SCIENCE AND PRACTICE OF PHARMACOTHERAPY I

BIOSTATISTICS

Scott A. Strassels, Pharm.D., Ph.D., BCPS

Reviewed by Madeline McCarren, Ph.D., M.P.H., and Todd P. Semla, Pharm.D., M.S., FCCP, BCPS, AGSF

Learning Objectives

1. Describe the strengths and limitations of measures of central tendency and measures of variability.
2. Classify common statistical tests and tools.
3. Interpret results of confidence intervals (CIs).
4. Interpret commonly used statistical tests.
5. Distinguish between p-values and CIs as measures of statistical significance.
6. Evaluate commonly used statistical and epidemiologic measures.

Introduction

Health care changes at a blistering pace. Interventions that are state-of-the art today may be discredited tomorrow. As a result, clinicians must be able to interpret the flood of new information available every day, and to communicate it across disciplines in meaningful ways.

These goals in health care practice and research are accomplished by the use of statistics. These analytic tools are not perfect, but without them, attempts to answer clinical and research questions would quickly become nothing more than decision-making by opinion and consensus. Clinicians would become inundated with data, and some questions would be time-consuming or in extreme cases, impossible to answer. Despite the importance of understanding how to use statistics and ongoing changes in pharmacy education to increase the emphasis on understanding research tools in the clinical setting, many clinicians remain uncomfortable with interpreting statistics found in the biomedical literature. This chapter helps reduce this anxiety by reviewing a few important statistical principles, and discussing some biostatistics that pharmacists are likely to encounter in the biomedical literature.

Descriptive Statistics

Descriptive estimates provide a general view of data. For example, the mean, median, and mode each provide different information about middle data values. They are referred to as measures of central tendency. Although the purpose of this chapter is not to repeat the basics of how these statistics are calculated, the mean is sensitive to outliers. Comparing the median, or the central value for a variable, to the mean can tell the reader about the distribution of the data. If the mean is different than the median, the data are likely to be skewed. Most of the time we talk about the arithmetic mean, but there are variations on this theme. For example, in regression analyses, the dependent variable is sometimes transformed to its logarithm to improve how well the model satisfies underlying assumptions. But this approach also makes interpretation of the transformed variable more difficult, because the log values are not very meaningful to the reader.

Measures of variability, including the interquartile range (IQR), range, standard deviation (SD), standard error of the mean (SEM), and variance are also useful to determine whether it is reasonable to infer from the sample to the population. The range is the interval between the minimum and maximum value, and the IQR is the interval between the 75th percentile and 25th percentile values. The IQR describes the middle 50% of the data in the sample. The variance, SD, and SEM are also commonly encountered measures of spread in data. The variance is the measure of variation in a dataset for one variable, and the SD is the square root of the variance. The mean plus or minus two SDs will include the central 95% of values. The SEM is an estimate of certainty that a calculated sample mean represents the true mean of the population. As such, the SEM is an inferential statistic, and is not interchangeable with the SD, even though it is sometimes used in this way. In addition, because the SEM is calculated by dividing the SD by the square root of the number of individuals in a dataset $\frac{SD}{\sqrt{n}}$

Abbreviations in This Chapter

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
CI	Confidence interval
DF	Degrees of freedom
H_0	Null hypothesis
H_1 or H_A	Alternative hypothesis
IQR	Interquartile range
SD	Standard deviation
SEM	Standard error of the mean

it will always be smaller than the SD, and can be erroneously interpreted as indicating that a set of observations is less variable than it really is.

Data Distributions

Variables may be discrete or continuous. Distributions generally describe how variables are spread out. The normal distribution is well known to clinicians, but many types of data encountered in clinical practice are not normally distributed. This section introduces some commonly encountered discrete and continuous distributions.

Discrete Distributions

The binomial and Poisson distributions are two discrete probability distributions. The binomial distribution is used when considering a sample of some number of independent trials that have only two possible outcomes, such as some indeterminate measure of success or failure. Imagine that a coin is flipped 1000 times. Each time the coin is flipped, there is some probability of success or failure. The Poisson distribution is used most commonly when rare events are being considered. An example is the number of serious adverse drug reactions due to Drug A over some time period. Approximations to the binomial or Poisson distributions are used when use of the distribution under question would be onerous and when specific conditions exist. For example, the binomial distribution is useful when considering some number of independent trials. But if the number of trials performed is large and the probability of success is either very high or very low, the distribution will be skewed. If, on the other hand, the number of trials undertaken is at least moderately large, and the probability of success is not too extreme, then the distribution will be symmetric and will be approximated by the normal distribution. Similarly, when the expected number of events over a time interval of interest is large, the Poisson distribution is unwieldy, and a similar normal approximation may apply.

Continuous Distributions

The most well-known example of a continuous distribution is the normal distribution. Most clinicians are familiar with at least some of the features of this distribution, which is also referred to as a Gaussian distribution or a bell-shaped curve. It is symmetric around the mean, and when the mean is 0 and has a variance (and, thus, an SD, as well) of 1, it is referred to as a standard normal distribution. The area under the standard normal curve from one SD below the mean to one SD above the mean includes about 68% of the distribution while 95% of the distribution lies in the area from 2 SD below the mean to 2 SD above, and 99% of the area is from 2.5 SD below the mean to 2.5 SD above the mean.

Data from many samples are not normally distributed, but when the underlying distribution is itself normally distributed, the central-limit theorem applies. The central-limit theorem states that in considering a random sample, when the number of observations is large, the distribution of the mean is approximately normally distributed, even if the distribution of the observations in the sample being studied is not normally distributed. Often, 30 observations are used as a rule-of-thumb minimum cutoff for defining what is meant by "large." Normally distributed data are rare in real life, but if there is reason to believe that the central-limit theorem applies, inferential statistics can be used. In practice, this is important because the means of most physiologic measurements, such as blood pressure, are considered to be normally distributed.

Degrees of Freedom

Many statistical tests depend on a factor called the degrees of freedom (DF). This term refers to the number of independent comparisons that can be made when calculating a given statistic. For example, in a dataset that contains 50 persons, 49 of those individuals can be compared with the first person. As a result, there are 49 DF for this calculation.

Hypothesis Testing

The interpretation of statistical results in the biomedical literature while using the log transformation is useful, it also makes interpreting the transformed variable more difficult because log values are often not meaningful to reader. In addition, transforming the log of the variable back (using the antilog) results in the geometric mean, rather than the arithmetic mean. This gives rise to the null hypothesis (H_0), and the alternative hypothesis (H_A or H_1). If there is a difference between the study groups, the H_0 is rejected. If no difference between groups is found, the result is failure to reject the H_0. The latter form of interpreting H_0 is used because every study is limited, and it is impossible to know if there are truly no differences in the factor of interest between groups. The H_0 and H_A or H_1 are typically expressed by saying that the groups are equal (or that there is no difference between them), and the groups are not equal (or that there are some differences between groups), respectively, though more specific ways of expressing those ideas are also used. For example, the H_A is often stated as a general inequality (i.e., the mean of group 1 ≠ the mean of group 2), because the direction of the inequality is uncertain. In the uncommon case where there is certainty about the direction of differences between groups, the researcher may choose to express the H_A in that direction (e.g., the mean of group 1 greater than the mean of group 2).

The potential for making errors when evaluating the H_0 and H_A is well known. A Type I error happens when the H_0 is rejected when it is true, thus finding differences where none exist. A Type II error is failure to reject the H_0 although a difference is present. The probability of a Type I error, or the significance level of a test, is denoted by alpha (α), and the probability of a Type II error is referred to as beta (β). Beta is a component of the power of a test, which contributes to determining the sample size needed to detect a difference between groups, if there is such a difference. Power is calculated as $1 - \beta$. The importance of the estimated power of a study is highlighted mainly when the observed difference between groups is not statistically significant. In such trials, sometimes called negative studies, sufficient power indicates that the groups are similar, but in underpowered analyses, no conclusion can be drawn because there is not enough information to answer the question.

Biomedical researchers often set $\alpha = 0.05$ and $\beta = 0.20$ by convention, though there is no reason that different values of each cannot be used. For example, because making an error can have important consequences, there may be compelling reasons to avoid making a Type I error, even if that means making a Type II error, or vice-versa. For example, imagine that researchers develop a new screening test to detect colorectal cancer. In a clinical trial, the investigators find that the new test is more sensitive and specific than the existing test, despite being more invasive and expensive. If the investigators have made a Type I error, and there is really no difference between the tests, the consequences include exposing the person to unneeded risks of adverse effects due to the invasiveness of the test. Alternately, if a Type II error is found, and the tests are found to be equivalent when they are not, patients may not be able to take advantage of the improved technology.

The power of a test is affected by several factors: the significance level (α), the difference between groups, the SD of an observation, and the sample size. Power increases as the difference between groups and as the sample size increase. Power decreases as α falls and as the SD of an observation increases.

Another important consideration is the sample size needed to conduct the study. This consideration must take into account the significance level, desired power, and the difference between means under the H_0 and alternative hypothesis. The number of study participants needed to detect some difference between groups increases with the variance in the data, as the chosen α decreases, and as power increases (or β decreases). For example, suppose that researchers want to conduct a study in which they are willing to accept a 1% probability that observed differences between groups are due solely to chance. If the investigators are willing to accept a 5% probability that differences are due to chance instead of just 1%, the needed sample size will decrease. The implications of these types of decisions are important because a larger sample increases the time and expense needed to conduct a clinical study. For example, an odds ratio of 2.0 would indicate that the independent variable is associated with a 100% higher risk of the outcome, while an odds ratio of 75% would indicate that the independent variable is associated with a 25% decrease in the likelihood of the outcome. In addition, when a confidence interval (CI) for an odds ratio includes 1.0, the effect of the independent variable is interpreted as not being statistically different from 0, that is, there is no difference in the outcome between persons who received the intervention and individuals who did not.

One-Tailed and Two-Tailed Tests

A one-tailed statistical test refers to one in which the parameter being studied (such as the mean of some variable) under H_A is allowed to be either less than or greater than the values under the H_0, but not both. In contrast, a two-tailed test is one where values of the parameter of interest under H_A are allowed to be less than or greater than the values under the H_0.

Deciding which approach to use requires some thought before analyzing the data. It is acceptable to test for differences in either direction, in which case the H_0 will be rejected if the value of the test statistic is above or below the critical point. This is tantamount to saying that researchers will conclude that an observed difference is statistically significant if the mean value for group 1 is above or below that for group 2. In general, however, use of a one-tailed test is analogous to stating that one has no interest in changes in the other direction. In addition, justifying the decision to use a one-tailed approach often requires information that is not available. As a result, this discussion focuses on the two-tailed approach.

P-Values and Confidence Intervals

A p-value is the probability of obtaining a result at least as extreme as the one observed, given that the H_0 is true. This definition is problematic, however. P-values are conditional on H_0, but whether the H_0 is actually true is unknown. Furthermore, p-values are calculated using models that correspond to the type of data, and most models assume that observations are independent. Yet, depending on the type of study, this assumption may not be true. As a result, a p-value is generally not a meaningful estimate of probability, but more an indication of consistency between the H_0 and the data. The result is that a large p-value suggests that the data are consistent with the H_0 and a small p-value suggests that the data are not consistent with the H_0. Neither, however, tells us whether the H_0 is true.

There are many ways to interpret p-values. Assuming that the significance level for a study was 0.05, one approach is to view p-values between 0.01 and 0.05 as being significant, those between 0.001 and 0.01 as being highly significant, those less than 0.001 as very highly significant, and p-values higher than 0.05 as not statistically significant. The p-values higher than 0.05 but lower than 0.10 are referred to as trending toward significance, but trends represent value judgments. In other words, a trend toward significance can also easily be interpreted as a trend away from significance. Because the significance level is arbitrary, statistical significance tells only that the p-value is less than the cutoff value chosen. Furthermore, significance testing, with its dichotomous outcome, provides no information about the size of the effect. Similarly, a statistically nonsignificant p-value does not indicate that there is no association in the data.

Along with p-values, CIs are used to account for the random error. As with significance levels, the desired level of confidence is chosen arbitrarily. A 95% CI, often used by convention in biomedical research, indicates that, if the experiment were to be repeated many times, at least 95% of the resulting CIs constructed would include the true (unobservable) population mean. But this interpretation assumes that the statistical model being used is correct and that bias is negligible. These assumptions may not hold in all types of medical research. As a result, it is important to consider the CI as only a general and minimum estimate of uncertainty in the estimate. In addition, the higher the confidence level, the wider the interval will be, because we need to include a wider range of values to be more certain that the true value of the variable of interest is included. There are several ways to calculate CIs. For example, we are interested in constructing a 95% CI for a mean. One way of estimating the limits of a 95% CI by multiplying 1.96 by the SEM, then adding and subtracting that quantity to the mean, respectively. The factor, 1.96, comes from the probability that a value will fall within about two SDs to either side of the mean under a standard normal distribution.

The connection between p-values and confidence levels is obvious, but represents an important pitfall. Although CIs can be viewed as analogous to hypothesis tests of significance, doing so is pointless and ignores the information that CIs offer above what hypothesis testing provides. For example, the CI estimates the effect size as well as the variability in the estimate, whereas a p-value provides only an estimate of the consistency between the data and the hypothesis.

Incidence Rates, Prevalence Rates, Odds, and Odds Ratios

Incidence rates, prevalence rates, odds, and odds ratios are commonly encountered in epidemiologic research. The incidence rate is an estimate of the instantaneous rate of developing disease. It is calculated by dividing the number of persons who develop disease in a population over a given time by the summed amount of time that people in the study were at risk of developing the disease,

$$\frac{\textit{number of subjects developing disease}}{\textit{total time at risk of disease for subjects followed}}$$

If 100 people are observed for a year, and are at risk of disease for that time, the denominator of the incidence rate would be 100 person-years.

Although incidence reflects the rate at which people develop disease, prevalence estimates the number of people who have a condition at a particular point in time. Like incidence, prevalence is an important epidemiologic measure. In addition, by dividing the proportion of persons with a disease by the proportion of persons without the disease

$$\frac{P}{1-P}$$

it is possible to estimate the prevalence odds. This relation holds for other measures, as well. In a case-control study, the odds ratio is the ratio of cases to controls among the exposed persons, divided by the ratio of cases to controls among the unexposed persons. An odds ratio of 1.0 indicates that the independent variable is not associated with outcome. Values greater than 1.0 indicate that the independent variable is associated with higher risk of the outcome, while values lower than 1.0 indicate that exposure to the independent variable is associated with lower risk of outcome. In case-control studies (discussed in detail in the Pharmacoepidemiology chapter), control subjects can be selected using different methods. One such method is called incidence-density, or risk-set, sampling in which control subjects are selected from among all persons at risk of the event at the time the event occured. When this method is used, the control subjects are chosen from among individuals at risk of experiencing the outcome when the event of interest occurs for a person. Under this condition, the odds ratio provides a valid estimate of the incidence rate ratio.

Contingency Tables

Contingency tables are useful to estimate the association between variables. When constructed to examine the relation between two variables, the table includes two rows and two columns, though any number of rows or columns can be used to accommodate variables with more than two categories (Figure 1-1). Within this 2 X 2 table, the variables must be categorized so that each term has only two possible results. One variable is arbitrarily assigned to the rows of the table, and the other to the columns. Each cell of the table contains the number of individuals who meet the criteria for both variables, such as exposed, with outcome, or not exposed, no outcome. By convention, row and column totals are calculated and written in the right and bottom margins, respectively. The grand total, or sum of all individuals in the table, is also written in the lower right-hand corner of the table.

Before performing statistical tests using a contingency table of any size, the expected values for each of the cells are calculated, which represent expected counts if the H_0 is true. This information permits comparisons of expected and observed cell totals, which provides an opportunity to visually evaluate how close the two types of data are.

The expected value for each cell is the product of the row and column totals, divided by the grand total. In addition, two other measures can be estimated from observed and expected values. First, by cross-multiplying and dividing the observed estimates

$$\frac{AD}{BC}$$

we can estimate the odds ratio. In a case-control study where risk-set sampling has been used, the odds ratio is an estimate of the incidence rate ratio Second, the standardized mortality (or morbidity) ratio can be estimated by dividing the observed number of events by the expected number of events, and multiplying that number by 100%. Values of this ratio that are less than, equal to, or greater than 100% indicate that the risk of the outcome in the study population is reduced, the same as, or exceeds that of the general population, respectively.

Common Statistical Tests

There are many statistical tests used to evaluate study data. To further complicate matters, multiple tests may be appropriate for a given situation. It is critically important to be familiar with the assumptions underlying a specific test, the number of groups being compared, whether the samples are independent or paired, and whether the data are nominal, ordinal, or continuous. Statistical tests are

Figure 1-1. 2 X 2 Contingency Table

	Outcome yes	Outcome no	
Exposure yes	A	B	Row 1 total (R_1)
Exposure no	C	D	Row 2 total (R_2)
	Column 1 total (C_1)	Column 2 total (C_2)	Grand Total (GT)

The odds ratio is estimated by calculating $\frac{AD}{BC}$.

The expected value for a particular cell is calculated as the product of the row total and the column total, divided by the grand total. Using the table above, the expected value for cell A is. $\frac{R_1C_1}{GT}$

Expected values for the other cells are calculated in the same way.

sometimes described as being parametric or nonparametric. These categories refer to whether the data follow a known distribution or not. Continuous data are generally analyzed using parametric tests, such as the t-test, whereas categorical data are often analyzed using nonparametric tests, such as the chi-square test (χ^2).

The z-test is used to make an inference about the mean of the sampling when the underlying distribution is normal or the central-limit theorem applies, and the variance is known. Yet, the variance is often not known or the analyst may have reason to believe that the sample and population variances are different. In such a case, a t-test, rather than the z-test, is used.

Proportions can be evaluated using the binomial distribution if there are only two possible outcomes, such as success and failure, or present and absent for simplicity. For example, a study might compare the percentage of persons with breast cancer in a sample compared with the general population. As mentioned earlier, when the number of trials is large, the binomial distribution is hard to use. If the number of trials is moderately large and the probability of success in each trial is not too extreme in either direction, the central-limit theorem applies, and a normal approximation to the binomial distribution can be used.

The Poisson test is used when considering uncommon conditions, because the expected number of events per unit of time follows a Poisson distribution. When using the binomial distribution, the focus is on a finite number of trials, in which the number of events is limited to the number of trials. Under the Poisson distribution, however, the potential number of trials is infinite and, as a result, the number of events can also be unlimited. As with the binomial distribution, when the expected number of events per unit time is large, the Poisson distribution becomes difficult to use. When the number of expected events per unit time is at least 10, the normal approximation to the Poisson distribution is used.

Paired tests, such as the paired t-test, apply when estimates from one sample are compared with estimates in a matched sample. An example of related data is a pre-post design in which study participants act as their own experimental control. This particular design is often used in clinical drug trials.

Some tests require that the analyst assume the variances between samples are equivalent. To test this assumption, we can assess the ratio of sample variances using the F test. This test is used because the ratio of sample variances follows an F distribution with n_1-1 and n_2-1 DF. When the F test is used for this purpose, it is sensitive to departures from normality.

Analysis of variance (ANOVA) methods permit comparison of more than two groups. For example, to examine the effect of smoking on the severity of pulmonary disease, rather than examining only the two categories of either no exposure to tobacco smoke or any exposure, the study could assess the relation between degree of pulmonary disease and never smokers, passive smokers, former smokers, current light smokers, current moderate smokers, and current heavy smokers. When the effect of one variable on the outcome is analyzed, a one-way ANOVA is used.

When ANOVA indicates a difference exists among the several groups, further analysis must be done to determine which groups are different from one another. Although t-tests are used to compare pairs of groups, making many comparisons increases the chance of finding a statistically significant difference between groups. To account for this issue, multiple-comparison procedures are used to ensure that the chance of finding significant differences between all possible groups is held constant. There are many multiple-comparison procedures, such as the Bonferroni adjustment, Scheffé test, and the Honest Significant Difference methods.

In the Bonferroni adjustment, the initial significance level is divided by the possible number of independent two-group comparisons. For example, if there are 10 groups, there are 45 possible two-group combinations. If α is set *a priori* at 0.05, 0.05 divided by 45 gives an adjusted α of 0.0011. To reject or fail to reject the H_0 for each comparison would be based on the adjusted α. Note that some of the comparisons may not be independent, in which case, the Bonferroni adjustment will be conservative.

The one-way ANOVA is used to estimate the effect of one factor on the dependant variable. In this model, we are interested in estimating the mean of all groups considered together, the difference between the mean of a specific group and the overall mean, and the random error between the overall mean plus the group mean and a single observation. This is also called a one-way ANOVA fixed-effects model because the groups being compared have been fixed by the study design. An alternative to the fixed-effects model approach is the random-effects model. In this variation, an assessment is made on overall differences between groups and the general breakdown of total variation into between-groups and within-group components. Like the fixed-effects model, a random-effects one-way ANOVA model is assessed using the F test. In addition to the one-way model are two-way and multiple ANOVA approaches. The two-way ANOVA is used when the effects of one variable are analyzed while controlling for the effects of the other variable. This model also allows us to examine whether the effects of a variable on the outcome differ by the level of a second variable. This type of assessment is done by including an interaction term—the product of the variables of interest—in the model, and interpreting the results of the statistical test for the two variables. For example, the effects of age and sex on some outcome can be assessed using a two-way ANOVA. To do this, the effect of age and sex (also called the main effects), and a third (interaction) term, age X sex, on the outcome are modeled. By generalizing this approach in an n- (or multi-) way ANOVA, it follows that the effect of higher order interaction terms on an outcome can also be estimated, such as a three-term product (e.g., age X sex X race).

Analysis of covariance (ANCOVA) is another variation on the ANOVA theme. Like multiway ANOVA, an ANCOVA models the effects of more than one independent variable on the outcome. However, an ANOVA model compares the effect of categorical groups on the mean outcome, whereas ANCOVA provides the flexibility to estimate and control for the effect of continuous independent variables, or covariates, on the outcome. Similarly, it is possible to model more than one outcome in an ANOVA or ANCOVA setting with the multiple ANOVA or multiple ANCOVA.

Nonparametric Tests

When data are distributed normally or normal approximations apply, techniques like those discussed above are commonly used. When assumptions of normality cannot be made, and the data do not follow a known parametric distribution, or are categorical, nonparametric methods are used to test hypotheses. A few nonparametric tests are the Wilcoxon signed-rank and rank-sum tests, the χ^2 test, and the Kruskal-Wallis test, among many others.

The Wilcoxon signed-rank test is analogous to the paired t-test. The Wilcoxon signed-rank test considers the difference between the observation and the H_0, taking into account the direction and relative size of the observed differences. Instead of the precise magnitude of the difference, the relative magnitude is considered, with greater weight, or higher rank, given to the larger differences. A similar situation exists when data are collected from two independent samples. If the data measures are continuous, the t-test for independent samples can be used. When the outcome data are ordinal, however, a nonparametric approach, such as the Wilcoxon rank-sum test, should be used.

Chi-square tests (and variations of it) are commonly used to test hypotheses when the data are categorical, but they can also be used for other purposes, such as testing for the variance of a normal distribution. When using the χ^2 test to evaluate hypotheses pertaining to variances, deviations from normality are important. If the underlying distribution is not normal, the results of the tests are likely to be invalid. In addition, all expected values must be at least 5 in order to use the χ^2 test, otherwise, the Fisher's exact test is used. McNemar's test is a nonparametric test used to evaluate categorical data from matched pairs.

Just as the ANOVA is a generalization of the t-test, the Kruskal-Wallis test is a generalization of the Wilcoxon rank-sum test and serves as a nonparametric approach to ANOVA. The Kruskal-Wallis test allows hypothesis testing when we have more than two samples and ordinal data.

Two methods commonly seen in survival, or time-to-event, analyses are the log-rank test and life table analysis. The log-rank test allows comparison of the events of two or more groups, with the H_0 that there are no differences between groups at any point along the times curves. This test has g-1 DF, where g is equal to the number of groups being compared. If the analyst believes that the early part of the survival curve should be more heavily weighted, the Peto test may be used instead of the log-rank test. For example, in considering the effects of two drugs on survival of persons with cancer, there may be evidence that early diagnosis and treatment with one of the drugs is more important than treatment at some later point. In this case, the Peto test would be appropriate, or at least the results of the two tests should be compared.

Life tables provide an estimate of the rate of an event of interest, by comparing the occurrence in a group between adjacent, small time periods. One of the advantages of using life tables is that they can be used to summarize large amounts of data without sacrificing statistical detail. Imagine that a group of individuals has been observed over some time period, which is subdivided into mutually exclusive, contiguous intervals. Estimating the percentage of individuals in a group who are expected to survive (or not have the event of interest) to the end of one of the time intervals is done by multiplying the probabilities of surviving to the end of each of the previous time intervals. If the goal is to estimate the proportion of individuals in a group expected to survive to the end of the fourth time interval, the estimate is calculated by multiplying the proportion of the group that survives to the end of the first interval by the proportion who survive to the end of the second interval by the proportion who survive to the end of the third interval.

Correlation

Correlation methods are used to show the general linear relation betweenship variables. Values of the correlation coefficient, r, vary from positive one to negative one (+1 to -1), where the respective extremes indicate perfect agreement and disagreement. Positive correlation values indicate that as one variable increases, so does the other. Conversely, a negative correlation indicates that as one variable increases, the other decreases. If the correlation coefficient is 0, the variables are not related, though lack of an observed association may also reflect limits in the data collected. For example, if inclusion criteria are particularly restrictive, an artificial association or lack of one may be observed. It is critically important to recognize that variables may be related without a causal relationship existing, that is, correlation is not causation. Similarly, simply saying that variables are correlated is not informative. It is preferable to indicate how strongly the linear relation is and in what direction.

There are several types of correlation measures, including Kendall's rank-correlation, Pearson's product-moment correlation, and Spearman's rank-order correlation. Kendall's correlation measures the relationship between ordinal variables, Pearson's correlation assesses the association between approximately normally distributed continuous variables, and the Spearman method is used when at least one of the variables is not normally distributed.

Other types of correlation coefficients estimate the degree of agreement within or between raters. These measures are referred to as interclass or intraclass correlation. An example of this type of measure is the kappa (κ) statistic, which is used to describe the degree of agreement by several observers of the same subject. Specifically, if we are interested in how reproducibly a variable is measured by different surveys or different tools or observers, κ is used to compare the observed and expected probabilities of agreement between the different measurements. In general, κ is estimated using a one-tailed test, because negative values typically do not provide useful information.

When the correlation coefficient is squared (r^2), the result is the coefficient of determination. This value represents the percentage of the variance in the dependent variable that is explained by the independent variable. Note that in the context of a regression model, the r^2 increases with the number of covariates in the model, but building such a model is not a headlong pursuit of highest r^2. To help construct a model that maximizes the variance in the dependent variable, but that is also relatively parsimonious, the adjusted r^2 is used. This measure includes a penalty for including unnecessary independent variables. For example, we expect r^2 to rise with each independent term added to the model, but the adjusted r^2 may increase or decrease depending on whether and how much a specific term contributes to explaining the total model variation.

Regression

Regression methods are used to estimate the relation between variables. Simple regression describes the relation between a single independent variable and the dependent variable; multiple regression is used when there are more than one independent variable. These methods provide explanation and prediction of expected relations within the range of the data. Extrapolation of model results outside of the observed range of data is not recommended.

A variety of regression methods exist, including linear, logistic, survival, and Poisson models. Linear regression methods are used when there is assumed to be a straight-line relation between the dependent and independent variables. When the dependent variable is binomial, logistic regression is used. Survival data are analyzed using proportional hazards regression (and related) methods. Major assumptions underlie each of these models, though there are methods that provide some flexibility when assumptions are not met. For example, the logistic model typically requires that the outcome have only two possibilities, but the ordered logistic model can be used when the dependent variable has more than two categories. Similarly, the proportional hazards assumption is an important part of survival analysis, but when this assumption does not hold, allowing the effect of variables to vary over time (time-varying covariates) may be useful.

Linear Regression

Like all statistical methods, appropriate use of the regression models depends on whether certain assumptions are met. The assumptions underlying the linear regression model are shown in Table 1-1.

The general form of the linear regression line is:

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

where $\mu\{Y|X\}$ refers to the mean of the dependent variable, given the values of the independent variable(s), X_1 - X_n, β_0 is the intercept of the line, measured in the same units as the dependent variable, and β_1-β_n are the slopes of the lines. Another name for β is coefficient. The slope of the line (or coefficient) is the change in the value of the dependent variable resulting from a 1-unit change in the independent variable. For example, in a simple regression model, if the coefficient is 1.0, the value of the dependent variable increases by one for each 1-unit increase in the independent variable**.** Of course, this interpretation is dependent on the scale of the independent variable. If, for example, the independent variable has been transformed to its logarithm, the coefficient would still be the change in the mean of dependant variable for each 1-unit change in independant variable, but the unit change would be interpreted as a change on the particular log scale, such as doubling or a 10-fold increase. In multiple linear regression, any given coefficient represents the change in the mean value of dependant variable for each 1-unit change in independant variable, assuming the other variables are held constant, or adjusting for the other variables.

Dummy, or indicator, variables can be used when an independent variable has more than one level. For example, to examine the effect of sex on blood pressure, one way of coding sex is as a single variable that has two different values, one for males and one for females. Another way to approach this issue is to split sex into two separate variables, which have a value of 0 if the characteristic is not present or a value of 1 if it is present. As a result, in a simple linear regression model where dummy variables are used to model the relation of sex on the outcome, the model would look like this:

$$\mu\{Y|X\} = \beta_0 + \beta_1\text{Male} + \beta_2\text{ Female}.$$

In this model, for Male study participants, Female will equal 0, and β_2Female will drop out of the equation. Similarly, for female study participants, Male will equal 0, and β_1Male will drop out of the equation.

The slope and intercept (also called the constant) in a regression model can be evaluated by testing the hypothesis that these terms are equal to 0. If we accept the H_0 that the slope is 0, this observation suggests that the independent variable does not contribute to explaining or predicting the dependent variable, or that the relation between the parameters is not linear (in a linear model). When the H_0 is rejected, we generally conclude the opposite, although other models with linear components (curvilinear graphs, for example) may also fit the data well. Under the test for the intercept, if we fail to reject the H_0 that the intercept is 0, we can remove the intercept from the model. This step is analogous to forcing the line through the origin, but observations for some variables are generally not available when some independent variables, like age, equal 0. As a result, we are generally not interested in the statistical significance of the intercept.

Logistic Regression

In logistic regression, the goal is to estimate the relation between study variables and an outcome that can be categorized into two groups, such as disease or no disease, or severe pain or no severe pain. There are generalized logistic methods, such as ordered logistic regression, that allow modeling an outcome with more than two categories; however, this discussion is focused on the dichotomous-outcome logistic model.

As with other regression models, the β coefficients in a logistic regression model indicate how much the dependent variable changes with a 1-unit change in the independent variable. In multiple logistic regression, the coefficient represents the relationship between each independent variable and the dependent variable when all other independent variables are held constant. The result of the logistic regression model is the odds ratio, regardless of the type of underlying study. The odds ratio quantifies the odds of the outcome in those exposed to the independent variable divided by the odds of the outcome (the dependent variable) in unexposed persons. For example, an odds ratio of 2.0 would indicate that the independent variable is associated with a 100% higher risk of the outcome, while an odds ratio of 75% would indicate that the independent variable is associated with a 25% decrease in the likelihood of the outcome. Additionally, when a confidence interval for an odds ratio includes 1.0, the effect of the independent variable is interpreted as not being statistically different from 0, that is, there is no difference in the outcome between persons who received the intervention and individuals who did not.

Note that the interpretation of the β coefficients may differ based on the type of variable and how it is coded. When the independent variable of interest has only two possible outcomes, coded as 0 and 1, for example, the odds ratio provides an estimate of the likelihood of the outcome between individuals with the variable equal to 1 and those with the variable equal to 0. When the independent variable has more than two possible categories, the odds ratio is derived from comparing persons with the variable equal to one level to persons in the reference, or baseline, group. Similarly, when the independent variable is continuous, the odds ratio represents the change in the likelihood of the outcome associated with a 1-unit increase in the value of the independent variable. It is worth noting that the scale of the independent variable is very important here. For example, a 1 mm of mercury increase in blood pressure may not be clinically significant. On the other hand, if the variable is measured on a small scale, 1 unit may be far too large.

Survival Analysis

These methods apply to questions where the outcome is time until an event takes place. Events of potential interest may include progression of disease, recurrence of a condition, and, of course, death, among many others. One advantage of survival analysis techniques is that these methods account for incomplete observation, or censoring. Censoring refers to situations in which the analyst knows something about an individual's experience, but the exact time until the outcome is unknown. For example, an individual may withdraw from a study, be lost to follow-up, or may not have the event being studied by the time the study ends. In addition, we may consider survival data as being right- or left-censored. Right censoring occurs when the observation period ends and some persons have not yet had the event of interest. Left censoring happens when the event of interest happens at some time before the start of observation. Observations may also be interval-censored, which occurs when the event happens at some unknown time between scheduled observation points. For example, if a subject's disease progresses some time between the scheduled follow-up appointments at 3 and 12 months, interval censoring has occurred. Another mechanism of incomplete observation is called truncation, but censoring is inherently different from truncation. Specifically, censoring occurs at the individual level while truncation is a design issue. Thus, if observation does not begin until some specified time after the start of exposure, the dataset is left-truncated. If all the people in a study experience the event before the start of the study, the dataset is right-truncated.

Survival data are often analyzed using Cox proportional hazards models. The major assumption underlying this model is the proportional hazards assumption. Although hazard generally refers to the chance that some event will occur, this assumption states that the hazard—defined as the instantaneous potential per unit time for a person to experience the event of interest, given that the person survives until that time—is proportional to that for any other individual, and is independent of time. The proportional hazards assumption can be tested statistically and by examining curves on a survival graph. For example, if survival curves cross, the proportional hazards assumption does not hold.

Methods for addressing situations when the proportional hazards assumption does not hold include examining the model more closely to determine which covariates contribute nonproportionally, stratifying on the exposure variable, or using an extended model in which some variables are modeled to permit them to vary over time.

Interpreting Model Results

Once a survival analysis model is constructed, it is important to consider how to interpret the results. When the independent variable is measured on a nominal scale with two possible categories, the β coefficient for the variable represents the change in the log hazard for a 1-unit change in the covariate. Categorical variables that have more than two levels are interpreted as in the logistic model. That is, the β coefficient represents the change in log hazard for the group compared with the reference or baseline category. Similarly, for covariates measured on a continuous scale, the β coefficient represents the change in the log hazard for a 1-unit change in the value of the variable. The key here, though, is to make sure that the scale is appropriate. When there is more than one independent covariate in the model, the effect of the covariate is interpreted as that for a 1-unit change in the variable, assuming all other terms are held constant.

Poisson Regression

Poisson regression models the number of events, with the following assumptions that the incidence rate reflects how often events occur, that the incidence rate multiplied by the exposure provides an estimate of the expected number of events, that during very small exposure periods, the probability of more than one event taking place is small, and that nonoverlapping exposures are mutually independent. Coefficients from Poisson models represent the change in the log incidence rate for a 1-unit change in the independent variable.

Missing Data

For many reasons, data are often missing or unusable. Study participants may not understand a question, or refuse to answer. Data collection may rely on different people, and as a result, vary in completeness. Unexpected events may prevent a person from completing a questionnaire or survey. If missing information obscures data and relations that are important for the analysis, bias is introduced, which can result in the use of the data being significantly hampered.

To better understand the effect of missing data, it is necessary to consider the patterns and mechanism of missingness. Patterns of missing data refer to which variables and values are present and which are not. For example, in univariate missingness, data for a single variable are not available. A related pattern of missingness, called unit and item nonresponse, occurs when study participants do not complete a survey. Longitudinal studies can suffer from attrition or other loss-to-followup. Drug safety studies can provide an example of this pattern of missingness. When large amounts of data are missing, some variables may not be observed together. This limitation results in an inability to estimate the association between affected variables or even if estimates can be derived, they may be misleading. Last, situations in which variables that were not observed at all may present missing-data problems.

In addition to the patterns described above, the mechanisms that result in missing data are important to consider. The central issue is whether the values in the database are related to missing data. If missingness does not depend on the values of the data, missing or observed, the mechanism is called missing completely at random. Despite its name, this mechanism does not imply that the pattern of missingness is itself random, just that it is independent of the data. The next step on this continuum is called missing at random, and this mechanism applies when the missing data depend on the data values that are observed, and not on those that are unobserved. The third mechanism is called not missing at random, and it applies when the missing data depend on the missing values. When data are missing completely at random, the observed data represent a random sample of all the data. On the other hand, when the missing data are not missing at random, the subsample analyses for the missing variables may be biased. An example is when final clinical measures are missing because of death. Simply ignoring the missingness will bias the findings toward more favorable outcomes. The ANOVA methods are not well-suited to dealing with missing data; missed regression models are preferred. However, good study design and conduct are the first line of defense against missing data. Analytic strategies are a distant second.

Summary Measures of Effect

In addition to literacy in statistics, clinicians must also be comfortable using the vocabulary of epidemiology. A few of the terms and concepts from this field are discussed briefly below.

Absolute and Relative Differences

The terms “absolute” and “relative” are commonly used in the biomedical literature, typically when discussing the rate of some event. For example, in a clinical trial comparing the effects of a new drug with a placebo, the relative risk reduction is estimated by subtracting the percentage of persons in the treatment group who have the outcome from the percentage of persons with that event from the control group, divided by the percentage of persons who have the event in the control group. In contrast, the absolute risk reduction is simply the numerator of the above ratio. That is, the percentage of persons in the control group who have the outcome, less that in the active comparator group. Relative measures are often larger than absolute measures, and as a result, are more commonly reported in the literature. Yet, a large relative reduction may translate into few events, and absolute measures may be more meaningful to consumers and purchasers of health care. Similarly, relative measures are generally viewed as being more relevant for etiologic questions, whereas absolute measures are more applicable for policy questions. This observation makes some sense because it is easy to imagine that employers and payers are likely to be interested in the absolute number of events like injuries, cases of disease, or number of missed days or work.

When the absolute risk reduction is expressed as a decimal (i.e., 1% = 0.01), its inverse is called the number-needed-to-treat. This estimate refers to the number of persons who must receive a treatment for some amount of time to prevent one undesirable outcome or to achieve one good result. For example, in the Oxford league table of analgesics in acute pain, the numbers-needed-to-treat to provide 50% or greater acute pain relief are listed for a variety of drugs. An analogous measure, the number-needed-to-harm, refers to the number of persons who must receive a treatment to cause one death or other serious injury. The number-needed-to-harm is calculated in exactly the same way as the number-needed-to-treat, except that the outcome being considered is some undesirable outcome. Although a small number-needed-to-treat indicates that a drug is highly efficacious, a large number needed to harm is preferable because it indicates greater safety.

Sensitivity and specificity are concepts often used in discussions of diagnostic tests. Sensitivity of a new test indicates the percentage of persons who will have a positive result using the new test and who really have the condition according to the gold standard method. Specificity refers to persons who do not have the condition according to the standard who also have a negative test using the new test. These two concepts are closely linked, and while we would like to have high values for each measure, many times, it is a careful trade-off between them. Receiver-operator characteristic curves are a tool used to show the performance of a new test graphically. The true-positive rate (sensitivity) is plotted along the Y-axis and the false-positive rate (1-specificity) along the X-axis. A 45-degree line is drawn to show where the test is no better than chance, and the area under the curve indicates how the test performs, with higher values indicating better characteristics.

Conclusion: Avoiding Common Pitfalls

Last, it is also important to understand how to recognize and avoid common problems, errors, and barriers to understanding statistics in the biomedical literature. Some of these potential pitfalls are discussed in this chapter; however, the reader is encouraged to seek out discussions devoted to these topics. A few common errors include relying on statistical software to make decisions, inattention to detail when collecting data, assuming that one statistical method fits all questions, dumping poorly organized information into unclear charts and graphs, and failure to consult a statistician before data collection. In addition, while post-hoc analyses are useful for generating hypotheses, it is important to distinguish these data from the results of a priori hypothesis testing. Each problem is the foundation for numerous other issues, and addressing them early will benefit the researcher and the reader.

Annotated Bibliography

1. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text, 2nd ed. New York: Springer, 2005.

   In this self-directed text, the authors present a clear, understandable, and common-sense approach to understanding survival analysis methods. As with the authors' other book on logistic regression listed below (Reference 2), the material is presented in a straightforward manner, and includes learning objectives, examples, summaries after major points, detailed outlines, and practice exercises. For example, in the first chapter, the authors define survival analysis and the terms used commonly in this type of analysis (e.g., time, event, competing risk, and failure), and describe how those common words apply specifically in this arena. The authors then discuss censoring, how to read the symbols and terminology used in survival analysis (e.g., survivor function or hazard function). The authors also provide useful discussions on statistical tests, the Cox proportional hazards model, how the proportional hazards assumption is tested, and how the Cox model can be adapted to include time-dependent covariates.

2. Kleinbaum DG, Klein M, Pryor ER. Logistic Regression: A Self-Learning Text, 2nd ed. New York: Springer, 2005.

   This text is well suited for people who want to learn about logistic regression at their own pace. The authors do not assume that the reader has extensive experience in biostatistics. Each topic is presented using clear language, with learning objectives listed before the chapter, summaries placed after major points (e.g., Why is the logistic model popular?), examples of how logistic regression can be used, detailed outlines of the material, key formulas, and practice exercises. In addition to introducing the logistic model, the authors cover special cases of this model, estimating the odds ratio, use of maximum likelihood techniques, approaches to modeling, and studies using matched data.

3. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Multivariable Methods, 3rd ed. New York: Duxbury Press, 1998.

   This graduate-level textbook provides a clear, understandable, and detailed discussion of simple and multiple linear regression, correlation, and analysis of variance methods. The authors also cover multiple correlation, confounding and interaction, use of dummy (or indicator) variables, and regression diagnostics. One of the most useful parts of this book is the importance the authors place on understanding the assumptions underlying a linear model and their discussion on assessing how well these assumptions are met using analyses of the residuals, outliers, and influential points, and assessing the presence of colinearity. The authors also provide a useful discussion of the correlation coefficient.

4. Hosmer DW Jr, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time to Event Data. New York: John Wiley & Sons, 1999.

   This textbook presents an advanced discussion of statistical theory and methods. The authors assume that the reader is familiar with the concept underlying survival analysis as a set of methods designed to examine time to event data. Their discussion is detailed, and includes issues such as mechanisms of censoring, fitting the model under differing circumstances, assessing the model, and variations on the survival model. The discussion of how fitted survival analysis models are interpreted and how those explanations differ based on the type of independent variable is extremely helpful and straightforwardly presented.

5. Hosmer DW Jr, Lemeshow S. Applied Logistic Regression, 2nd ed. New York: John Wiley & Sons, 2000.

   This advanced text covers the theoretical underpinnings, model building, interpretation of output, and model refinement of the logistic model. It also includes discussion of variations of this type of statistical analysis, such as the ordinal logistic model. In this type of model, the outcome has more than two categories, and these levels are ordered. Unlike the introductory text in Reference 6 or the self-directed text in Reference 2, this text assumes the reader is familiar with biostatistics and logistic regression analysis in general. One of the most useful parts of this book is the in-depth and clearly explained discussion of how to interpret the fitted model, and how that explanation differs by the type of independent variable.

6. Rosner B. Fundamentals of Biostatistics, 6th ed. New York: Duxbury Press, 2005.

   Of the many available introductory biostatistics textbooks, this one excels because it is clear and concise, and provides many examples. The author does not assume the reader has any background in biostatistics, and progresses through the material in a way that builds on the material discussed previously. Topics particularly useful include the discussion of measures of central tendency, discrete and continuous probability distributions, hypothesis testing, analysis of variance, and correlation. The author also provides a useful introduction into regression methods. The text includes several detailed flowcharts that walk the reader through choosing statistical tests and techniques. For example, the first of these figures asks if there is only one variable of interest, if the problem is one-sample, if the data are (or can be assumed to be) normally distributed, and so on, until the correct test is identified.

7. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer 2003;89:232–38.

8. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. Br J Cancer 2003;89:431–6.

9. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part III: multivariate data analysis—choosing a model and assessing its adequacy and fit. Br J Cancer 2003;89:605–11.

10. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. Br J Cancer 2003;89:781–6.

These four references introduce the use of survival analysis using examples from cancer care and research. The authors discuss events of interest in oncology—not simply death or absolute survival; clinicians are also often interested in time from response to treatment to recurrence, also known as disease-free survival time. The introduction also includes a discussion of the mechanisms of censoring, the concepts of survival and hazard, the Kaplan-Meier survival estimate, cumulative hazard, and statistical tests that compare the survival experience between groups. In Reference 8, the authors discuss multivariate survival analysis. They describe the usefulness of multivariate modeling, and the use of Cox proportional hazards model and its interpretation. In Reference 9, the authors expand their discussion of multivariate survival modeling to issues of modeling. The discussion focuses on how covariates are chosen and assessing how well the model fits the data. Of importance, they discuss overall goodness-of-fit tests, as well as evaluating whether the proportional hazards assumption holds. Last, the authors in Reference 10 address questions that often arise in survival (and other types of regression) modeling, including how to categorize variables, use of time-dependent covariates, and addressing missing data.

11. Altman DG, Royston P. The cost of dichotomizing continuous variables. BMJ 2006;332:1080.

To interpret clinical, economic, and human outcomes data, clinicians are often tempted to categorize data that were originally measured on a continuous scale. In this one-page essay, the authors cut to the heart of why dichotomizing continuous data may not be a good idea. The authors argue that dichotomizing continuous variables results in a loss of statistical power to identify a relation between the exposure and the outcome, increases the chance of a false-positive result (a Type I error), may obscure important fluctuations within the groups, and, similarly, makes the relation between outcome and exposure appear linear when it may not actually be. In addition, it is often unclear where to divide the group in question and splitting a variable in regression analyses may result in incomplete adjustment for confounding.

12. Altman DG, Bland M. Standard deviations and standard errors. BMJ 2005;331:903.

This discussion of the major distinctions between the SD and the standard error helps clarify an issue that comes up frequently in the biomedical literature: how to best express the variability in a sample estimate. The authors help eliminate confusion over these measures by emphasizing that the SD is a measure of variability of a population, and that this estimate can be used with any distribution. In contrast, the standard error refers to how repeated sample means differ from each other. In essence, then, the SD is analogous to error within a sample while the standard error refers to variation between samples. The issue is important because the standard error is always smaller than the SD. As a result, if the standard error is used incorrectly, differences between groups look smaller than they actually are.

13. Centre for Health Evidence and JAMA and Archives. Users' Guides to the Medical Literature. Available at *http://pubs.ama-assn.org/misc/usersguides.dtl.* Accessed August 21, 2007.

Published between 1993 and 2000, this series of 25 essays (some divided into multiple parts) provides useful tools for busy clinicians to make sense of the biomedical literature. Topics covered include how to decide whether study results are valid and applicable to patients, how to understand articles about diagnostic tests, prognostic estimates, clinical decision analyses, and clinical practice guidelines. Other topics include use of health outcomes data, including health-related quality of life and economic analyses, screening tests, and, applying these guides to individual patient care. The Users' Guides are available individually through an interactive Web site and published as textbooks on the essentials of evidence-based clinical practice.