

cake diagram– Same as *pie chart*.

canned program– An old term used to describe a **computer program** written and documented so that the user needs only a data deck and the proper calling cards to access and run the particular program of interest.

canonical correlation– See *canonical correlation analysis*.

canonical correlation analysis– A **multivariate statistical technique** for examining the relationships between two sets of numerical measurements made on the same set of subjects. The technique involves grouping the two sets of **independent** and **dependent variables** into linear composites which are a weighted combination of **predictor variables** and a weighted combination of **criterion variables**. It then calculates a **bivariate correlation** known as a canonical correlation between the two composites. The technique can be considered an extension of **multiple regression analysis** to situations involving more than a single dependent variable. It can also be viewed as an analogue of **principal components analysis** where a **correlation** rather than a **variance** is maximized. Canonical correlation analysis is a useful and powerful technique for exploring the relationships among multiple predictor (independent) and multiple criterion (dependent) variables.

capture–recapture sampling– A **sampling** scheme especially designed for the **estimation** of size of a wildlife population such as fish in a lake or birds in a sanctuary. The procedure involves the selection of an initial **sample** of animals that are marked and then released and allowed to mix with the population. Subsequently, a second sample is taken and the **proportion** of marked animals is determined. From this proportion the total number of animals is estimated by using the relation between the **parameters** of a **hypergeometric distribution**. For example, let n_1 be the size of the first sample, n_2 be the size of the second sample, and m the number of marked animals in the second sample. Then an estimator of the total number of animals is given by $\hat{N} = n_1 n_2 / m$. The estimator is sometimes known as the Petersen estimator.

carryover effect– In **crossover studies**, a carryover effect occurs when the **treatment** given in one period of the **trial** continues to exert its effect into the following period.

Carryover effects may lead to **treatment-period interactions**, and it is generally important to assess the relative importance of the effects attributable to the treatment given in a period compared to the period given in the previous period. In order to minimize the influence of carryover effects, **washout periods** of appropriate length must be allowed between two consecutive treatments.

Cartesian coordinate— A point that is located by measuring the distances from the coordinate axes (**x axis** and **y axis**) on a two-dimensional graph.

Cartesian graph— A graph drawn in a *cartesian plane*.

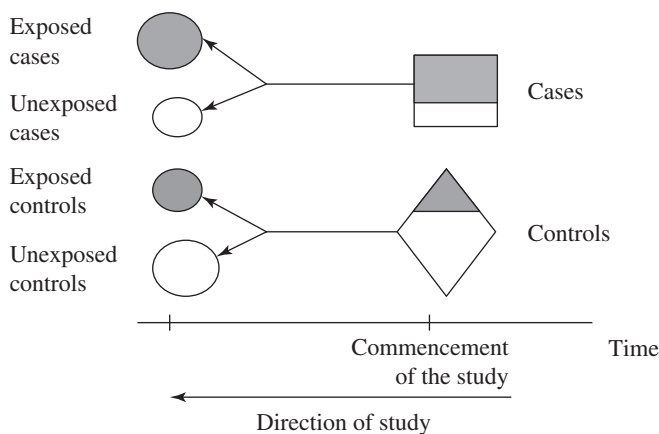
Cartesian plane— A plane whose points are labeled with **Cartesian coordinates**.

Cartesian product— The set of ordered pairs (x, y) of real numbers.

Cartesian space— Same as *cartesian plane*.

case— A term used most frequently in **epidemiology** to designate an individual in a **study population** having a certain disease or condition of interest.

case-control study— An **observational study** that entails **patient cases** who have a certain outcome or disease under investigation and comparable **control subjects** who do not have the outcome or disease. It then examines backward to identify possible **etiologic** or **risk factors**. Such a study is also called a **retrospective study** because it starts after the onset of disease and looks retrospectively to identify **risk** or **causal factors**. Case-control studies are often used to investigate the relationship between an **exposure** or risk factor and one or more **outcomes**. They are particularly useful in the study of rare disorders and infectious disease outbreaks. However, case-control studies are prone to some common sources of **bias**, such as **selection bias** and **recall bias**, among others. See also *Berkson's fallacy*, *cohort study*, *cross-sectional study*, *prospective study*.



Schematic diagram of a case-control study

case-fatality rate— This **rate** is designed to measure the **probability** of death among diagnosed **cases** of a disease. It is obtained as the **proportion** of cases of the disease who die during the same time period. More specifically, it is given by the number of deaths from a disease in a given period divided by the number of diagnosed cases of that disease in the same period.

case report– Published report describing a detailed clinical case history of **cases** which are unique or rare in certain aspects.

case-series study– A simple narrative description or **case report** of certain interesting or intriguing observations that occurred in a small group of patients. Case-series studies frequently lead to generation of **hypotheses** that are subsequently tested in a **case-control**, **cross-sectional**, or **cohort study**. See also *retrospective study*.

categorical data– See *categorical variable*.

categorical observations– Same as *categorical data*.

categorical variable– A **variable** whose values are categories or groups of objects as **measurements**. Examples of categorical variables are sex (male or female), marital status (married, single, divorced, etc.), and blood group (A, B, AB, O). For convenience of data collection and analysis, the categories are often assigned numerical labels, but they have no quantitative significance whatsoever. The values of a categorical variable are known as categorical data or observations. See also *qualitative variable*, *quantitative variable*.

Cauchy–Schwartz inequality– Given two **random variables** X and Y having finite second **moments**, the Cauchy–Schwartz inequality states that

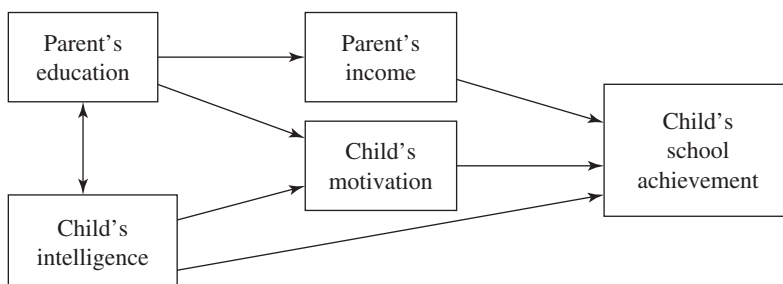
$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

As a corollary to the above inequality it follows that $|\rho| \leq 1$ where ρ is the **correlation coefficient** between X and Y . In general, if a_i 's and b_i 's are real integers, then it follows that

$$\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) \geq \left(\sum_{i=1}^n a_i b_i\right)^2$$

causal analysis– A method, such as **path analysis** or **latent variable modeling**, that analyzes **correlations** among a group of **variables** in terms of predicted patterns of **causal relations** among them.

causal diagram– A **graphical representation** of the **cause–effect relationship** between **variables**. In a causal diagram, paths in the form of unidirectional or bidirectional arrows are drawn from the variables taken as causes (independent) to the variables taken as effects (dependent). The **correlation** between two **exogenous variables** is depicted by a curved line with an arrowhead at both ends.



Schematic illustration of a causal diagram of child's school achievement via links between the child's intelligence, child's motivation, parent's education, and parent's income

causal factor– Same as *causal variable*.

causal inference– A form of **inference** used for assessing a **causal relationship** by designing a valid **experiment**.

causality– The term is most commonly used to describe a **cause–effect relationship** between **variables**. Many investigations in social, medical, and health sciences purport to establish causality between certain **events**; for example, cigarette smoking and lung cancer. See also *causal analysis, causal diagram, causal model, causal modeling, causal variable*.

causal model– A **mathematical model** describing **causal relations** among sets of **exogenous** and **endogenous variables**. See also *path analysis, structural equation model*.

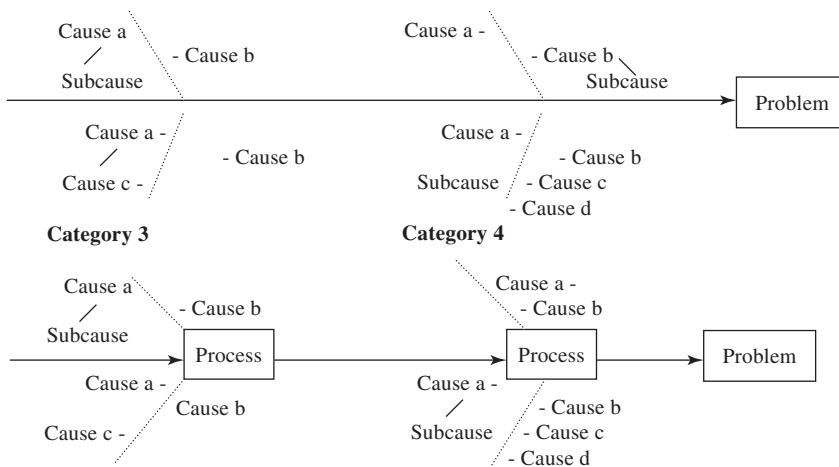
causal modeling– A method of analysis of **causal relations** among sets of **exogenous** and **endogenous variables**. **Path analysis** and **structural equation models** are examples of causal modeling.

causal relation– Same as *cause–effect relationship*.

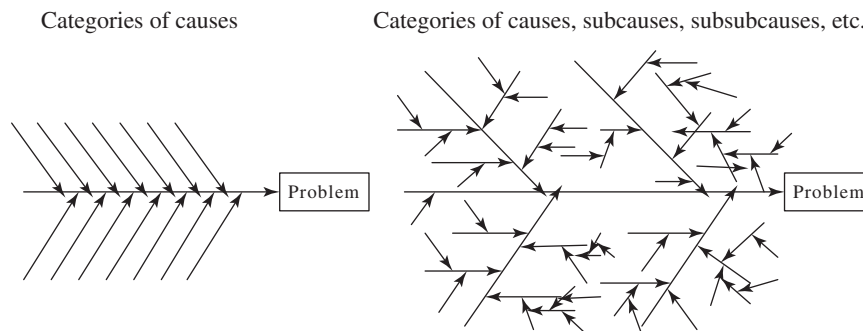
causal relationship– Same as *cause–effect relationship*.

causal variable– A **variable** that brings about changes in a given variable. A causal variable is treated as an **independent variable**. See also *causal diagram, causal model, causal modeling*.

cause-and-effect diagram– A **graphical device** that is used to identify, display, and examine possible causes of a poor quality or an undesirable condition present in a system or process. It is also known as an Ishikawa diagram, after K. Ishikawa, who first popularized its use during the mid-40s. The five common causes of a poor quality are environment, materials, manpower, machines, and methods. The steps in constructing a cause-and-effect diagram can be summarized as follows: (1) Identify the quality characteristic for which a cause-and-effect relationship is to be established. (2) Using the experience of knowledgeable people, generate several major categories of causes that can affect the quality. (3) For each of those major categories of causes, identify the possible causes that fall within that category and insert these subcauses into the diagram via horizontal lines emanating from the major category names.



Schematic illustration of two hypothetical cause-and-effect diagrams



Schematic illustration of a simple and a complex cause-and-effect diagram

cause-effect relationship— A term used to describe the **association** between two **variables** whenever it can be established that one of the variables causes the other. Statistical analysis has a long way to go toward establishing the existence of a cause-effect relationship between any two variables. It cannot establish the nature of any causal relationship, nor can it be used for proving that any two variables are not causally related. Statistical analysis may show that two variables X and Y are related; however, it cannot show X causes Y or that Y causes X . It is just possible that the relationship shown to exist may be the effect of a third variable Z ; i.e., it may be that X and Y represent joint effects of Z . There are several criteria, such as biological plausibility, **dose-response relationship**, temporal relationship, consistency with other studies, lack of **bias**, and **confounding** effect, among others, that must be met before reaching such a conclusion.

cause-specific death rate— The **death rate** in a specified period of time and place due to a specific disease, source, or cause. This **rate** is designed to measure the **probability** of death from a particular disease. It is obtained as the total number of deaths due to the specified cause during a calendar year divided by the midyear population of the region (expressed per 1000). See also *standardized death rate*.

**Death rates and percent of total deaths for the 15 leading causes of death:
United States 1980 (rates per 100,000 population)**

Rank	Cause of death	Rate	Percent of total deaths
	All causes	878.3	100.0
1.	Heart diseases	336.0	38.2
2.	Malignant neoplasm, including neoplasm of lymphatic and hematopoietic tissues	183.9	20.9
3.	Cerebrovascular diseases	75.1	8.6
4.	Accidents and adverse effects	46.7	5.3
5.	Chronic obstructive pulmonary diseases and allied conditions	24.7	2.8
6.	Pneumonia and influenza	24.1	2.7
7.	Diabetes mellitus	15.4	1.8
8.	Chronic liver disease and cirrhosis	13.5	1.5
9.	Atherosclerosis	13.0	1.5
10.	Suicide	11.9	1.4

(Continued)

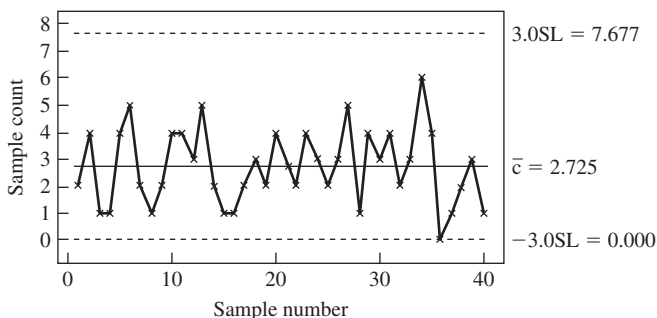
(Continued)

Rank	Cause of death	Rate	Percent of total deaths
11.	Homicide and legal intervention	10.7	1.2
12.	Certain conditions originating in the perinatal period	10.1	1.1
13.	Nephritis, nephrotic syndrome, and nephrosis	7.4	0.8
14.	Congenital anomalies	6.2	0.7
15.	Septicemia	4.2	0.5
	All other causes	95.6	10.9

Source: National Center of Health Statistics, U.S. Department of Health, Education & Welfare.

cause-specific mortality rate— Same as *cause-specific death rate*.

***c* chart**— A **graphical device** used to control a process by inspecting the number of defectives (*c*) taken from various batches or subgroups. The values of *c* computed from each batch are plotted on the **vertical axis** and can then be used to control the quality of the batch. The **center line** of the *c* chart is the **average** number of defectives (\bar{c}) taken from a pilot set (about 20 rational subgroups). **Control lines** are fixed at three **standard deviations** from the center line (based on the **normal approximation** to the **Poisson distribution**, i.e., $\bar{c} \pm 3\sqrt{\bar{c}}$). See also *control chart*, *p chart*, *run chart*, *s chart*, *x-bar chart*.



An example of a *c* chart

cell— A category of counts or values in a **contingency table**. It is formed by the intersection of a row and column in a **statistical table**. In an **analysis of variance** design, a cell represents any single group.

cell count— Same as *cell frequency*.

cell frequency— **Frequency counts** relating to a particular **cell** in a **contingency table**.

cell mean— The **mean** of all the **observations** in a particular **cell** or **level** of a **factor**.

censored data— Same as *censored observations*.

censored observations— An **observation** whose value is unknown simply because the subject or item has not been in the study a sufficient time for the outcome of interest, such

as death or breakdown, to occur or the observation is less than the **measurement** limit of detection (LOD) or it is purposely ignored. Censored observations frequently arise in many **longitudinal studies** where the **event** of interest has not occurred to a number of subjects at the completion of the study. Moreover, the **loss to follow-up** often leads to censoring since the outcomes remain unknown.

censored regression analysis— A form of **regression** where the values of the **dependent variable** are censored or truncated.

censored sample— A **sample** that has some of its values, usually the largest and/or smallest, censored because they are unobservable.

censoring— See *censored observations*.

census— The complete count (enumeration) or **survey** involving the observation of every member of a **population** or a group of items at a point in time with respect to certain well-defined characteristics of interest. A census of the human population is a counting of the people within the boundaries of a country. More generally, it is the total process of collecting, compiling, and publishing demographic, economic, and social **data** pertaining, at given time period, to all persons in a country or delimited territory. The use of information derived from a census has become indispensable to any modern government. In modern times, censuses have come to include many topics other than just counting people. Some of the areas independent of the census of human population are agriculture, housing, business establishments, and industries.

census area— The well-defined geographical area in which the **census** is undertaken.

census unit— The smallest geographical area into which the entire **census area** is divided for administrative and data collection purposes.

center line— See *control charts*.

centile charts— Same as *percentile charts*.

centiles— In a series of **observations** arranged in ascending order of magnitude, centiles are those values that divide the observations into 100 equal parts. It is an abbreviated form of **percentile** not commonly used but is frequently encountered in psychological and educational testing literature. See also *deciles*, *quartiles*.

centralized database— In a **multicenter clinical trial**, a term sometimes used to refer to a **database** that is located and maintained in a central coordinating office.

central limit theorem— A mathematical theorem that states that, regardless of the **distribution** form of the **parent population**, the **sampling distribution** of the **sample mean** approaches the **normal distribution** as the **sample size** n becomes very large. More specifically, if a **random variable** X has **population mean** μ and **population variance** σ^2 , then the sample mean \bar{X} , based on n **observations**, has an approximate normal distribution with **mean** μ and **variance** σ^2/n for sufficiently large n . It enables us to use the normal probability distribution to approximate the sampling distribution of the mean whenever the sample size is large. The central limit theorem generally applies whenever the sample size exceeds 30. This theorem is of great importance in **probability** and **statistics** since it justifies the use of normal distribution for a great variety of statistical applications.

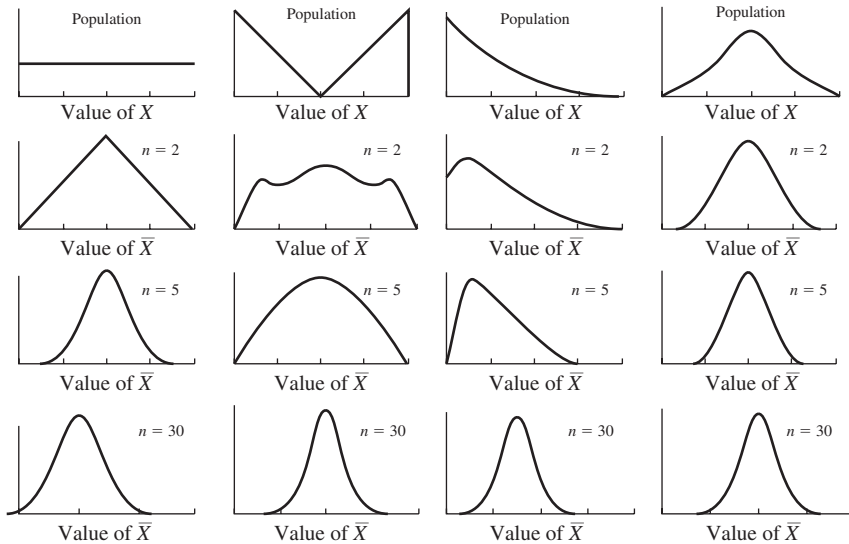


Diagram illustrating central limit theorem

central location— Same as *central tendency*.

central moments— See *moments*.

central range— The range of values that contains the central 90 percent of **observations** of a **data set**.

central tendency— Central tendency refers to the property of clustering of the data points in a distribution around a more or less central value. It is central or typical value of a given **data set** and provides an indication of the center or middle of a **distribution**. It is also referred to as **location**. See also *measures of central tendency*.

centroid method— A method of **factor analysis** developed by L. L. Thurstone, which mathematically designates a center point, from which all reference axes extend. **Principal components analysis**, the most common method of factor analysis, employs this method.

certainty equivalent— The figure that a decision maker would be indifferent to receiving for certain as compared to participating in a particular gamble.

chain-base index number— A type of **index number** which changes its base and its pattern of weights from one period to the other. Both **Laspeyres'** and **Paasche's index** numbers can be easily converted to chain-base index.

chance— A complex system of cause and effect leading to the occurrence of an **event** or phenomenon which cannot be explained otherwise. The term is loosely used as a synonym for **probability**.

chance agreement— A measure of the **proportion** of times two or more observers would agree in their **measurement** or assessment of a phenomenon under investigation simply by **chance**. See also *kappa statistic*.

chance error– Same as *random error*.

chance variable– Same as *random variable*.

chaos theory– A term coined to designate a scientific discipline concerned with investigating the apparently **random** and chaotic behavior of a system or phenomenon by use of **deterministic models**.

Chapman's estimator– In **capture–recapture sampling**, a modification of the **Petersen estimator** made to avoid the possibility of zero in the denominator. More specifically, the Chapman estimator of the total number of animals is given by

$$\hat{N} = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1$$

where n_1 and n_2 are the sizes of the first and second samples, respectively, and m is the number of marked animals in the second sample.

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

$$\phi_X(t) = E(e^{itX})$$

where $i = \sqrt{-1}$. If $\phi_X(t)$ is expanded as a power series in t , the coefficient of $(it)^k/k!$ gives the k th **moment** of X about the origin. For some **distributions**, the **moment generating function** does not exist. However, the characteristic function always exists and plays an important role in the characterization of a probability distribution.

Chebyshev's inequality– Same as *Chebyshev's theorem*.

Chebyshev's theorem– A theorem in **probability theory** that allows the use of the knowledge of the **standard deviation** and **mean** to determine the fraction of a **population** within k standard deviations of the mean. It states that, regardless of the **shape** of a population's frequency distribution, the **proportion** of **observations** falling within k standard deviation of the mean is at least $(1 - 1/k^2)$ given that k is 1 or more. Thus, according to this theorem, at least $(1 - 1/2^2)$, i.e., 75% of the observations fall within two standard deviations of the mean.

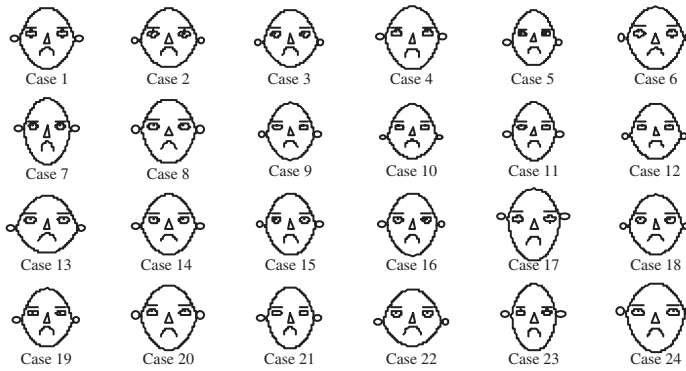
Chernoff's faces– A statistical technique for representing **multivariate data** in which each **data point** is represented by a computer-generated graphic resembling a human face, and the shape, size, and feature of each face is determined by the values taken by particular **variables**. The **sample data** are then arranged or grouped according to similarities among faces and thus may be used to assess similarities or differences between **observations**.

chi distribution– The **probability distribution** of a **random variable** $\chi = +\sqrt{X}$ where X has a **chi-square distribution**.

child death rate– The number of deaths of children aged 1 to 4 years observed in a given year divided by the total number of children in this age group (expressed per 1000).

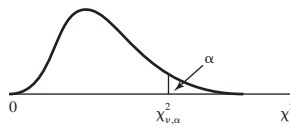
child mortality rate– Same as *child death rate*.

chi (random) variable– A **random variable** that has a **chi distribution**.



Schematic diagram illustrating Chernoff's faces

chi-square distribution– The **distribution** may be considered as a sum of squares of k independent variables, where each variable follows a **normal distribution** with **mean 0** and **standard deviation 1**. The **parameter** k is known as the number of **degrees of freedom**. The distribution is frequently used in many applications of **statistics**, for example, in testing the **goodness of fit of models** and in analyzing count data in **frequency tables**. The **chi-square test** is based on it. The following table gives the **critical values** of a **chi-square variable**, which denotes the value for which the area to its right under the chi-square distribution with ν degrees of freedom is equal to α . The entries in this table are values of $\chi^2_{\nu,\alpha}$ for which the area to their right under the chi-square distribution with ν degrees of freedom is equal to α .



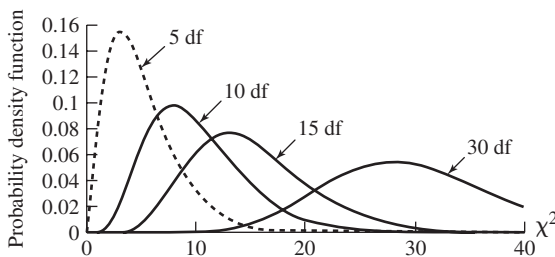
Chi-square table

$\alpha \rightarrow$ $\nu \downarrow$	0.995	0.99	0.975	0.95	.05	0.025	0.01	0.005	$\leftarrow \alpha$ $\downarrow \nu$
1	.0000393	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	0.011	0.0201	0.051	0.103	5.991	7.378	9.210	10.597	2
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12

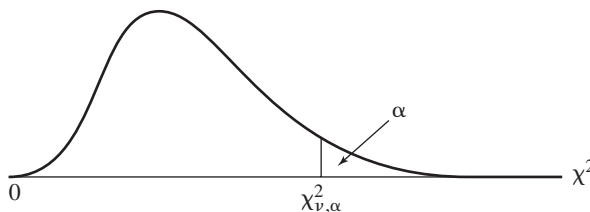
(Continued)

$\alpha \rightarrow$ $v \downarrow$	0.995	0.99	0.975	0.95	.05	0.025	0.01	0.005	$\leftarrow \alpha$ $\downarrow v$
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.719	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559	24
25	10.520	11.524	13.120	14.611	37.653	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30

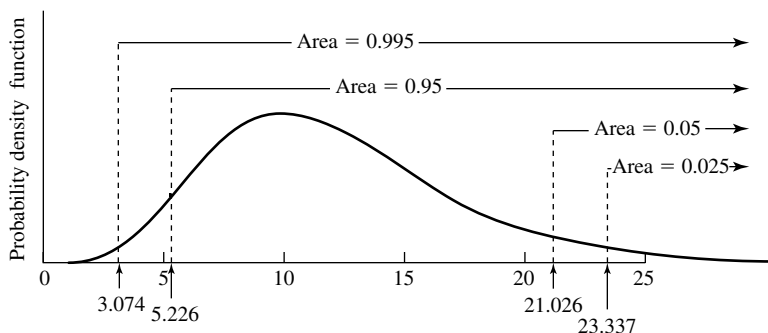
Source: Computed by using software.



Probability density curves for chi-square distributions with 5, 10, 15, and 30 degrees of freedom



For v degrees of freedom, χ^2 value such that area in the right tail is α



Probability density curve of the chi-square distribution with 12 degrees of freedom (area to the right of 5.226 is 0.95 and so on)

chi-square (random) variable— A **random variable** that has a **chi-square distribution**.

chi-square statistic— In general, any **statistic** that has a **chi-square distribution**. There are many statistical applications of the chi-square statistic. One of the most common procedures involves testing the **hypothesis of independence** for the **two-way classifications** of a **contingency table**. In this case, the chi-square statistic is obtained as the sum of all the quantities obtained by taking the difference between each **observed** and **expected frequency**, squaring the difference, and dividing this squared deviation by the expected frequency. See also *goodness of fit statistic*.

chi-square test— A test of **statistical significance** based on the **chi-square distribution**. This test is used in many situations. Some of the more common uses are: (1) an overall **goodness-of-fit test** for comparing the **frequencies of events** that are classified in **nominal categories** with hypothetical frequencies falling into specified categories; (2) testing the **association** in a **contingency table** by comparing the observed **cell frequencies** with the frequencies that would be expected under the **null hypothesis** of no association; (3) testing the **hypothesis** that a **sample** comes from a hypothetical **normal population** with known **variance**. For the validity of a chi-square test, it is generally assumed that **expected frequencies** of all the **cells** be greater than 1 and at least 80% of the cells have expected frequencies greater than 5. When these assumptions are not met, other tests, such as **Fisher's exact test**, are more appropriate.

chi-square test for independence— A **chi-square test** used in a **contingency table** by comparing the observed **cell frequencies** with the **frequencies** that would be obtained under the **null hypothesis of independence** of row and column categories.

chi-square test for trend— A **chi-square test** used in a $2 \times k$ **contingency table** with k ordered categories to test the **hypothesis** of a difference in the trend of the k **proportions** in the two groups. The test is generally more powerful than the usual **chi-square test for independence**.

circle chart— Same as *pie chart*.

circular distribution— The **probability distribution** of a **random variable** defined as the value of an angle confined to be on the unit circle. It ranges in value from 0 to 2π . It is used to model the phenomena that have a period of 2π so that the **probability density** at any

point θ is the same that any point $\theta + 2\pi k$ for any integral value of k . The **probability mass** may be regarded as distributed around the circumference of a circle.

class boundary– Same as *class limit*.

class frequency– Same as *absolute class frequency*.

classical inference– Same as *classical statistical inference*.

classical probability– A definition of **probability** that assumes that all the experimental **outcomes** of a **random phenomenon** are equally likely or is based on some other objective or theoretical considerations. It is equal to the number of equally likely outcomes favorable to the occurrence of an **event** of interest divided by the total number of equally likely basic outcomes possible. See also *empirical probability, objective probability, subjective probability*.

classical statistical inference– Same as *statistical inference*. The term is sometimes used to distinguish it from the so-called **Bayesian inference**.

classical statistics– Same as *classical statistical inference*.

classical time-series model– A **time-series model** that attempts to explain the pattern or **variation** observed in an actual **time-series data** by the sum/product of the four components: **trend, cyclical, seasonal, and irregular components**. See also *additive time-series model, multiplicative time-series model, mixed time-series model*.

classification– The process of subdividing the range of values of a **variable** into classes or groups.

classification errors– **Errors** in assigning or classifying persons, objects, or **events** into separate classes, categories or groups.

classification techniques– A general term applied to any of the techniques used in **cluster and discriminant analysis**.

class interval– One of the intervals into which the entire range of the **variable** values has been divided. It represents the length of a class or the range of values covered by a class of a **frequency distribution**.

class limits– In a **frequency distribution**, the **variable** values that demarcate each **class interval**. For example, 2.1 and 2.4 are, respectively, the lower and upper class limits of the class interval 2.1–2.4.

class mark– Same as *midpoint*.

class midpoint– Same as *midpoint*.

class midvalue– Same as *midpoint*.

class width– The length or difference between the numerical values of the **upper real limit** of a class and the **lower real limit** of that class.

clinical decision making– Same as *medical decision making*.

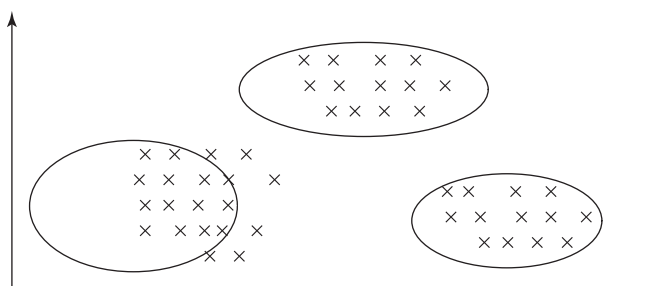
clinical significance– Same as *practical significance*.

clinical trial– An **experimental study** of a medical treatment or procedure on human beings, designed to investigate the efficacy of the treatment. It generally entails comparison

between two or more **study groups**, by administering treatments/interventions to at least one of the study groups to assess the relative efficacy of treatments. The paradigm of a clinical trial is a **randomized controlled trial**. See also *phase I trial*, *phase II trial*, *phase III trial*, *phase IV trial*.

cluster— A cluster is a subset of the **population** of objects. Generally, the clusters consist of natural groupings of the individuals or objects such as residents in a city block, a family, hospital, school, etc.

cluster analysis— An advanced statistical technique in **multivariate analysis** that determines a **classification** or taxonomy from multiple measures of an initially unclassified set of individual or objects. The procedure is designed to determine whether individuals or objects are similar enough to belong to the same or separate groups or **clusters**. The sets of **measurements** pertaining to individuals being studied, known as profiles, are compared and individuals that are close or similar are classified as being in the same cluster or group. During recent decades, applications of cluster analysis have grown at rapid pace. Programs for carrying out cluster analyses are now included in much of the widely used **statistical software**.



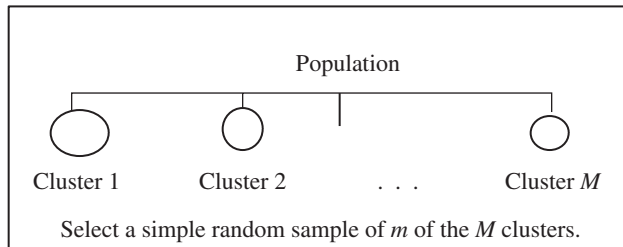
Schematic diagram for clusters of data

clustering— The division of a **population** into a number of subpopulations commonly known as **clusters**. Clustering makes use of natural groupings; for example, employees in a firm may be divided into work groups, children in a school may be grouped into designated classes, and dwellings of a city may be organized into blocks.

cluster randomization— A method of **randomization** in which groups or **clusters** of individuals rather than individuals themselves are randomly assigned to **treatment groups**. Although the method is not as efficient as the individual randomization, it is useful in terms of certain economic, ethical, and practical considerations.

cluster sampling— A **two-stage sampling** procedure in which the **population** is divided into groups of units known as **clusters**. A **random sample** of clusters is drawn, and then random samples of subjects within the clusters are selected. A one-stage cluster random sample entails a complete enumeration of all randomly chosen clusters. Typically, entire households, schools, or hospitals are sampled. Generally, the clusters consist of natural groupings of individuals or objects. Cluster sampling is usually employed when the researcher cannot obtain a complete list of the elements of a population under study but can get a complete list of groups of individuals (all persons in a city block, a family, hospital, school, etc.) of the population. In the determination of the **sample size** required in a study,

where the clusters are the **sampling units**, it is necessary to modify the commonly used formulas for this purpose. See also *multistage sampling*.



Schematic diagram for cluster sampling

COBOL– An acronym for Common Business Oriented Language. A business-oriented programming language used for writing programs.

Cochran's C test– A **test procedure** for testing three or more **independent samples** for **homogeneity of variances** before using an **analysis of variance** procedure. It is based on the ratio of the largest **sample variance** to the sum of all the sample variances and was proposed by W. G. Cochran in 1941. See also *Bartlett's test*, *Box's test*, *Hartley's test*.

Cochran's Q test– A **nonparametric procedure** for comparing several correlated **proportions** arising from **dependent** or **matched groups** to determine whether **frequencies** or **correlations** differ significantly among themselves. It is a generalization of **McNemar's chi-square test** for more than two matched groups and is best suited for nominal or dichotomized **ordinal data**.

coding– A term used to refer to a **variable** with arbitrary origin or possibly transforming it in some other unit.

coefficient– A **constant** multiplier that measures some property of a **variable** or functions of a variable.

coefficient of alienation– A term sometimes used to denote a measure of the **proportion** of **variability** in the **response variable** that is not explained by the **estimated regression equation**. It is obtained as the **ratio** of the **sum of squares due to residuals** to the **total sum of squares**. It can be considered as a measure of the lack of fit of the estimated regression equation. It is interpreted as the amount of error in predicting values of the **dependent variable** that could not be eliminated by using values of the **independent variables**. It is equivalent to $1 - R^2$ where R^2 is the **coefficient of multiple determination**.

coefficient of concordance– Same as *Kendall's coefficient of concordance*.

coefficient of contingency– Same as *contingency coefficient*.

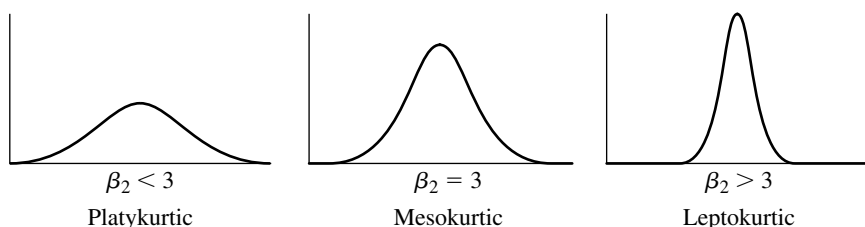
coefficient of correlation– Same as *correlation coefficient*.

coefficient of cross-elasticity– The mathematical relationship between a percentage change in the price of a certain commodity or service and the resulting percentage change in the sales of a substitute commodity or service.

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

coefficient of elasticity– The mathematical relationship between the percentage change in the quantity of a commodity or service acquired or offered and the percentage change in the price.

coefficient of kurtosis– A measure of **kurtosis** of a **distribution** defined by $\beta_2 = \mu_4/\mu_2^2$ where μ_2 and μ_4 denote the second and fourth **central moments** of the distribution. For a **normal** or **mesokurtic distribution** $\beta_2 = 3$, for a **leptokurtic distribution** $\beta_2 > 3$, and for a **platykurtic distribution** $\beta_2 < 3$.



The relationship between the coefficient of kurtosis and the degree of peakedness

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

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

coefficient of multiple determination– It is a measure of the **proportion of variability** in the **response variate** that is explained by the **estimated regression equation**. It is obtained as the **ratio** of the **sum of squares due to regression** to the **total sum of squares**. It can be interpreted as a measure of how well the estimated regression equation fits the **data** or explains the **variation** in the data. It is equivalent to R^2 where R is the **multiple correlation coefficient**. See also *adjusted sample coefficient of multiple determination*, *sample coefficient of multiple determination*.

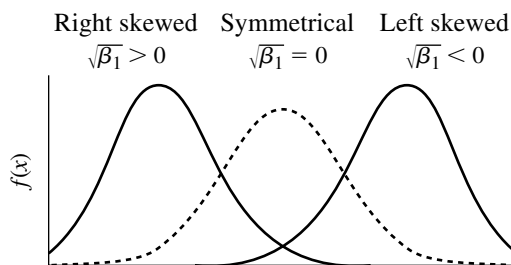
coefficient of part correlation– See *part correlation*.

coefficient of partial determination– In **multiple regression analysis**, a **measure of association** between the **dependent variable** and one of the **independent variables**, after adjusting for the **effects** of one or more other independent variables. See also *sample coefficient of partial determination*.

coefficient of regression– Same as *regression coefficient*.

coefficient of relative variation– Same as *coefficient of variation*.

coefficient of skewness– A measure of **skewness** of a **distribution** defined by $\beta_1 = \mu_3^2/\mu_2^3$ where μ_2 and μ_3 denote the second and third **central moments** of the distribution. The size of μ_3 relative to $\mu_2^{3/2}$ indicates the extent to which the distribution departs from **symmetry**. Thus $\sqrt{\beta_1}$ gives a measure of the relative skewness of a distribution, or its skewness normalized by its **spread**. It can be used to compare the symmetry of two distributions with different values of scales. For a **symmetrical distribution**, $\sqrt{\beta_1} = 0$. For a distribution having right tails, $\sqrt{\beta_1} > 0$. For a distribution having left tails, $\sqrt{\beta_1} < 0$. See also *coefficient of kurtosis*, *kurtosis*.



The relationship between the coefficient of skewness and concentration of tails of a distribution

coefficient of variation— A measure of relative **dispersion** for a **data set**. It is calculated by dividing the **standard deviation** by the **mean**. It is generally represented as percentage by multiplying it by 100. It expresses the magnitude of the **variation** relative to its **average** size and is used for comparing the **variability** in different **distributions**. The standard deviation provides an absolute **measure of dispersion** of a data set expressed in the same units of **measurements**, for example, tons, yards, or pounds. However, the coefficient of variation provides a means of comparing the variability in two or more data sets measured in different units and can be considered a statistical measure of the relative dispersion, variability, or **scatter** of a data set or **frequency distribution**. It is a purely statistical entity free of any units of measurement. Coefficient of variation is also often used as a measure of the repeatability of a measurement method by taking repeated measurements with the method in question and calculating its coefficient of variation.

cofactor of a matrix— The ij th cofactor of an $n \times n$ **matrix** A denoted by A_{ij} is given by

$$A_{ij} = (-1)^{i+j} |M_{ij}|$$

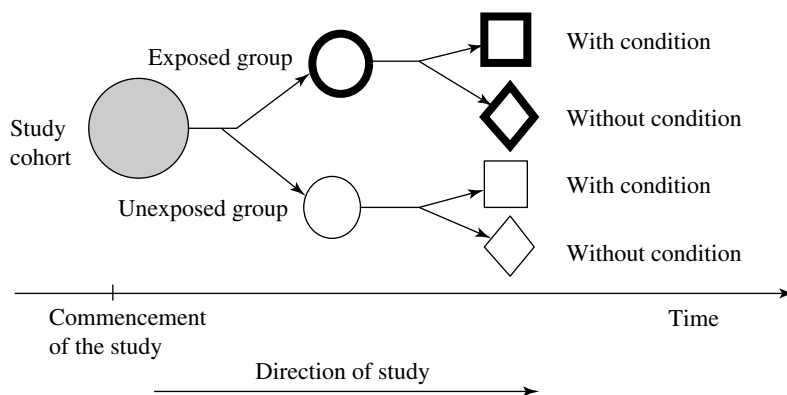
where M_{ij} is the ij th minor of A .

cohort— A group composed of individuals of the same generation, age, occupation, geographical area; or any designated group of persons with some common characteristics who are followed or traced over a period of time, as in a **cohort analysis** or **study**.

cohort analysis— The study of the same cohort over an extended period of time. See also *cohort study*.

cohort study— An **observational study** that includes a group of subjects who have a **risk factor** or have been exposed to an agent and a second group of subjects who do not have the risk factor or **exposure**. Both groups are followed prospectively through time to determine and compare the **outcomes** of interest in the two groups. The alternative terms for a cohort study are **follow-up**, **longitudinal**, and **prospective study**. In investigating the relationship between an exposure or risk factor and the **incidence** of disease, cohort studies generally yield more precise results and are less prone to **biases** of different sources than **case-control studies**. However, the cohort study generally entails the study of a large population for a prolonged period of time. Since the cohort studies can take a long period of time to complete, they may be very costly to conduct. They are usually unsuitable for investigating rare outcomes since it would require that an extremely large number of

subjects be followed in order to get an adequate number of **events** of interest. Moreover, in many cohort studies, some subjects may not be followed for the full length of the study since they may move to another area or may even die. Thus, **loss to follow-up** and surveillance bias are two common sources of bias in this type of study. See also *cohort analysis*, *cross-sectional study*.



Schematic diagram of a cohort study

collectively exhaustive events– Same as *exhaustive events*.

collinearity– Same as *multicollinearity*.

column chart– Same as *bar chart*.

column marginals– In a **cross-tabulation**, the **frequencies** of the **variable** appearing across the columns. Compare *row marginals*.

column sum of squares– Same as *sum of squares for columns*.

combination– A combination is a nonrepeating arrangement or selection of distinguishable elements or objects in which the order is ignored. Thus, the arrangement ABC is the same combination as BCA, CAB, CBA, or ACB. The number of possible combinations, each containing r objects, that can be formed from a set of n distinct objects is given by

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

community controls– In **case-control studies**, the selection of **controls** from the same population from which **cases** are drawn. The use of community controls is appropriate if the source population is well defined and the cases in the **study sample** are considered representative of all the cases in this population. See also *hospital controls*.

comparative experiment– An **experimental study** designed to make comparisons between a **control group** and one or more **treatment groups**. In a **clinical trial**, the term is synonymous with **phase III trial**.

comparative study– A study designed to make comparisons between one or more groups of subjects.

comparative treatment trial– Same as *phase III trial*.

comparative trial– Same as *controlled trial*.

comparison group– Same as *control group*.

comparisonwise error rate– In a **multiple comparison** procedure, one is concerned with individual comparisons as well as sets of such comparisons. In individual comparisons, the **significance level** is referred to as comparisonwise error rate. See also *experimentwise error rate*.

compatible events– Different **random events** that have at least some basic **outcomes** in common.

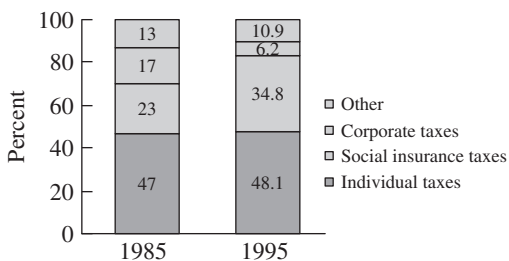
complementary event– Same as *complement of an event*.

complement of an event– The complement of an **event** A is the event containing all **sample points** that are not in A . It is an event contrary to the one of interest. It is denoted by \bar{A} , A' , or A^C .

completely randomized design– An **experimental design** in which the **treatments** are allocated to the **experimental units** randomly without any restriction. This type of design controls **extraneous variables** by creating one **treatment group** for each treatment and assigning each experimental unit to one of these groups by a **random process**. Thus, a completely randomized design assigns the experimental units to the treatments in such a way that any one allocation of experimental units to the treatments is just as probable as any other. See also *block design*, *blocking*, *randomized block design*.

compliance– A term used in **clinical trial** to indicate the extent of adherence of patients to the study **protocol**.

component bar chart– A **bar graph** in which each bar is divided into sections proportional in size to the components of the total they represent. The various components are usually colored or shaded to enhance the overall appearance and effectiveness of the graph.



Component bar charts showing percent distributions for the sources of federal income for hypothetical data: 1985 and 1995

component bar graph— Same as *component bar chart*.

composite event— Any **event** comprising two or more **basic outcomes**.

composite hypothesis— A **hypothesis** that specifies a range of values for an unknown **parameter**—for example, the hypothesis that the **mean** of a **population** is different than some given value.

composite sampling— A relatively inexpensive method of **sampling** used for items that require an expensive and time-consuming process of **measurements**. For example, to estimate the moisture content of a trainload of corn, one could take a bushel of corn from each wagon, mix these in a blender, and measure the moisture content of the resulting composite sample. In this manner, every part of the trainload can be sampled but only one expensive measurement need to be taken.

compound distribution— A type of **probability distribution** where a **parameter** of the **distribution** is also a **random variable** having a given probability distribution. For example, the **negative binomial distribution** can be expressed as a **Poisson distribution** where the **mean** is a random variable having a **gamma distribution**.

compound event— An **event** that comprises two or more **simple events** that are not necessary **mutually exclusive**.

computational formula— An algebraic formula that is mathematically equivalent to the **definitional formula** and is easier to use for manual computations but does not directly display the meaning of the procedure it symbolizes. Compare *definitional formula*.

computer-aided diagnosis— The use of computers to assist clinicians in approaching the diagnostic task by compiling available **data** and developing a list of one or more diagnostic possibilities. The basic idea behind computer-aided diagnosis is to use the historical data gathered from the clinical study of previously examined patients to determine the likely diagnosis in a new patient exhibiting another set of data on symptoms, signs, or laboratory results. A computer-aided diagnosis thus requires a **mathematical** and **statistical model** and the use of a computer to store, organize, and process vast quantities of information related to symptoms, common clinical findings, and laboratory results. Several mathematical and statistical models have been proposed to assist the computer-aided diagnosis and prognosis. Among the procedures employed are **Bayes' theorem**, **discriminant analysis**, **likelihood ratio statistic**, **logistic regression**, and **numerical taxonomy**.

computer-assisted survey— The use of a computer to aid the interview and **data** collection process during a **survey**. Typically, the computer presents the question text on the screen, along with the available response categories, and the interviewer or respondent answers directly into the computer. The computer can also be used to undertake various forms of data processing at the time of interview, including electronic transfer of data files.

computer-intensive statistical methods— Statistical methods that require recomputing the **test statistic** for many (typically 100 to 5000) artificially constructed **data sets**. Examples of these methods include **randomization tests**, **bootstrap**, and other **resampling** procedures. However, these methods are very general; for example, practically every **nonparametric procedure** is a special case of one of these methods. Computer-intensive methods

are easy to use, do not make the usual assumptions about the data set, and can be used to assess the significance in a **hypothesis test**.

computer package– A set of **computer programs** for storing, retrieving, and analyzing data using commonly used statistical procedures and techniques. Some widely used computer packages are **SAS, SPSS, BMDP**, and **MINITAB**, among others.

computer program– A set of instructions written in a language that a computer can read.

computer simulation– See *Monte Carlo method*.

computer software– Same as *computer package*.

conceptual model– The process of conceiving or defining **outcomes** of a phenomenon on the basis of theoretical considerations.

concordant pairs– See *Kendall's tau*.

concurrent control group– Same as *concurrent controls*.

concurrent controls– In a **clinical trial**, concurrent controls are subjects assigned to a **placebo** or **control group**. The most widely used method of assigning subjects to a **treatment** or control group is to use **random allocation** to determine which treatment each patient receives.

conditional distribution– Same as *conditional probability distribution*.

conditional logistic regression– A type of **logistic regression** used for paired **binary data**. It is commonly used in the analysis of **case-control studies** where **cases** and **controls** have been individually matched.

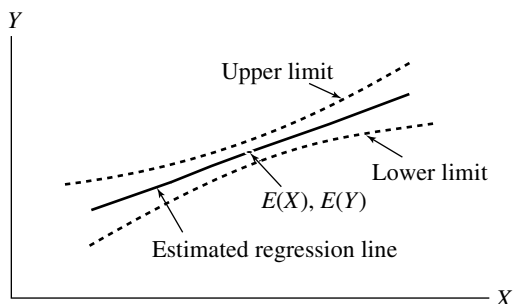
conditional mean of Y– In a **regression analysis**, the mean $\mu_{Y|X}$ of a **conditional probability distribution** of the **dependent variable** Y , for a given value of the **independent variable** X . For example, if two **random variables** X and Y with means μ_1 and μ_2 , **variances** σ_1^2 and σ_2^2 , and **correlation** ρ have a **bivariate normal distribution**, then the conditional probability distribution of Y given X is normal with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$ and variance $(1 - \rho^2)\sigma_2^2$.

conditional probability– The **probability** of an **event** given that another event has occurred. The conditional probability of A given that another event B has occurred is denoted as $P(A|B)$. It is calculated by the formula $P(A|B) = P(A \cap B)/P(B)$, where $P(A \cap B)$ is the probability of intersection of A and B . The formula assumes that $P(B) > 0$. The conditional probability is a measure of the likelihood that a particular event will occur, given that another event has already occurred. The notion of conditional probability plays a fundamental role in the postulation of **Bayes' theorem**. Compare *unconditional probability*.

conditional probability distribution– In a **bivariate** or **multivariate distribution**, the **probability distribution** of a **random variable** (or the **joint distribution** of several random variables) when the values of one or several other random variables are held constant.

conditional standard deviation– In a **bivariate analysis**, the **standard deviation** of a **conditional probability distribution** of Y given X .

confidence bands– In **regression analysis**, dashed lines on each side of an **estimated regression** line or curve that have a specified **probability** of including the line or curve in the **population**. The confidence bands can be constructed by determining **confidence intervals** for the **regression line** for the entire range of X values. One can then plot the upper and lower **confidence limits** obtained for several specified values of X and sketch the two curves that connect these points. Confidence bands are also known as confidence belts.



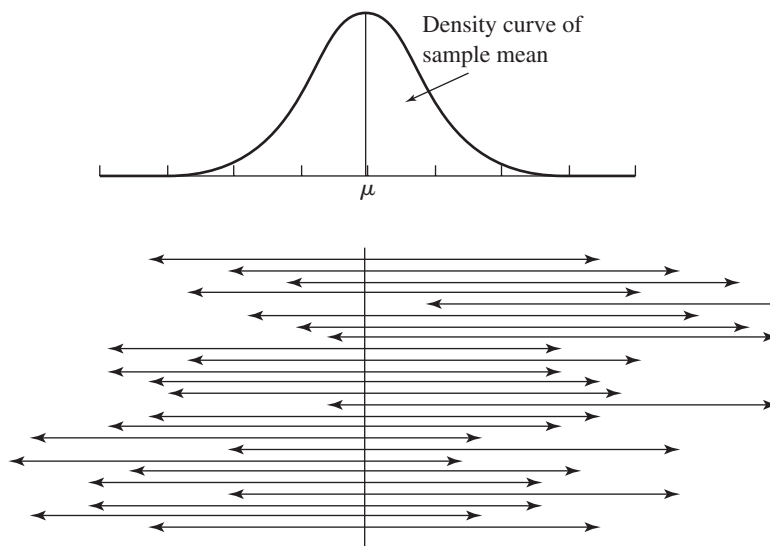
Confidence bands showing how the confidence intervals for $\mu_{\hat{y}}$ become larger as the distance between X and $E(X)$ increases

confidence belts– Same as *confidence bands*.

confidence coefficient– The confidence coefficient of a **confidence interval** for a **parameter** is the **probability** that the interval contains the value of the parameter of interest. It is the percentage of intervals (obtained from repeated **samples**, each of size n , taken from a given **population**) that can be expected to include the actual value of the parameter being estimated. For example, if an **interval estimation** procedure yields an interval such that 95% of the time the value of the **population mean** is included within the interval, the **interval estimate** is said to be constructed at the 95% confidence coefficient and 0.95 is referred to as the confidence coefficient.

confidence interval– The interval computed from **sample data** that has a specified **probability** that the unknown **parameter** of interest is contained within the interval. For example, a $1 - \alpha$ confidence interval for an unknown parameter μ is an interval computed from the sample data having the property that, in repeated **sampling**, $100(1 - \alpha)$ percent of the intervals obtained will contain the value μ . Thus, a 95% confidence interval implies that in repeated sampling 95% of the intervals would be expected to contain the true parameter value. It should be noted that the stated probability level refers to the property of the interval in repeated sampling and not to that of the parameter. Some common confidence intervals are 90%, 95%, and 99%. Note that a 99% confidence interval will be wider than the 95% confidence interval, which in turn will be wider than the corresponding 90% confidence interval. The width of a confidence interval is also related to **sample size** and measurement **variability**. The width is decreased by increasing the sample size, but is increased with the increasing variability. Wide confidence intervals reflect considerable uncertainty about the true parameter values and stem from small sample sizes, large variability, and a high **confidence coefficient**. The confidence intervals are very useful in assessing the **practical significance** of a given result.

confidence level– Same as *confidence coefficient*.



Empirical illustration of confidence intervals: Twenty-five samples from the same population generated these 95% confidence intervals. In the long run, 95% of all samples produce an interval that contains μ

confidence limits– The lower and upper limits of a **confidence interval** that define the interval within which a **population parameter** being estimated presumably lies. These limits are computed from **sample data** and have a known **probability** that the unknown parameter of interest is contained between them.

confirmatory data analysis– A term used to designate statistical procedures of **inferential statistics** in contrast to the methods and techniques of **exploratory data analysis**.

confirmatory factor analysis– See *factor analysis*.

confluent hypergeometric function– The confluent hypergeometric function denoted by $M(\alpha, \beta, x)$ is defined as

$$M(\alpha, \beta, x) = 1 + \frac{\alpha}{\beta \cdot 1!}x + \frac{\alpha(\alpha + 1)}{\beta(\beta + 1) \cdot 2!}x^2 + \frac{\alpha(\alpha + 1)(\alpha + 2)}{\beta(\beta + 1)(\beta + 2) \cdot 3!}x^3 + \dots$$

Confluent hypergeometric functions have been found very useful in the solution of many statistical problems.

confounded– A term used to describe an **experiment** or study that has one or more **extra-neous variables** present that may lead to biased **estimates** and incorrect interpretations of the results. The term is also used to refer to two or more processes whose separate **effects** cannot be determined.

confounder– Same as *confounding variable*.

confounding— A term used to describe a condition in a **factorial design** where certain comparisons can be made only for **treatments** in combinations and not for separate treatments; for example, **main effects** and **interactions** cannot be estimated separately. This is so since the **contrast** that measures one of the effects is exactly the same that measures the other. The two effects that are **confounded** are usually referred to as **aliases**. In **epidemiology**, the term is used to refer to **bias** arising from comparing groups that are different with regard to important **risk** or **prognostic factors** other than the **factor** under investigation. For example, in comparing the **incidence** of heart disease between smokers and nonsmokers any observed difference between the two groups could well be due to one group being older than the other. Here, age is acting as a **confounder** and the effect of smoking on heart disease cannot be properly assessed, as a result of important age differences between the two groups.

confounding factor— Same as *confounding variable*.

confounding variable— A **variable** more likely to be present in one group of subjects than another that is related to the **outcome** of interest and thus potentially confuses or “confounds” the results. A confounding variable is associated with both **treatment** and outcome and can affect both. The term is generally used in the context of epidemiologic and other **observational studies**.

confounding variate— Same as *confounding variable*.

congruential method— A method for generating **random numbers** based on a congruence relationship. Although the method is found to generate a good sequence of random numbers with satisfactory statistical properties, in certain cases its behavior is too erratic.

Conover test— A **nonparametric test** procedure for testing the equality of **variances** of two **populations** having different **medians**. The test has rather a low **power**; its **asymptotic relative efficiency** compared to the traditional **F test** for **normal distribution** is only 76 percent, which is slightly higher than the Siegel-Tukey efficiency measure of 0.61. See also *Ansari-Bradley test, Barton-David test, F test for two population variances, Klotz test, Mood test, Rosenbaum test, Siegel-Tukey test*.

conservative confidence interval— A term used to describe a **confidence interval** in which the actual **confidence coefficient** exceeds the nominal or stated level.

conservative test— A term used to describe a **statistical test** in which the **probability** of a **Type I error** is smaller than the nominal or stated level. Conservative tests are often preferred when only **approximate tests** are available. See also *exact test, liberal test*.

consistency— A term used to describe the property of a **consistent estimator**.

consistency checks— A term sometimes used to describe the checks being performed to assess the internal consistency of a set of **observations** in a **database**.

consistent estimator— A **sample estimator** or **statistic** such that the **probability** of its being close to the **parameter** being estimated gets ever larger (and, therefore, approaches unity) as the **sample size** increases. A consistent estimator is said to converge in probability, as the sample size increases, to the parameter being estimated.

consistent test— A test of a **hypothesis** is said to be consistent with respect to a particular **alternative hypothesis** if the **power** of the test approaches unity as the **sample size** tends to infinity.

constant– A mathematical term or a value that does not change; that is, it remains the same for all units of analysis. There are the universal mathematical constants such as π and e , and the so-called physical constants such as the velocity of light. The opposite of a constant is **variable**.

consumer price index– An **index number** designed to measure the **variations** in prices of the goods and services. It includes changes in prices of a fixed market basket of hundred of goods and services, including such items as milk, lettuce, rent, and doctor's visit, among others. The index is compiled by the U.S. Bureau of Labor Statistics and is based on about 125,000 monthly quotation prices.

contingency– A chance occurrence, i.e., an **event** incidental to another. In a **contingency table**, it is the difference between the **observed frequency** and the **expected frequency** under the assumption that the two characteristics are independent.

contingency coefficient– In a **contingency table**, a measure of the strength of the **association** between two **categorical** or **qualitative variables**. The contingency coefficient is a function of the **chi-square statistic** and is never negative, but has a maximum value less than one. It is calculated by the formula

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

where χ^2 is the usual chi-square statistic for testing the **independence** of the two **variables** and n is the **sample size**. See also *phi* (ϕ) *coefficient*, *Sakoda coefficient*, *Tschuprov coefficient*.

contingency table– A contingency table is a table that cross-classifies **bivariate data** where two **variables** are **nominal** or **categorical**. The **cells** in the table contain the **observed frequencies** of the combinations of the **levels** of two variables. The cells are mutually exclusive where each **observation** can be included in one and only one of the cells. In general, a contingency table classifies **data** according to two or more categories associated with each of two **qualitative variables**. For example, if the characteristic A is r -fold and the characteristic B is c -fold, the contingency table will have r rows and c columns. It is then often called an $r \times c$ contingency table, or simply an $r \times c$ table. The objective of an analysis of a contingency table is to determine whether two directions of classifications are dependent on each other.

Column \ Row	1	2	...	c	Row totals
1	n_{11}	n_{12}	...	n_{1c}	n_1
2	n_{21}	n_{22}	...	n_{2c}	n_2
\vdots	\vdots	\vdots		\vdots	\vdots
r	n_{r1}	n_{r2}	...	n_{rc}	n_r
Column totals	$n_{.1}$	$n_{.2}$...	$n_{.c}$	n

General $r \times c$ contingency table

contingency table analysis– Methods and techniques for analyzing relationships between **categorical variables** forming a **contingency table** using the familiar **chi-square test**. Three- and higher-dimensional tables are analyzed by using **log-linear models** and related procedures.

continuity correction— Same as *correction for continuity*.

continuous data— **Data** obtained on measures of a **continuous variable**, i.e., using interval and ratio **scales of measurement**. See also *discrete data, nominal data, numerical data, qualitative data*.

continuous distribution— Same as *continuous probability distribution*.

continuous probability distribution— It is the **probability distribution** of a **continuous random variable**. A continuous probability distribution is represented by a continuous function called a **probability density function**. Compare *discrete probability distribution*.

continuous quantitative variable— Same as *continuous variable*.

continuous scale— A scale used to measure a numerical characteristic with values that occur on an entire continuum.

continuous stochastic process— See *stochastic process*.

continuous (random) variable— A (**random**) **variable** that can theoretically assume any real value between the two points on a **measurement scale** with no gaps or spaces between possible values. When recording an **observation** on a continuous variable, it is not restricted to a particular value, except by the accuracy of the **measurement**, and a refinement of the measuring instrument yields a more precise observation. Some examples of continuous variables are height and weight. See also *categorical variable, discrete variable, ordinal variable*.

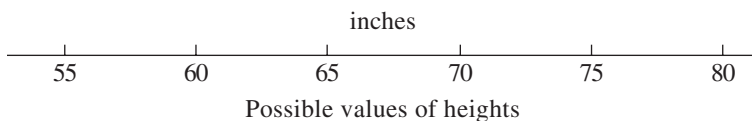


Illustration of a continuous random variable

contrast (in population means)— A **linear combination** of the **population means** such that the **coefficients** of the population means sum to zero. Thus, the statement that

$$\sum_{i=1}^k \ell_i \mu_i = \ell_1 \mu_1 + \ell_2 \mu_2 + \cdots + \ell_k \mu_k$$

is a contrast in the k population means $\mu_1, \mu_2, \dots, \mu_k$ if the ℓ_i 's sum to zero; that is, if

$$\sum_{i=1}^k \ell_i = \ell_1 + \ell_2 + \cdots + \ell_k = 0$$

Two such contrasts are said to be orthogonal if the sum of the pairwise products of their coefficients is equal to zero. Contrasts are used in making **post-hoc comparisons** of population means.

contrast (in sample means)— A **linear combination** of the **sample means** such that the **coefficients** of the sample means sum to zero. Thus, the statement that

$$\sum_{i=1}^k \ell_i \bar{x}_i = \ell_1 \bar{x}_1 + \ell_2 \bar{x}_2 + \cdots + \ell_k \bar{x}_k$$

is a contrast in the k sample means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ if the ℓ_i 's sum to zero; that is, if

$$\sum_{i=1}^k \ell_i = \ell_1 + \ell_2 + \dots + \ell_k = 0$$

Two such contrasts are said to be orthogonal if the sum of the pairwise products of their coefficients is equal to zero.

control— In a **case-control study**, the term is used to refer to an individual who does not have the disease or condition of interest. In a **clinical trial**, the term is used for a subject assigned to the **placebo** or **control condition**. See also *control group*.

control charts— **Graphs** that highlight the **average** performance values and the **variation** around this average so that average and variation of the past become standards for controlling performance in the present. A control chart is made up of three horizontal lines; one, called the center line, is drawn at the **mean** value, and the other two, called action lines or control lines, are drawn at appropriate and equal distance above and below the center line. The center line corresponds to the mean value of the characteristic under investigation. Control charts are used to decide whether a process is in statistical control. The process is judged to be “in control” as long as the plotted points lie between the two lines, and is considered “out of control” if any one of the points falls outside the control limits. Central to the idea of a control chart is the concept of **variance**. Walter Shewhart, an engineer working at Bell Laboratories, devised control charts. See also *c-chart*, *p-chart*, *R-chart*, *run chart*, *statistical quality control*, *x-bar chart*.

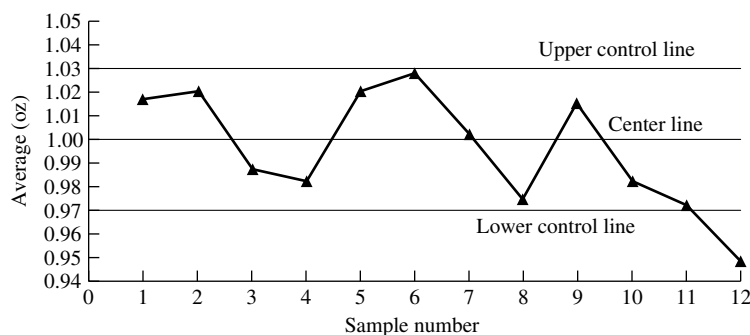


Figure showing a control chart

control condition— **Placebo** or any other standard **treatment** assigned to a **control group**.

control group— The subjects in an experiment that do not receive an **intervention**. In a **clinical trial**, these are subjects assigned to the **placebo** or any other **control condition**. A control group can be concurrent or historical, depending on whether subjects are investigated concurrently or taken from some historical records. In crossover trials, there is usually a single group of subjects where each individual acts as its own control. In a **case-control study**, the subjects without the disease or outcome are called a control group. See also *community controls*, *controlled clinical trial*, *crossover study*, *historical controls*, *hospital controls*.

controlled (for)– A term used to describe an extraneous **factor** or **variate** that is adjusted for its **confounding** effect either in the design or the analysis of the study.

controlled clinical trial– A **phase III clinical trial** in which subjects are allocated to a **control group** as well as to an experimental **treatment group**. A control group may be either the current standard **treatment** or a **placebo**. The most widely used method of unbiased treatment allocation is to use **random allocation** to determine which treatment each patient receives. Controlled trials provide direct comparison between the treatment and control groups. See also *clinical trial, phase I trial, phase II trial, phase III trial, phase IV trial, randomized controlled clinical trial*.

controlled trial– Same as *controlled clinical trial*.

control lines– See *control charts*.

controls– Same as *control group*.

control subjects– Same as *control group*.

control treatment– The **placebo** or any other **control condition** being assigned to the **control group**.

control variable– Same as *covariate*.

convenience sample– A **sample** selected in such a manner that convenience and expediency is the main consideration in selecting **elementary units** for **observation**, and usually the most easily accessible units are taken in the study. Some examples of convenience samples are workers in an office, houses in block, a group of people interviewed on a street corner, or the top items in a carton. Since **probability theory** is not employed in drawing a convenience sample, **standard errors** of the **sample estimates** cannot be determined. See also *judgment sample, nonprobability sample, probability sample, random sample*.

convenience sampling– See *convenience sample*.

conventional levels of significance– The **levels of significance** ($p < 0.05$, $p < 0.01$) that are widely used in scientific research and other statistical applications.

convolution– A mathematical procedure used to determine the **probability distribution** of the sum of two or more **random variables**.

Cook's distance– A diagnostic measure commonly used in **regression analysis** to detect the presence of an **outlier**. It is designed to measure the shift (change) in the estimated **parameter** values from fitting a **regression model** when a particular **observation** is omitted. The values of measure greater than 1 suggest the undue influence of the observation on the corresponding **regression coefficients**. See also *DFBETA, DFFIT*.

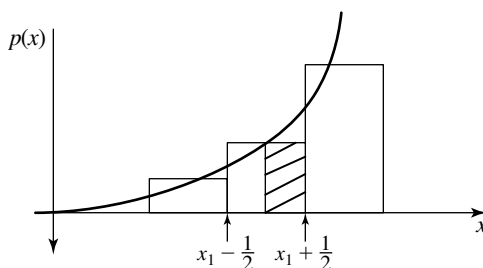
cooperative clinical trial– Same as *multicenter clinical trial*.

corner test– A **graphical procedure** designed to measure the **association** between two **variables**. The procedure involves drawing a **scatter plot** for the pairs of **observations**, and dividing it into four quadrants by lines parallel to **x** and **y axes**, passing through the

medians of the **bivariate data sets**. The **test statistic** is based on the outlying members in each quadrant.

corrected chi-square test– A **chi-square test** for a 2×2 **table** that uses **Yates' correction for continuity**. The corrected test, however, results in a more **conservative test**.

correction for continuity– When a **statistic** is discrete, but its **distribution** is being approximated by a **continuous distribution** (such as the **normal distribution**), **probabilities** can sometimes be more accurately obtained by using the tables of the continuous distribution, not with the actual values of the statistic, but with slightly corrected values. The corrected values are obtained generally by adding or subtracting a value $\frac{1}{2}$. The correction is known as 'correction for continuity.' See also *Yates' correction for continuity*.



Schematic diagram illustrating correction for continuity

correlated groups– Same as *dependent groups*.

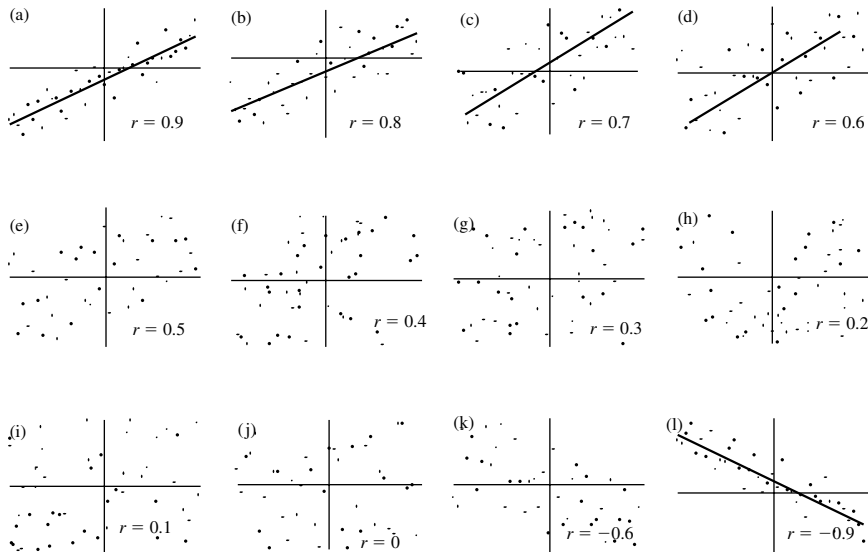
correlated samples– Same as *dependent samples*.

correlated samples t test– Same as *paired t test*.

correlation– A general term denoting **association** or relationship between two or more **variables**. More generally, it is the extent or degree to which two or more quantities are associated or related. It is measured by an index called **correlation coefficient**. See also *intraclass correlation*, *Kendall's rank correlation*, *Spearman's rank correlation*.

correlation analysis– A technique for measuring the **association** or relationship between sets of **data** involving two or more **variables**. When the two sets of scores increase and decrease simultaneously (or vary directly), the variables are said to be positively correlated. Conversely, when the sets of scores change in opposite directions so that one set decreases as the other set increases (or vary inversely), the variables are said to be negatively correlated. See also *correlation*, *correlation coefficient*.

correlation coefficient– A numerical measure of the **linear relationship** between two sets of **measurements** made on the same set of subjects. It is also known as the Pearson product moment correlation coefficient. It is denoted by the letter r and its value ranges from -1 to $+1$. A value of $+1$ denotes that two sets are perfectly related in a positive sense and a value of -1 indicates that two sets are perfectly related in a negative sense. A value close to zero indicates that they are not linearly related. See also *rank correlation coefficient*.



Bivariate data with correlation coefficient r of various magnitudes

correlation difference test— A statistical test for testing the hypothesis concerning the difference between two population correlation coefficients.

correlation for attenuation— Same as *attenuation*.

correlation matrix— A square array that represents all pairs of correlations of a set of random variables. The correlation matrix is a square matrix with as many rows as columns. Each cell of the matrix is occupied by a correlation coefficient between the variables x_i and x_j . The diagonal elements, those going from the upper left-hand corner to the lower right-hand corner of the matrix, are each equal to 1, i.e., $r_{ii} = 1$ for all i . Moreover, the correlation matrix is symmetrical about the diagonal, i.e., $r_{ij} = r_{ji}$ for $i \neq j$.

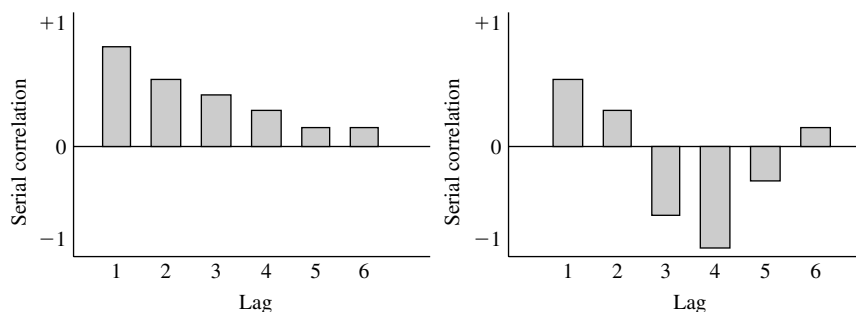
$$\begin{array}{c}
 \text{Variables} \\
 \begin{array}{cccc}
 & x_1 & x_2 & \cdots & x_p \\
 \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_p \end{array} & \left[\begin{array}{cccc}
 1 & r_{12} & \cdots & r_{1p} \\
 r_{21} & 1 & \cdots & r_{2p} \\
 \vdots & \vdots & \ddots & \vdots \\
 r_{p1} & r_{p2} & \cdots & 1
 \end{array} \right]
 \end{array}
 \end{array}$$

Correlation matrix

correlation ratio— Same as *eta*.

correlation research— Studies that do not control and manipulate variables. Correlation research examines the covariation among variables.

correlogram— A plot of the sample values of the autocorrelation against the lag.



Sample correlograms of the serial correlation coefficient

correspondence analysis— A **multivariate statistical technique** used to describe the relationship between two **variables** measured on a **nominal scale**. The method uses a set of coordinate values to represent the rows and columns of a **contingency table** and thus allows the **association** in the table to be displayed graphically. For each variable, the distance between category points in a plot reflects the relationship between categories with similar ones plotted in proximity to each other. The horizontal and vertical coordinates are analogous to those derived from **principal components analysis**. The technique, however, differs from principal components analysis in that it involves a partition of a **chi-square statistic** rather than the total **variance**.

cost–benefit analysis— An economic analysis in which costs and benefits of various alternative decisions and actions (treatments/interventions/procedures, etc.) and the associated risks and uncertainties (loss of net earnings due to illness, death or disabilities, etc.) are evaluated. The preferred action is one that provides the greatest benefit for a given cost or requires the least cost for a given level of benefit.

cost–effectiveness analysis— An economic analysis of costs and effectiveness of alternative decisions and actions.

cost–minimization analysis— An economic analysis of the costs and outcomes of alternative actions when these actions can be shown to have comparable results or impact.

cost–utility analysis— An economic analysis of costs and outcomes of alternative actions in which outcomes are measured in terms of their personal or social utility.

count data— **Data** relating to **frequency counts** of occurrences of certain **random events** or phenomena in contrast to **continuous data** that are obtained by taking **measurements** on some scale. Count data arise frequently in demographic **sampling**, in **survey research**, in learning experiments, and in almost every other branch of social, engineering, and life sciences.

covariance— The first product moment of two **variables** about their **mean** values. It is calculated as the sum of the product of **deviations** of the x 's and y 's about their respective means divided by $n - 1$ in a **sample** and N in the **population**. It is a measure of the joint **variance** of two variables. It ranges from $-\infty$ to $+\infty$. A positive value indicates that two variables are directly related and a negative value indicates that they are inversely related. See also *correlation*, *covariance matrix*, *sample covariance*.

covariance matrix— A square array that represents all pairs of **covariances** of a set of **random variables**. A covariance matrix is a **square matrix** in which main diagonal

elements represent **variances** of the **variables** and off-diagonal elements are the covariances. Moreover, like the **correlation matrix**, a covariance matrix is also symmetrical about the diagonal.

$$\begin{array}{c} \text{Variables} \\ \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{matrix} \end{array} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

Covariance matrix

covariance structure model– Same as *structural equation model*.

covariate– The term used for a **confounding variate** as a source of possible explanation of **variation** in the **dependent variable**. This is a **variable** that the researcher seeks to control by use of techniques such as **analysis of covariance** and **regression**. The value of a covariate is held constant in an analysis in order to observe its **effect** on the original **association** between two or more variables. The term is also used simply as an alternative name for an **explanatory variable**. It is also sometimes employed to refer to a variable that is not of primary interest in an investigation but is thought to be related to the **response variable** of interest and probably should be taken into account in any analysis and **model building**. It is also known as a control variable.

covariation– Joint variation in **observations** involving a **bivariate data set**. See also *covariance*.

Cox regression– Same as *proportional hazards regression*.

Cox–Mantel test– A **nonparametric statistical test** for comparing two **survival curves**. If the survival experience of the two groups is the same, then the **test statistic** can be approximated by a **standard normal distribution**.

Cramér–Rao inequality– An inequality giving a lower bound of the **variance** of any **unbiased estimator** of a **parameter** θ , or more generally a given parametric function $g(\theta)$, in the **probability density function** $f(x, \theta)$ of the observed **random variable**. The inequality states that

$$\text{Var}(T) \geq \frac{[g'(\theta)]^2}{nE \left[\left(\frac{\partial}{\partial \theta} \log_e f(x, \theta) \right)^2 \right]}$$

where T is an unbiased estimator of $g(\theta)$, $g'(\theta)$ is the derivative of $g(\theta)$ with respect to θ , and n is the **sample size**.

Cramér–Rao lower bound– See *Cramér–Rao inequality*.

Cramér’s V– Same as *Cramer’s V coefficient*.

Cramér’s V coefficient– A **measure of the association** or relationship between two **nominal** or **categorical variables** whose **data** are cross-classified in a 2×2 or higher-order

contingency table. It is based on the usual **chi-square statistic** for testing the **independence** and is calculated by the formula

$$V = \sqrt{\frac{\chi^2}{n \times \min(r - 1, c - 1)}}$$

where χ^2 is the usual chi-square statistic, r and c are the number of rows and columns of the table and n is the **sample size**. It is related to the **phi coefficient** by the formula, $V = \phi / \sqrt{\min(r - 1, c - 1)}$.

Cramér–von Mises statistic– A **goodness-of-fit statistic** for testing the **hypothesis** that the **cumulative distribution** of a **random variable** has a specified form. It was proposed by Harold Cramér in 1928 and independently by von Mises in 1931.

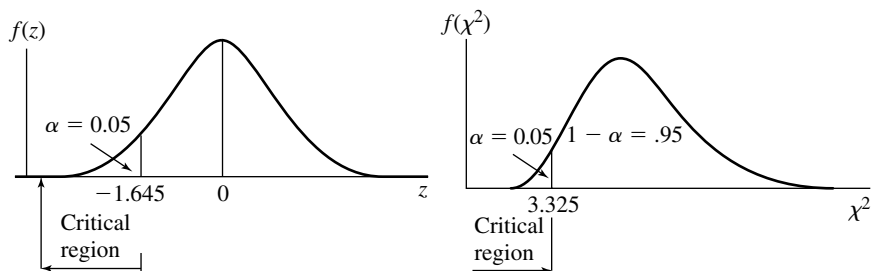
Cramér–von Mises test– A test of **normality** based on **order statistics** from **sample data**. See also *Anderson–Darling test*, *D’Agostino’s test*, *Michael’s test*, *Shapiro–Francis test*, *Shapiro–Wilk W test*.

criterion variable– The **dependent variable** that is being predicted in a **regression analysis**. In such usage the **independent variable** is known as the **predictor variable**.

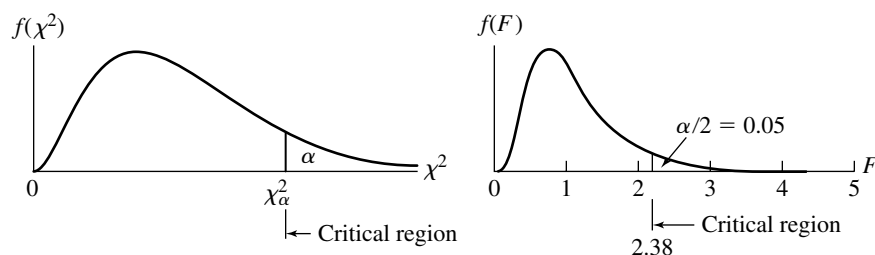
critical bounds– Same as *critical values*.

critical ratio– The term for the z or t score and other **test statistics** that define the **critical region** of a **statistical test**.

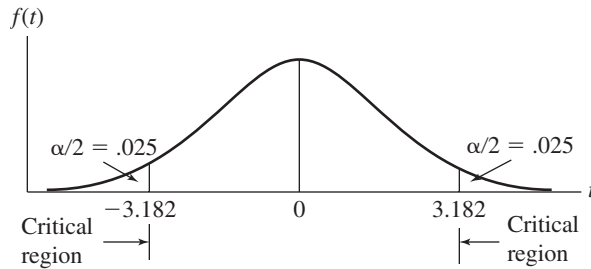
critical region– In **hypothesis testing**, the range of possible values of the area in the **sampling distribution** of a **test statistic** that leads to rejection of the **null hypothesis**. It is also known as the rejection region. The value of the test statistic must fall in this region in order for the null hypothesis to be rejected. Compare **region of acceptance**.



Examples of left-tailed critical regions



Examples of right-tailed critical regions



Example of a two-tailed critical region

critical value— The theoretical value of a **test statistic** that leads to rejection of the **null hypothesis** at a given **level of significance**. It provides a cut off point for the **region of rejection** and the **region of acceptance** of the null hypothesis. Thus, in a **statistical test**, the critical value divides the rejection and the acceptance regions. The **decision rule** for the test can be stated in terms of the critical value or values. The critical value is related to the level of significance chosen.

Cronbach's alpha— A measure of **reliability** or internal consistency of the items or **variables** in a composite index developed on a summation scale. For binary test items, it is calculated by the formula

$$\alpha = \frac{n}{n-1} \left[1 - \frac{1}{\sigma^2} \sum_{j=1}^n \sigma_j^2 \right]$$

where n is the number of items, σ^2 is the **variance** of the total score, and σ_j^2 is the variance of binary score (0 or 1) on item j . It is commonly used to measure the reliability of multiple item scales employed in psychological and mental health tests. A multiple item instrument is internally consistent if its items are highly intercorrelated, and Cronbach's alpha measures this internal consistency.

crossbreak table— Same as *cross-tabulation*.

cross-classification— Same as *cross-tabulation*.

crossed model— An **analysis of variance** model in which the **levels** of one or more **factors** cut across the levels of one or more other factors. Compare *crossed-nested model*, *nested model*.

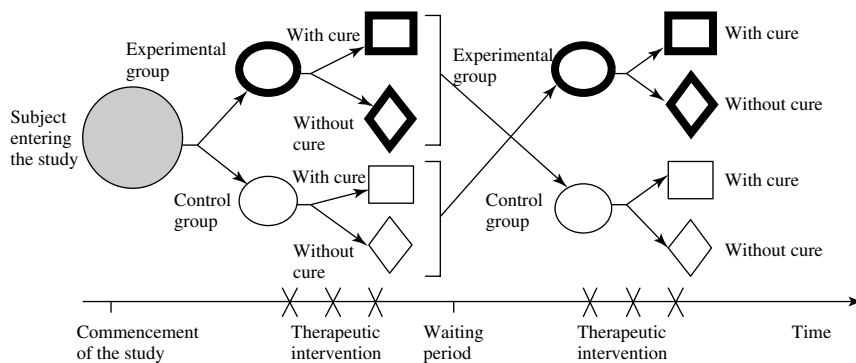
crossed-nested model— An **analysis of variance** model in which the **levels** of some **factors** are crossed while of some other factors are nested. Compare *crossed model*, *nested model*.

crossover design— See *crossover study*.

crossover rate— The **proportion** or percent of subjects who switch over from the **treatment** to which they were initially allocated to the alternative treatment. See also **crossovers**, **intention-to-treat analysis**.

crossovers— In **clinical trials**, the term is used for patients who, for some reason, do not take or receive the **treatment** to which they were allocated, but instead take or receive the alternate treatment. See also *intention-to-treat analysis*.

crossover study– A **study design** in which patients act as their own **controls** by receiving both the **treatment** being assessed and the **control treatment** in an alternate random sequence. The study uses two groups of subjects where one group is assigned to **experimental treatment** and the other to **placebo** or **control group**. After a certain period of time, both groups are withdrawn for a waiting or **washout period** without receiving any treatment. After the washout period, the **experimental group** receives the placebo and the control group receives the experimental treatment. The analysis of a crossover design is complicated because of the possibility of **carryover effects**, that is, the residual effects of the treatment administered on the first occasion that may remain present into the second occasion. Thus, it is important to introduce appropriate washout periods. In the presence of a strong **treatment period interaction**, the data for the second period are usually discarded, resulting in a **parallel design** trial lacking in sufficient **power**. The use of this type of design is not recommended if there is the possibility of strong carryover effects. In addition, this type of design is not appropriate for studies involving acute conditions or when treatment periods are too long, since patients are prone to drop out.



Schematic diagram of a crossover study

crossover trial– Same as *crossover study*.

cross-product ratio– Same as *odds ratio*.

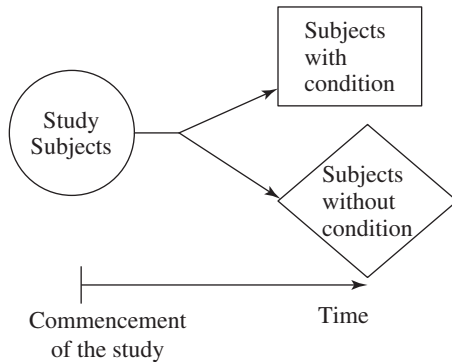
cross ratio– An abbreviated form for the cross-product ratio.

cross-sectional data– **Data** relating to units of different subjects that have been observed simultaneously at a particular point in time or during a particular period of time. See also *cross-sectional study*.

cross-sectional design– See *cross-sectional study*.

cross-sectional study– An **observational study** that explores the characteristics of interest in a group of subjects at a single point in time. In contrast to a **follow-up study**, a cross-sectional study gathers data on subjects on just one occasion. A cross-sectional study provides a “snapshot” of the characteristics or conditions of interest. In epidemiological studies, a cross-sectional design yields **estimates of prevalence** rather than **incidence**. A cross-sectional study offers only indirect evidence about the effects of time and must be interpreted with extreme caution concerning any inference regarding change. However, such a study may be suggestive of an **association** that should be investigated more

thoroughly later, say, by a **prospective** or **retrospective study**. It is also called a **survey** or **poll** in social science research. Some common problems with this type of study are the selection of an adequate **sampling design** and **nonresponse** and **volunteer bias**. See also cross-sectional data.



Schematic diagram of a cross-sectional study

cross-section series— A series that relates to different things or places at the same time, as distinct from a **time series** which relates to the same thing or place at different times.

cross-tabulation— A **frequency table** involving at least two **variables** that have been cross-classified. It is a way of presenting **data** about two variables in a table so that their relations are more clearly understood. It is also called a **contingency table** or crossbreak table. See also *cross-tabulation analysis*.

cross-tabulation analysis— Analysis of **data sets** involving two or more **qualitative variables** by cross-classifying in the form of **contingency tables**. See also *cross-tabulation*.

cross-validation— A procedure for applying the results of statistical analysis from one **sample** of subjects to a new sample of subjects in order to assess the **reliability** of the estimated **parameters**. It is frequently used in **regression** and other **multivariate statistical procedures**.

crude annual death rate— Same as *crude death rate*.

crude birth rate— Same as *birth rate*.

crude death rate— A measure or **rate** of **mortality** in which no adjustments are made to take into account social, demographic, economic, or other factors that may contribute to mortality. It is calculated as the number of deaths actually observed divided by the population of the region as estimated at the middle of particular time period, usually the calendar year (expressed per 100,000 of population). See also *age-specific death rate*, *cause-specific death rate*, *standardized mortality rate*.

crude estimates— A term used for **estimates** obtained from a **study population** without taking into account the **effects** of **confounding factors**. If the study involves some strong confounding effects, then the results obtained from the crude estimates will be biased and must be adjusted for the effects of confounding factors.

crude mortality rate— Same as *crude death rate*.

crude rate— A **rate** for the total population that is not specific for any given segment of the population or adjusted to take into account other factors. If different populations have different age structures, a direct comparison of crude rates will be biased if age is not taken into account.

cumulant generating function— The function $\Psi_X(t) = \log_e \phi_X(t)$ is known as the cumulant generating function, where $\phi_X(t)$ is the **characteristic function** of a **random variable** X . If $\Psi_X(t)$ is expressed as a power series in t , the coefficient of $(it)^k/k!$ gives the k th cumulant of X . See also *moment generating function*.

cumulants— The cumulants of a **probability distribution** are defined by the following identity in t :

$$\exp\left(\sum_{r=1}^{\infty} \frac{\kappa_r t^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{\mu'_r t^r}{r!}$$

where κ_r is the r th cumulant and μ'_r is the r th **moment about the origin**. Like **moments**, cumulants are used to characterize the **distribution** of a **random variable**. However, cumulants have certain mathematical properties that make them more useful for theoretical work.

cumulative class frequency— The number of **observations** belonging to a particular class or the ones below it. It is obtained by summing all the **frequencies** (absolute or relative) of previous classes including the class in question. See also *absolute class frequency*, *cumulative frequency*.

cumulative distribution— Same as *distribution function*.

cumulative distribution function— Same as *distribution function*.

cumulative frequency— For a given **value** or **outcome**, the total number of cases in a **data set** that are less than or equal to that value. See also *cumulative class frequency*.

cumulative frequency distribution— A **tabular representation** of a **frequency distribution** that shows the total number of **data values** with a value less than or equal to the real upper limit for the class. See also *cumulative relative frequency distribution*.

**Cumulative frequency/percentage distribution for student grades:
hypothetical data**

Class	Frequency	Cumulative Frequency	Cumulative Percentage
50–54	4	4	4.0
55–59	8	12	12.0
60–64	11	23	23.0
65–69	20	43	43.0
70–74	18	61	61.0
75–79	15	76	76.0
80–84	11	87	87.0
85–89	6	93	93.0
90–94	5	98	98.0
95–99	2	100	100.0

cumulative frequency polygon— A **frequency polygon** expressed in terms of the **cumulative class frequency**. At the right-hand endpoint of each **class interval**, at a height equal

to the cumulative class frequency of that interval, a dot is placed on a graph. Then the successive dots or points are joined by straight-line segments to form the cumulative frequency polygon. The term is more or less synonymous with **ogive curve**.

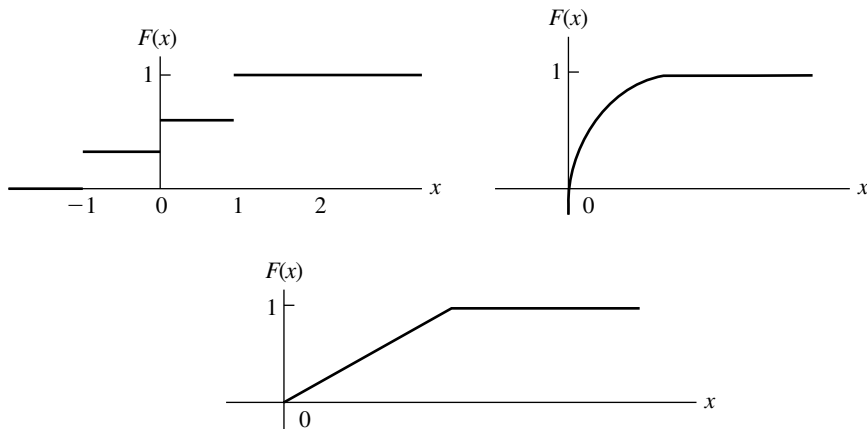
cumulative hazard– In **survival analysis**, the **risk** of an **event** over a specified period of time.

cumulative meta-analysis– A special type of **meta-analysis** which combines the results from individual studies, as these studies are carried out and the results gradually become available.

cumulative percentage– See *cumulative relative frequency*.

cumulative percentage distribution– See *cumulative relative frequency distribution*.

cumulative probability distribution– A **probability distribution** that shows the **probability** of a **random variable** being less than or equal to any given value of the random variable.



Some examples of cumulative probability distributions

cumulative relative class frequency– The **cumulative class frequency** expressed as a **proportion** or percentage of the total number of values.

cumulative relative frequency– The **cumulative frequency** expressed as a **proportion** or percentage of the total number of values.

cumulative relative frequency distribution– A **cumulative frequency distribution** expressed in terms of **proportions** or percentages of **cumulative relative frequency**.

cumulative relative frequency polygon– A **cumulative frequency polygon** expressed in terms of the **cumulative relative class frequency**.

Current Population Survey– The **sample survey** conducted annually by the **U.S. Bureau of the Census** to obtain estimates of income, employment, and other characteristics of the general labor force and of the population as a whole or of various subgroups of the population. The survey is based on about 60,000 households, which are sampled by a complex multistage stratified cluster design.

curvilinear regression– Same as *nonlinear regression*.

curvilinear relationship– A relationship between two variables that forms a curve rather than a straight line.

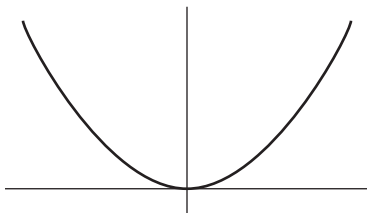


Figure showing a curvilinear relationship

cutoff level– Same as *significance level*.

cutoff point– See *critical value*.

cycle– A term used in **time-series analysis** to denote the period of the series resulting in one complete up-and-down and down-and-up movement. See also *cyclical component*, *trend*.

cycle plot– A method of **graphical representation** for investigating the behavior of a seasonal **time series**. It provides a powerful visual aid for assessing the overall pattern of the seasonal change.

cyclical component– In a **time-series analysis**, up-and-down fluctuations of the **variable** of interest around the **trend**, with the swings lasting from one to several years each and typically of different length and amplitude from one to the next. These are long-term periodic **variations** caused by forces generating a **business cycle**, as distinct from **seasonal components**. There are a number of statistical procedures currently available for estimating cyclical components. See also *cycle*, *time series*.

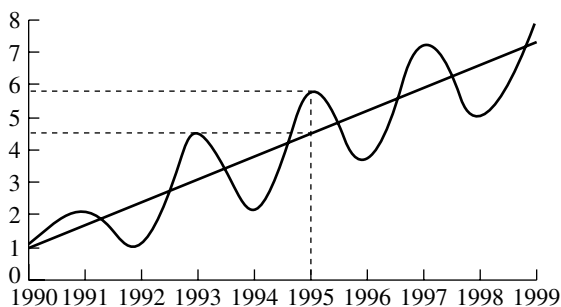


Figure showing a cyclical component in a time series: hypothetical data

cyclical fluctuation– Same as *cyclical component*.

cyclical variation– Same as *cyclical component*.