

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 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.

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. 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.

bias: A systematic error. Examples: a mis-calibrated 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 antihypertensive properties.

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 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 40% of the time it rained when the predicted probability of rain was 0.4, the rain forecast is perfectly calibrated.

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*.

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.

clinical trial: An experiment with human subjects in which there is control over treatment assignment.

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.

confidence limits: To say that the 95% confidence limits for an unknown quantity are $[a, b]$ means that 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 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. 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.

confounder: A variable which is correlated with the response variable and with the treatment assignment (or exposure variable). A confounder, when properly controlled for, can explain away an apparent association between the treatment and the response.

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.

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.

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.

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. 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 errors, power, P -values, confidence limits, and adjustments of P -values

for testing multiple hypotheses from the same study. Final 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.

Gaussian distribution: See *normal distribution*.

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.

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 (25th and 75th percentiles).

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.

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* [7, 6] 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.

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 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).

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.

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 50th percentile or the 0.5 quantile. The 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 median is not as precise as the mean in describing the central tendency.

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 a false positive finding increases as the number of tests increase. This is called "inflation of type I error." When one wishes to control the *overall* type I error, 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."

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 error for the study, the probability of at least one false positive result, will be greater than α .

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 (typically 4%) 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.

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.

observational study: Study in which no experimental condition (e.g., treatment) is manipulated by the investigator, i.e., randomization is not used.

odds: The probability an event occurs divided by the probability that it doesn't occur. An event that occurs 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.

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.

***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 tests'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 0.08. 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 confidence intervals would be helpful.

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.

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: Postmarketing 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. 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.

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 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].

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: 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” [5, P. 55]. 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. $\text{Pr}\{\text{death} \mid \text{male, age}=70\}$) is what we would check the probability against. In other words, we do not ask whether $\text{Pr}(\text{death} \mid \text{male, age}=70)$ is accurate when compared against $\text{Pr}(\text{death} \mid \text{male, age}=70, \text{meanbp}=45, \text{patient on downhill course})$.

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 25th and 75th 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.

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.

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 $> 190\text{mg}\%$ to qualify for a study, and the median LDL cholesterol for qualifying subjects at the screening visit was $230\text{mg}\%$. The median LDL cholesterol value at their second visit might be $200\text{mg}\%$, 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 lowers mortality by 60%. 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 ratio*, not all risk ratios are possible, depending on the probability of event for the reference group. For example, the relative risk must be < 2 if the reference group has an event probability of 0.5. In other words, relative risks cannot be constant over large intervals of probabilities. Also unlike the odds ratio, the risk ratio depends on which of two outcomes is labeled as the “event”; a mortality ratio does not equal the survival ratio. Relative risk can be a confusing term as it sometimes is taken to mean a hazard ratio (‘relative risk’ over the whole survival experience) and sometimes an odds ratio (‘relative risk’ for a discrete event)[2].

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 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.

semi-parametric: ‘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 non-parametric estimate of baseline hazard[2].

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 95% of the distribution of the measurement. Standard deviation is the square root of the *variance*.

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 $\rightarrow \infty$.

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 25th and 50th percentiles equals the distance between the 50th and 75th percentiles.

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].

type I error: False positive rate – the probability of rejecting H_0 (i.e., declaring “statistical significance”) when the null hypothesis is in fact true. The type I error is often called α .

type II error: Failing to detect an effect that is real, i.e., the false negative rate. The type II error is referred to as β , which is one minus the power of the test. In other words, the power of the test is $1 - \beta$.

variance: A measure of the spread or variability of a distribution, equalling 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*.

References

- [1] S. R. Brazer, F. S. Pancotto, T. T. Long III, F. E. Harrell, K. L. Lee, M. P. Tyor, and D. B. Pryor. Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia. *Journal of Clinical Epidemiology*, 44:1263–1270, 1991.
- [2] K. Bull and D. Spiegelhalter. Survival analysis in observational studies. *Statistics in Medicine*, 16:1041–1074, 1997.
- [3] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972.
- [4] M. A. Hlatky, D. B. Pryor, F. E. Harrell, R. M. Califf, D. B. Mark, and R. A. Rosati. Factors affecting the sensitivity and specificity of the exercise electrocardiography. Multivariable analysis. *American Journal of Medicine*, 77:64–71, 1984.
- [5] S. Kotz and N. L. Johnson, editors. *Encyclopedia of Statistical Sciences*, volume 9. Wiley, New York, 1988.
- [6] A. Spanos, F. E. Harrell, and D. T. Durack. Differential diagnosis of acute meningitis: An analysis of the predictive value of initial observations. *Journal of the American Medical Association*, 262:2700–2707, 1989.
- [7] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–178, 1967.

Acknowledgements: Richard Goldstein (Qualitas Inc.) provided valuable additions and clarifications to the glossary and additional medical statistics citations. As noted in the glossary, several definitions came from [2].