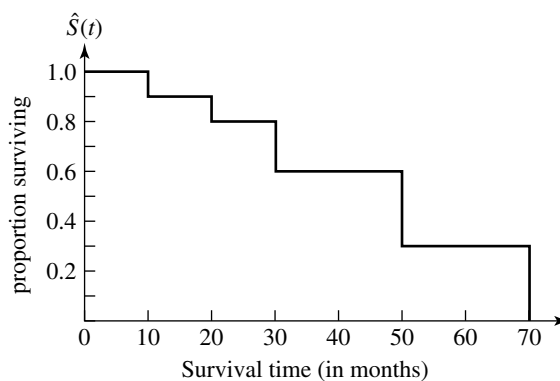
survival analysis– The statistical methods for analyzing survival data when there are **censored observations**. Survival analysis focuses on how long subjects persist (survive) in a given state. It has been used in demography to study life expectancy and in medical research to study the duration of illnesses and making inferences about the effects on it of **treatments, prognostic factors, exposures** and other **covariates**. The main aim of survival analysis is to make **inference** about the distribution of survival time. Survival data usually involve **censoring** in which the **outcome** for some individuals is not known at the end of the study period. In addition, the **follow-up period** may also vary from one subject to the other. The usual **summary statistics** such as **proportion** (people cured or died) and **mean** (survival time) are not appropriate in analyzing survival data. The appropriate methods of analyzing survival data include **life table analysis, Kaplan-Meier estimator, Cox regression, and logrank test** among others.

survival curve– See *survival function*.

survival data– See *survival analysis*.

survival function– In **survival analysis**, if X denotes the period of time for a specified event (such as death or relapse) to occur, then the survival function gives an **estimate** of $S(t) = \Pr\{X > t\}$ for each time period t . A plot of this **probability** against time is called a **survival curve**. See also *hazard function*.

survival probability– Same as *probability of survival*.



Example of a survival curve

survivor function– Same as *survival function*.

syllogism– A kind of **deductive reasoning** involving a formal argument that consists of two premises, followed by a conclusion. For example, all men are mortal, all kings are men; therefore all kings are mortal.

symmetrical distribution– A **distribution** is said to be symmetrical if a vertical line drawn from the center divides it into two equal halves. In a symmetrical distribution, the values having equal distance from the **mean** have the same **frequencies, probabilities, and probability densities**. Thus, a symmetrical distribution has the same **shape** on both sides of the mean. Compare *asymmetrical distribution*. See also *skewed distribution*.

symmetrical population– A **population** or theoretical **distribution** that is symmetrical.

symmetric matrix– A **square matrix** that is symmetrical about its leading diagonal. Thus, a matrix **A** is symmetric if $A' = A$, that is $a_{ij} = a_{ji}$ for all pairs i and j . **Correlation matrix** and **covariance matrix** are examples of a symmetric matrix.

symmetric measure of association– A **measure of association** that does not depend on the choice of **independent** and **dependent variables**. Compare *asymmetric measure of association*.

symmetry– The property of the **shape** of a **frequency** or **probability distribution** that exhibits similarity of form or arrangement on either side of a dividing line or plane. Compare *asymmetry*. See also *symmetrical distribution*.

synergism– A term often used in medical and epidemiological studies when the combined **effect** of two **treatments** is greater or less than the sum of their separate individual effects. When the combined effect is greater than the sum of their effects, it is called positive synergism; when it is less than the sum of their effects, it is called negative synergism or antagonism.

synergistic effect– See *interaction*.

synthetic birth cohort– An artificial **birth cohort**, composed of a cross-sectional **sample** of the **population**. A real birth cohort is a group of births occurring at the same time.

SYSTAT– A general-purpose **statistical software package** for personal computers. It offers an extremely flexible language and contains an extensive list of statistical procedures with powerful graphical capabilities.

systematic allocation– A procedure for assigning **treatments** to subjects by using some systematic scheme such as assigning the **active treatment** to those with even birth dates and **control treatment** to those with odd dates. A systematic allocation is not the same as **random allocation**.

systematic error– A nonrandom **error** that introduces a **bias** into all the **observations**. As opposed to a **random error**, a systematic error is the same (or **constant**) over all the observations. It is usually caused by faulty or poorly adjusted measuring instruments.

systematic review– A review of individual research studies in terms of design, data collection, and results, performed to answer a particular research question. The term is also commonly used as a synonym for **meta-analysis**.

systematic sample– A **sample** obtained by using a systematic method of **sampling**.

systematic sampling— A **sampling** procedure in which a **sample** is obtained by selecting from a list of sampling units every k th subject or object at equally spaced intervals. The size of k , called the sampling interval, is obtained by dividing the **population size** N by the desired **sample size** n ; i.e., $k = N/n$. A systematic sampling from an area is carried out by determining a pattern of points on a map and then selecting the desired sample of points in a systematic manner. For example, suppose one wants to select a **systematic sample** of 100 cases from a list of 10,000 items. One would first divide 10,000 by 100 to get 100, and then randomly select a number between 1 and 100, say 27. Finally, one would select the 27th item from the list and every 100th item thereafter, i.e., the 127th, the 227th, the 327th, and so on. Systematic sampling provides a useful alternative to **simple random sampling** because it is easier to perform in the field and can provide greater information per unit cost than **simple random sampling**. It is commonly used in a wide variety of contexts, e.g., sampling of dwellings from a list of city blocks, sampling of manufactured items moving along an assembly line, sampling from a list of accounts to check compliance with accounting procedures, sampling customers at checkout counters for their opinion on food products, and so forth. If the elements of a population are distributed in a **random** order, then systematic sampling gives results that are equivalent to simple random sampling. If the elements of a population are ordered in magnitude according to some scheme, then systematic sampling provides more information per unit cost than does a simple random sampling. Finally, if the elements of the population have **cyclical variation**, then systematic sampling provides less information per unit cost than does simple random sampling.