

# Glossary of Statistical Terms

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

## A

**Age Distribution** The frequency of different ages or age groups in a given population. The distribution may refer to either how many or what proportion of the group. The population is usually patients with a specific disease but the concept is not restricted to humans and is not restricted to medicine.

**Age Factors** Age as a constituent element or influence contributing to the production of a result. It may be applicable to the cause or the effect of a circumstance. It is used with human or animal concepts but should be differentiated from AGING, a physiological process, and time factors which refers only to the passage of time.

**Alternative Hypothesis** In hypothesis testing, a null hypothesis (typically, that there is no effect) is compared with an alternative hypothesis (typically, that there is an effect, or that there is an effect of a particular sign). For example, in evaluating whether a new cancer remedy works, the null hypothesis typically would be that the remedy does not work, while the alternative hypothesis would be that the remedy does work. When the data are sufficiently improbable under the assumption that the null hypothesis is true, the null hypothesis is rejected in favor of the alternative hypothesis. (This does not imply that the data are probable under the assumption that the alternative hypothesis is true, nor that the null hypothesis is false, nor that the alternative hypothesis is true.)

**Analysis of Variance (ANOVA)** An analytical method that compares the means of groups by analyzing each group's contribution to the overall uncertainty of the data, the variance.

**Arachnid Vectors** Members of the class Arachnida, especially spiders, scorpions, mites, and ticks, which transmit infective organisms from one host to another or from an inanimate reservoir to an animate host.

**Area Under Curve (AUC)** A statistical means of summarizing information from a series of measurements on one individual. It is frequently used in clinical pharmacology where the AUC

from serum levels can be interpreted as the total uptake of whatever has been administered. As a plot of the concentration of a drug against time, after a single dose of medicine, producing a standard shape curve, it is a means of comparing the bioavailability of the same drug made by different companies.

**Arthropod Vectors** Arthropods, other than insects and arachnids, which transmit infective organisms from one host to another or from an inanimate reservoir to an animate host.

**Association** Two variables are associated if some of the variability of one can be accounted for by the other. In a scatterplot of the two variables, if the scatter in the values of the variable plotted on the vertical axis is smaller in narrow ranges of the variable plotted on the horizontal axis (*i.e.*, in vertical "slices") than it is overall, the two variables are associated. The correlation coefficient is a measure of linear association, which is a special case of association in which large values of one variable tend to occur with large values of the other, and small values of one tend to occur with small values of the other (positive association), or in which large values of one tend to occur with small values of the other, and *vice versa* (negative association).

**Average** *Average* usually denotes the arithmetic mean, but it can also denote the median, the mode, the geometric mean, and weighted means, among other things.

**Axioms of Probability** There are three axioms of probability: (1) Chances are always at least zero. (2) The chance that *something* happens is 100%. (3) If two events cannot both occur at the same time (if they are disjoint or mutually exclusive), the chance that either one occurs is the sum of the chances that each occurs. For example, consider an experiment that consists of tossing a coin once. The first axiom says that the chance that the coin lands heads, for instance, must be at least zero. The second axiom says that the chance that the coin either lands heads or lands tails or lands on its edge or doesn't land at all is 100%. The third axiom says that the chance that the coin either lands heads or lands tails is the sum of the chance that the coin lands heads and the chance that the coin lands tails, because both cannot occur in the same coin toss. All other mathematical facts about probability can be derived from these three axioms. For example, it is true that the chance that an event does not occur is (100% - the chance that the event occurs). This is a consequence of the second and third axioms.

---

[top](#)

## B

**Balance** The condition in a study in which all subgroups being analyzed have equal numbers of patients.

**Bias** A measurement procedure or estimator is said to be biased if, on the average, it gives an answer that differs from the truth. The bias is the average (expected) difference between the measurement and the truth. For example, if you get on the scale with clothes on, that

biases the measurement to be larger than your true weight (this would be a positive bias). The design of an experiment or of a survey can also lead to bias. Bias can be deliberate, but it is not necessarily so.

**Bimodal** Having two modes.

**Binomial Distribution** A random variable has a binomial distribution (with parameters  $n$  and  $p$ ) if it is the number of "successes" in a fixed number  $n$  of independent random trials, all of which have the same probability  $p$  of resulting in "success." Under these assumptions, the probability of  $k$  successes (and  $n-k$  failures) is  ${}_nC_k p^k(1-p)^{n-k}$ , where  ${}_nC_k$  is the number of combinations of  $n$  objects taken  $k$  at a time:  ${}_nC_k = n!/(k!(n-k)!)$ . The expected value of a random variable with the Binomial distribution is  $nxp$ , and the standard error of a random variable with the Binomial distribution is  $(nxp(1-p))^{1/2}$ .

**Bivariate** Having or having to do with two variables. For example, bivariate data are data where we have two measurements of each "individual." These measurements might be the heights and weights of a group of people (an "individual" is a person), the heights of fathers and sons (an "individual" is a father-son pair), the pressure and temperature of a fixed volume of gas (an "individual" is the volume of gas under a certain set of experimental conditions), *etc.* Scatterplots, the correlation coefficient, and regression make sense for bivariate data but not univariate data. *C.f.* univariate.

**Blind, Blind Experiment** In a blind experiment, the subjects do not know whether they are in the treatment group or the control group. In order to have a blind experiment with human subjects, it is usually necessary to administer a placebo to the control group.

**Binomial Distribution** The probability distribution associated with two mutually exclusive outcomes; used to model cumulative incidence rates and prevalence rates. The Bernoulli distribution is a special case of binomial distribution.

**Biometry** The use of statistical methods to analyze biological observations and phenomena.

**Birth Certificates** Official certifications by a physician recording the individual's birth date, place of birth, parentage and other required identifying data which are filed with the local registrar of vital statistics.

**Birth Order** The sequence in which children are born into the family.

**Birth Rate** The number of births in a given population per year or other unit of time.

**Bootstrap estimate of Standard Error** The name for this idea comes from the idiom "to pull oneself up by one's bootstraps," which connotes getting out of a hole without anything to stand on. The idea of the bootstrap is to assume, for the purposes of estimating uncertainties, that the sample is the population, then use the SE for sampling from the sample to estimate the SE of sampling from the population. For sampling from a box of numbers, the SD of the sample is

the bootstrap estimate of the SD of the box from which the sample is drawn. For sample percentages, this takes a particularly simple form: the SE of the sample percentage of  $n$  draws from a box, with replacement, is  $SD(\text{box})/n^{1/2}$ , where for a box that contains only zeros and ones,  $SD(\text{box}) = ((\text{fraction of ones in box}) \times (\text{fraction of zeros in box}))^{1/2}$ . The bootstrap estimate of the SE of the sample percentage consists of estimating  $SD(\text{box})$  by  $((\text{fraction of ones in sample}) \times (\text{fraction of zeros in sample}))^{1/2}$ . When the sample size is large, this approximation is likely to be good.

---

[top](#)

## C

**Catchment Area (Health)** A geographic area defined and served by a health program or institution.

**Causality** The relating of causes to the effects they produce. Causes are termed necessary when they must always precede an effect and sufficient when they initiate or produce an effect. Any of several factors may be associated with the potential disease causation or outcome, including predisposing factors, enabling factors, precipitating factors, reinforcing factors, and risk factors.

**Cause of Death** Factors which produce cessation of all vital bodily functions. They can be analyzed from an epidemiologic viewpoint.

**Censuses** Enumerations of populations usually recording identities of all persons in every place of residence with age or date of birth, sex, occupation, national origin, language, marital status, income, relation to head of household, information on the dwelling place, education, literacy, health-related data (e.g., permanent disability), etc.

**Chi-Square Distribution** A distribution in which a variable is distributed like the sum of the the squares of any given independent random variable, each of which has a normal distribution with mean of zero and variance of one. The chi-square test is a statistical test based on comparison of a test statistic to a chi-square distribution. The oldest of these tests are used to detect whether two or more population distributions differ from one another.

**Chi-Square Methods** A group of qualitative variable techniques whose results are compared to values found in a theoretical Chi-square distribution table.

**Clinical Trials** Pre-planned studies of the safety, efficacy, or optimum dosage schedule (if appropriate) of one or more diagnostic, therapeutic, or prophylactic drugs, devices, or techniques selected according to predetermined criteria of eligibility and observed for predefined evidence of favorable and unfavorable effects. This concept includes clinical trials conducted both in the U.S. and in other countries.

**Clinical Trials, Phase I** Studies performed to evaluate the safety of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques in healthy subjects and to determine the safe dosage range (if appropriate). These tests also are used to determine pharmacologic and pharmacokinetic properties (toxicity, metabolism, absorption, elimination, and preferred route of administration). They involve a small number of persons and usually last about 1 year. This concept includes phase I studies conducted both in the U.S. and in other countries.

**Clinical Trials, Phase II** Studies that are usually controlled to assess the effectiveness and dosage (if appropriate) of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques. These studies are performed on several hundred volunteers, including a limited number of patients with the target disease or disorder, and last about two years. This concept includes phase II studies conducted in both the U.S. and in other countries.

**Clinical Trials, Phase III** Comparative studies to verify the effectiveness of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques determined in phase II studies. During these trials, patients are monitored closely by physicians to identify any adverse reactions from long-term use. These studies are performed on groups of patients large enough to identify clinically significant responses and usually last about three years. This concept includes phase III studies conducted in both the U.S. and in other countries.

**Clinical Trials, Phase IV** Planned post-marketing studies of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques that have been approved for general sale. These studies are often conducted to obtain additional data about the safety and efficacy of a product. This concept includes phase IV studies conducted in both the U.S. and in other countries.

**Cochran-Mantel-Haenzel Method** A Chi-square method that permits statistical comparison of odds ratios across subgroups and also allows differences in those ratios to be adjusted.

**Controlled Clinical Trials** Clinical trials involving one or more test treatments, at least one control treatment, specified outcome measures for evaluating the studied intervention, and a bias-free method for assigning patients to the test treatment. The treatment may be drugs, devices, or procedures studied for diagnostic, therapeutic, or prophylactic effectiveness. Control measures include placebos, active medicines, no-treatment, dosage forms and regimens, historical comparisons, etc. When randomization using mathematical techniques, such as the use of a random numbers table, is employed to assign patients to test or control treatments, the trials are characterized as randomized controlled trials. However, trials employing treatment allocation methods such as coin flips, odd-even numbers, patient social security numbers, days of the week, medical record numbers, or other such pseudo- or quasi-random processes, are simply designated as controlled clinical trials.

**Correlation Coefficient** In linear regression, a measure of the closeness of data points to the best-fit line. It can assume a value between -1 and +1; the nearer the value to either -1 or +1, the nearer are the points to the line.

**Cox Regression Method** An analytical method in which event data for each group under comparison are transformed to fit a linear model. Models for each group are then compared to determine whether they are equal. This method assumes that hazard rates for each group are at least proportional to each other.

**Cluster Analysis** A set of statistical methods used to group variables or observations into strongly inter-related subgroups. In epidemiology, it may be used to analyze a closely grouped series of events or cases of disease or other health-related phenomenon with well-defined distribution patterns in relation to time or place or both.

**Confidence Intervals** A range of values for a variable of interest, e.g., a rate, constructed so that this range has a specified probability of including the true value of the variable.

**Confounding Factors (Epidemiology)** Factors that can cause or prevent the outcome of interest, are not intermediate variables, and are not associated with the factor(s) under investigation. They give rise to situations in which the effects of two processes are not separated, or the contribution of causal factors cannot be separated, or the measure of the effect of exposure or risk is distorted because of its association with other factors influencing the outcome of the study.

**Comorbidity** The presence of co-existing or additional diseases with reference to an initial diagnosis or with reference to the index condition that is the subject of study. Comorbidity may affect the ability of affected individuals to function and also their survival; it may be used as a prognostic indicator for length of hospital stay, cost factors, and outcome or survival.

**Cross Sectional Study** In survey research, a study in which data are obtained only once. Contrast with longitudinal studies in which a panel of individuals is interviewed repeatedly over a period of time. Note that a cross sectional study can ask questions about previous periods of time, though.

**Categorical Variable** A variable whose value ranges over categories, such as {red, green, blue}, {male, female}, {Arizona, California, Montana, New York}, {short, tall}, {Asian, African-American, Caucasian, Hispanic, Native American, Polynesian}, {straight, curly}, etc. Some categorical variables are ordinal. The distinction between categorical variables and qualitative variables is a bit blurry. *C.f.* quantitative variable.

**Causation, causal relation** Two variables are causally related if changes in the value of one cause the other to change. For example, if one heats a rigid container filled with a gas, that causes the pressure of the gas in the container to increase. Two variables can be associated without having any causal relation, and even if two variables have a causal relation, their correlation can be small or zero.

**Central Limit Theorem** The central limit theorem states that the probability histograms of the sample mean and sample sum of  $n$  draws with replacement from a box of labeled tickets converge to a normal curve as the sample size  $n$  grows, in the following sense: As  $n$  grows, the area of the probability histogram for any range of values approaches the area under the

normal curve for the same range of values, converted to standard units. See also the normal approximation.

**Certain Event** An event is *certain* if its probability is 100%. Even if an event is certain, it might not occur. However, by the complement rule, the chance that it does not occur is 0%.

**Chance variation, chance error** A random variable can be decomposed into a sum of its expected value and chance variation around its expected value. The expected value of the chance variation is zero; the standard error of the chance variation is the same as the standard error of the random variable---the size of a "typical" difference between the random variable and its expected value. See also sampling error.

**Chebychev's Inequality** *For lists:* For every number  $k > 0$ , the fraction of elements in a list that are  $k$  SD's or further from the arithmetic mean of the list is at most  $1/k^2$ . *For random variables:* For every number  $k > 0$ , the probability that a random variable  $X$  is  $k$  SEs or further from its expected value is at most  $1/k^2$ .

**Chi-square curve** The chi-square curve is a family of curves that depend on a parameter called degrees of freedom (*d.f.*). The chi-square curve is an approximation to the probability histogram of the *chi-square statistic* for multinomial model if the expected number of outcomes in each category is large. The chi-square curve is positive, and its total area is 100%, so we can think of it as the probability histogram of a random variable. The balance point of the curve is *d.f.*, so the expected value of the corresponding random variable would equal *d.f.*. The standard error of the corresponding random variable would be  $(2 \times d.f.)^{1/2}$ . As *d.f.* grows, the shape of the chi-square curve approaches the shape of the normal curve.

**Chi-square Statistic** The *chi-square* statistic is used to measure the agreement between categorical data and a multinomial model that predicts the relative frequency of outcomes in each possible category. Suppose there are  $n$  independent trials, each of which can result in one of  $k$  possible outcomes. Suppose that in each trial, the probability that outcome  $i$  occurs is  $p_i$ , for  $i = 1, 2, \dots, k$ , and that these probabilities are the same in every trial. The expected number of times outcome 1 occurs in the  $n$  trials is  $n \times p_1$ ; more generally, the expected number of times outcome  $i$  occurs is  $\text{expected}_i = n \times p_i$ . If the model is correct, we would expect the  $n$  trials to result in outcome  $i$  about  $n \times p_i$  times, give or take a bit. Let  $\text{observed}_i$  denote the number of times an outcome of type  $i$  occurs in the  $n$  trials, for  $i = 1, 2, \dots, k$ . The *chi-squared statistic* summarizes the discrepancies between the expected number of times each outcome occurs (assuming that the model is true) and the observed number of times each outcome occurs, by summing the squares of the discrepancies, normalized by the expected numbers, over all the categories:

*chi-squared* =

$$\frac{(\text{observed}_1 - \text{expected}_1)^2}{\text{expected}_1} + \frac{(\text{observed}_2 - \text{expected}_2)^2}{\text{expected}_2} + \dots + \frac{(\text{observed}_k - \text{expected}_k)^2}{\text{expected}_k}$$

As the sample size  $n$  increases, if the model is correct, the sampling distribution of the *chi-squared statistic* is approximated increasingly well by the chi-squared curve with (#categories -

1) =  $k - 1$  degrees of freedom (*d.f.*), in the sense that the chance that the *chi-squared statistic* is in any given range grows closer and closer to the area under the Chi-Squared curve over the same range.

**Class Boundary** A point that is the left endpoint of one class interval, and the right endpoint of another class interval.

**Class Interval** In plotting a histogram, one starts by dividing the range of values into a set of non-overlapping intervals, called *class intervals*, in such a way that every datum is contained in some class interval. See the related entries class boundary and endpoint convention.

**Cluster Sample** In a cluster sample, the sampling unit is a collection of population units, not single population units. For example, techniques for adjusting the U.S. census start with a sample of geographic blocks, then (try to) enumerate all inhabitants of the blocks in the sample to obtain a sample of people. This is an example of a cluster sample. (The blocks are chosen separately from different strata, so the overall design is a stratified cluster sample.)

**Combinations** The number of combinations of  $n$  things taken  $k$  at a time is the number of ways of picking a subset of  $k$  of the  $n$  things, without replacement, and without regard to the order in which the elements of the subset are picked. The number of such combinations is  ${}_nC_k = n!/(k!(n-k)!)$ , where  $k!$  (pronounced " $k$  factorial") is  $k \times (k-1) \times (k-2) \times \dots \times 1$ . The numbers  ${}_nC_k$  are also called the *Binomial coefficients*. From a set that has  $n$  elements one can form a total of  $2^n$  subsets of all sizes. For example, from the set  $\{a, b, c\}$ , which has 3 elements, one can form the  $2^3 = 8$  subsets  $\{\}$ ,  $\{a\}$ ,  $\{b\}$ ,  $\{c\}$ ,  $\{a,b\}$ ,  $\{a,c\}$ ,  $\{b,c\}$ ,  $\{a,b,c\}$ . Because the number of subsets with  $k$  elements one can form from a set with  $n$  elements is  ${}_nC_k$ , and the total number of subsets of a set is the sum of the numbers of possible subsets of each size, it follows that  ${}_nC_0 + {}_nC_1 + {}_nC_2 + \dots + {}_nC_n = 2^n$ . The calculator has a button ( $nCm$ ) that lets you compute the number of combinations of  $m$  things chosen from a set of  $n$  things. To use the button, first type the value of  $n$ , then push the  $nCm$  button, then type the value of  $m$ , then press the "=" button.

**Complement** The complement of a subset of a given set is the collection of all elements of the set that are not elements of the subset.

**Complement rule** The probability of the complement of an event is 100% minus the probability of the event:  $P(A^c) = 100\% - P(A)$ .

**Conditional Probability** Suppose we are interested in the probability that some event  $A$  occurs, and we learn that the event  $B$  occurred. How should we update the probability of  $A$  to reflect this new knowledge? This is what the conditional probability does: it says how the additional knowledge that  $B$  occurred should affect the probability that  $A$  occurred quantitatively. For example, suppose that  $A$  and  $B$  are mutually exclusive. Then if  $B$  occurred,  $A$  did not, so the *conditional probability that  $A$  occurred given that  $B$  occurred* is zero. At the other extreme, suppose that  $B$  is a subset of  $A$ , so that  $A$  must occur whenever  $B$  does. Then if we learn that  $B$  occurred,  $A$  must have occurred too, so the *conditional probability that  $A$  occurred given that  $B$  occurred* is 100%. For in-between cases, where  $A$  and  $B$  intersect, but  $B$  is not a subset of  $A$ , the conditional probability of  $A$  given  $B$  is a number between zero and

100%. Basically, one "restricts" the outcome space  $\mathbf{S}$  to consider only the part of  $\mathbf{S}$  that is in B, because we know that B occurred. For A to have happened given that B happened requires that AB happened, so we are interested in the event AB. To have a legitimate probability requires that  $P(\mathbf{S}) = 100\%$ , so if we are restricting the outcome space to B, we need to divide by the probability of B to make the probability of this new  $\mathbf{S}$  be 100%. On this scale, the probability that AB happened is  $P(AB)/P(B)$ . This is the definition of the conditional probability of A given B, provided  $P(B)$  is not zero (division by zero is undefined). Note that the special cases  $AB = \{\}$  (A and B are mutually exclusive) and  $AB = B$  (B is a subset of A) agree with our intuition as described at the top of this paragraph. Conditional probabilities satisfy the axioms of probability, just as ordinary probabilities do.

**Confidence Interval** A confidence interval for a parameter is a random interval constructed from data in such a way that the probability that the interval contains the true value of the parameter can be specified before the data are collected.

**Confidence Level** The confidence level of a confidence interval is the chance that the interval that will result once data are collected will contain the corresponding parameter. If one computes confidence intervals again and again from independent data, the long-term limit of the fraction of intervals that contain the parameter is the confidence level.

**Confounding** When the differences between the treatment and control groups other than the treatment produce differences in response that are not distinguishable from the effect of the treatment, those differences between the groups are said to be *confounded* with the effect of the treatment (if any). For example, prominent statisticians questioned whether differences between individuals that led some to smoke and others not to (rather than the act of smoking itself) were responsible for the observed difference in the frequencies with which smokers and non-smokers contract various illnesses. If that were the case, those factors would be confounded with the effect of smoking. Confounding is quite likely to affect observational studies and experiments that are not randomized. Confounding tends to be decreased by randomization. See also Simpson's Paradox.

**Continuity Correction** In using the normal approximation to the binomial probability histogram, one can get more accurate answers by finding the area under the normal curve corresponding to half-integers, transformed to standard units. This is clearest if we are seeking the chance of a particular number of successes. For example, suppose we seek to approximate the chance of 10 successes in 25 independent trials, each with probability  $p = 40\%$  of success. The number of successes in this scenario has a binomial distribution with parameters  $n = 25$  and  $p = 40\%$ . The expected number of successes is  $np = 10$ , and the standard error is  $(np(1-p))^{1/2} = 6^{1/2} = 2.45$ . If we consider the area under the normal curve at the point 10 successes, transformed to standard units, we get zero: the area under a point is always zero. We get a better approximation by considering 10 successes to be the range from  $9 \frac{1}{2}$  to  $10 \frac{1}{2}$  successes. The only possible number of successes between  $9 \frac{1}{2}$  and  $10 \frac{1}{2}$  is 10, so this is exactly right for the binomial distribution. Because the normal curve is continuous and a binomial random variable is discrete, we need to "smear out" the binomial probability over an appropriate range. The lower endpoint of the range,  $9 \frac{1}{2}$  successes, is  $(9.5 - 10)/2.45 = -0.20$  standard units. The upper endpoint of the range,  $10 \frac{1}{2}$  successes, is  $(10.5 - 10)/2.45$

= +0.20 standard units. The area under the normal curve between -0.20 and +0.20 is about 15.8%. The true binomial probability is  ${}_{25}C_{10} \times (0.4)^{10} \times (0.6)^{15} = 16\%$ . In a similar way, if we seek the normal approximation to the probability that a binomial random variable is in the range from  $i$  successes to  $k$  successes, inclusive, we should find the area under the normal curve from  $i - 1/2$  to  $k + 1/2$  successes, transformed to standard units. If we seek the probability of more than  $i$  successes and fewer than  $k$  successes, we should find the area under the normal curve corresponding to the range  $i + 1/2$  to  $k - 1/2$  successes, transformed to standard units. If we seek the probability of more than  $i$  but no more than  $k$  successes, we should find the area under the normal curve corresponding to the range  $i + 1/2$  to  $k + 1/2$  successes, transformed to standard units. If we seek the probability of at least  $i$  but fewer than  $k$  successes, we should find the area under the normal curve corresponding to the range  $i - 1/2$  to  $k - 1/2$  successes, transformed to standard units. Including or excluding the half-integer ranges at the ends of the interval in this manner is called the continuity correction.

**Continuous Variable** A quantitative variable is *continuous* if its set of possible values is uncountable. Examples include temperature, exact height, exact age (including parts of a second). In practice, one can never measure a continuous variable to infinite precision, so continuous variables are sometimes approximated by discrete variables. A random variable  $X$  is also called *continuous* if its set of possible values is uncountable, and the chance that it takes any particular value is zero (in symbols, if  $P(X = x) = 0$  for every real number  $x$ ). A random variable is continuous if and only if its cumulative probability distribution function is a continuous function (a function with no jumps).

**Contrapositive** If  $p$  and  $q$  are two logical propositions, then the *contrapositive* of the proposition ( $p$  **IMPLIES**  $q$ ) is the proposition **((NOT  $q$ ) IMPLIES (NOT  $p$ ))**. The contrapositive is logically equivalent to the original proposition.

**Control** There are at least three senses of "control" in statistics: a member of the control group, to whom no treatment is given; a controlled experiment, and to control for a possible confounding variable.

**Controlled experiment** An experiment that uses the method of comparison to evaluate the effect of a treatment by comparing treated subjects with a control group, who do not receive the treatment.

**Controlled, randomized experiment** A controlled experiment in which the assignment of subjects to the treatment group or control group is done at random, for example, by tossing a coin.

**Control for a variable** To control for a variable is to try to separate its effect from the treatment effect, so it will not confound with the treatment. There are many methods that try to control for variables. Some are based on matching individuals between treatment and control; others use assumptions about the nature of the effects of the variables to try to model the effect mathematically, for example, using regression.

**Control group** The subjects in a controlled experiment who do not receive the treatment.

**Convenience Sample** A sample drawn because of its convenience; not a probability sample. For example, I might take a sample of opinions in Columbus (where I live) by just asking my 10 nearest neighbors. That would be a sample of convenience, and would be unlikely to be representative of all of Columbus. Samples of convenience are not typically representative, and it is not typically possible to quantify how unrepresentative results based on samples of convenience will be.

**Converge, convergence** A sequence of numbers  $x_1, x_2, x_3 \dots$  *converges* if there is a number  $x$  such that for any number  $E > 0$ , there is a number  $k$  (which can depend on  $E$ ) such that  $|x_j - x| < E$  whenever  $j > k$ . If such a number  $x$  exists, it is called the limit of the sequence  $x_1, x_2, x_3 \dots$ .

**Convergence in probability** A sequence of random variables  $X_1, X_2, X_3 \dots$  converges in probability if there is a random variable  $X$  such that for any number  $E > 0$ , the sequence of numbers  $P(|X_1 - X| < e), P(|X_2 - X| < e), P(|X_3 - X| < e), \dots$  converges to 100%.

**Converse** If  $p$  and  $q$  are two logical propositions, then the *converse* of the proposition ( $p$  **IMPLIES**  $q$ ) is the proposition ( $q$  **IMPLIES**  $p$ ).

**Correlation** A measure of linear association between two (ordered) lists. Two variables can be strongly correlated without having any causal relationship, and two variables can have a causal relationship and yet be uncorrelated.

**Correlation coefficient** The correlation coefficient  $r$  is a measure of how nearly a scatterplot falls on a straight line. The correlation coefficient is always between -1 and +1. To compute the correlation coefficient of a list of pairs of measurements  $(X, Y)$ , first transform  $X$  and  $Y$  individually into standard units. Multiply corresponding elements of the transformed pairs to get a single list of numbers. The correlation coefficient is the mean of that list of products.

**Countable Set** A set is countable if its elements can be put in one-to-one correspondence with a subset of the integers. For example, the sets  $\{0, 1, 7, -3\}$ ,  $\{\text{red, green, blue}\}$ ,  $\{\dots, -2, -1, 0, 1, 2, \dots\}$ ,  $\{\text{straight, curly}\}$ , and the set of all fractions, are countable. If a set is not countable, it is uncountable. The set of all real numbers is uncountable.

**Cover** A confidence interval is said to *cover* if the interval contains the true value of the parameter. Before the data are collected, the chance that the confidence interval will contain the parameter value is the coverage probability, which equals the confidence level after the data are collected and the confidence interval is actually computed.

**Coverage probability** The *coverage probability* of a procedure for making confidence intervals is the chance that the procedure produces an interval that covers the truth.

**Critical value** The *critical value* in an hypothesis test is the value of the test statistic beyond which we would reject the null hypothesis. The critical value is set so that the probability that

the test statistic is beyond the critical value is at most equal to the significance level if the null hypothesis be true.

**Cross-sectional study** A cross-sectional study compares different individuals to each other at the same time--it looks at a cross-section of a population. The differences between those individuals can confound with the effect being explored. For example, in trying to determine the effect of age on sexual promiscuity, a cross-sectional study would be likely to confound the effect of age with the effect of the mores the subjects were taught as children: the older individuals were probably raised with a very different attitude towards promiscuity than the younger subjects. Thus it would be imprudent to attribute differences in promiscuity to the aging process. *C.f.* longitudinal study.

**Cumulative Probability Distribution Function (cdf)** The cumulative distribution function of a random variable is the chance that the random variable is less than or equal to  $x$ , as a function of  $x$ . In symbols, if  $F$  is the cdf of the random variable  $X$ , then  $F(x) = P(X \leq x)$ . The cumulative distribution function must tend to zero as  $x$  approaches minus infinity, and must tend to unity as  $x$  approaches infinity. It is a positive function, and increases monotonically: if  $y > x$ , then  $F(y) \geq F(x)$ . The cumulative distribution function completely characterizes the probability distribution of a random variable.

---

[top](#)

## D

**Data Collection** Systematic gathering of data for a particular purpose from various sources, including questionnaires, interviews, observation, existing records, and electronic devices. The process is usually preliminary to statistical analysis of the data.

**Data Interpretation, Statistical** Application of statistical procedures to analyze specific observed or assumed facts from a particular study.

**Death Certificates** Official records of individual deaths including the cause of death certified by a physician, and any other required identifying information.

### **Demography**

Statistical interpretation and description of a population with reference to distribution, composition, or structure.

**Density, Density Scale** The vertical axis of a histogram has units of percent per unit of the horizontal axis. This is called a density scale; it measures how "dense" the observations are in each bin. See also probability density.

**Dental Health Surveys** A systematic collection of factual data pertaining to dental or oral health and disease in a human population within a given geographic area.

**Dependent Events, Dependent Random Variables** Two events or random variables are dependent if they are not independent.

**Dependent Variable** In regression, the variable whose values are supposed to be explained by changes in the other variable (the independent or explanatory variable). Usually one regresses the dependent variable on the independent variable.

**Deviation** A deviation is the difference between a datum and some reference value, typically the mean of the data. In computing the SD, one finds the rms of the deviations from the mean, the differences between the individual data and the mean of the data.

**Diet Surveys** Systematic collections of factual data pertaining to the diet of a human population within a given geographic area.

**Discrete Variable** A quantitative variable whose set of possible values is countable. Typical examples of discrete variables are variables whose possible values are a subset of the integers, such as Social Security numbers, the number of people in a family, ages rounded to the nearest year, *etc.* Discrete variables are "chunky." *C.f.* continuous variable. A discrete random variable is one whose set of possible values is countable. A random variable is discrete if and only if its cumulative probability distribution function is a stair-step function; *i.e.*, if it is piecewise constant and only increases by jumps.

**Discriminant Analysis** A statistical analytic technique used with discrete dependent variables, concerned with separating sets of observed values and allocating new values. It is sometimes used instead of regression analysis.

**Disease-Free Survival** Period after successful treatment in which there is no appearance of the symptoms or effects of the disease.

**Disease Notification** Notification or reporting by a physician or other health care provider of the occurrence of specified contagious diseases such as tuberculosis and HIV infections to designated public health agencies. The United States system of reporting notifiable diseases evolved from the Quarantine Act of 1878, which authorized the US Public Health Service to collect morbidity data on cholera, smallpox, and yellow fever; each state in the U.S. (as well as the USAF) has its own list of notifiable diseases and depends largely on reporting by the individual health care provider.

**Disease Outbreaks** Sudden increase in the incidence of a disease. The concept includes epidemics.

**Disease Transmission** The transmission of infectious disease or pathogens. When transmission is within the same species, the mode can be horizontal or vertical.

**Disease Transmission, Horizontal** The transmission of infectious disease or pathogens from one individual to another in the same generation.

**Disease Transmission, Patient-to-Professional** The transmission of infectious disease or pathogens from patients to health professionals or health care workers. It includes transmission via direct or indirect exposure to bacterial, fungal, parasitic, or viral agents.

**Disease Transmission, Professional-to-Patient** The transmission of infectious disease or pathogens from health professional or health care worker to patients. It includes transmission via direct or indirect exposure to bacterial, fungal, parasitic, or viral agents

**Disease Transmission, Vertical** The transmission of infectious disease or pathogens from one generation to another. It includes transmission in utero or intrapartum by exposure to blood and secretions, and postpartum exposure via breastfeeding.

**Disease Vectors** Invertebrates or non-human vertebrates which transmit infective organisms from one host to another.

**Disjoint or Mutually Exclusive Events** Two events are disjoint or mutually exclusive if the occurrence of one is incompatible with the occurrence of the other; that is, if they can't both happen at once (if they have no outcome in common). Equivalently, two events are disjoint if their intersection is the empty set.

**Distribution** The distribution of a set of numerical data is how their values are distributed over the real numbers. It is completely characterized by the empirical distribution function. Similarly, the probability distribution of a random variable is completely characterized by its probability distribution function. Sometimes the word "distribution" is used as a synonym for the empirical distribution function or the probability distribution function.

**Distribution Function, Empirical** The empirical (cumulative) distribution function of a set of numerical data is, for each real value of  $x$ , the fraction of observations that are less than or equal to  $x$ . A plot of the empirical distribution function is an uneven set of stairs. The width of the stairs is the spacing between adjacent data; the height of the stairs depends on how many data have exactly the same value. The distribution function is zero for small enough (negative) values of  $x$ , and is unity for large enough values of  $x$ . It increases monotonically: if  $y > x$ , the empirical distribution function evaluated at  $y$  is at least as large as the empirical distribution function evaluated at  $x$ .

**Distribution (or Probability Distribution)** A mathematical function characterized by constants, called parameters, that relate the values that a variable can assume to the probability that a particular value will occur.

**Double-Blind, Double-Blind Experiment** In a double-blind experiment, neither the subjects nor the people evaluating the subjects knows who is in the treatment group and who is in the control group. This mitigates the placebo effect and guards against conscious and unconscious prejudice for or against the treatment on the part of the evaluators.

**Double-Blind Method** A method of studying a drug or procedure in which both the subjects and investigators are kept unaware of who is actually getting which specific treatment.

---

[top](#)

# E

**Ecological Correlation** The correlation between averages of groups of individuals, instead of individuals. Ecological correlation can be misleading about the association of individuals.

**Effect Modifiers (Epidemiology)** Factors that modify the effect of the putative causal factor(s) under study.

**Empirical Law of Averages** The Empirical Law of Averages lies at the base of the frequency theory of probability. This law, which is, in fact, an assumption about how the world works, rather than a mathematical or physical law, states that if one repeats a random experiment over and over, independently and under "identical" conditions, the fraction of trials that result in a given outcome converges to a limit as the number of trials grows without bound.

**Empty Set** The empty set, denoted  $\{\}$  or  $\emptyset$ , is the set that has no members.

**Endpoint Convention** In plotting a histogram, one must decide whether to include a datum that lies at a class boundary with the class interval to the left or the right of the boundary. The rule for making this assignment is called an *endpoint convention*. The two standard endpoint conventions are (1) to include the left endpoint of all class intervals and exclude the right, except for the rightmost class interval, which includes both of its endpoints, and (2) to include the right endpoint of all class intervals and exclude the left, except for the leftmost interval, which includes both of its endpoints.

**Estimator** An estimator is a rule for "guessing" the value of a population parameter based on a random sample from the population. An estimator is a random variable, because its value depends on which particular sample is obtained, which is random. A canonical example of an estimator is the sample mean, which is an estimator of the population mean.

**Event** An *event* is a subset of outcome space. An *event* determined by a random variable is an event of the form  $A = \{X \text{ is in } A\}$ . When the random variable  $X$  is observed, that *determines* whether or not  $A$  occurs: if the value of  $X$  happens to be in  $A$ ,  $A$  occurs; if not,  $A$  does not occur.

**Exhaustive** A collection of events  $\{A_1, A_2, A_3, \dots\}$  is *exhaustive* if at least one of them must occur; that is, if  $\mathbf{S} = A_1 \cup A_2 \cup A_3 \cup \dots$  where  $\mathbf{S}$  is the outcome space. A collection of subsets *exhausts* another set if that set is contained in the  $\cup$  union of the collection.

**Expectation, Expected Value** The expected value of a random variable is the long-term limiting average of its values in independent repeated experiments. The expected value of the random variable  $X$  is denoted  $EX$  or  $E(X)$ . For a discrete random variable (one that has a countable number of possible values) the expected value is the weighted average of its possible values, where the weight assigned to each possible value is the chance that the random variable takes that value. One can think of the expected value of a random variable as the point at which its probability histogram would balance, if it were cut out of a uniform material. Taking the expected value is a linear operation: if  $X$  and  $Y$  are two random variables, the expected value of their sum is the sum of their expected values ( $E(X+Y) = E(X) + E(Y)$ ), and the expected value of a constant  $a$  times a random variable  $X$  is the constant times the expected value of  $X$  ( $E(ax) = aE(X)$ ).

**Experiment** What distinguishes an experiment from an observational study is that in an experiment, the experimenter decides who receives the treatment.

**Explanatory Variable** In regression, the explanatory or independent variable is the one that is supposed to "explain" the other. For example, in examining crop yield versus quantity of fertilizer applied, the quantity of fertilizer would be the explanatory or independent variable, and the crop yield would be the dependent variable. In experiments, the explanatory variable is the one that is manipulated; the one that is observed is the dependent variable.

**Extrapolation** See interpolation.

---

[top](#)

## F

**Factor Analysis, Statistical** A set of statistical methods for analyzing the correlations among several variables in order to estimate the number of fundamental dimensions that underlie the observed data and to describe and measure those dimensions. It is used frequently in the development of scoring systems for rating scales and questionnaires.

**Factorial** For an integer  $k$  that is greater than or equal to 1,  $k!$  (pronounced " $k$  factorial") is  $k \times (k-1) \times (k-2) \times \dots \times 1$ . By convention,  $0! = 1$ . There are  $k!$  ways of ordering  $k$  distinct objects. For example,  $9!$  is the number of batting orders of 9 baseball players, and  $52!$  is the number of different ways a standard deck of playing cards can be ordered. The calculator above has a button to compute the factorial of a number. To compute  $k!$ , first type the value of  $k$ , then press the button labeled "!".

**False Discovery Rate** In testing a collection of hypotheses, the false discovery rate is the fraction of rejected null hypotheses that are rejected erroneously (the number of Type I errors

divided by the number of rejected null hypotheses), with the convention that if no hypothesis is rejected, the false discovery rate is zero.

**Family Characteristics** Size and composition of the family.

**Fatal Outcome** Death resulting from the presence of a disease in an individual, as shown by a single case report or a limited number of patients. This should be differentiated from death, the physiological cessation of life and from mortality, an epidemiological or statistical concept.

**Finite Population Correction** When sampling without replacement, as in a simple random sample, the SE of sample sums and sample means depends on the fraction of the population that is in the sample: the greater the fraction, the smaller the SE. Sampling with replacement is like sampling from an infinitely large population. The adjustment to the SE for sampling without replacement is called the finite population correction. The SE for sampling without replacement is smaller than the SE for sampling with replacement by the finite population correction factor  $((N - n)/(N - 1))^{1/2}$ . Note that for sample size  $n=1$ , there is no difference between sampling with and without replacement; the finite population correction is then unity. If the sample size is the entire population of  $N$  units, there is no variability in the result of sampling without replacement (every member of the population is in the sample exactly once), and the SE should be zero. This is indeed what the finite population correction gives (the numerator vanishes).

**Fisher's exact test (for the equality of two percentages)** Consider two populations of zeros and ones. Let  $p_1$  be the proportion of ones in the first population, and let  $p_2$  be the proportion of ones in the second population. We would like to test the null hypothesis that  $p_1 = p_2$  on the basis of a simple random sample from each population. Let  $n_1$  be the size of the sample from population 1, and let  $n_2$  be the size of the sample from population 2. Let  $G$  be the total number of ones in both samples. If the null hypothesis be true, the two samples are like one larger sample from a single population of zeros and ones. The allocation of ones between the two samples would be expected to be proportional to the relative sizes of the samples, but would have some chance variability. Conditional on  $G$  and the two sample sizes, under the null hypothesis, the tickets in the first sample are like a random sample of size  $n_1$  without replacement from a collection of  $N = n_1 + n_2$  units of which  $G$  are labeled with ones. Thus, under the null hypothesis, the number of tickets labeled with ones in the first sample has (conditional on  $G$ ) a hypergeometric distribution with parameters  $N$ ,  $G$ , and  $n_1$ . Fisher's exact test uses this distribution to set the ranges of observed values of the number of ones in the first sample for which we would reject the null hypothesis.

**Football-Shaped Scatterplot** In a football-shaped scatterplot, most of the points lie within a tilted oval, shaped more-or-less like a football. A football-shaped scatterplot is one in which the data are homoscedastically scattered about a straight line.

**Frame, sampling frame** A *sampling frame* is a collection of units from which a sample will be drawn. Ideally, the frame is identical to the population we want to learn about; more typically, the frame is only a subset of the population of interest. The difference between the frame and the population can be a source of bias in sampling design, if the parameter of interest has a different value for the frame than it does for the population. For example, one might desire to

estimate the current annual average income of 1998 graduates of the University of California at Berkeley. I propose to use the sample mean income of a sample of graduates drawn at random. To facilitate taking the sample and contacting the graduates to obtain income information from them, I might draw names at random from the list of 1998 graduates for whom the alumni association has an accurate current address. The population is the collection of 1998 graduates; the frame is those graduates who have current addresses on file with the alumni association. If there is a tendency for graduates with higher incomes to have up-to-date addresses on file with the alumni association, that would introduce a positive bias into the annual average income estimated from the sample by the sample mean.

**Frequency theory of probability** See Probability, Theories of.

**Frequency table** A table listing the frequency (number) or relative frequency (fraction or percentage) of observations in different ranges, called class intervals.

**Fundamental Rule of Counting** If a sequence of experiments or trials  $T_1, T_2, T_3, \dots, T_k$  could result, respectively, in  $n_1, n_2, n_3, \dots, n_k$  possible outcomes, and the numbers  $n_1, n_2, n_3, \dots, n_k$  do not depend on which outcomes actually occurred, the entire *sequence* of  $k$  experiments has  $n_1 \times n_2 \times n_3 \times \dots \times n_k$  possible outcomes.

---

[top](#)

## G

**Genetic Screening** Searching a population or individuals for persons possessing certain genotypes or karyotypes that: (1) are already associated with disease or predispose to disease; (2) may lead to disease in their descendants; or (3) produce other variations not known to be associated with disease. Genetic screening may be directed toward identifying phenotypic expression of genetic traits. It includes prenatal genetic screening.

**Geometric Distribution** The geometric distribution describes the number of trials up to and including the first success, in independent trials with the same probability of success. The geometric distribution depends only on the single parameter  $p$ , the probability of success in each trial. For example, the number of times one must toss a fair coin until the first time the coin lands heads has a geometric distribution with parameter  $p = 50\%$ . The geometric distribution assigns probability  $p \times (1 - p)^{k-1}$  to the event that it takes  $k$  trials to the first success. The expected value of the geometric distribution is  $1/p$ , and its SE is  $(1-p)^{1/2}/p$ .

**Geometric Mean** The geometric mean of  $n$  numbers  $\{x_1, x_2, x_3, \dots, x_n\}$  is the  $n$ th root of their product:  $(x_1 \times x_2 \times x_3 \times \dots \times x_n)^{1/n}$ .

**Geriatric Assessment** Evaluation of the level of physical, physiological, or mental functioning in the older population group.

**Graph of Averages** For bivariate data, a graph of averages is a plot of the average values of one variable (say  $y$ ) for small ranges of values of the other variable (say  $x$ ), against the value of the second variable ( $x$ ) at the midpoints of the ranges.

**Gravidity** The number of pregnancies, complete or incomplete, experienced by a female. It is different from parity, which is the number of offspring born.

---

[top](#)

# H

**Health Status** The level of health of the individual, group, or population as subjectively assessed by the individual or by more objective measures.

**Health Status Indicators** The measurement of the health status for a given population using a variety of indices, including morbidity, mortality, and available health resources.

**Health Surveys** A systematic collection of factual data pertaining to health and disease in a human population within a given geographic area.

**Health Transition** Demographic and epidemiologic changes that have occurred in the last five decades in many developing countries and that are characterized by major growth in the number and proportion of middle-aged and elderly persons and in the frequency of the diseases that occur in these age groups. The health transition is the result of efforts to improve maternal and child health via primary care and outreach services and such efforts have been responsible for a decrease in the birth rate; reduced maternal mortality; improved preventive services; reduced infant mortality, and the increased life expectancy that defines the transition.

**Heteroscedasticity** "Mixed scatter." A scatterplot or residual plot shows heteroscedasticity if the scatter in vertical slices through the plot depends on where you take the slice. Linear regression is not usually a good idea if the data are heteroscedastic.

**Histogram** A histogram is a kind of plot that summarizes how data are distributed. Starting with a set of class intervals, the histogram is a set of rectangles ("bins") sitting on the horizontal axis. The bases of the rectangles are the class intervals, and their heights are such that their areas are proportional to the fraction of observations in the corresponding class intervals. That is, the height of a given rectangle is the fraction of observations in the corresponding class interval, divided by the length of the corresponding class interval. A histogram does not need a vertical scale, because the total area of the histogram must equal 100%. The units of the vertical axis are percent per unit of the horizontal axis. This is called the *density scale*. The horizontal axis of a histogram needs a scale. If any observations coincide with the endpoints of class intervals, the endpoint convention is important.

**Historical Controls** Sometimes, the a treatment group is compared with individuals from another epoch who did not receive the treatment; for example, in studying the possible effect of fluoridated water on childhood cancer, we might compare cancer rates in a community before and after fluorine was added to the water supply. Those individuals who were children before fluoridation started would comprise an historical control group. Experiments and studies with historical controls tend to be more susceptible to confounding than those with contemporary controls, because many factors that might affect the outcome other than the treatment tend to change over time as well. (In this example, the level of other potential carcinogens in the environment also could have changed.)

**Homoscedasticity** "Same scatter." A scatterplot or residual plot shows homoscedasticity if the scatter in vertical slices through the plot does not depend much on where you take the slice. *C.f.* heteroscedasticity.

**Hospital Mortality** A vital statistic measuring or recording the rate of death from any cause in hospitalized populations.

**Hospital Records** Compilations of data on hospital activities and programs; excludes patient medical records.

**Hypergeometric Distribution** The hypergeometric distribution with parameters  $N$ ,  $G$  and  $n$  is the distribution of the number of "good" objects in a simple random sample of size  $n$  (*i.e.*, a random sample without replacement in which every subset of size  $n$  has the same chance of occurring) from a population of  $N$  objects of which  $G$  are "good." The chance of getting exactly  $g$  good objects in such a sample is:  $\frac{{}_G C_g \times {}_{N-G} C_{n-g}}{{}_N C_n}$ , provided  $g \leq n$ ,  $g \leq G$ , and  $n - g \leq N - G$ . (The probability is zero otherwise.) The expected value of the hypergeometric distribution is  $n \times G/N$ , and its standard error is:  $\left(\frac{(N-n)(N-1)}{N}\right)^{1/2} \times \left(n \times \frac{G}{N} \times \left(1 - \frac{G}{N}\right)\right)^{1/2}$ .

**Hypothesis testing** Statistical hypothesis testing is formalized as making a decision between rejecting or not rejecting a null hypothesis, on the basis of a set of observations. Two types of errors can result from any decision rule (test): rejecting the null hypothesis when it is true (a Type I error), and failing to reject the null hypothesis when it is false (a Type II error). For any hypothesis, it is possible to develop many different decision rules (tests). Typically, one specifies ahead of time the chance of a Type I error one is willing to allow. That chance is called the significance level of the test or decision rule. For a given significance level, one way of deciding which decision rule is best is to pick the one that has the smallest chance of a Type II error when a given alternative hypothesis is true. The chance of correctly rejecting the null hypothesis when a given alternative hypothesis is true is called the power of the test against that alternative hypothesis.

**IFF, if and only if** If  $p$  and  $q$  are two logical propositions, then  $(p \text{ IFF } q)$  is a proposition that is true when both  $p$  and  $q$  are true, and when both  $p$  and  $q$  are false. It is logically equivalent to the proposition:  $((p \text{ IMPLIES } q) \text{ AND } (q \text{ IMPLIES } p))$  and to the proposition  $((p \text{ AND } q) \text{ OR } ((\text{NOT } p) \text{ AND } (\text{NOT } q)))$ .

**Implies, logical implication** Logical implication is an operation on two logical propositions. If  $p$  and  $q$  are two logical propositions,  $(p \text{ IMPLIES } q)$  is a logical proposition that is true if  $p$  is false, or if both  $p$  and  $q$  are true. The proposition  $(p \text{ IMPLIES } q)$  is logically equivalent to the proposition  $((\text{NOT } p) \text{ OR } q)$ .

**Infant Mortality** Perinatal, neonatal, and infant deaths in a given population.

**Incidence** The number of new cases of a given disease during a given period in a specified population. It also is used for the rate at which new events occur in a defined population. It is differentiated from **prevalence**; which refers to all cases, new or old, in the population at a given time.

**Independent and identically distributed (iid)** A collection of two or more random variables  $\{X_1, X_2, \dots, \}$  is *independent and identically distributed* if the variables have the same probability distribution, and are independent.

**Independent, independence** Two events  $A$  and  $B$  are (statistically) independent if the chance that they both happen simultaneously is the product of the chances that each occurs individually; *i.e.*, if  $P(AB) = P(A)P(B)$ . This is essentially equivalent to saying that learning that one event occurs does not give any information about whether the other event occurred too: the conditional probability of  $A$  given  $B$  is the same as the unconditional probability of  $A$ , *i.e.*,  $P(A|B) = P(A)$ . Two random variables  $X$  and  $Y$  are independent if all events they determine are independent, for example, if the event  $\{a < X \leq b\}$  is independent of the event  $\{c < Y \leq d\}$  for **all** choices of  $a, b, c,$  and  $d$ . A collection of more than two random variables is independent if for every proper subset of the variables, every event determined by that subset of the variables is independent of every event determined by the variables in the complement of the subset. For example, the three random variables  $X, Y,$  and  $Z$  are independent if every event determined by  $X$  is independent of every event determined by  $Y$  and every event determined by  $X$  is independent of every event determined by  $Y$  and  $Z$  and every event determined by  $Y$  is independent of every event determined by  $X$  and  $Z$  and every event determined by  $Z$  is independent of every event determined by  $X$  and  $Y$ .

**Independent Variable** In regression, the independent variable is the one that is supposed to explain the other; the term is a synonym for "explanatory variable." Usually, one regresses the "dependent variable" on the "independent variable." There is not always a clear choice of the independent variable. The independent variable is usually plotted on the horizontal axis. Independent in this context does not mean the same thing as statistically independent.

**Indicator Random Variable** The indicator [random variable] of the event A, often written  $1_A$ , is the random variable that equals unity if A occurs, and zero if A does not occur. The expected value of the indicator of A is the probability of A,  $P(A)$ , and the standard error of the indicator of A is  $(P(A) \times (1-P(A)))^{1/2}$ . The sum  $1_A + 1_B + 1_C + \dots$  of the indicators of a collection of events  $\{A, B, C, \dots\}$  counts how many of the events  $\{A, B, C, \dots\}$  occur in a given trial. The product of the indicators of a collection of events is the indicator of the intersection of the events (the product equals one if and only if all of indicators equal one). The maximum of the indicators of a collection of events is the indicator of the union of the events (the maximum equals one if any of the indicators equals one).

**Insect Vectors** Insects that transmit infective organisms from one host to another or from an inanimate reservoir to an animate host.

**Inter-quartile Range (IQR)** The inter-quartile range of a list of numbers is the upper quartile minus the lower quartile.

**Interpolation** Given a set of bivariate data  $(x, y)$ , to impute a value of  $y$  corresponding to some value of  $x$  at which there is no measurement of  $y$  is called interpolation, if the value of  $x$  is within the range of the measured values of  $x$ . If the value of  $x$  is outside the range of measured values, imputing a corresponding value of  $y$  is called extrapolation.

**Intersection** The intersection of two or more sets is the set of elements that all the sets have in common; the elements contained in every one of the sets. The intersection of the events A and B is written "A and B" and "AB." *C.f.* union. See also Venn diagrams.

**Intervention Studies** Epidemiologic investigations designed to test a hypothesized cause-effect relation by modifying the supposed causal factor(s) in the study population.

**Interviews** Conversations with an individual or individuals held in order to obtain information about their background and other personal biographical data, their attitudes and opinions, etc. It includes school admission or job interviews.

---

[top](#)

**J**

**Joint Probability Distribution** If  $X_1, X_2, \dots, X_k$  are random variables, their *joint probability distribution* gives the probability of events determined by the collection of random variables: for any

collection of sets of numbers  $\{A_1, \dots, A_k\}$ , the joint probability distribution determines  $P(X_1 \text{ is in } A_1)$  and  $(X_2 \text{ is in } A_2)$  and  $\dots$  and  $(X_k \text{ is in } A_k)$ .

---

[top](#)

## K

**Kaplan-Meier Method (or Product Limit Method)** A method for analyzing survival data, based on the distribution of variable time periods between events (or deaths).

**Karnofsky Performance Status** A performance measure for rating the ability of a person to perform usual activities, evaluating a patient's progress after a therapeutic procedure, and determining a patient's suitability for therapy. It is used most commonly in the prognosis of cancer therapy, usually after chemotherapy and customarily administered before and after therapy.

---

[top](#)

## L

**Law of Averages** The Law of Averages says that the average of independent observations of random variables that have the same probability distribution is increasingly likely to be close to the expected value of the random variables as the number of observations grows. More precisely, if  $X_1, X_2, X_3, \dots$ , are independent random variables with the same probability distribution, and  $E(X)$  is their common expected value, then for every number  $E > 0$ ,  $P\{|(X_1 + X_2 + \dots + X_n)/n - E(X)| < E\}$  converges to 100% as  $n$  grows. This is equivalent to saying that the sequence of sample means  $X_1, (X_1+X_2)/2, (X_1+X_2+X_3)/3, \dots$  converges in probability to  $E(X)$ .

**Law of Large Numbers** The Law of Large Numbers says that in repeated, independent trials with the same probability  $p$  of success in each trial, the percentage of successes is increasingly likely to be close to the chance of success as the number of trials increases. More precisely, the chance that the percentage of successes differs from the probability  $p$  by more than a fixed positive amount,  $E > 0$ , converges to zero as the number of trials  $n$  goes to infinity, for every number  $e > 0$ . Note that in contrast to the difference between the *percentage* of successes and the probability of success, the difference between the *number* of successes and the expected number of successes,  $n \times p$ , tends to grow as  $n$  grows. The following tool illustrates the law of large numbers; the button toggles between displaying the difference

between the number of successes and the expected number of successes, and the difference between the percentage of successes and the expected percentage of successes.

**Life Expectancy** A figure representing the number of years, based on known statistics, to which any person of a given age may reasonably expect to live.

**Life Tables** Summarizing techniques used to describe the pattern of mortality and survival in populations. These methods can be applied to the study not only of death, but also of any defined endpoint such as the onset of disease or the occurrence of disease complications.

**Life Table Method** A method for analyzing survival data, based on the proportion of study subjects surviving to fixed time intervals after treatment or study initiation.

**Least-Squares Analysis** A principle of estimation in which the estimates of a set of parameters in a statistical model are those quantities minimizing the sum of squared differences between the observed values of a dependent variable and the values predicted by the model.

**Likelihood Functions** Functions constructed from a statistical model and a set of observed data which give the probability of that data for various values of the unknown model parameters. Those parameter values that maximize the probability are the maximum likelihood estimates of the parameters.

**Limit** See *converge*.

**Linear association** Two variables are linearly associated if a change in one is associated with a proportional change in the other, with the same constant of proportionality throughout the range of measurement. The correlation coefficient measures the degree of linear association on a scale of -1 to 1.

**Linear Models** Statistical models in which the value of a parameter for a given value of a factor is assumed to be equal to  $a + bx$ , where  $a$  and  $b$  are constants. The models predict a linear regression.

**Linear Operation** Suppose  $f$  is a function or operation that acts on things we shall denote generically by the lower-case Roman letters  $x$  and  $y$ . Suppose it makes sense to multiply  $x$  and  $y$  by numbers (which we denote by  $a$ ), and that it makes sense to add things like  $x$  and  $y$  together. We say that  $f$  is *linear* if for every number  $a$  and every value of  $x$  and  $y$  for which  $f(x)$  and  $f(y)$  are defined, (i)  $f(ax)$  is defined and equals  $axf(x)$ , and (ii)  $f(x + y)$  is defined and equals  $f(x) + f(y)$ . *C.f.* affine.

**Linear Regression Method** For a single item, a method for determining the best-fit line through points representing the paired values of two measurement systems (one representing a dependent variable and the other representing an independent variable). Under certain conditions, statistical tests of the slope and intercept can be made, and confidence intervals about the line can be computed.

**Location, Measure of** A measure of location is a way of summarizing what a "typical" element of a list is---it is a one-number summary of a distribution. See also arithmetic mean, median, and mode.

**Log-Linear Modeling Techniques** Methods for analyzing qualitative data in which a function of the probability that a particular event will occur is logarithmically transformed to fit a linear model.

**Logistic Models** Statistical models which describe the relationship between a qualitative dependent variable (that is, one which can take only certain discrete values, such as the presence or absence of a disease) and an independent variable. A common application is in epidemiology for estimating an individual's risk (probability of a disease) as a function of a given risk factor.

**Logistic Regression Method** A specialized log-linear modeling technique in which the logarithm of the proportion of a group having a particular characteristic, divided by one minus that proportion, is fit into a multiple regression linear model.

**Longitudinal study** A study in which individuals are followed over time, and compared with themselves at different times, to determine, for example, the effect of aging on some measured variable. Longitudinal studies provide much more persuasive evidence about the effect of aging than do cross-sectional studies.

---

[top](#)

## M

**Margin of error** A measure of the uncertainty in an estimate of a parameter; unfortunately, not everyone agrees what it should mean. The *margin of error* of an estimate is typically one or two times the estimated standard error of the estimate.

**Markov's Inequality** *For lists:* If a list contains no negative numbers, the fraction of numbers in the list at least as large as any given constant  $a > 0$  is no larger than the arithmetic mean of the list, divided by  $a$ . *For random variables:* if a random variable  $X$  must be nonnegative, the chance that  $X$  exceeds any given constant  $a > 0$  is no larger than the expected value of  $X$ , divided by  $a$ .

**Mass Screening** Organized periodic procedures performed on large groups of people for the purpose of detecting disease.

**Matched-Pair Analysis** A type of analysis in which subjects in a study group and a comparison group are made comparable with respect to extraneous factors by individually pairing study subjects with the comparison group subjects (e.g., age-matched controls).

**Maternal Mortality** Maternal deaths resulting from complications of pregnancy and childbirth in a given population.

**Maximum Likelihood Estimate (MLE)** The maximum likelihood estimate of a parameter from data is the possible value of the parameter for which the chance of observing the data largest. That is, suppose that the parameter is  $p$ , and that we observe data  $x$ . Then the maximum likelihood estimate of  $p$  is: estimate  $p$  by the value  $q$  that makes  $P(\text{observing } x \text{ when the value of } p \text{ is } q)$  as large as possible. For example, suppose we are trying to estimate the chance that a (possibly biased) coin lands heads when it is tossed. Our data will be the number of times  $x$  the coin lands heads in  $n$  independent tosses of the coin. The distribution of the number of times the coin lands heads is binomial with parameters  $n$  (known) and  $p$  (unknown). The chance of observing  $x$  heads in  $n$  trials if the chance of heads in a given trial is  $q$  is  ${}_n C_x q^x (1-q)^{n-x}$ . The maximum likelihood estimate of  $p$  would be the value of  $q$  that makes that chance largest. We can find that value of  $q$  explicitly using calculus; it turns out to be  $q = x/n$ , the fraction of times the coin is observed to land heads in the  $n$  tosses. Thus the maximum likelihood estimate of the chance of heads from the number of heads in  $n$  independent tosses of the coin is the observed fraction of tosses in which the coin lands heads.

**Mean, Arithmetic mean** The sum of a list of numbers, divided by the number of numbers. See also average.

**Mean Squared Error (MSE)** The mean squared error of an estimator of a parameter is the expected value of the square of the difference between the estimator and the parameter. In symbols, if  $X$  is an estimator of the parameter  $t$ , then  $MSE(X) = E((X-t)^2)$ . The MSE measures how far the estimator is off from what it is trying to estimate, on the average in repeated experiments. It is a summary measure of the accuracy of the estimator. It combines any tendency of the estimator to overshoot or undershoot the truth (bias), and the variability of the estimator (SE). The MSE can be written in terms of the bias and SE of the estimator:  $MSE(X) = (\text{bias}(X))^2 + (\text{SE}(X))^2$ .

**Median** "Middle value" of a list. The smallest number such that at least half the numbers in the list are no greater than it. If the list has an odd number of entries, the median is the middle entry in the list after sorting the list into increasing order. If the list has an even number of entries, the median is the smaller of the two middle numbers after sorting. The median can be estimated from a histogram by finding the smallest number such that the area under the histogram to the left of that number is 50%.

**Medical Records** Recording of pertinent information concerning patient's illness or illnesses.

**Member of a set** Something is a member (or element) of a set if it is one of the things in the set.

**Method of Comparison** The most basic and important method of determining whether a treatment has an effect: compare what happens to individuals who are treated (the treatment group) with what happens to individuals who are not treated (the control group).

**Mode** For lists, the mode is a most common (frequent) value. A list can have more than one mode. For histograms, a mode is a relative maximum ("bump").

**Models, Statistical** Statistical formulations or analyses which, when applied to data and found to fit the data, are then used to verify the assumptions and parameters used in the analysis. Examples of statistical models are the linear model, binomial model, polynomial model, two-parameter model, etc.

**Moment** The  $k$ th moment of a list is the average value of the elements raised to the  $k$ th power; that is, if the list consists of the  $N$  elements  $x_1, x_2, \dots, x_N$ , the  $k$ th moment of the list is:  $(x_1^k + x_2^k + \dots + x_N^k)/N$ . The  $k$ th moment of a random variable  $X$  is the expected value of  $X^k$ ,  $E(X^k)$ .

**Morbidity** The proportion of patients with a particular disease during a given year per given unit of population.

**Mortality** All deaths reported in a given population.

**Multimodal Distribution** A distribution with more than one mode.

**Multinomial Distribution** Consider a sequence of  $n$  independent trials, each of which can result in an outcome in any of  $k$  categories. Let  $p_j$  be the probability that each trial results in an outcome in category  $j$ ,  $j = 1, 2, \dots, k$ , so  $p_1 + p_2 + \dots + p_k = 100\%$ . The number of outcomes of each type has a *multinomial distribution*. In particular, the probability that the  $n$  trials result in  $n_1$  outcomes of sub> outcomes of type 2,  $\dots$ , and  $n_k$  outcomes of type  $k$  is...  $n!/(n_1! \times n_2! \times \dots \times n_k!) \times p_1^{n_1} \times p_2^{n_2} \times \dots \times p_k^{n_k}$ , if  $n_1, \dots, n_k$  are nonnegative integers that sum to  $n$ ; the chance is zero otherwise.

**Multiphasic Screening** The simultaneous use of multiple laboratory procedures for the detection of various diseases. These are usually performed on groups of people.

**Multiple Regression Analysis** A multivariate extension of linear regression in which two or more independent variables are fit into a best linear model of a dependent variable.

**Multiplication rule** The chance that events A and B both occur (*i.e.*, that event AB occurs), is the conditional probability that A occurs given that B occurs, times the unconditional probability that B occurs.

**Multiplicity in hypothesis tests** In hypothesis testing, if more than one hypothesis is tested, the actual significance level of the combined tests is not equal to the nominal significance level of the individual tests.

**Multivariate Analysis** A set of techniques used when variation in several variables has to be studied simultaneously. In statistics, multivariate analysis is interpreted as any analytic method that allows simultaneous study of two or more dependent variables

**Multivariate Data** A set of measurements of two or more variables per individual. See bivariate.

**Mutually Exclusive** Two events are mutually exclusive if the occurrence of one is incompatible with the occurrence of the other; that is, if they can't both happen at once (if they have no outcome in common). Equivalently, two events are disjoint if their intersection is the empty set.

---

[top](#)

## N

**Nearly normal distribution** A population of numbers (a list of numbers) is said to have a *nearly normal distribution* if the histogram of its values in standard units nearly follows a normal curve. More precisely, suppose that the mean of the list is  $\mu$  and the standard deviation of the list is SD. Then the list is nearly normally distributed if, for every two numbers  $a < b$ , the fraction of numbers in the list that are between  $a$  and  $b$  is approximately equal to the area under the normal curve between  $(a - \mu)/SD$  and  $(b - \mu)/SD$ .

**Negative Binomial Distribution** Consider a sequence of independent trials with the same probability  $p$  of success in each trial. The number of trials up to and including the  $r$ th success has the negative Binomial distribution with parameters  $n$  and  $r$ . If the random variable  $N$  has the negative binomial distribution with parameters  $n$  and  $r$ , then  $P(N=k) = {}_{k-1}C_{r-1} \times p^r \times (1-p)^{k-r}$ , for  $k = r, r+1, r+2, \dots$ , and zero for  $k < r$ , because there must be at least  $r$  trials to have  $r$  successes. The negative binomial distribution is derived as follows: for the  $r$ th success to occur on the  $k$ th trial, there must have been  $r-1$  successes in the first  $k-1$  trials, and the  $k$ th trial must result in success. The chance of the former is the chance of  $r-1$  successes in  $k-1$  independent trials with the same probability of success in each trial, which, according to the Binomial distribution with parameters  $n=k-1$  and  $p$ , has probability  ${}_{k-1}C_{r-1} \times p^{r-1} \times (1-p)^{k-r}$ . The chance of the latter event is  $p$ , by assumption. Because the trials are independent, we can find the chance that both events occur by multiplying their chances together, which gives the expression for  $P(N=k)$  above.

**Neonatal Screening** The identification of selected parameters in newborn infants by various tests, examinations, or other procedures. Screening may be performed by clinical or laboratory measures. A screening test is designed to sort out healthy neonates from those not well, but the screening test is not intended as a diagnostic device, rather instead as epidemiologic.

**Nonlinear Association** The relationship between two variables is nonlinear if a change in one is associated with a change in the other that is depends on the value of the first; that is, if the change in the second is not simply proportional to the change in the first, independent of the value of the first variable.

**Nonparametric Statistics** A class of statistical methods applicable to a large set of probability distributions used to test for correlation, location, independence, etc. In most nonparametric statistical tests, the original scores or observations are replaced by another variable containing less information. An important class of nonparametric tests employs the ordinal properties of the data. Another class of tests uses information about whether an observation is above or below some fixed value such as the median, and a third class is based on the frequency of the occurrence of runs in the data.

**Nonparametric Tests** Hypothesis tests that do not require data to be consistent with any particular theoretical distribution, such as normal distribution.

**Nonresponse** In surveys, it is rare that everyone who is "invited" to participate (everyone whose phone number is called, everyone who is mailed a questionnaire, everyone an interviewer tries to stop on the street . . .) in fact responds. The difference between the "invited" sample sought, and that obtained, is the nonresponse.

**Nonresponse bias** In a survey, those who respond may differ from those who do not, in ways that are related to the effect one is trying to measure. For example, a telephone survey of how many hours people work is likely to miss people who are working late, and are therefore not at home to answer the phone. When that happens, the survey may suffer from nonresponse bias. Nonresponse bias makes the result of a survey differ systematically from the truth.

**Nonresponse rate** The fraction of nonresponders in a survey: the number of nonresponders divided by the number of people invited to participate (the number sent questionnaires, the number of interview attempts, etc.) If the nonresponse rate is appreciable, the survey suffer from large nonresponse bias.

**Normal approximation** The normal approximation to data is to approximate areas under the histogram of data, transformed into standard units, by the corresponding areas under the normal curve. Many probability distributions can be approximated by a normal distribution, in the sense that the area under the probability histogram is close to the area under a corresponding part of the normal curve. To find the corresponding part of the normal curve, the range must be converted to standard units, by subtracting the expected value and dividing by the standard error. For example, the area under the binomial probability histogram for  $n = 50$  and  $p = 30\%$  between 9.5 and 17.5 is 74.2%. To use the normal approximation, we transform the endpoints to standard units, by subtracting the expected value (for the Binomial random variable,  $np = 15$  for these values of  $n$  and  $p$ ) and dividing the result by the standard error (for a Binomial,  $(n \times p \times (1-p))^{1/2} = 3.24$  for these values of  $n$  and  $p$ ). The area normal approximation is the area under the normal curve between  $(9.5 - 15)/3.24 = -1.697$  and  $(17.5 - 15)/3.24 = 0.772$ ; that area is 73.5%, slightly smaller than the corresponding area under the binomial histogram. See also the continuity correction.

**Normal curve** The normal curve is the familiar "bell curve." The mathematical expression for the normal curve is  $y = (2\pi)^{-1/2} e^{-x^2/2}$ , where  $\pi$  is the ratio of the circumference of a circle to its

diameter (3.14159265 . . . ), and  $E$  is the base of the natural logarithm (2.71828 . . . ). The normal curve is symmetric around the point  $x=0$ , and positive for every value of  $x$ . The area under the normal curve is unity, and the SD of the normal curve, suitably defined, is also unity. Many (but not most) histograms, converted into standard units, approximately follow the normal curve.

**Normal distribution** A random variable  $X$  has a normal distribution with mean  $m$  and standard error  $s$  if for every pair of numbers  $a \leq b$ , the chance that  $a < (X-m)/s < b$  is...  $P(a < (X-m)/s < b) = \text{area under the normal curve between } a \text{ and } b$ . If there are numbers  $m$  and  $s$  such that  $X$  has a normal distribution with mean  $m$  and standard error  $s$ , then  $X$  is said to have a normal distribution or to be normally distributed. If  $X$  has a normal distribution with mean  $m=0$  and standard error  $s=1$ , then  $X$  is said to have a standard normal distribution. The notation  $X \sim N(m, s^2)$  means that  $X$  has a normal distribution with mean  $m$  and standard error  $s$ ; for example,  $X \sim N(0, 1)$ , means  $X$  has a standard normal distribution.

**Normal Distribution** Continuous frequency distribution of infinite range. Its properties are as follows: 1) continuous, symmetrical distribution with both tails extending to infinity; 2) arithmetic mean, mode, and median identical; and 3) shape completely determined by the mean and standard deviation.

**NOT, Negation, Logical Negation** The negation of a logical proposition  $p$ , **NOT**  $p$ , is a proposition that is the logical opposite of  $p$ . That is, if  $p$  is true, **NOT**  $p$  is false, and if  $p$  is false, **NOT**  $p$  is true. Negation takes precedence over other logical operations.

**Number Needed to Treat (NNT)** The number of patients who need to be treated to prevent 1 adverse outcome.

**Null hypothesis** In hypothesis testing, the hypothesis we wish to falsify on the basis of the data. The null hypothesis is typically that something is not present, that there is no effect, or that there is no difference between treatment and control.

---

[top](#)

## O

**Observational Study** *C.f.* controlled experiment.

**Observer Variation** The failure by the observer to measure or identify a phenomenon accurately, which results in an error. Sources for this may be due to the observer's missing an abnormality, or to faulty technique resulting in incorrect test measurement, or to misinterpretation of the data. Two varieties are inter-observer variation (the amount observers vary from one another when reporting on the same material) and intra-observer variation (the

amount one observer varies between observations when reporting more than once on the same material).

**Odds** The *odds in favor of an event* is the ratio of the probability that the event occurs to the probability that the event does not occur. For example, suppose an experiment can result in any of  $n$  possible outcomes, all equally likely, and that  $k$  of the outcomes result in a "win" and  $n-k$  result in a "loss." Then the chance of winning is  $k/n$ ; the chance of not winning is  $(n-k)/n$ ; and the odds in favor of winning are  $(k/n)/((n-k)/n) = k/(n-k)$ , which is the number of favorable outcomes divided by the number of unfavorable outcomes. Note that odds are not synonymous with probability, but the two can be converted back and forth. If the odds in favor of an event are  $q$ , then the probability of the event is  $q/(1+q)$ . If the probability of an event is  $p$ , the odds in favor of the event are  $p/(1-p)$  and the odds against the event are  $(1-p)/p$ .

**One-sided Test** C.f. two-sided test. A hypothesis test of the null hypothesis that the value of a parameter,  $\mu$ , is equal to a null value,  $\mu_0$ , designed to have power against either the alternative hypothesis that  $\mu < \mu_0$  or the alternative  $\mu > \mu_0$  (but not both). For example, a significance level 5%, one-sided  $z$  test of the null hypothesis that the mean of a population equals zero against the alternative that it is greater than zero, would reject the null hypothesis for values of

$$z = \frac{\text{sample mean}}{\text{SE}(\text{sample mean})} > 1.64.$$

**OR, Disjunction, Logical Disjunction** An operation on two logical propositions. If  $p$  and  $q$  are two propositions,  $(p \text{ OR } q)$  is a proposition that is true if  $p$  is true or if  $q$  is true (or both); otherwise, it is false. That is,  $(p \text{ OR } q)$  is true unless both  $p$  and  $q$  are false. C.f. exclusive disjunction, **XOR**.

**Ordinal Variable** A variable whose possible values have a natural order, such as {short, medium, long}, {cold, warm, hot}, or {0, 1, 2, 3, . . .}. In contrast, a variable whose possible values are {straight, curly} or {Arizona, California, Montana, New York} would not naturally be ordinal. Arithmetic with the possible values of an ordinal variable does not necessarily make sense, but it does make sense to say that one possible value is larger than another.

**Outcome Space** The outcome space is the set of all possible outcomes of a given random experiment. The outcome space is often denoted by the capital letter **S**.

**Outlier** An outlier is an observation that is many SD's from the mean. It is sometimes tempting to discard outliers, but this is imprudent unless the cause of the outlier can be identified, and the outlier is determined to be spurious. Otherwise, discarding outliers can cause one to underestimate the true variability of the measurement process.

# P

**P-value** Suppose we have a family of hypothesis tests of a null hypothesis that let us test the hypothesis at any significance level  $p$  between 0 and 100% we choose. The  $P$  value of the null hypothesis given the data is the smallest significance level  $p$  for which any of the tests would have rejected the null hypothesis. For example, let  $X$  be a test statistic, and for  $p$  between 0 and 100%, let  $x_p$  be the smallest number such that, under the null hypothesis,  $P(X \leq x) \geq p$ . Then for any  $p$  between 0 and 100%, the rule reject the null hypothesis if  $X < x_p$  tests the null hypothesis at significance level  $p$ . If we observed  $X = x$ , the  $P$ -value of the null hypothesis given the data would be the smallest  $p$  such that  $x < x_p$ .

**Paired t-Test** A test in which two related samples (such as before and after measurements) arise from a study; the test is based on the difference between the sample values, and the test statistic is called a Student's  $t$ .

**Parameter** A numerical property of a population, such as its mean.

**Parametric Test** A hypothesis test that requires data to conform to some well-known theoretical distribution, such as normal distribution.

**Parity** The number of offspring a female has borne. It is contrasted with **gravidity**; which refers to the number of pregnancies, regardless of outcome.

**Partition** A *partition* of an event  $B$  is a collection of events  $\{A_1, A_2, A_3, \dots\}$  such that the events in the collection are disjoint, and their union is  $B$  (they exhaust  $B$ ). That is,  $A_j A_k = \{\}$  unless  $j = k$ , and  $B = A_1 \cup A_2 \cup A_3 \cup \dots$ . If the event  $B$  is not specified, it is assumed to be the entire outcome space  $S$ .

**Percentile** The  $p$ th percentile of a list is the smallest number such that at least  $p\%$  of the numbers in the list are no larger than it. The  $p$ th percentile of a random variable is the smallest number such that the chance that the random variable is no larger than it is at least  $p\%$ . *C.f.* quantile.

**Permutation** A permutation of a set is an arrangement of the elements of the set in some order. If the set has  $n$  things in it, there are  $n!$  different orderings of its elements. For the first element in an ordering, there are  $n$  possible choices, for the second, there remain  $n-1$  possible choices, for the third, there are  $n-2$ , *etc.*, and for the  $n$ th element of the ordering, there is a single choice remaining. By the fundamental rule of counting, the total number of sequences is thus  $n \times (n-1) \times (n-2) \times \dots \times 1$ . Similarly, the number of orderings of length  $k$  one can form from  $n \geq k$  things is  $n \times (n-1) \times (n-2) \times \dots \times (n-k+1) = n! / (n-k)!$ . This is denoted  ${}_n P_k$ , the number of permutations of  $n$  things taken  $k$  at a time. *C.f.* combinations.

**Placebo** A "dummy" treatment that has no pharmacological effect; *e.g.*, a sugar pill.

**Placebo effect** The belief or knowledge that one is being treated can itself have an effect that confounds with the real effect of the treatment. Subjects given a placebo as a pain-killer report statistically significant reductions in pain in randomized experiments that compare them with subjects who receive no treatment at all. This very real psychological effect of a placebo, which has no direct biochemical effect, is called the placebo effect. Administering a placebo to the control group is thus important in experiments with human subjects; this is the essence of a blind experiment.

**Point of Averages** In a scatterplot, the point whose coordinates are the arithmetic means of the corresponding variables. For example, if the variable  $X$  is plotted on the horizontal axis and the variable  $Y$  is plotted on the vertical axis, the point of averages has coordinates (mean of  $X$ , mean of  $Y$ ).

**Poisson Distribution** The Poisson distribution is a discrete probability distribution that depends on one parameter,  $m$ . If  $X$  is a random variable with the Poisson distribution with parameter  $m$ , then the probability that  $X = k$  is  $E^{-m} \times m^k/k!$ ,  $k = 0, 1, 2, \dots$ , where  $E$  is the base of the natural logarithm and  $!$  is the factorial function. For all other values of  $k$ , the probability is zero. The expected value the Poisson distribution with parameter  $m$  is  $m$ , and the standard error of the Poisson distribution with parameter  $m$  is  $m^{1/2}$ .

**Population** A collection of units being studied. Units can be people, places, objects, epochs, drugs, procedures, or many other things. Much of statistics is concerned with estimating numerical properties (parameters) of an entire population from a random sample of units from the population.

**Population Control** Includes mechanisms or programs which control the numbers of individuals in a population of humans or animals.

**Population Density** Number of individuals in a population relative to space.

**Population Dynamics** The pattern of any process, or the interrelationship of phenomena, which affects growth or change within a population.

**Population Growth** Increase, over a specific period of time, in the number of individuals living in a country or region.

**Population Mean** The mean of the numbers in a numerical population. For example, the population mean of a box of numbered tickets is the mean of the list comprised of all the numbers on all the tickets. The population mean is a parameter. *C.f.* sample mean.

**Population Percentage** The percentage of units in a population that possess a specified property. For example, the percentage of a given collection of registered voters who are registered as Republicans. If each unit that possesses the property is labeled with "1," and each unit that does not possess the property is labeled with "0," the population percentage is the same as the mean of that list of zeros and ones; that is, the population percentage is the population mean for a population of zeros and ones. The population percentage is a parameter. *C.f.* sample percentage.

**Population Standard Deviation** The standard deviation of the values of a variable for a population. This is a parameter, not a statistic. *C.f.* sample standard deviation.

**Population Surveillance** Ongoing scrutiny of a population (general population, study population, target population, etc.), generally using methods distinguished by their practicability, uniformity, and frequently their rapidity, rather than by complete accuracy.

**Power** Refers to an hypothesis test. The power of a test against a specific alternative hypothesis is the chance that the test correctly rejects the null hypothesis when the alternative hypothesis is true.

**Pregnancy Outcome** Results of conception and ensuing pregnancy, including live birth, stillbirth, spontaneous abortion, induced abortion. The outcome may follow natural or artificial insemination or any of the various reproduction techniques, such as embryo transfer or fertilization in vitro.

**Pregnancy Rate** Ratio of the number of conceptions that occur during a period to the mean number of women of reproductive age.

**Prevalence** The total number of cases of a given disease in a specified population at a designated time. It is differentiated from **incidence**; which refers to the number of new cases in the population at a given time.

**Probability** The probability of an event is a number between zero and 100%. The meaning (interpretation) of probability is the subject of theories of probability, which differ in their interpretations. However, any rule for assigning probabilities to events has to satisfy the axioms of probability.

**Probability density function** The chance that a continuous random variable is in any range of values can be calculated as the area under a curve over that range of values. The curve is the probability density function of the random variable. That is, if  $X$  is a continuous random variable, there is a function  $f(x)$  such that for every pair of numbers  $a \leq b$ ,  $P(a \leq X \leq b) =$  (area under  $f$  between  $a$  and  $b$ );  $f$  is the probability density function of  $X$ . For example, the probability density function of a random variable with a standard normal distribution is the normal curve. Only continuous random variables have probability density functions.

**Probability Distribution** The probability distribution of a random variable specifies the chance that the variable takes a value in any subset of the real numbers. (The subsets have to satisfy some technical conditions that are not important for this course.) The probability distribution of a random variable is completely characterized by the cumulative probability distribution function; the terms sometimes are used synonymously. The probability distribution of a discrete random variable can be characterized by the chance that the random variable takes each of its possible values. For example, the probability distribution of the total number of spots  $S$  showing on the roll of two fair dice can be written as a table:

s	P(S=s)
2	1/36

3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

The probability distribution of a continuous random variable can be characterized by its probability density function.

**Probability Histogram** A probability histogram for a random variable is analogous to a histogram of data, but instead of plotting the area of the bins proportional to the relative frequency of observations in the class interval, one plots the area of the bins proportional to the probability that the random variable is in the class interval.

**Probability Sample** A sample drawn from a population using a random mechanism so that every element of the population has a known chance of ending up in the sample.

**Probability, Theories of** A *theory of probability* is a way of assigning meaning to probability statements such as "the chance that a thumbtack lands point-up is 2/3." That is, a theory of probability connects the mathematics of probability, which is the set of consequences of the axioms of probability, with the real world of observation and experiment. There are several common theories of probability. According to the *frequency theory of probability*, the probability of an event is the limit of the percentage of times that the event occurs in repeated, independent trials under essentially the same circumstances. According to the *subjective theory of probability*, a probability is a number that measures how strongly we believe an event will occur. The number is on a scale of 0% to 100%, with 0% indicating that we are completely sure it won't occur, and 100% indicating that we are completely sure that it will occur. According to the theory of *equally likely outcomes*, if an experiment has  $n$  possible outcomes, and (for example, by symmetry) there is no reason that any of the  $n$  possible outcomes should occur preferentially to any of the others, then the chance of each outcome is  $100\%/n$ . Each of these theories has its limitations, its proponents, and its detractors.

**Proportional Hazards Models** Statistical models used in survival analysis that assert that the effect of the study factors on the hazard rate in the study population is multiplicative and does not change over time.

**Proposition, logical proposition** A logical proposition is a statement that can be either true or false. For example, "the sun is shining in Berkeley right now" is a proposition.

---

[top](#)

## Q

**Qualitative Variable** A qualitative variable is one whose values are adjectives, such as colors, genders, nationalities, *etc.* *C.f.* quantitative variable and categorical variable.

**Quantile** The  $q$ th quantile of a list ( $0 < q \leq 1$ ) is the smallest number such that the fraction  $q$  or more of the elements of the list are less than or equal to it. *I.e.*, if the list contains  $n$  numbers, the  $q$ th quantile, is the smallest number  $Q$  such that at least  $n \times q$  elements of the list are less than or equal to  $Q$ .

**Quantitative Variable** A variable that takes numerical values for which arithmetic makes sense, for example, counts, temperatures, weights, amounts of money, *etc.* For some variables that take numerical values, arithmetic with those values does not make sense; such variables are not quantitative. For example, adding and subtracting social security numbers does not make sense. Quantitative variables typically have units of measurement, such as inches, people, or pounds.

**Quartiles** There are three quartiles. The first or lower quartile (LQ) of a list is a number (not necessarily a number in the list) such that at least  $1/4$  of the numbers in the list are no larger than it, and at least  $3/4$  of the numbers in the list are no smaller than it. The second quartile is the median. The third or upper quartile (UQ) is a number such that at least  $3/4$  of the entries in the list are no larger than it, and at least  $1/4$  of the numbers in the list are no smaller than it. To find the quartiles, first sort the list into increasing order. Find the smallest integer that is at least as big as the number of entries in the list divided by four. Call that integer  $k$ . The  $k$ th element of the sorted list is the lower quartile. Find the smallest integer that is at least as big as the number of entries in the list divided by two. Call that integer  $l$ . The  $l$ th element of the sorted list is the median. Find the smallest integer that is at least as large as the number of entries in the list times  $3/4$ . Call that integer  $m$ . The  $m$ th element of the sorted list is the upper quartile.

**Questionnaires** Predetermined sets of questions used to collect data - clinical data, social status, occupational group, *etc.* The term is often applied to a self-completed survey instrument.

---

[top](#)

# R

**Random Allocation** A process involving chance used in therapeutic trials or other research endeavor for allocating experimental subjects, human or animal, between treatment and control groups, or among treatment groups. It may also apply to experiments on inanimate objects.

**Random Error** All measurements are subject to error, which can often be broken down into two components: a bias or systematic error, which affects all measurements the same way; and a random error, which is in general different each time a measurement is made, and behaves like a number drawn with replacement from a box of numbered tickets whose average is zero.

**Random Event** See random experiment.

**Random Experiment** An experiment or trial whose outcome is not perfectly predictable, but for which the long-run relative frequency of outcomes of different types in repeated trials is predictable. Note that "random" is different from "haphazard," which does not necessarily imply long-term regularity.

**Random Sample** A random sample is a sample whose members are chosen at random from a given population in such a way that the chance of obtaining any particular sample can be computed. The number of units in the sample is called the *sample size*, often denoted  $n$ . The number of units in the population often is denoted  $N$ . Random samples can be drawn with or without replacing objects between draws; that is, drawing all  $n$  objects in the sample at once (a random sample without replacement), or drawing the objects one at a time, replacing them in the population between draws (a random sample with replacement). In a random sample with replacement, any given member of the population can occur in the sample more than once. In a random sample without replacement, any given member of the population can be in the sample at most once. A random sample without replacement in which every subset of  $n$  of the  $N$  units in the population is equally likely is also called a simple random sample. The term *random sample with replacement* denotes a random sample drawn in such a way that every  $n$ -tuple of units in the population is equally likely. See also probability sample.

**Random Variable** A random variable is an assignment of numbers to possible outcomes of a random experiment. For example, consider tossing three coins. The number of heads showing when the coins land is a random variable: it assigns the number 0 to the outcome {T, T, T}, the number 1 to the outcome {T, T, H}, the number 2 to the outcome {T, H, H}, and the number 3 to the outcome {H, H, H}.

**Randomized Controlled Experiment** An experiment in which chance is deliberately introduced in assigning subjects to the treatment and control groups. For example, we could write an identifying number for each subject on a slip of paper, stir up the slips of paper, and draw slips without replacement until we have drawn half of them. The subjects identified on the

slips drawn could then be assigned to treatment, and the rest to control. Randomizing the assignment tends to decrease confounding of the treatment effect with other factors, by making the treatment and control groups roughly comparable in all respects but the treatment.

**Randomized Controlled Trials** Clinical trials that involve at least one test treatment and one control treatment, concurrent enrollment and follow-up of the test- and control-treated groups, and in which the treatments to be administered are selected by a random process, such as the use of a random-numbers table. Treatment allocations using coin flips, odd-even numbers, patient social security numbers, days of the week, medical record numbers, or other such pseudo- or quasi-random processes, are not truly randomized and trials employing any of these techniques for patient assignment are designated simply controlled clinical trials.

**Range** The range of a set of numbers is the largest value in the set minus the smallest value in the set. Note that as a statistical term, the range is a single number, not a range of numbers.

**Records** The commitment in writing, as authentic evidence, of something having legal importance. The concept includes certificates of birth, death, etc., as well as hospital, medical, and other institutional records.

**Registries** The systems and processes involved in the establishment, support, management, and operation of registers, e.g., disease registers.

**Regression Analysis** Procedures for finding the mathematical function which best describes the relationship between a dependent variable and one or more independent variables. In linear regression the relationship is constrained to be a straight line and least-squares analysis is used to determine the best fit. In logistic regression the dependent variable is qualitative rather than continuously variable and likelihood functions are used to find the best relationship. In multiple regression the dependent variable is considered to depend on more than a single independent variable.

**Regression, Linear Regression** Linear regression fits a line to a scatterplot in such a way as to minimize the sum of the squares of the residuals. The resulting regression line, together with the standard deviations of the two variables or their correlation coefficient, can be a reasonable summary of a scatterplot if the scatterplot is roughly football-shaped. In other cases, it is a poor summary. If we are regressing the variable  $Y$  on the variable  $X$ , and if  $Y$  is plotted on the vertical axis and  $X$  is plotted on the horizontal axis, the regression line passes through the point of averages, and has slope equal to the correlation coefficient times the SD of  $Y$  divided by the SD of  $X$ .

**Regression Fallacy** The regression fallacy is to attribute the regression effect to an external cause.

**Regression Toward the Mean, Regression Effect** Suppose one measures two variables for each member of a group of individuals, and that the correlation coefficient of the variables is positive (negative). If the value of the first variable for that individual is above average, the

value of the second variable for that individual is likely to be above (below) average, but by fewer standard deviations than the first variable is. That is, the second observation is likely to be closer to the mean in standard units. For example, suppose one measures the heights of fathers and sons. Each individual is a (father, son) pair; the two variables measured are the height of the father and the height of the son. These two variables will tend to have a positive correlation coefficient: fathers who are taller than average tend to have sons who are taller than average. Consider a (father, son) pair chosen at random from this group. Suppose the father's height is 3SD above the average of all the fathers' heights. (The SD is the standard deviation of the fathers' heights.) Then the son's height is also likely to be above the average of the sons' heights, but by fewer than 3SD (here the SD is the standard deviation of the sons' heights).

**Relative Risk Assessment** An evaluation of the risk of disease in a patient who possesses a certain characteristic relative to one who does not possess that characteristic. Relative risk can be assessed as a property of a clinical test.

**Relative Risk Reduction (RRR)** The proportional reduction in outcome rates between control and experimental patients in a trial.

**Repeated Measures Analysis of Variance** An ANOVA that analyzes two or more related measurements of the same variable.

**Reproducibility of Results** The statistical reproducibility of measurements (often in a clinical context), including the testing of instrumentation or techniques to obtain reproducible results. The concept includes reproducibility of physiological measurements, which may be used to develop rules to assess probability or prognosis, or response to a stimulus; reproducibility of occurrence of a condition; and reproducibility of experimental results.

**Reproductive History** An important aggregate factor in epidemiological studies of women's health. The concept usually includes the number and timing of pregnancies and their outcomes, the incidence of breast feeding, and may include age of menarche and menopause, regularity of menstruation, fertility, gynecological or obstetric problems, or contraceptive usage.

**Residence Characteristics** Elements of residence that characterize a population. They are applicable in determining need for and utilization of health services

**Residential Mobility** Frequent change of residence, either in the same city or town, or between cities, states or communities.

**Residual** The difference between a datum and the value predicted for it by a model. In linear regression of a variable plotted on the vertical axis onto a variable plotted on the horizontal axis, a residual is the "vertical" distance from a datum to the line. Residuals can be positive (if the datum is above the line) or negative (if the datum is below the line). Plots of residuals can reveal computational errors in linear regression, as well as conditions under which linear regression is inappropriate, such as nonlinearity and heteroscedasticity. If linear regression is

performed properly, the sum of the residuals from the regression line must be zero; otherwise, there is a computational error somewhere.

**Residual Plot** A residual plot for a regression is a plot of the residuals from the regression against the explanatory variable.

**Resistant** A statistic is said to be resistant if corrupting a datum cannot change the statistic much. The mean is not resistant; the median is.

**Risk** The probability that an event will occur. It encompasses a variety of measures of the probability of a generally unfavorable outcome.

**Risk Assessment** The qualitative or quantitative estimation of the likelihood of adverse effects that may result from exposure to specified health hazards or from the absence of beneficial influences.

**Risk Factors** An aspect of personal behavior or lifestyle, environmental exposure, or inborn or inherited characteristic, which, on the basis of epidemiologic evidence, is known to be associated with a health-related condition considered important to prevent.

**Root-mean-square (RMS)** The RMS of a list is the square-root of the mean of the squares of the elements in the list. It is a measure of the average "size" of the elements of the list. To compute the RMS of a list, you square all the entries, average the numbers you get, and take the square-root of that average.

**Root-mean-square error (RMSE)** The RMSE of an estimator of a parameter is the square-root of the mean squared error (MSE) of the estimator. In symbols, if  $X$  is an estimator of the parameter  $t$ , then  $RMSE(X) = (E((X-t)^2))^{1/2}$ . The RMSE of an estimator is a measure of the expected error of the estimator. The units of RMSE are the same as the units of the estimator. See also mean squared error.

**rms Error of Regression** The rms error of regression is the rms of the vertical residuals from the regression line. For regressing  $Y$  on  $X$ , the rms error of regression is equal to  $(1 - r^2)^{1/2} \times SD_Y$ , where  $r$  is the correlation coefficient between  $X$  and  $Y$  and  $SD_Y$  is the standard deviation of the values of  $Y$ .

---

[top](#)

## S

**Sample** A sample is a collection of units from a population. See also random sample.

**Sample Mean** The arithmetic mean of a random sample from a population. It is a statistic commonly used to estimate the population mean. Suppose there are  $n$  data,  $\{x_1, x_2, \dots, x_n\}$ . The sample mean is  $(x_1 + x_2 + \dots + x_n)/n$ . The expected value of the sample mean is the population mean. For sampling with replacement, the SE of the sample mean is the population standard deviation, divided by the square-root of the sample size. For sampling without replacement, the SE of the sample mean is the finite-population correction  $((N-n)/(N-1))^{1/2}$  times the SE of the sample mean for sampling with replacement, with  $N$  the size of the population and  $n$  the size of the sample.

**Sample Percentage** The percentage of a random sample with a certain property, such as the percentage of voters registered as Democrats in a simple random sample of voters. The sample mean is a statistic commonly used to estimate the population percentage. The expected value of the sample percentage from a simple random sample or a random sample with replacement is the population percentage. The SE of the sample percentage for sampling with replacement is  $(p(1-p)/n)^{1/2}$ , where  $p$  is the population percentage and  $n$  is the sample size. The SE of the sample percentage for sampling without replacement is the finite-population correction  $((N-n)/(N-1))^{1/2}$  times the SE of the sample percentage for sampling with replacement, with  $N$  the size of the population and  $n$  the size of the sample. The SE of the sample percentage is often estimated by the bootstrap.

**Sample Size** The number of units (persons, animals, patients, specified circumstances, etc.) in a population to be studied. The sample size should be big enough to have a high likelihood of detecting a true difference between two groups.

**Sample Standard Deviation,  $S$**  The sample standard deviation  $S$  is an estimator of the standard deviation of a population based on a random sample from the population. The sample standard deviation is a statistic that measures how "spread out" the sample is around the sample mean. It is quite similar to the standard deviation of the sample, but instead of averaging the squared deviations (to get the rms of the deviations of the data from the sample mean) it divides the sum of the squared deviations by (number of data - 1) before taking the square-root. Suppose there are  $n$  data,  $\{x_1, x_2, \dots, x_n\}$ , with mean  $M = (x_1 + x_2 + \dots + x_n)/n$ . Then  $s = ((x_1 - M)^2 + (x_2 - M)^2 + \dots + (x_n - M)^2)/(n-1)^{1/2}$ . The square of the sample standard deviation,  $S^2$  (the sample variance) is an unbiased estimator of the square of the SD of the population (the variance of the population).

**Sample Sum** The sum of a random sample from a population. The expected value of the sample sum is the sample size times the population mean. For sampling with replacement, the SE of the sample sum is the population standard deviation, times the square-root of the sample size. For sampling without replacement, the SE of the sample sum is the finite-population correction  $((N-n)/(N-1))^{1/2}$  times the SE of the sample sum for sampling with replacement, with  $N$  the size of the population and  $n$  the size of the sample.

**Sample Survey** A survey based on the responses of a sample of individuals, rather than the entire population.

**Sample Variance** The sample variance is the square of the sample standard deviation  $S$ . It is an unbiased estimator of the square of the population standard deviation, which is also called the variance of the population.

**Sampling distribution** The sampling distribution of an estimator is the probability distribution of the estimator when it is applied to random samples.

**Sampling error** In estimating from a random sample, the difference between the estimator and the parameter can be written as the sum of two components: bias and sampling error. The bias is the average error of the estimator over all possible samples. The bias is not random. Sampling error is the component of error that varies from sample to sample. The sampling error is random: it comes from "the luck of the draw" in which units happen to be in the sample. It is the chance variation of the estimator. The average of the sampling error over all possible samples (the expected value of the sampling error) is zero. The standard error of the estimator is a measure of the typical size of the sampling error.

**Sampling Studies** Studies in which a number of subjects are selected from all subjects in a defined population. Conclusions based on sample results may be attributed only to the population sampled.

**Sampling unit** A sample from a population can be drawn one unit at a time, or more than one unit at a time (one can sample clusters of units). The fundamental unit of the sample is called the *sampling unit*. It need not be a unit of the population.

**Scatterplot** A scatterplot is a way to visualize bivariate data. A scatterplot is a plot of pairs of measurements on a collection of "individuals" (which need not be people). For example, suppose we record the heights and weights of a group of 100 people. The scatterplot of those data would be 100 points. Each point represents one person's height and weight. In a scatterplot of weight *against* height, the  $x$ -coordinate of each point would be height of one person, the  $y$ -coordinate of that point would be the weight of the same person. In a scatterplot of height against weight, the  $x$ -coordinates would be the weights and the  $y$ -coordinates would be the heights.

**SD line** For a scatterplot, a line that goes through the point of averages, with slope equal to the ratio of the standard deviations of the two plotted variables. If the variable plotted on the horizontal axis is called  $X$  and the variable plotted on the vertical axis is called  $Y$ , the slope of the SD line is the SD of  $Y$ , divided by the SD of  $X$ .

**Secular Trend** A linear association (trend) with time.

**Selection Bias** A systematic tendency for a sampling procedure to include and/or exclude units of a certain type. For example, in a quota sample, unconscious prejudices or predilections on the part of the interviewer can result in selection bias. Selection bias is a potential problem whenever a human has latitude in selecting individual units for the sample; it tends to be eliminated by probability sampling schemes in which the interviewer is told exactly whom to contact (with no room for individual choice).

**Self-Selection** Self-selection occurs when individuals decide for themselves whether they are in the control group or the treatment group. Self-selection is quite common in studies of human behavior. For example, studies of the effect of smoking on human health involve self-selection: individuals choose for themselves whether or not to smoke. Self-selection precludes an experiment; it results in an observational study. When there is self-selection, one must be wary of possible confounding from factors that influence individuals' decisions to belong to the treatment group.

**Sensitivity** Measures for assessing the results of diagnostic and screening tests. Sensitivity represents the proportion of truly diseased persons in a screened population who are identified as being diseased by the test. It is a measure of the probability of correctly diagnosing a condition.

**Sentinel Surveillance** Monitoring of rate of occurrence of specific conditions to assess the stability or change in health levels of a population. It is also the study of disease rates in a specific cohort, geographic area, population subgroup, etc. to estimate trends in larger population.

**Set** A set is a collection of things, without regard to their order.

**Significance, Significance level, Statistical significance** The significance level of an hypothesis test is the chance that the test erroneously rejects the null hypothesis when the null hypothesis is true.

**Simple Random Sample** A simple random sample of  $n$  units from a population is a random sample drawn by a procedure that is equally likely to give every collection of  $n$  units from the population; that is, the probability that the sample will consist of any given subset of  $n$  of the  $N$  units in the population is  $1/N C_n$ . Simple random sampling is sampling at random without replacement (without replacing the units between draws). A simple random sample of size  $n$  from a population of  $N \geq n$  units can be constructed by assigning a random number between zero and one to each unit in the population, then taking those units that were assigned the  $n$  largest random numbers to be the sample.

**Simpson's Paradox** What is true for the parts is not necessarily true for the whole. See also confounding.

**Single-Blind Method** A method in which either the observer(s) or the subject(s) is kept ignorant of the group to which the subjects are assigned.

**Skewed Distribution** A distribution that is not symmetrical.

### **Specificity**

Measures for assessing the results of diagnostic and screening tests. Specificity is the proportion of truly non-diseased persons who are so identified by the screening test. It is a measure of the probability of correctly identifying a non-diseased person.

**Severity of Illness Index** Levels of severity of illness within a diagnostic group which are established by various measurement criteria.

**Sex Distribution** The number of males and females in a given population. The distribution may refer to how many men or women or what proportion of either in the group. The population is usually patients with a specific disease but the concept is not restricted to humans and is not restricted to medicine.

**Sickness Impact Profile** A quality-of-life scale developed in the United States in 1972 as a measure of health status or dysfunction generated by a disease. It is a behaviorally based questionnaire for patients and addresses activities such as sleep and rest, mobility, recreation, home management, emotional behavior, social interaction, and the like. It measures the patient's perceived health status and is sensitive enough to detect changes or differences in health status occurring over time or between groups.

**Small-Area Analysis** A method of analyzing the variation in utilization of health care in small geographic or demographic areas. It often studies, for example, the usage rates for a given service or procedure in several small areas, documenting the variation among the areas. By comparing high- and low-use areas, the analysis attempts to determine whether there is a pattern to such use and to identify variables that are associated with and contribute to the variation.

**Space-Time Clustering** A statistically significant excess of cases of a disease, occurring within a limited space-time continuum.

**Square-Root Law** The Square-Root Law says that the standard error (SE) of the sample sum of  $n$  random draws with replacement from a box of tickets with numbers on them is...  $SE(\text{sample sum}) = n^{1/2} \times SD(\text{box})$ , and the standard error of the sample mean of  $n$  random draws with replacement from a box of tickets is...  $SE(\text{sample mean}) = n^{-1/2} \times SD(\text{box})$ , where  $SD(\text{box})$  is the standard deviation of the list of the numbers on all the tickets in the box (including repeated values).

**Standard Deviation (SD)** The standard deviation of a set of numbers is the rms of the set of deviations between each element of the set and the mean of the set. See also sample standard deviation.

**Standard Error (SE)** The Standard Error of a random variable is a measure of how far it is likely to be from its expected value; that is, its scatter in repeated experiments. The SE of a random variable  $X$  is defined to be...  $SE(X) = [E((X - E(X))^2)]^{1/2}$ . That is, the standard error is the square-root of the expected squared difference between the random variable and its expected value. The SE of a random variable is analogous to the SD of a list.

**Standard Units** A variable (a set of data) is said to be in standard units if its mean is zero and its standard deviation is one. You transform a set of data into standard units by subtracting the mean from each element of the list, and dividing the results by the standard deviation. A random variable is said to be in standard units if its expected value is zero and its standard

error is one. You transform a random variable to standard units by subtracting its expected value then dividing by its standard error.

**Standardize** To transform into standard units.

**Statistic** A number that can be computed from data, involving no unknown parameters. As a function of a random sample, a statistic is a random variable. Statistics are used to estimate parameters, and to test hypotheses.

**Statistical Distributions** The complete summaries of the frequencies of the values or categories of a measurement made on a group of items, a population, or other collection of data. The distribution tells either how many or what proportion of the group was found to have each value (or each range of values) out of all the possible values that the quantitative measure can have.

**Stratified Sample** In a stratified sample, subsets of sampling units are selected separately from different strata, rather than from the frame as a whole.

**Stratified sampling** The act of drawing a stratified sample.

**Stratum** In random sampling, sometimes the sample is drawn separately from different disjoint subsets of the population. Each such subset is called a *stratum*. (The plural of *stratum* is *strata*.) Samples drawn in such a way are called stratified samples. Estimators based on stratified random samples can have smaller sampling errors than estimators computed from simple random samples of the same size, if the average variability of the variable of interest within strata is smaller than it is across the entire population; that is, if stratum membership is associated with the variable. For example, to determine average home prices in the U.S., it would be advantageous to *stratify* on geography, because average home prices vary enormously with location. We might divide the country into states, then divide each state into urban, suburban, and rural areas; then draw random samples separately from each such division.

**Studentized score** The observed value of a statistic, minus the expected value of the statistic, divided by the estimated standard error of the statistic.

**Student's *t* curve** Student's *t* curve is a family of curves indexed by a parameter called the *degrees of freedom*, which can take the values 1, 2, . . . Student's *t* curve is used to approximate some probability histograms. Consider a population of numbers that are nearly normally distributed and have population mean is  $\mu$ . Consider drawing a random sample of size  $n$  with replacement from the population, and computing the sample mean  $M$  and the sample standard deviation  $S$  define the random variable...  $T = (M - \mu)/(S/n^{1/2})$ . If the sample size  $n$  is large, the probability histogram of  $T$  can be approximated accurately by the normal curve. However, for small and intermediate values of  $n$ , Student's *t* curve with  $n - 1$  *degrees of freedom* gives a better approximation. That is.....  $P(a < T < b)$  is approximately the area under Student's  $T$  curve with  $n - 1$  degrees of freedom, from  $a$  to  $b$ . Student's *t* curve can be used to

test hypotheses about the population mean and construct confidence intervals for the population mean, when the population distribution is known to be nearly normally distributed.

**Subject, Experimental Subject** A member of the control group or the treatment group.

**Subset** A subset of a given set is a collection of things that belong to the original set. Every element of the subset must belong to the original set, but not every element of the original set need be in a subset (otherwise, a subset would always be identical to the set it came from).

**Survival Analysis** A class of statistical procedures for estimating the survival function (function of time, starting with a population 100% well at a given time and providing the percentage of the population still well at later times). The survival analysis is then used for making inferences about the effects of treatments, prognostic factors, exposures, and other covariates on the function.

**Survival Rate** The proportion of survivors in a group, e.g., of patients, studied and followed over a period, or the proportion of persons in a specified group alive at the beginning of a time interval who survive to the end of the interval. It is often studied using life table methods.

**Symmetric Distribution** The probability distribution of a random variable  $X$  is symmetric if there is a number  $a$  such that the chance that  $X \geq a+b$  is the same as the chance that  $X \leq a-b$  for every value of  $b$ . A list of numbers has a symmetric distribution if there is a number  $a$  such that the fraction of numbers in the list that are greater than or equal to  $a+b$  is the same as the fraction of numbers in the list that are less than or equal to  $a-b$ , for every value of  $b$ . In either case, the histogram or the probability histogram will be symmetrical about a vertical line drawn at  $x=a$ .

**Systematic error** An error that affects all the measurements similarly. For example, if a ruler is too short, everything measured with it will appear to be longer than it really is (ignoring random error). If your watch runs fast, every time interval you measure with it will appear to be longer than it really is (again, ignoring random error). Systematic errors do not tend to average out.

**Systematic random sample** A systematic sample starting at a random point in the listing of units in the of frame, instead of starting at the first unit. Systematic random sampling is better than systematic sampling, but typically not as good as simple random sampling.

**Systematic sample** A systematic sample from a frame of units is one drawn by listing the units and selecting every  $k$ th element of the list. For example, if there are  $N$  units in the frame, and we want a sample of size  $N/10$ , we would take every tenth unit: the first unit, the eleventh unit, the 21st unit, *etc.* Systematic samples are not random samples, but they often behave essentially as if they were random, if the order in which the units appears in the list is haphazard. Systematic samples are a special case of cluster samples.

---

[top](#)

# T

**t-Test** A hypothesis test based on approximating the probability histogram of the test statistic by Student's  $t$  curve.  $t$  tests usually are used to test hypotheses about the mean of a population when the sample size is intermediate and the distribution of the population is known to be nearly normal.

**Test Statistic** A statistic used to test hypotheses. An hypothesis test can be constructed by deciding to reject the null hypothesis when the value of the test statistic is in some range or collection of ranges. To get a test with a specified significance level, the chance when the null hypothesis is true that the test statistic falls in the range where the hypothesis would be rejected must be at most the specified significance level. The  $Z$  statistic is a common test statistic.

**Transformation** Transformations turn lists into other lists, or variables into other variables. For example, to transform a list of temperatures in degrees Celsius into the corresponding list of temperatures in degrees Fahrenheit, you multiply each element by  $9/5$ , and add 32 to each product. This is an example of an affine transformation: multiply by something and add something ( $y = ax + b$  is the general affine transformation of  $x$ ; it's the familiar equation of a straight line). In a linear transformation, you only multiply by something ( $y = ax$ ). Affine transformations are used to put variables in standard units. In that case, you subtract the mean and divide the results by the SD. This is equivalent to multiplying by the reciprocal of the SD and adding the negative of the mean, divided by the SD, so it is an affine transformation. Affine transformations with positive multiplicative constants have a simple effect on the mean, median, mode, quartiles, and other percentiles: the new value of any of these is the old one, transformed using exactly the same formula. When the multiplicative constant is negative, the mean, median, mode, are still transformed by the same rule, but quartiles and percentiles are reversed: the  $q$ th quantile of the transformed distribution is the transformed value of the  $1-q$ th quantile of the original distribution (ignoring the effect of data spacing). The effect of an affine transformation on the SD, range, and IQR, is to make the new value the old value times the absolute value of the number you multiplied the first list by: what you added does not affect them.

**Treatment** The substance or procedure studied in an experiment or observational study. At issue is whether the treatment has an effect on the outcome or variable of interest.

**Treatment Effect** The effect of the treatment on the variable of interest. Establishing whether the treatment has an effect is the point of an experiment.

**Treatment group** The individuals who receive the treatment, as opposed to those in the control group, who do not.

**Two-sided Hypothesis test** C.f. one-sided test. An hypothesis test of the null hypothesis that the value of a parameter,  $\mu$ , is equal to a null value,  $\mu_0$ , designed to have power against the alternative hypothesis that either  $\mu < \mu_0$  or  $\mu > \mu_0$  (the alternative hypothesis contains values on both sides of the null value). For example, a significance level 5%, two-sided z test of the null hypothesis that the mean of a population equals zero against the alternative that it is greater than zero would reject the null hypothesis for values of

$$|z| = \frac{|(\text{sample mean})|}{\text{SE}(\text{sample mean})} > 1.96.$$

**Type I and Type II errors** These refer to hypothesis testing. A Type I error occurs when the null hypothesis is rejected erroneously when it is in fact true. A Type II error occurs if the null hypothesis is not rejected when it is in fact false.

**Trauma Severity Indices** Systems for assessing, classifying, and coding injuries. These systems are used in medical records, surveillance systems, and state and national registries to aid in the collection and reporting of trauma.

**Twin Studies** Methods of detecting genetic etiology in human traits. The basic premise of twin studies is that monozygotic twins, being formed by the division of a single fertilized ovum, carry identical genes, while dizygotic twins, being formed by the fertilization of two ova by two different spermatozoa, are genetically no more similar than two siblings born after separate pregnancies.

---

[top](#)

## U

**Unbiased** Not biased; having zero bias.

**Uncontrolled Experiment** An experiment in which there is no control group; *i.e.*, in which the method of comparison is not used: the experimenter decides who gets the treatment, but the outcome of the treated group is not compared with the outcome of a control group that does not receive treatment.

**Uncorrelated** A set of bivariate data is uncorrelated if its correlation coefficient is zero. Two random variables are uncorrelated if the expected value of their product equals the product of their expected values. If two random variables are independent, they are uncorrelated. (The converse is not true in general.)

**Uncountable** A set is uncountable if it is not countable.

**Unimodal** Having exactly one mode.

**Union** The union of two or more sets is the set of objects contained by at least one of the sets. The union of the events A and B is denoted "A" plus "B", "A or B", and "AUB". *C.f.* intersection.

**Unit** A member of a population.

**Univariate** Having or having to do with a single variable. Some univariate techniques and statistics include the histogram, IQR, mean, median, percentiles, quantiles, and SD. *C.f.* bivariate.

**Upper Quartile (UQ)** See quartiles.

---

[top](#)

## V

**Variable** A numerical value or a characteristic that can differ from individual to individual.

**Variance, population variance** The variance of a list is the square of the standard deviation of the list, that is, the average of the squares of the deviations of the numbers in the list from their mean. The variance of a random variable X,  $\text{Var}(X)$ , is the expected value of the squared difference between the variable and its expected value:  $\text{Var}(X) = E((X - E(X))^2)$ . The variance of a random variable is the square of the standard error (SE) of the variable.

**Venn Diagram** A pictorial way of showing the relations among sets or events. The universal set or outcome space is usually drawn as a rectangle; sets are regions within the rectangle. The overlap of the regions corresponds to the intersection of the sets. If the regions do not overlap, the sets are disjoint. The part of the rectangle included in one or more of the regions corresponds to the union of the sets.

**Vital Statistics** Used for general articles concerning statistics of births, deaths, marriages, etc.

---

[top](#)

## X

**XOR, exclusive disjunction** XOR is an operation on two logical propositions. If  $p$  and  $q$  are two propositions,  $(p \text{ XOR } q)$  is a proposition that is true if either  $p$  is true or if  $q$  is true, but not both.  $(p \text{ XOR } q)$  is logically equivalent to  $((p \text{ OR } q) \text{ AND NOT } (p \text{ AND } q))$ .

---

[top](#)

# Z

**z-score** The observed value of the  $Z$  statistic.

**Z statistic** A  $Z$  statistic is a test statistic whose distribution under the null hypothesis has expected value zero and can be approximated well by the normal curve. Usually,  $Z$  statistics are constructed by standardizing some other statistic. The  $Z$  statistic is related to the original statistic by...  $Z = (\text{original} - \text{expected value of original}) / \text{SE}(\text{original})$ .

**z-test** An hypothesis test based on approximating the probability histogram of the  $Z$  statistic under the null hypothesis by the normal curve.