Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
hbiostat.org/doc/glossary.pdf
hbiostat.org/bbr
To request or improve a definition:
bit.ly/datamethods-glossary
December 17, 2020

Glossary of Statistical Terms

adjusting or controlling for a variable: Assessing the effect of one variable while accounting for the effect of another (confounding) variable. Adjustment for the other variable can be carried out by stratifying the analysis (especially if the variable is categorical) or by statistically estimating the relationship between the variable and the outcome and then subtracting out that effect to study which effects are "left over." For example, in a non-randomized study comparing the effects of treatments A and B on blood pressure reduction, the patients' ages may have been used to select the treatment. It would be advisable in that case to control for the effect of age before estimating the treatment effect. This can be done using a regression model with blood pressure as the dependent variable and treatment and age as the independent variables (controlling for age using subtraction) or crudely and approximately (with some residual confounding) by stratifying by deciles of age and averaging the treatment effects estimated within the deciles. Adjustment results in adjusted odds ratios, adjusted hazard ratios, adjusted slopes, etc.

allocation ratio: In a parallel group randomized trial of two treatments, is the ratio of sample sizes of the two groups.

ANCOVA: Analysis of covariance is just multiple regression (i.e., a *linear model*) where one variable is of major interest and is categorical (e.g., treatment group). In classic ANCOVA there is a treatment variable and a continuous covariate used to reduce unexplained variation in the dependent variable, thereby increasing power.

ANOVA: Analysis of variance usually refers to an analysis of a continuous dependent variable where all the predictor variables are categorical. One-way ANOVA, where there is only one predictor variable (factor; grouping variable), is a generalization of the 2-sample t-test. ANOVA with 2 groups is identical to the t-test. Two-way ANOVA refers to two predictors, and if the two are allowed to interact in the model, two-way ANOVA involves cross-classification of observations simultaneously by both factors. It is not appropriate to refer to repeated measures within subjects as two-way ANOVA (e.g., treatment $\times$ time). An ANOVA table sometimes refers to statistics for more complex models, where explained variation from partial and total effects are displayed and continuous variables may be included.

artificial intelligence: Frequently confused with *machine learning*, AI is a procedure for flexibly learning from data, which may be built from elements of machine learning, but is distinguished by the underlying algorithms being created so that the "machine" can accept new inputs after the developer has completed the initial algorithm. In that way the machine can continue to update, refine, and teach itself.

Bayes' rule or theorem: $\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$, read as the probability that event A happens given that event B has happened equals the probability that B happens given that A has happened multiplied by the (unconditional) probability that A happens and divided by the (unconditional) probability that B happens. Bayes' rule follows immediately from the law of conditional probability which states that $\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$.

Bayesian inference: A branch of statistics based on Bayes' theorem. Bayesian inference doesn't use P-values and generally does not test hypotheses. It requires one to formally specify a probability distribution encapsulating the prior knowledge about, say, a treatment effect. The state of prior knowledge can be specified as "no knowledge" by using a flat distribution, although this can lead to wild and nonsensical estimates. Once the prior distribution is specified, the data are used to modify the prior state of knowledge to obtain the post-experiment state of knowledge. Final probabilities computed in the Bayesian framework are probabilities of various treatment effects. The price of being able to compute subjective probabilities about the data generating process is the necessity of specifying a prior distribution to anchor the calculations.

bias: A systematic error. Examples: a miscalibrated machine that reports cholesterol too high by 20mg% on the average; a satisfaction questionnaire that leads patients to never report that they are dissatisfied with their medical care; using each patient's lowest blood pressure over 24 hours to describe a drug's antihyptertensive properties. Bias typically pertains to the discrepency between the average of many estimates over repeated sampling and the true value of a parameter. Therefore bias is more related to frequentist statistics than to Bayesian statistics.

big data: A dataset too large to fit on an ordinary workstation computer.

binary variable: A variable whose only two possible values, usually zero and one.

bootstrap: A simulation technique for studying properties of statistics without the need to have the infinite population available. The most common use of the bootstrap involves taking random samples (with replacement) from the original dataset and studying how some quantity of interest varies. Each random sample has the same number of observations as the original dataset. Some of the original subjects may be omitted from the random sample and some may be sampled more than once. The bootstrap can be used to compute standard deviations and confidence limits (compatibility limits) without assuming a model. For example, if one took 200 samples with replacement from the original dataset, computed the sample median from each sample, and then computed the sample standard deviation of the 200 medians, the result would be a good estimate of the true standard deviation of the original sample median. The bootstrap can also be used to internally validate a predictive model without holding back patient data during model development.

calibration: Reliability of predicted values, i.e., extent to which predicted values agree with observed values. For a predictive model a calibration curve is constructed by relating predicted to observed values in some smooth manner. The calibration curve is judged against a 45° line. Miscalibration could be called bias. Calibration error is frequently assessed for predicted event probabilities. If for example 0.4 of the time it rained when the predicted probability of rain was 0.4, the rain forecast is perfectly calibrated. There are specific classes of calibration. *Calibration in the large* refers to being accurate on the average. If the average daily rainfall probability in your region was $\frac{1}{7}$ and it rained on $\frac{1}{7}^{\text{th}}$ of the days each year, the probability estimate would be perfectly calibrated in the large. *Calibration in the small* refers to each level of predicted probability being accurate. On days in which the rainfall probability was $\frac{1}{5}$ did it rain $\frac{1}{5}^{\text{th}}$ of the time? One could go further and define *calibration in the tiny* as the extent to which a given type of subject (say a 35 year old male) and a given outcome probability for that subject is accurate. Or is an 0.4 rainfall forecast accurate in the spring?

case-control study: A study in which subjects are selected on the basis of their outcomes, and then exposures (treatments) are ascertained. For example, to assess the association between race and

operative mortality one might select all patients who died after open heart surgery in a given year and then select an equal number of patients who survived, matching on several variables other than race so as to equalize (control for) their distributions between the cases and non-cases.

categorical variable: A variable having only certain possible values for which there is no logical ordering of the values. Also called a *nominal*, *polytomous*, *discrete categorical* variable or *factor*.

causal inference: The study of how/whether outcomes vary across levels of an exposure when that exposure is manipulated. Done properly, the study of causal inference typically concerns itself with defining target parameters, precisely defining the conditions under which causality may be inferred, and evaluation of sensitivity to departures from such conditions. In a randomized and properly blinded experiment in which all experimental units adhere to the experimental manipulation called for in the design, most experimentalists are willing to make a causal interpretation of the experimental effect without further ado. In more complex situations involving observational data or imperfect adherence, things are more nuanced. See Pearl Sections 2.1–2.3 for more information[6].

censoring: When the response variable is the time until an event, subjects not followed long enough for the event to have occurred have their event times *censored* at the time of last follow-up. This kind of censoring is *right censoring*. For example, in a follow-up study, patients entering the study during its last year will be followed a maximum of 1 year, so they will have their time until event censored at 1 year or less. *Left censoring* means that the time to the event is known to be less than some value. In *interval censoring* the time is known to be in a specified interval. Most statistical analyses assume that what causes a subject to be censored is independent of what would cause her to have an event. If this is not the case, *informative censoring* is said to be present. For example, if a subject is pulled off of a drug because of a treatment failure, the censoring time is indirectly reflecting a bad clinical outcome and the resulting analysis will be biased.

classification and classifier: When considering patterns of associations between inputs and categorical outcomes, classification is the act of assigning a predicted outcome on the basis of all the inputs. A *classifier* is an algorithm developed for classification. Classification is a forced choice and the result is not a probability. It could be deemed a premature decision, or a decision based on optimizing an implicit or explicit utility/loss/cost function. When the utility function is not specified by the end-user, classification may not be consistent with good decision making. Classification ignores close calls. Logistic regression is frequently mislabeled as a *classifier*; it is a direct probability estimator. The term *classification* is frequently used improperly when the outcome variable is categorical (i.e., represents *classes*) and a probability estimator is used to analyze the data to make probability predictions. The correct term for this situation is *prediction*.

clinical trial: Though almost always used to denote a randomized experiment, a clinical trial may be any type of prospective study of human subjects in which therapies or clinical strategies are compared. Treatments may be assigned to individual patients or to groups, the latter including cluster randomized trials. For a randomized clinical trial or randomized controlled trial (RCT), the choice and timing of treatments is outside of the control of the physician and patient but is (usually) set in advance by a randomization device. This may be used for traditional parallel group designs, or using a randomized crossover design. Randomization is used to remove the connection between patient characteristics and treatment assignment so that treatment selection bias due to both *known* and *unknown* (at the time of randomization) factors is avoided. RCTs do not require representative patients but do require representative treatment effects. If a patient characteristic interacts with the treatment effect, and a

wide spectrum of patients over the distribution of the interacting factor is not included in the trial, the trial results may not apply to patients outside (with respect to the interacting factor) of those studied. For example, if age is an effect modifier for treatment and a trial included primarily patients aged 40-65, the relative benefit of a treatment for those older than 65 may not be estimable. RCTs may involve more than two therapies. The "controlled" in *randomized controlled trial* often refers to having a reference treatment arm that is a placebo or standard of care. But the comparison group can be anything including active controls (as in head-to-head comparisons of drugs). The RCT is the gold standard for establishing causality. An RCT may be mechanistic as in a pure efficacy study, a *policy* or *strategy* study, or an *effectiveness* study. The latter pertains to the attempt to mimic clinical practice in the field.

cohort study: A study in which all subjects meeting the entry criteria are included. Entry criteria are defined at baseline, e.g., at time of diagnosis or treatment.

comparative trial: Trials with two or more treatment groups, designed with sufficient power or precision to detect relevant clinical differences in treatment efficacy among the groups.

conditioning: Conditioning on something means to assume it is true, or in more statistical terms, to set its value to some constant or assume it belongs to some set of values. We might say that the mean systolic blood pressure conditional on the person being female is 125mmHg, which is concisely stated as "of females, the mean SBP is 125mmHg." Conditioning statements are "if statements." The notation used for conditioning in statistics is to place the qualifying condition after a vertical bar.

conditional probability: The probability of the veracity of a statement or of an event A given that a specific condition B holds or that an event B has already occurred, denoted by $P(A|B)$. This is a probability in the presence of knowledge captured by B. For example, if the condition B is that a person is male, the conditional probability is the probability of A for males. It could be argued that there is no such thing as a completely *un*conditional probability. In this example one is implicitly conditioning on humans even if not considering the person's sex.

confidence limits: To say that the 0.95 confidence limits for an unknown quantity are $[a, b]$ means that 0.95 of similarly constructed confidence limits in repeated samples from the same population *would* contain the unknown quantity. Very loosely speaking one could say that she is 0.95 "confident" that the unknown value is in the interval $[a, b]$, although in the frequentist school unknown parameters are constants, so they are either inside or outside intervals and there are no probabilities associated with these events. The interpretation of a single confidence interval in frequentist statistics is highly problematic, and in fact the word *confidence* is poorly defined and was just an attempt to gloss over this problem. Note that a confidence interval should be symmetric about a point estimate only when the distribution of the point estimate is symmetric. Many confidence intervals are asymmetric, e.g., intervals for probabilities, odds ratios, and other ratios. Another way to define a confidence interval is the set of all values that if null hypothesized would not be rejected at one minus the confidence level by a specific statistical test. For that reason, confidence intervals are better called *compatibility intervals*.

confounder: A variable measured before the exposure (treatment) that is a common cause of (or is just associated with) the response and the exposure variable. A confounder, when properly controlled for, can explain away an apparent association between the exposure and the response. A formal definition is: a "pre-exposure covariate C for which there exists a set of other covariates X such that effect of the

exposure on the outcome is unconfounded conditional on (X, C) but such that for no proper subset of (X, C) is the effect of the exposure on the outcome unconfounded given the subset."

continuous variable: A variable that can take on any number of possible values. Practically speaking, when a variable can take on at least, say, 10 values, it can be treated as a continuous variable. For example, it can be plotted on a scatterplot and certain meaningful calculations can be made using the variable.

covariate: See *predictor*

Cox model: The Cox proportional hazards regression model[3] is a model for relating a set of patient descriptor variables to time until death or other event. Cox analyses are based on the entire survival curve. The time-to-event may be *censored* due to loss to follow-up or by another event, as long as the censoring is independent of the risk of the event under study. Descriptor variables may be used in two ways: as part of the regression model and as stratification factors. For variables that enter as regressors, the model specifies the relative effect of a variable through its impact on the hazard or instantaneous risk of death at any given time since enrollment. For stratification factors, no assumption is made about how these factors affect survival, i.e., the proportional hazards assumption is not made. Separately shaped survival curves are allowed for these factors. The *logrank test* for comparing two survival distributions is a special case of the Cox model. Also see *survival analysis*. Cox models are used to estimate adjusted hazard ratios.

critical value: The value of a test statistic (e.g., t, F, χ^2, z) that if exceeded by the observed test statistic would result in statistical significance at a chosen α level or better. For a z-test (normal deviate test) the critical level of z is 1.96 when $\alpha = 0.05$ for a two-sided test. For t and F tests, critical values decrease as the sample size increases, as one requires less penalty for having to estimate the population variance as n gets large.

cross-validation: This technique involves leaving out m patients at a time, fitting a model on the remaining $n - m$ patients, and obtaining an unbiased evaluation of predictive accuracy on the m patients. The estimates are averaged over $\geq n/m$ repetitions. Cross-validation provides estimates that have more variation than those from bootstrapping. It may require > 200 model fits to yield precise estimates of predictive accuracy.

data science: A same-sex marriage between statistics and computer science.

degrees of freedom: The has somewhat different meanings depending on the context. In general, d.f. is the number of "free floating" parameters or the number of opportunities a statistical estimator or method was given. For a continuous variable Y, there are two types of d.f.: numerator d.f. and denominator d.f. Denominator d.f. is also called *error d.f.* and is the sample size minus the number of parameters needing to be estimated. It is the denominator of a variance estimator. Numerator d.f. is more aligned with opportunities and is the number of parameters currently being considered/tested. For example, in a "chunk" test for testing whether either height or weight is associated with blood pressure, the test has 2 d.f. if linearity and absence of interaction are assumed. In a traditional ANOVA comparing 4 groups, the comparisons have 3 d.f. because any 3 differences involving the 4 means or combinations of means will uniquely define all possible differences in the 4. One can say that the d.f. for a hypothesis is the number of opportunities one gives associations to be present (relationships to be non-flat), which is the same as the number of restrictions one needs to place on parameters so that the null hypothesis of no association (flat relationships) holds.

detectable difference: The value of a true population treatment effect (difference between two treatments) that if held would result in a statistical test having exactly the desired power.

discrimination: A variable or model's discrimination ability is its ability to separate subjects having a low responses from subjects having high responses. One way to quantify discrimination is the *ROC curve* area.

dummy variable: A device used in a multivariable regression model to describe a categorical predictor without assuming a numeric scoring. *Indicator variable* might be a better term. For example, treatments A, B, C might be described by the two dummy predictor variables X_1 and X_2, where X_1 is a binary variable taking on the value of 1 if the treatment for the subject is B and 0 otherwise, and X_2 takes on the value 1 if the subject is under treatment C and 0 otherwise. The two dummy variables completely define 3 categories, because when $X_1 = X_2 = 0$ the treatment is A.

entry time: The time when a patient starts contributing to the study. In randomized studies or observational studies where all patients have come under observation before the study starts (for example, studies of survival after surgery) the entry time and time origin of the study will be identical. However, for some observational studies, the patient may not start follow-up until after the time origin of the study and these patients contribute to the study group only after their 'late entry.'[2]

estimate: A statistical estimate of a parameter based on the data. See *parameter*. Examples include the sample mean, sample median, and estimated regression coefficients.

frequentist statistical inference: Currently the most commonly used statistical philosophy. It uses hypothesis testing, type I and II assertion probabilities, power, P-values, confidence limits (compatibility intervals), and adjustments of P-values for testing multiple hypotheses from the same study. Probabilities computed using frequentist methods, P-values, are probabilities of obtaining values of *statistics*. The frequentist approach is also called the *sampling* approach as it considers the distribution of statistics over hypothetical repeated samples from the same population. The frequentist approach is concerned with long-run operating characteristics of statistics and estimates. Because of this and because of the backwards time/information ordering of P-values, frequentist testing requires complex multiplicity adjustments but provides no guiding principles for exactly how those adjustments should be derived. Frequentist statistics involves confusion of two ideas: (1) the apriori probability that an experiment will generate misleading information (e.g., the chance of an assertian of an effect when there is no effect, i.e., type I assertion probability α) and (2) the evidence for an assertion after the experiment is run. The latter should not involve a multiplicity adjustment, but because the former does, frequentists do not know how to interpret the latter when multiple hypotheses are tested or when a single hypothesis is tested sequentially. Frequentist statistics as typically practiced places emphasis on hypothesis testing rather than estimation.

Gaussian distribution: See *normal distribution*.

generalizability: See *replication, reproduction, robust, generalizable*

generalized linear model: A model that has the same right-hand side form as a *linear regression model* but whose dependent variable can be categorical or can have a continuous distribution that is not normal. Examples of GLMs include binary logistic regression, probit regression, Poisson regression, and models for continuous Y having a γ distribution, plus the Gaussian distribution special case of

the linear regression model. GLMs can be fitted by maximum likelihood, quasi-likelihood, or Bayesian methods.

Gini's mean difference: A measure of variability (dispersion) that is much more interpretable than the standard deviation and more robust to outliers, and also applies to non-symmetric distributions. It is the mean absolute difference between all possible pairs of observations. There is a fast computing formula for the index, and the index is highly statistical efficient.

goodness of fit: Assessment of the agreement of the data with either a hypothesized pattern (e.g., independence of row and column factors in a contingency table or the form of a regression relationship) or a hypothesized distribution (e.g., comparing a histogram with expected frequencies from the normal distribution).

hazard rate: The instantaneous risk of a patient experiencing a particular event at each specified time[2]. The instantaneous rate with which an event occurs at a single point in time. It is the probability that the event occurs between time t and time $t + \delta$ given that it has not yet occurred by time t, divided by δ, as δ becomes vanishingly small. Note that rates, unlike probabilities, can exceed 1.0 because they are quotients.

hazard ratio: The ratio of hazard rates at a single time t, for two types of subjects. Hazard ratios are in the interval $[0, \infty)$, and they are frequently good ways to summarize the relative effects of two treatments at a specific time t. Like odds ratios, hazard ratios can apply to any level of outcome probability for the reference group. Note that a hazard ratio is distinct from a *risk ratio*, the latter being the ratio of two simple probabilities and not the ratio of two rates.

Hawthorne effect: A change in a subject response that results from the subject knowing she is being observed.

heterogeneity of treatment effect: Variation of the effect of a treatment on a scale for which it is mathematically possible for a treatment that has a nonzero effect on the average to have the same effect for different types of subjects. HTE should not be considered on the absolute risk scale (see *risk magnification*) but rather on a relative scale such as log odds or log hazard. HTE is best thought of as something due to a particular combination of treatment and patient that is mechanistic and not just related to the generalized risk that sicker patients are operating under. For example, patients with more severe coronary artery disease may get more relative benefit from revascularization, and patients who are poor metabolizers of a drug may get less relative benefit of the drug. Variation in the absolute risk reduction (ARR) due to a treatment is often misstated as HTE. Since ARR must vary by subject when risk factors exist and when the overall treatment effect is nonzero, variation in ARR is a mathematical necessity. It is dominated by subjects' baseline risk so is more accurately termed *heterogeneity in subjects* rather than *heterogeneity of treatment effects*.

intention-to-treat: Subjects in a randomized clinical trial are analyzed according to the treatment group to which they were assigned, even if they did not receive the intended treatment or received only a portion of it. If in a randomized study an analysis is done which does not classify all patients to the groups to which they were randomized, the study can no longer be strictly interpreted as a randomized trial, i.e., the randomization is "broken". Intention-to-treat analyses are pragmatic in that they reflect real-world non-adherence to treatment.

inter-quartile range: The range between the outer quartiles (25^{th} and 75^{th} percentiles). It is a measure of the spread of the data distribution (dispersion), i.e., a central interval containing half the sample.

least squares estimate: The value of a regression coefficient that results in the minimum sum of squared errors, where an error is defined as the difference between an observed and a predicted dependent variable value.

likelihood function: The probability of the observed data as a function of the unknown parameters for the data distribution. Here we use "probability" in a loose sense (and call it *likelihood*) so that it can apply to both discrete and continous[1] outcome variables. When the outcome variable Y can take on only discrete values (e.g., Y is binary or categorical), given a statistical model one can compute the exact probability that any given possible value of Y can be observed. In this case, the joint probability of a set of such occurrences can easily be computed. When the observations are independent, this joint probability is the product of all the individual probabilities. The likelihood function is then the joint probability that all the *observed* values of Y *would have occurred*[2], as a function of the unknown parameters that create the entire distribution of an individual observation's Y. When Y is continuous, the probability elements making up the likelihood function are the *probability density function* values evaluated at the observed data. Because joint probabilities of many observations are very small, and for another reason about to be given, it is customary to state natural logs of likelihoods rather than using the original scale. The log likelihood achieved by a model, that is, the log likelihood at the maximum likelihood estimates of the unknown parameters, is a gold standard information measure and is used to compute various statistics including R^2, AIC, and likelihood ratio χ^2 tests of association. See *maximum likelihood estimate*, which is the set of parameter values making the observed data most likely to have been observed.

linear regression model: This is also called OLS or ordinary least squares and refers to regression for a continuous dependent variable, and usually to the case where the residuals are assumed to be Gaussian. The linear model is sometimes called *general linear model*, not to be confused with *generalized linear model* where the distribution can take on many non-Gaussian forms.

logistic regression model: A multivariable regression model relating one or more predictor variables to the probabilities of various outcomes. The most commonly used logistic model is the *binary logistic model*[8,7] which predicts the probability of an event as a function of several variables. There are several types of *ordinal logistic models* for predicting an ordinal outcome variable, and there is a *polytomous logistic model* for categorical responses. The binary and polytomous models generalize the χ^2 test for testing for association between categorical variables. One commonly used ordinal model, the proportional odds model[1], generalizes the Wilcoxon 2-sample rank test. Binary logistic models are useful for predicting events in which time is not very important. They can be used to predict events by a specified time, but this can result in a loss of information. Logistic models are used to estimate adjusted odds ratios as well as probabilities of events.

machine learning: An algorithmic procedure for prediction or classification that tends to be empirical, nonparametric, flexible, and does not capitalize on additivity of predictors. Machine learning does not use a data model, i.e., a probability distribution for the outcome variable given the inputs, and does not place emphasis on interpretable parameters. Examples of machine learning algorithms include neural

[1] The actual probability of a specific value for a continuous variable is zero.
[2] The probability that they actually occur is now moot since the Y values have already been observed.

networks, support vector machines, bagging, boosting, recursive partitioning, and random forests. Ridge regression, the *lasso*, *elastic net*, and other penalized regression techniques (which have identified parameters and make heavy use of additivity assumptions) fall under *statistical models* rather than machine learning. By allowing high-order interactions to be potentially as important as main effects, machine learning is "data hungry", as sample sizes needed to estimate interaction effects are much larger than sample sizes needed to estimate additive main effects. Machine learning is not to be confused with *artificial intelligence*.

masking: Preventing the subject, treating physician, patient interviewer, study director, or statistician from knowing which treatment a patient is given in a comparative study. A single-masked study is one in which the patient does not know which treatment she's getting. A double-masked study is one in which neither the patient nor the treating physician or other personnel involved in data collection know the treatment assignment. A triple-masked study is one in which the statistician is unaware of which treatment is which. Masking is also known as *blinding*.

maximum likelihood estimate: An estimate of a statistical parameter (such as a regression coefficient, mean, variance, or standard deviation) that is the value of that parameter making the data most likely to have been observed. MLEs have excellent statistical properties in general, such as converging to population values as the sample size increases, and having the best precision from among all such competing estimators, when the statistical model is correctly specified. When the data are normally distributed, maximum likelihood estimates of regression coefficients and means are equivalent to least squares estimates. When the data are not normally distributed (e.g. binary outcomes, or survival times), maximum likelihood is the standard method to estimate the regression coefficients (e.g. logistic regression, Cox regression). Unlike Bayesian estimators, MLEs cannot take extra-study information into account. MLEs can be overfitted when the data's information content does not allow reliable estimation of the number of parameters involved (see *overfitting*). Penalized MLEs can solve this problem, by maximizing a penalized log likelihood function. When extra-study information is not allowed to be utilized, MLE is considered a gold standard estimation technique. See *likelihood function*.

mean: Arithmetic average, i.e., the sum of all the values divided by the number of observations. The mean of a binary variable is equal to the proportion of ones because the sum of all the zero and one values equals the number of ones. The mean can be heavily influenced by outliers. When the tails of the distribution are not heavy, this influence of more extreme values is what gives the mean its efficiency compared to other estimators such as the *median*. When the data distribution is symmetric, the population mean and median are the same. The sample mean is a better estimator of the population median than is the sample median, when the data distribution is symmetric and Gaussian-like.

median: Value such that half of the observations' values are less than and half are greater than that value. The median is also called the 50^{th} percentile or the 0.5 quantile. The sample median is not heavily influenced by outliers so it can be more representative of "typical" subjects. When the data happen to be normally (Gaussian) distributed, the sample median is not as precise as the mean in describing the central tendency, its efficiency being $\frac{2}{\pi} \approx 0.64$.

multiple comparisons: It is common for one study to involve the calculation of more than one P-value. For example, the investigator may wish to test for treatment effects in 3 groups defined by disease etiology, she may test the effects on 4 different patient response variables, or she may look for a significant difference in blood pressure at each of 24 hourly measurements. When multiple statistical tests are done, the chances of at least one of them resulting in an assertion of an effect when there

are no effects increases as the number of tests increase. This is called "inflation of type I assertion probability α." When one wishes to control the *overall* type I probability, individual tests can be done using a more stringent α level, or individual P-values can be adjusted upward. Such adjustments are usually dictated when using frequentist statistics, as P-values mean the probability of getting a result this impressive if there is really no effect, and "this impressive" can be taken to mean "this impressive given the large number of statistics examined." Multiple comparisons and related inflation of type I probability are solely the result of chances that a frequentist gives data to be more extreme. In Bayesian inference, one deals with the (prior) chances that the true unknown multiple effects are large, and multiplicity per se does not apply.

multivariable model: A model relating multiple predictor variables (risk factors, treatments, etc.) to a single response or dependent variable. The predictor variables may be continuous, binary, or categorical. When a continuous variable is used, a linearity assumption is made unless the variable is expanded to include nonlinear terms. Categorical variables are modeled using *dummy variables* so as to not assume numeric assignments to categories.

multivariate model: A model that simultaneously predicts more than one dependent variable, e.g. a model to predict systolic and diastolic blood pressure or a model to predict systolic blood pressure 5 min. and 60 min. after drug administration.

nominal significance level: In the context of multiple comparisons involving multiple statistical tests, the apparent significance level α of each test is called the nominal significance level. The overall type I assertion probability for the study, the probability of at least one positive assertion when the true effect is zero, will be greater than α.

non-inferiority study: A study designed to show that a treatment is not clinically significantly worse than another treatment. Regardless of the significance/non-significance of a traditional superiority test for comparing the two treatments (with H_0 at a zero difference), the new treatment would be accepted as non-inferior to the reference treatment if the confidence interval (compatibility interval) for the unknown true difference between treatments excludes a clinically meaningful worsening of outcome with the new treatment.

nonparametric estimator: A method for estimating a parameter without assuming an underlying distribution for the data. Examples include sample quantiles, the empirical cumulative distribution, and the Kaplan-Meier survival curve estimator.

nonparametric tests: A test that makes minimal assumptions about the distribution of the data or about certain parameters of a statistical model. Nonparametric tests for ordinal or continuous variables are typically based on the ranks of the data values. Such tests are unaffected by any one-one transformation of the data, e.g., by taking logs. Even if the data come from a normal distribution, rank tests lose very little efficiency (they have a relative efficiency of $\frac{3}{\pi} = 0.955$ if the distribution is normal) compared with parametric tests such as the t-test and the linear correlation test. If the data are not normal, a rank test can be much more efficient than the corresponding parametric test. For these reasons, it is not very fruitful to test data for normality and then to decide between the parametric and nonparametric approaches. In addition, tests of normality are not always very powerful. Examples of nonparametric tests are the 2-sample Wilcoxon-Mann-Whitney test, the 1-sample Wilcoxon signed-rank test (usually used for paired data), and the Spearman, Kendall, or Somers rank correlation tests. Even though nonparametric tests do not assume a specific distribution for a group, they assume a connection between

the distributions of any two groups. For example, the logrank test assumes proportional hazards, i.e., that the survival curve for group A is a power of the survival curve for group B. The Wilcoxon test, for optimal power, assumes that the cumulative distributions are in proportional odds.

normal distribution: A symmetric, bell-shaped distribution that is most useful for approximating the distribution of statistical estimators. Also called the *Gaussian distribution*. The normal distribution cannot be relied upon to approximate the distribution of raw data. The normal distribution's bell shape follows a rigid mathematical equation of the form e^{-x^2}. For a normal distribution the probability that a measurement will fall within ± 1.96 standard deviations of the mean is 0.95.

null hypothesis: Customarily but not necessarily a hypothesis of no effect, e.g., no reduction in mean blood pressure or no correlation between age and blood pressure. The null hypothesis, labeled H_0, is often used in the *frequentist* branch of statistical inference as a "straw person"; classical statistics often assumes what one hopes doesn't happen (no effect of a treatment) and attempts to gather evidence against that assumption (i.e., tries to reject H_0). H_0 usually specifies a single point such as 0mmHg reduction in blood pressure, but it can specify an interval, e.g., H_0: blood pressure reduction is between -1 and +1 mmHg. "Null hypotheses" can also be e.g. H_0: correlation between X and Y is 0.5.

number needed to treat: A quantity that applies to an extremely oversimplified and unrealistic situation where (1) there is a special time horizon t and (2) all patients have the same absolute risk of having an outcome by time t, i.e., risk factors do not exist (otherwise NNT cannot be a single number). Specifically, NNT is the number of patients needed to be treated to prevent one bad outcome by time t, which is the reciprocal of the absolute outcome risk difference between two treatments. When there are risk factors, absolute risk difference varies tremendously over patient types, so an NNT may not apply to anyone in the patient population. Typically, sicker patients get more benefit of treatment, so the risk difference magnifies and NNT falls for them. There are a huge number of serious problems with NNT, detailed here. Confidence intervals for NNT are problematic but whether done correctly or incorrectly are often so wide as to cast doubt on the use of the point estimate.

observational study: Study in which no experimental condition (e.g., treatment) is manipulated by the investigator, i.e., randomization is not used. Such studies are frequently used to estimate characteristics of subjects (means, proportions, etc.) and to assess associations between variables. They have known limitations for therapeutic comparisons, because of unknown confounders.

odds: The probability an event occurs divided by the probability that it doesn't occur. An event that occurs 0.90 of the time has 9:1 odds of occurring since $\frac{0.9}{1-0.9} = 9$.

odds ratio: The odds ratio for comparing two groups (A, B) on their probabilities of an outcome occurring is the odds of the event occurring for group A divided by the odds that it occurs for group B. If P_A and P_B represent the probability of the outcome for the two groups of subjects, the $A:B$ odds ratio is $\frac{P_A}{1-P_A} / \frac{P_B}{1-P_B}$. Odds ratios are in the interval $[0, \infty)$. An odds ratio for a treatment is a measure of relative effect of that treatment on a binary outcome. As summary measures, odds ratios have advantages over *risk ratios*: they don't depend on which of two possible outcomes is labeled the "event", and any odds ratio can apply to any probability of outcome in the reference group. Because of this, one frequently finds that odds ratios for comparing treatments are relatively constant across different types of patients. The same is not true of risk ratios or risk differences; these depend on the level of risk in the reference group.

one-sided test: A test designed to test a directional hypothesis, yielding a one-sided P-value. For example, one might test the null hypothesis H_0 that there is no difference in mortality between two treatments, with the alternative hypothesis being that the new drug lowers mortality. See also *two-sided test*.

ordinal variable: A categorical variable for which there is a definite ordering of the categories. For example, severity of lower back pain could be ordered as none, mild, moderate, severe, and coded using these names or using numeric codes such as 0,1,2,10. Spacings between codes are not important.

overfitting: In the context of a prediction tool developed using a statistical model or using an algorithmic procedure such as machine learning, the tendency for the predicted values to be too extreme. Too-extreme predictions make the *calibration* curve show symptoms of regression to the mean: a flattening of the curve to be less steep than the 45° line of identify. When overfitting is present, low predicted values are too low and/or high predicted values are too high. Overfitting is synonymous with over-interpretation caused by slicing the data into pieces that do not have huge denominators. The cause of overfitting is typically having too many candidate features in a supervised learning (informed by Y) feature selection setting or estimating too many parameters in a pre-specified model. Each parameter estimated may be unbiased, but predictions are formed by putting all the parameters together, and unless the model is intentionally underfitted using penalization (shrinkage; regularization), the combination of parameters exhibits the "low values too low or high values too high" phenomenon. This is due in part to sorting predicted values or to selecting subjects with extreme predictions. It is possible that the overall mean predicted value be unbiased even with extreme overfitting. That is why it is important to estimate the entire *calibration* curve.

P-value: The probability of getting a result (e.g., t or χ^2 statistics) as or more extreme than the observed statistic had H_0 been true. An α-level test would reject H_0 if $P \leq \alpha$. However, the P-value can be reported instead of choosing an arbitrary value of α. Examples: (1) An investigator compared two randomized groups for differences in systolic blood pressure, with the two mean pressures being 134.4 mmHg and 138.2 mmHg. She obtained a two-tailed $P = 0.03$. This means that if there is truly no difference in the population means, one would expect to find a difference in means exceeding 3.8 mmHg in absolute value 0.03 of the time. The investigator might conclude there is evidence for a treatment effect on mean systolic blood pressure if the statistical test's assumptions are true. (2) An investigator obtained $P = 0.23$ for testing a correlation being zero, with the sample correlation being 0.08. The probability of getting a correlation this large or larger in absolute value if the population correlation is zero is 0.23. No conclusion is possible other than (a) more data are needed and (b) there is no convincing evidence for or against a zero correlation. For both of these examples compatibility (confidence) intervals would be helpful. The P-value is **not** the probability that the null hypothesis is true, and is **not** the probability that the results are due to chance. P is computed under the assumption that the results **are** due to chance.

paired data: When each subject has two response measurements, there is a natural pairing to the data and the two responses are correlated. The correlation results from the fact that generally there is more variation between subjects than there is within subjects. Sometimes one can take the difference or log ratio of the two responses for each subject, and then analyze these "effect measures" using an unpaired one-sample approach such as the Wilcoxon signed-rank test or the paired t-test. One must be careful that the effect measure is properly chosen so that it is independent of the baseline value.

parameter: An unknown quantity such as the population mean, population variance, difference in two means, or regression coefficient.

parametric model: A model based on a mathematical function having a few unknown parameters. Typically the number of parameters in a parametric model does not grow with the sample size, and a specific distribution is assumed for the dependent variable Y, conditional on X. See also *semiparametric model*.

parametric test: A test which makes specific assumptions about the distribution of the data or specific assumptions about model parameters. Examples include the t-test and the Pearson product-moment linear correlation test.

percentile: The p-th percentile is the value such that $\frac{np}{100}$ of the observations' values are less than that value. The p-th *quantile* is the value such that np of the observations' values are less.

phase I: Studies to obtain preliminary information on dosage, absorption, metabolism, and the relationship between toxicity and the dose-schedule of treatment.

phase II: Studies to determine feasibility and estimate treatment activity and safety in diseases (or for example tumor types) for which the treatment appears promising. Generates hypotheses for later testing.

phase III: Comparative trial to determine the effectiveness and safety of a new treatment relative to standard therapy. These trials usually represent the most rigorous proof of treatment efficacy (pivotal trials) and are the last stage before product licensing..

phase IV: Post-marketing studies of licensed products.

posterior probability: In a *Bayesian* context, this is the probability of an event after making use of the information in the data. In other words, it is the *prior probability* of an event after updating it with the data. Posterior probability can also be called post-test probability if one equates a diagnostic test with "data" (see also *ROC curve*).

power: Probability of rejecting the null hypothesis for a set value of the unknown effect. Power could also be called the sensitivity of the statistical test in detecting that effect. Power increases when the sample size and true unknown effect increase and when the inter-subject variability decreases. In a two-group comparison, power generally increases as the allocation ratio gets closer to 1:1. For a given experiment it is desirable to use a statistical test expected to have maximum power (sensitivity). A less powerful statistical test will have the same power as a better test that was applied after discarding some of the observations. For example, testing for differences in the proportion of patients with hypertension in a 500-patient study may yield the same power as a 350-patient study which used blood pressure as a continuous variable. See *type II probability*.

precision: Degree of absence of random error. The precision of a statistical estimator is related to the expected error that occurs when approximating the infinite-data value. In other words, when you try to estimate some measure in a population, the precision is related to the error in the estimate. So precision can be thought of as a "margin of error" in estimating some unknown value. Precision can be quantified by the width of a confidence (compatibility) interval and sometimes by a standard deviation of the estimator (standard error). For the confidence intervals, a "margin for error" is computed so that the quoted interval has a certain probability of containing the true value (e.g., population mean difference). Some authors define precision as the reciprocal of the variance of an estimate. By that definition, precision increases linearly as the sample size increases. If instead one defines precision on

the original scale of measurement instead of its square (i.e., if one uses the standard error or width of a confidence interval), precision increases as the square root of the sample size.

predictor, explanatory variable, risk factor, covariate, covariable, independent variable: quantities which may be associated with better or worse outcome[2]. Without further information, predictors (covariates) are taken to be measured at baseline. Time-dependent covariates are updated with post-baseline measurements. An external time-dependent covariate is one whose future values were already known at baseline. For example, in a crossover study, the new treatment assignment at one month (the crossover time) was already known at the time of randomization. Effects of external time-dependent covariates are easy to interpret. Internal time-dependent covariates (e.g., updated cholesterol measurements) may reflect changing subject condition. An especially difficult-to-interpret situation is a randomized trial in which one estimates the (supposedly constant) treatment effect after adjusting for internal time-dependent covariates.

prior probability: The probability of an event as it could best be assessed before the experiment. In diagnostic testing this is called the pre-test probability. The prior probability can come from an objective model based on previously available information, or it can be based on expert opinion. In some *Bayesian* analyses, prior probabilities are expressed as probability distributions which are flat lines, to reflect a complete absence of knowledge about an event. Such distributions are called non-informative, flat, or reference distributions, and analyses based on them fully let the data "speak for themselves."

probability: The probability that an event will occur, that an invisible event has already occurred, or that an assertion is true, is a number between 0 and 1 inclusive such that (1) of all possible outcomes (including non-events) the probability of some possible outcome occurring is 1, and (2) the probability of any of a set of mutually exclusive events (i.e., *union* of events) occurring is the sum of the individual event probabilities[3]. The *meaning* attached to the metric known as a probability is up to the user; it can represent long-run relative frequency of repeatable observations, a degree of belief, or a measure of veracity or plausibility. In the *frequentist* school, the probability of an event denotes the limit of the long-term fraction of occurrences of the event. This notion of probability implies that the same experiment which generated the outcome of interest can be repeated infinitely often. Even a coin will change after 100,000 flips. Likewise, some may argue that a patient is "one of a kind" and that repetitions of the same experiment are not possible. One could reasonably argue that a "repetition" does not denote the same patient at the same stage of the disease, but rather *any* patient with the same *severity* of disease (measured with current technology). There are other schools of probability that do not require the notion of replication at all. For example, the school of *subjective* probability (associated with the *Bayesian* school) "considers probability as a measure of the degree of belief of a given subject in the occurrence of an event or, more generally, in the veracity of a given assertion" (see P. 55 of[5]). de Finetti defined subjective probability in terms of wagers and odds in betting. A risk-neutral individual would be willing to wager $\$P$ that an event will occur when the payoff is \$1 and her subjective probability is P for the event. The domain of application of probability is all-important. We assume that the true event status (e.g., dead/alive) is unknown, and we also assume that the information the probability is conditional upon (e.g. Pr{death | male, age=70}) is what we would check the probability against. In other words, we do not ask whether Pr(death | male, age=70) is accurate when compared against Pr(death | male, age=70, meanbp=45, patient on downhill course).

[3]These are Kolmogorov's axioms of probability. All other probability rules can be derived from these axioms.

It is difficult to find a probability that is truly not conditional on anything. What is conditioned upon is all important. Probabilities are maximally useful when, as with Bayesian inference, they condition on what is known to provide a forecast for what is unknown. These are "forward time" or "forward information flow" probabilities. Forward time probabilities can meaningfully be taken out of context more often than backward-time probabilities, as they don't need to consider "what might have happened." In frequentist statistics, the P-value is a backward information flow probability, being conditional on the unknown effect size. This is why P-values must be adjusted for multiple data looks ("what might have happened") whereas the current Bayesian posterior probability merely override any posterior probabilities computed at earlier data looks, because they now condition on current data. As IJ Good has written, the axioms defining the "rules" under which probabilities must operate (e.g., a probability is between 0 and 1) do not define what a probability actually means. He also states that all probabilities are subjective, because they depend on the knowledge of the particular observer.

probability density function: When a random variable Y is continuous, i.e., it can take on every possible number within some interval, the probability density function is a function of y which is the limit, as the width δ of some interval goes to zero, of the probability that Y will be within the interval $[y, y+\delta]$, divided by δ. This is the first derivative (slope) of the cumulative probability distribution function for Y.

proper accuracy scoring rule: When applied to predicting categorical outcomes, a proper probability accuracy scoring rule is a measure that is optimized when the predicted probabilities are the true outcome probabilities. Examples of proper accuracy scores include the Brier score, the logarithmic probability score, and the log-likelihood from a correct statistical model. Examples of improper scoring rules, i.e., rules that are optimized by a bogus model, are proportion classified correctly, sensitivity, specificity, precision, recall, and the c-index (area under the receiver operating characteristic curve).

proportional hazards: This assumption is fulfilled if two categories of patient are being compared and their hazard ratio is constant over time (though the instantaneous hazards may vary)[2].

prospective study: One in which the study is first designed, then the subjects are enrolled. Prospective studies are usually characterized by intentional data collection.

quartiles: The 25^{th} and 75^{th} percentiles and the median. The three values divide a variables distributions into four intervals containing equal numbers of observations.

random error: An error caused by sampling from a group rather than knowing the true value of a quantity such as the mean blood pressure for the entire group, e.g., healthy men over age 80. One can also speak of random errors in single measurements for individual subjects, e.g., the error in using a single blood pressure measurement to represent a subject's long-term blood pressure.

random sample: A sample selected by a random device that ensures that the sample (if large enough) is representative of the infinite group. A *probability sample* is a kind of random sample in which each possible subject has a known probability of being sampled, but the probabilities can vary. For example, one may wish to over-sample African-Americans in a study to ensure good representation. In that case one could sample African-Americans with probability of 1.0 and others with a probability of 0.5.

randomized controlled trial: See *clinical trial*.

randomness: Absence of a systematic pattern. One might wish to examine whether some hormone level varies systematically over the day as opposed to having a random pattern, or whether events such as epileptic seizures tend to cluster or occur randomly in time. Sometimes the residuals in an ordinary regression model are plotted against the order in which subjects were accrued to make sure that the pattern is random (e.g., there was no learning trend for the investigators).

rate: A ratio such as a change per unit time. Rates are often limits, and shouldn't be confused with probabilities. The latter are constrained to be between 0 and 1 whereas there are no constraints on possible values for rates other than not being negative. A rate may also be a ratio such as "falls per distance walked" or "bacteria per unit of surface area."

regression to the mean: Tendency for a variable that has an extreme value on its first measurement to have a more typical value on its second measurement. For example, suppose that subjects must have LDL cholesterol > 190mg% to qualify for a study, and the median LDL cholesterol for qualifying subjects at the screening visit was 230 mg%. The median LDL cholesterol value at their second visit might be 200mg%, with several of the subjects having values below 190. This is the "sophomore slump" in baseball; second-year players are watched when they have phenomenal rookie years. Regression to the mean also takes many other forms, all arising because variables or subgroups are not examined at random but rather because they appear "impressive": (1) One might compare 5 treatments with a control and choose the treatment having the maximum difference. On a repeated study that treatment's average response will be found to be much closer to that of the control. (2) In a randomized controlled trial the investigators may wish to estimate the effect of treatment in multiple subgroups. They find that in 40 left-handed diabetics the treatment multiplies mortality by 0.4. If the study is replicated, they would find that the mortality reduction in left-handed diabetics is much closer to the mortality reduction in the overall sample of patients. (3) Researchers study the association between 40 possible risk factors and some outcome, and find that the factor with the strongest association had a correlation of 0.5 with the response. On replication, the correlation will be much lower. This result is very related to what happens in stepwise variable selection, where the most statistically significant variables selected will have their importance (regression coefficients) greatly overstated.

relative risk or risk ratio: The ratio of the probabilities of two events. Unlike the *odds ratios* and *hazard ratios*, risk ratios are not capable of being constant but instead must depend on the base risk (e.g., the risk for a subject who does not have a risk factor). For example, a risk ratio of 2 may apply only to subjects with base risks $< \frac{1}{2}$. Also unlike the odds ratio, the risk ratio depends greatly on which of two outcomes is labeled as the "event"; a mortality ratio does not equal the survival ratio. The term *relative risk* is often inappropriately used to describe an odds ratio or a hazard ratio.[2].

replication, reproduction, robust, generalization: Reproduction means to execute what is apparently the same data analysis used by the original authors, on their data. Replication means to do the original analysis on new data. A robust result is getting largely the same result with a different analysis on the original dataset. Generalization means to operationalize the experiment and analysis differently, use new data, and get largely the same result (e.g., using a different genetics, proteomics, or imaging platform or translating a questionnaire to a different language and doing a survey in a different country). Generalization is also taken to mean validating that a treatment works similarly on patients who are different from those in a clinical study. Potential reproducibility means that the investigators have provided data manipulation and analysis code that is fully self-contained and could be executed by another person to obtain all there analytical results obtained by the original researchers.

residual: A statistical quantity that should be unrelated to certain other variables because their effects should have already been subtracted out. In ordinary multiple regression, the most commonly used residual is the difference between predicted and observed values.

retrospective study: A study in which subjects were already enrolled before the study was designed, or the outcome of interest has occurred before the start of the study (an in a *case control study*). Such studies often have difficulties such as absence of needed adjustment (confounder) variables and missing data.

risk: Often used as another name for *probability* but a more accurate definition is the probability of an adverse event $\times$ the severity of the loss that experiencing that event would entail.

risk magnification: A treatment, even one for which there are no interactions with baseline covariates, that has a nonzero effect on a relative scale, by necessity must have different absolute effects. The variation of absolute differences is risk magnification due to baseline risk. Subjects having baseline risks near 0 or 1 have nowhere to go; absolute risk differences are less restricted in the middle of the baseline risk distribution. Treatments have greater absolute benefit for sicker patients, up to a point, even if their relative effects are universal.

risk set: The set of patients in the study at a specified time[2].

ROC curve: When an ordinal or continuous marker is used to diagnose a binary disease, a receiver operating characteristic or ROC curve can be drawn to study the discrimination ability of the marker. The ROC curve is a plot of *sensitivity* vs. one minus *specificity* of all possible dichotomizations of the marker as the cutpoints are varied. A major problem with the ROC curve is that it tempts the researcher to publish cutpoints to somewhat arbitrarily classify patients as "diseased" and "normal". In fact when the diagnostic analysis is based on a cohort study, the marker's value can be converted into a *post-test probability* of disease allowing different physicians to use different cutpoints when the need arises (e.g., depending on available resources). Another benefit of the latter approach is that the current probability of disease also defines the probability of an error. For example, if a physician elects not to treat when the probability of disease is 0.04, the false negative probability is 0.04. The area under the ROC curve is one way to summarize the diagnostic discrimination. This area is identical to another more intuitive and easily computed measure of discrimination, the probability that in a randomly chosen pair of patients, one with and one without disease, the one with disease is the one with a higher value of the marker or post-test probability. This is also called the probability of concordance between predicted and observed disease states. A frequently used index of rank correlation, Somers' D_{xy} equals $2 \times (c - \frac{1}{2})$ where c is the concordance (discrimination) probability. It is important to note that ROC curves play no role in formal decision making, as they ignore the utility (cost; loss) function or the cost of false positives and false negatives.

semiparametric model: 'Parametric' assumptions may be made about some aspects of a model, while other components may be estimated 'non-parametrically'. In the Cox regression procedure, a parametric model for the relative hazard is overlaid on a nonparametric estimate of baseline hazard[2]. Like the proportional odds ordinal logistic model, the Cox semiparametric (proportional hazards) model is fully parametric on the right hand side, and nonparametric on the left hand (dependent variable Y) side. These types of semiparametric models essentially have an intercept for each distinct value of Y occurring in the data, allowing for estimation of the distribution of Y in a way that is very similar to the empirical cumulative distribution function, a nonparametric distribution estimator.

sensitivity and specificity: One way to quantify the utility of a diagnostic test when both the disease and the test are binary. The sensitivity is the probability that a patient with disease will have a positive test, and the specificity is the probability that a patient without disease will have a negative test. In general, it is more natural and useful to study variations in post-test probabilities of disease given different test results and different patient pre-test characteristics because (1) in general both the sensitivity and specificity will vary with the type of patient being diagnosed, (2) sensitivity increases with the severity of the disease present unless the disease is all-or-nothing, (3) specificity can vary with gradations in pre-clinical amount of disease, and (4) many diagnostic tests are based on continuous rather than binary measurements[4]. *Multivariable models* are very useful for estimating post-test probabilities. The *calibration* and *discrimination* of the post-test probabilities can be quantified.

significance level: A preset value of α against which P-values are judged in order to reject H_0 (see *Type I error*). Sometimes a P-value itself is called the significance level.

standard deviation: A measure of the variability (spread) of measurements across subjects. The standard deviation has a simple interpretation only if the data distribution is Gaussian (normal), and in that restrictive case the mean ± 1.96 standard deviations is expected to cover 0.95 of the distribution of the measurement. Standard deviation is the square root of the *variance*. It does not apply very well to asymmetric (skewed) distributions, and is not robust to outliers.

standard error: The standard deviation of a statistical estimator. For example, the standard deviation of a *mean* is called the standard error of the mean, and it equals the standard deviation of individual measurements divided by the square root of the sample size. Standard errors describe the precision of a statistical summary, not the variability across subjects. Standard errors go to zero as the sample size $\rightarrow \infty$.

statistical model: A model with identified parameters that comprises a model for the data through a probability distribution and favors additivity of effects. Examples of statistical models include ordinary linear regression with an assumption of a Gaussian distribution for the residuals, logistic regression, Cox proportional hazards regression, longitudinal models, quantile regression, ridge regression, lasso, and elastic net.

survival analysis: A branch of statistics dealing with the analysis of the time until an event such as death. Survival analysis is distinguished by its emphasis on estimating the time course of events and in dealing with *censoring*. See *Cox model*.

survival function: The probability of being free of the event at a specified time[2].

survival time: Interval between the time origin and the occurrence of the event or censoring[2].

symmetric distribution: One in which values to the left of the mean by a certain amount are just as likely to be observed as values to the right of the mean by the same amount. For symmetric distributions, the population mean and median are identical and the distance between the 25^{th} and 50^{th} percentiles equals the distance between the 50^{th} and 75^{th} percentiles.

time-dependent covariate: See *predictor*

time origin: The beginning of the story the study aims at telling. In observational studies, the patients may come under observation before or after the time origin of the study[2], but one often attempts to

define time zero as date of diagnosis, initiation of exposure, or treatment. In randomized trials, the time origin is the date of randomization.

two-sided test: A test that is non-directional and that leads to a two-sided P-value. If the null hypothesis H_0 is that two treatments have the same mortality outcome, a two-sided alternative is that the mortality difference is nonzero. Two-sided P-values are larger than one-sided P-values (they are double if the distribution of the test statistic is symmetric). They can be thought of as a multiplicity adjustment that would allow a claim to be made that a treatment lowers **or** raises mortality. See also *one-sided test*.

type I assertion probability α: Frequently confusingly labeled as a false positive probability[4], this is the probability of rejecting H_0 (i.e., declaring "statistical significance" — not recommended) when the null hypothesis is assumed to be true. The type I assertion probability is often called α and is the probability of making an assertion of an effect when any assertion of effect is by definition false. It is usually called a *rate* but this is not accurate. In common use, the type I probability is the probability that the nominal P-value will be < 0.05 if there is no effect. This will be 0.05 when (1) only one P-value is computed, (2) all model and experimental design assumptions made by the P-value calculation are exactly true, and (3) the P-value is computed exactly. See here for a detailed discussion of the distinction between assertion probabilities and decision error probabilities.

type II assertion probability β: Frequently confusingly labeled as a false negative probability[5], this is the probability of not asserting an effect (i.e., failing to reject H_0) when there truly is a specific magnitude of effect. The type II probability is referred to as β, which is one minus the power of the test. In other words, the power of the test is $1 - \beta$. This probability β is often wrongly called a *rate*.

variance: A measure of the spread or variability of a distribution, equaling the average value of the squared difference between measurements and the population mean measurement. From a sample of measurements, the variance is estimated by the sample variance, which is the sum of squared differences from the sample mean, divided by the number of measurements minus 1. The minus 1 is a kind of "penalty" that corrects for estimating the population mean with the sample mean. Variances are typically only useful when the measurements follow a normal or at least a *symmetric distribution*.

Other Resources

- Glossary of Statistical Terms from the UC Berkely Statistics Department
- Glossary of Probability and Statistics in Wikipedia

[4] It is valid to say that α is the probability of indicating an effect when there is no effect, but this is much different from the probability of being wrong in asserting that an effect is present. This probability cannot be derived from a probability of asserting an effect given that the effect is zero. The probability of being wrong in asserting an effect is computed properly by taking one minus the Bayesian posterior probability of an effect being present.

[5] Type II probability may be called the probability of a false negative assertion, but this is very distinct from the probability that there is an effect when one does not assert an effect. This probability cannot be derived from a probability of failing to assert an effect given the effect is at a certain nonzero level. The Bayesian posterior probability of an effect is the unconditional (except for the data) probability of a nonzero effect.

References

[1] Scott R. Brazer et al. "Using Ordinal Logistic Regression to Estimate the Likelihood of Colorectal Neoplasia". In: *J Clin Epi* 44 (1991), pp. 1263–1270.

[2] Kate Bull and David Spiegelhalter. "Survival Analysis in Observational Studies". In: *Stat Med* 16 (1997), pp. 1041–1074.

[3] David R. Cox. "Regression Models and Life-Tables (with Discussion)". In: *J Roy Stat Soc B* 34 (1972), pp. 187–220.

[4] M. A. Hlatky et al. "Factors Affecting the Sensitivity and Specificity of Exercise Electrocardiography. Multivariable Analysis". In: *Am J Med* 77 (1984), pp. 64–71. URL: http://www.sciencedirect.com/science/article/pii/0002934384904376#.

[5] Samuel Kotz and Norman L. Johnson, eds. *Encyclopedia of Statistical Sciences*. Vol. 9. New York: Wiley, 1988.

[6] Pearl, Judea. *Causal Inference in Statistics: An Overview*. Sept. 2009, pp. 96–146. URL: http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf (visited on 08/14/2019).

DOI:10.1214/09-SS057 See Sections 2.1-2.3
.

[7] Alan Spanos, Frank E. Harrell, and David T. Durack. "Differential Diagnosis of Acute Meningitis: An Analysis of the Predictive Value of Initial Observations". In: *JAMA* 262 (1989), pp. 2700–2707. DOI: 10.1001/jama.262.19.2700.

[8] S. H. Walker and D. B. Duncan. "Estimation of the Probability of an Event as a Function of Several Independent Variables". In: *Biometrika* 54 (1967), pp. 167–178.

Acknowledgments: Richard Goldstein provided valuable additions and clarifications to the glossary and additional medical statistics citations. As noted in the glossary, several definitions came from[2]. Thanks to Sebastian Baumeister for the definition of confounder. Raphael Peter extended the definition of *rate*. Rob Zinkov and Raphael Peter provided input in the definition of clinical trials. Julia Rohrer provided the essence of the definitions of reproducibility, replicability, robustness, and generalizability. Ronan Conroy improved definitions of *inter-quartile range* and *observational study* and prompted improvements on *parametric model* and creation of a definition for *degrees of freedom*. Thanks to Bryan Shepherd for pointing out the best formal definition of confounding. Andrew Spieker provided the definition for *causal inference*.