UNDERSTANDING STATISTICS: AN APPROACH FOR THE CLINICIAN

G. Robert DeYoung, Pharm.D., BCPS

Reviewed by Jacqueline L. Olin, Pharm.D., BCPS; Eva M. Vasquez, Pharm.D., FCCP, BCPS; and Jennifer L. Waller, Ph.D.

Learning Objectives

1. Distinguish the roles for and types of descriptive and inferential statistics.
2. Classify data types found in the pharmacotherapy literature as nominal, ordinal, interval, or ratio.
3. Evaluate statistical significance using either p values or confidence intervals.
4. Judge the appropriateness of common statistical tests or techniques for a set of data.
5. Infer the reliability of trial results and conclusions based on an evaluation of statistical techniques.
6. Judge the clinical significance of statistical differences.
7. Interpret the results of correlation, regression, survival analysis, and meta-analysis in pharmacotherapy trials.
8. Design a plan to communicate statistical results to health care providers and to patients in a way that allows them to make well-informed decisions about the use of drugs.

Introduction

For many people, the mere thought of statistics conjures up disagreeable memories of long, complex calculations, tables in the back of textbooks, and a feeling of being only vaguely attached to the subject. However, a firm grounding in the science of statistics is an essential tool in the practice of pharmacotherapy. Just as an understanding of antibiotic, antiarrhythmic, antidepressant, or other drugs makes up a key component of the pharmacist's specialized knowledge and skills, so too does the ability to interpret the pharmacotherapy literature in an assured and accurate manner. As practicing clinicians realize, proficiency in the application and interpretation of statistics contributes to patient care in a crucial way. Just as the successful mastery of therapeutics arises from a combination of the basic and clinical sciences, knowledge of statistics should include both basic concepts and an understanding of their appropriate use and translation into practice.

At its most basic, the process of interpreting trial results includes two considerations: causality and certainty. Whether an intervention causes a particular outcome is a question that is best assessed within the context of how a trial is designed. The Literature Evaluation and Overview of Outcomes Research chapters in this book explain how matters related to trial design impact the ability to assess causation. How certain researchers are (or should be) that a particular outcome did not occur simply due to chance is the question that much of statistics is designed to answer.

Statistics is a science whose origins lie in the field of probability. Most statistical tests simply quantify the level of certainty that exists in the answer to the question: How likely is it that the difference observed between groups is simply the result of chance? Statistical tests that provide an answer to this question help to comprise the field of "inferential statistics". As the name implies, these tests (e.g., t test, chi-square, and analysis of variance) allow for the drawing of conclusions, or inferences, from the differences observed between two or more groups. A second area of statistics, descriptive statistics, provides another set of useful tools to the practicing clinician. These statistics describe data (e.g., mean and median) and certain types of descriptive statistics are used to calculate the values used to report inferential statistics. This chapter covers aspects of both of these types of statistics and also discusses issues related to data analysis, systematic reviews, correlation and regression, and to applying the results of statistical analysis to patients.

Types of Data

Although not generally considered by readers of statistical results, the underlying nature of collected data influences the type of statistical tests that are used to analyze

Abbreviations in this Chapter

CI	Confidence interval
H_0	Null hypothesis
OR	Odds ratio
SEM	Standard error of the mean

it. Assumptions about the distribution of the possible values for a variable of interest provide an expected distribution for any values that might be observed. More than just a set that encompasses all the possible values, the distribution provides information about the expected frequency of these possible values and, as such, is sometimes referred to as a frequency distribution. A common type of distribution, the normal distribution, is described by a line that takes on the appearance of a symmetric bell-shaped curve. Values near the center of the curve are more common (likely) than those at the ends of the curve. This type of distribution often is ascribed to continuous types of data (see the Continuous Variables section). Other types of data that can take on only certain quantities or conditions have expected frequencies of values that differ from the normal distribution. Examples of these include the binomial and Poisson distributions that can be used when conditions being observed are either present or absent (i.e., measured on a nominal scale). The type of distribution describing a certain variable is one of the basic assumptions for many statistical tests. For some types of data (e.g., ordinal data), statistical tests do not rely on assumptions about the expected distribution of values.

To evaluate observations using statistics, the data must be represented using numbers or some other symbols that allow the data to be sorted. Such a scheme is inherent when recording observations that are measured using numerical scales. For example, blood pressure, blood glucose, or pill counts are recorded with a numeric representation that can then be scrutinized mathematically. On the other hand, common variables, such as sex, smoking status, or presence of disease, are not measured using numbers. These types of variables can be represented with symbols (e.g., 1 = male, 2 = female) so that they may be sorted and counted (i.e., changed into a percentage or proportion) and then analyzed statistically.

Discrete Variables

Data can be broadly divided into two groups: discrete or continuous. Discrete variables are further distinguished as being either nominal or ordinal. Nominal data are those that exist in one of two or more conditions or states. Also referred to as categorical variables, these types of data are represented by fitting them into categories that do not have an associated ranking or magnitude compared to the other categories. Examples of nominal variables include sex, smoking status, race, marital status, and the presence of disease. Nominal variables, such as marital status, exist as one of several states (e.g., single, married, divorced, or widowed), whereas some nominal variables are present in only one of two states (e.g., smoker or nonsmoker). Variables of this latter type also are called dichotomous variables. It is not uncommon to see such data represented symbolically so that the number of observations in each "state" can be counted. Such counts often are reported as percentages, proportions, or rates. A common technique for dichotomous variables is to assign one possible condition a value of 0 and the alternative condition a value of 1. For example, patients who have diabetes in a study of lipid-lowering drugs would be assigned a value of 1, and those without diabetes would be assigned a value of 0. Comparing what percentage of study enrollees has diabetes to that percentage which does not allows for statistical analysis. Variables that have more than two possible conditions can be classified by using other representations (e.g., 1, 2, 3; A, B, C). Because they are summarized as proportions or probabilities, nominal data do not have normal distributions. These data can be subject to certain statistical tests that do not assume a normal distribution (see the Common Statistical Tests and Their Interpretation section).

Ordinal data measure some attribute using a finite number of ordered categories. These categories often are represented numerically, but this is arbitrary. Ordinal data can be represented by any type of symbol (e.g., Greek letters or colors), but numbers are used commonly because their order or ranking typically is understood. Likert-like scales, frequently used in studies of patient or consumer preferences, are a classic example of ordinal data. Traditionally, consisting of four or five possible values, such scales assign descriptions to each possible value in the scale (e.g., 0 = strongly disagree, 1 = disagree somewhat, 2 = neither agree nor disagree, 3 = agree somewhat, 4 = strongly agree). An important attribute of ordinal scales is that although the possible values are presented in order (ascending or descending), the amount of change represented by the difference between units in the scale (i.e., 1 to 2 vs. 2 to 3) is not constant. Using the example scale, it can be seen that the change from a value of 2 to a value of 4 does not imply a doubling in a respondent's agreement to a particular statement. Moreover, a change from 0 to 1, when compared with a change of 4 to 5, cannot be said to represent the same magnitude of change in the parameter that is being assessed. In clinical practice, many other types of data that pharmacists must evaluate are measured using a similar schema. For example, in congestive heart failure, studies of drug therapy usually classify patients into the New York Heart Association classes I–IV. Many assessment tools used in psychiatric research use ordinal data.

As with variables measured on a nominal scale, data derived from the use of ordinal scales are not necessarily normally distributed. Although meaningful values for ordinal variables also can be assigned to each category, reporting a mean for such data has no understandable meaning. For example, it is impossible to determine what a mean value of 3.5 might represent for a variable that, by definition, can only be assessed as 1, 2, 3, 4, or 5. As is discussed in the Parametric Versus Nonparametric Tests section, these distinctions, as well as other characteristics, require that ordinal data be analyzed using a distinct set of statistical tests referred to as nonparametric tests.

Abbreviations

Continuous Variables

Continuous variables include those types of measures that classify data as either interval or ratio. Analysis of both types of continuous variables and the results of such analyses are the same. Interval and ratio data can take on any possible value, limited by the techniques or instruments used to measure them. Continuous variables also can be thought of as "measuring" variables because of this quality. A distinguishing feature of interval data compared with ratio data is that items measured using an interval scale have an arbitrary 0 point, whereas ratio scale data have a 0 point that is absolute. The 0 point when measuring temperature in degrees Fahrenheit is set arbitrarily—it does not represent the point at which there is no longer any temperature. In contrast, a variable such as blood glucose is measured on a ratio scale with a 0 point that is absolute and indicates an absence of the property being measured. However, both temperature and blood glucose measurements can take on any value within a given range, including fractional values and, thus, are continuous. Moreover, in contrast to ordinal data, the magnitude of difference between units is constant (i.e., 64°F is twice as warm as 32°F).

As with discrete variables, such distinctions have importance in statistics because the properties of different data types determine the mathematical analyses to use appropriately in their evaluation. For instance, continuous variables can be normally distributed. This property of normality underpins the mathematical techniques, or statistical analysis, used to make sound inferences from trial results. Normality should be evaluated before proceeding with statistical analysis.

Descriptive Statistics

The terms used as descriptive statistics may be familiar to many people because of many years of exposure, often starting in grade school. Descriptive statistics simply explain or depict, in summary form, a set of observations (data) and can be applied to many different types of data. The measures of central tendency are one type of descriptive statistic; the mean, median, and mode reveal the numerical "center of gravity" for a set of observations. Another type of descriptive statistic includes measures of variability. Range, variance, standard deviation, and standard error of the mean (SEM) characterize this property.

Describing Central Tendency

Mean

The mean is the value that results from summing the values in a data set and dividing that sum by the number of observations in the set. Also called the average, or arithmetic mean, this value is used extensively in a multitude of everyday occasions (e.g., average miles/gallon, average price of a good, or average score on an examination). Because of the manner by which this statistic is calculated, it is susceptible to the influence of values in a data set that are distant from the mean. As a result, outlying or unusual values in a set of observations can impact the calculated mean disproportionately. This phenomenon is important to consider when interpreting many types of studies (e.g., drug interactions). A mean that is reported without some description of the other values in the sample can be misleading. A quick check of the range of values in the sample can give an idea if the mean has been unduly influenced by one or more outliers. When data in a sample cluster around extremes of values (e.g., bimodal), the mean and the range considered together may still be misleading on cursory examination.

Median

The median is that value in the data set located in the middle of all of the other values. It is the value in the sample that has an equal number of data points that are of higher value and of lower value. The median also can be described as demarcating the 50th percentile of data values because 50% of the data lie above and 50% lie below the median value. Because the determination of the median does not consider the values of the data above and below it, it is not susceptible to the influence of outliers. Median values often are used for data sets that span a wide range of values, or that have values that are concentrated away from the center or at one end of the values in the data set (i.e., skewed). The median also is used in the calculation of many nonparametric inferential statistics.

Mode

The mode is that value in a data set that occurs most often. The mode can provide additional important information about the distribution of the data not captured by the mean or median. For data that exhibit a normal (Gaussian) distribution, the mean, median, and mode all have the same value.

Describing Variability in Data

Considerations of the variability inherent in both a population and any sample of that population are essential to the understanding of statistics. In conducting studies of drug therapy, the intent is to extrapolate results from a representative sample of a population to the population itself. Two notable sources of variation that arise when using such a study approach are differences that exist between patients in the sample being studied and variations that exist between samples of the same population. The former is analyzed using the standard deviation and range, whereas the latter is reflected in the SEM. Variability among patients in a population is reported as the variance.

Standard Deviation

The standard deviation is a value that describes the distribution of values in a data set by comparing each measured value to the mean. For the population from which the data are sampled, variability is expressed by a parameter known as the variance. The value for standard deviation (sample variability) can be calculated by taking the square root of the variance (population variability). When the variance is not known, the standard deviation also may be determined using only the sample data by taking the square root of the sum of the squared values of each difference between an observation and the mean, and dividing this value by one less than the sample size (i.e., n-1). For a normally or near normally distributed set of data, the sample

mean ± 1 standard deviation will encompass 68% of the sample values. Similarly, mean ± 2 standard deviation will include 95% of the values measured in the sample. An analogous property exists with the SEM (see the Standard Error of the Mean section) and permits the calculation of confidence intervals. Values for standard deviations, as for means, are understandable only for data that are continuous. Even so, it is not uncommon to see summaries of ordinal data reported as means together with attendant standard deviations.

Range

The range represents the difference, or spread, between the lowest and the highest values of data in a set. It can be used with either population data or sample data. For example, when a range is presented with a median, a researcher has some impression as to whether the distribution of the data is skewed when the median lies conspicuously closer to one end of the range.

Standard Error of the Mean

The SEM often is less familiar to clinicians but conveys the magnitude of sampling variability. Each time a researcher samples a population, that sample likely will have a different mean value for the variable of interest than the previous sample. For example, if a researcher selected 10 samples of 200 people and measured their height, each of the 10 samples likely would produce a different mean value for the population that the samples represent. In other words, there is variability in the estimates of the population values for the variables under study, and this variability needs to be considered when assessing the results of statistical calculations. Methods of many statistical tests do not rely on differences in means to assess differences between groups, but instead examine whether the variability in the estimates of any difference between groups is consistent with that expected due to sampling variability. However, to arrive at the SEM, it is not required that researchers repeatedly sample a population. There exist mathematical ways, depending on the type of data, to calculate the SEM from a single sample. In the case of a mean value from a single sample, the SEM is calculated by dividing the sample standard deviation by the square root of the sample size (n). Similar to SD, ± 1 SEM encompasses 68% of all possible means for the population, and ± 2 SD encompasses 95% of the means for the population.

An application for the SEM is its use in deriving simple confidence intervals (CIs) around means for continuous data.

Testing Hypotheses Using Statistics

Conventions of Hypothesis Testing

At least two common points of confusion or elusiveness with regard to understanding the role of statistics and hypothesis testing exist. One involves a lack of clarity about what, exactly, is being tested. The second is unfamiliarity with the jargon used in describing these processes.

Hypothesis testing proceeds by using mathematical techniques to compare data from two or more groups and to determine the probability that the differences between the comparison groups occurred because of sampling variability. In other words, for inferential statistics, a calculated value answers the question: How likely is it that the differences being observed are simply because of chance? The terminology of hypothesis testing requires an understanding of what the null hypothesis (H_0) states and why.

The H_0

The H_0 puts forward the idea that there is no difference between the groups being compared. Such comparisons may be carried out on data from two different groups or on data from a single sample (e.g., as occurs when comparing the mean of a single sample to a known population mean). The hypothesis that Group A = Group B is the basis for statistical comparisons subject to inferential analysis. The study's objective is usually to determine a difference between groups. The idea that a difference exists is the alternative hypothesis. The alternative hypothesis postulates inequality between the estimates of the difference between groups (i.e., Group A ≠ group B). Other forms of the H_0 exist and can help in assessing whether an attribute of interest is different than 0 (i.e., H_0: mean = 0), or if some mean value exceeds a certain threshold of interest (i.e., H_0: mean < x). When a statistical test is performed, the H_0 is either rejected or not rejected. Rejecting the H_0 that posits no difference is consistent with accepting the alternative hypothesis with the conclusion that a difference exists between the groups being compared; that is, a difference is not likely due to chance observations. Of importance, failing to reject the H_0 is not sufficient to conclude that the groups are equal. As discussed further in the Decision Errors section, a sound determination of equivalence between comparison groups involves more than just the value of a test statistic.

Statistical Significance

Statistical significance is one of the most commonly cited results in studies of drug therapy. It also is subject to many misinterpretations. Whether observed differences are considered statistically significant can be determined by assessing the p values or the CIs determined for observed differences. P values, by convention, indicate statistical significance if they are less than a specified significance level. A significance level of 0.05 means that in deciding to reject the H_0 based on the data at hand, an error will result 5% of the time. In other words, the magnitude of difference observed between study groups would occur because of random variation (i.e., chance) in five out of each 100 times. A one in 20 chance of mistakenly rejecting the H_0 may be acceptable for many types of drug therapy decisions, but this

Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995;152:27–32.

level of uncertainty is better considered as a general rule. It might be argued that interventions carrying a high degree of risk and/or those with marginal benefit should be subject to more stringent criteria for rejecting the H_0. In some studies, researchers justify a cutoff point for rejecting H_0 that is less than 0.05 (e.g., less than 0.01). In contrast, it also must be acknowledged that for some questions, a p value marginally greater than 0.05 can be considered sufficient assurance that rejecting the H_0 should not result in harm. However, such an allowance usually is viewed with skepticism and should be justified thoroughly by the researchers.

Clinical Significance

Understanding and interpreting p values often is difficult. A common pitfall involves the misattribution of clinical importance to results that achieve the stated level of statistical significance. Small differences between groups may meet criteria for statistical significance but not surpass a threshold that clinicians (or patients) consider important for changes in care. Studies with large sample sizes often can demonstrate p values less than 0.05 for differences in outcomes that are small (i.e., less than 5%). Although such differences may be important in some circumstances, they are not necessarily so in others. Health care professionals have struggled over what course of action is appropriate in circumstances where disagreement has focused on what constitutes a clinically meaningful difference. The comparative efficacy of thrombolytics in myocardial infarction and, more recently, the use of thrombolytics in stroke, have been the subject of such debates. Given their expertise in clinical therapeutics, pharmacists should actively engage in discussions to resolve these conflicts. Table 1-1 provides some guidance on ascertaining clinical significance.

Some reports imply that smaller p values correspond to more important results. Authors, editors, and researchers contribute to this misperception by describing various magnitudes of p values as "significant" (e.g., $p<0.05$), "highly significant" (i.e., $p<0.01$), or sometimes even "very highly significant" (i.e., $p<0.001$). Such values have no relationship to the importance of findings; rather, they simply convey the estimated degree of certainty associated with a decision to reject the H_0. Then again, differences that do not achieve statistical significance cannot be dismissed, ipso facto, as unimportant or as demonstrating equivalence. Readers of such results must consider issues such as trial design, representativeness of the sample, previous studies, and statistical power before making such determinations. Considerations for interpreting p values can be found in Table 1-2.

Decision Errors

The Two Types of Errors

Conclusions about statistical test results are not inevitably correct. Typically, decisions about trial results are grouped into four types: two being correct and two being incorrect. Having decided whether to reject the H_0, researchers need to consider these possible decision errors.

Table 1-1. Judging the Clinical Significance of a Statistically Significant Difference

Determine what others think is clinically significant by:
- Considering the effect used in the sample size calculation (if reported)
- Considering existing evidence-based or expert consensus statements
- Considering any cost-effectiveness or decision analyses that have been performed

Absent such guidance, require that the minimum worthwhile effect be large when:
- The intervention is costly (e.g., in terms of time, money, or other resources)
- The intervention is high risk
- The outcome is unimportant, or has intermediate importance but with uncertain benefit to patients
- A patient is risk-averse

Accept the minimum worthwhile effect as small when:
- The intervention is low-cost
- The intervention is low-risk
- The intervention is important and has an unambiguous outcome (e.g., death)
- A patient is risk-taking

Adapted with permission from the American College of Physicians-American Society of Internal Medicine. Froehlich GW. What is the chance that this study is clinically significant? A proposal for Q values. Eff Clin Pract 1999;2:234–9.

Table 1-2. Frequent Misinterpretations of P Values and Details to Take into Account

Pitfall: Statistical significance (e.g., p less than 0.05) means that the results between the groups are different (not a chance variation)
- Think about—
- Are differences clinically important?
  - Large trials can easily demonstrate statistical differences that have no practical consequence
- A small p value does not correct for systematic error (bias)
  - Poorly designed studies can demonstrate statistical significance, but lead to erroneous inferences
- Small p values simply mean that differences are less likely due to random variation (chance)
  - A p value of less than 0.001 or less than 0.0001, indicating much lower likelihood of random variation, is not more important
  - The demarcation that a p value less than 0.05 is significant is a convention, but an arbitrary one

Pitfall: Lack of statistical significance means the results are unimportant
- Think about—
- The confidence interval may include mostly values that are valuable in patient care (and approach statistical significance); for low-risk interventions, this may be sufficient evidence
- Is this study an outlier?
  - Consider other data/studies that are available
- Evaluate the study design
  - Are there issues with the sample that make it different from the population? Are there other causes of bias?
  - Is there an adequate discussion of power?

Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ 2001;322:226–31.

Abbreviations

The two possible wrong decisions about the disposition of the H_0 are labeled type I and type II errors.

A type I error arises when a calculated p value leads to the rejection of the H_0 when in fact the H_0 is true. The value that represents the likelihood of making a type I error is known as the significance level and is represented by the symbol α. Concluding a difference between groups when no difference exists has a probability of ocurring one out of 20 times when α is set at 0.05 or one out of 10 times when α is 0.01. The significance level, as well as the p value, tells how often a type I error will be made. Determining α takes place during the design of a trial and is necessary for calculating sample size.

Type II errors require consideration when analyses of data do not allow for rejection of the H_0, and the possibility of no difference remains viable. Type II errors involve the inappropriate decision not to reject the H_0 (i.e., concluding that there is no difference) when a difference exists. Type II errors frequently are associated with declarations of equivalence between the effects of interventions. The threat of type II errors is driven by variables in trial design or analysis that impact the sensitivity to differences between the groups. As with type I errors, a Greek letter also is used to represent the chance of making a type II error, in this case the letter β. The conventional minimum value for β is less than or equal to 0.20.

Table 1-3 summarizes the nature of type I and type II errors.

Power

Power is a value calculated by subtracting β from 1 (i.e., 1-β) and represents the ability of a study to detect specified differences between groups. Researchers' discussions of power are essential to evaluating statistical results that fail to reject the H_0. In some instances, the inability to reject the H_0 may result from inadequate power to detect existing differences. Power is influenced by many factors of trial design with one of the most common being sample size. As sample size increases, so does the power to detect differences. This is one of the reasons smaller trials are more likely than larger trials not to find a difference between interventions. The existence of large differences between groups also improves the power of a trial to find differences. α and β are inversely related, so an increasing α also can increase power but risks a higher likelihood of a type I error.

Table 1-3. Types of Decisions Associated with Hypothesis Testing

	Epistemological Truth (the real answer)	
Decision made	H_0 is true	H_0 is false
Fail to reject H_0	No error, right decision	Type II error, wrong decision
Reject H_0	Type I error, wrong decision	No error, right decision

Approaches to Analyzing Clinical Trials

Power also is one of the considerations that plays a role in weighing the choices for a global approach to analyzing data from clinical trials. The best designed trials, when executed, may not transpire exactly as planned. This reality impacts the ability of statistics to present a reliable estimation of the differences between groups. Because of this influence, it is important that researchers specify beforehand the details of what data were to be included in the final analysis of a study. Three standard approaches have been developed by statisticians and researchers to identify the data for analysis: intention to treat, per protocol, and as treated.

Intention to Treat

Intention to treat refers to methodology accounting for data of all patients initially assigned to study treatments. For example, in a trial designed to assess the efficacy of a new drug, patients might be randomized to either active drug or to placebo. Ideally, all study patients would adhere fully to their allocated treatment. In reality, this does not happen. If a patient assigned to active treatment takes only one or two doses then stops because of side effects, intention-to-treat analysis includes their data in the active treatment group. Because the original assignment was to active drug, it can be said that the "intention" was "to treat" that person. Likewise for those allocated to placebo. Although this approach may seem misleading initially, it is methodologically sound.

Proponents of intention to treat point out that this approach mimics true clinical practice in as much as when patients are given a prescription, their future adherence cannot be known. Therefore, estimates of drug effects in trials should closely resemble those that clinicians should expect. As a consequence, intention to treat provides an idea of what has been termed "use effectiveness". Depending on perspective, this element of intention-to-treat analysis may be appealing. The trade-off is the absence of a reliable approximation of the true magnitude of effects when an intervention is used optimally (i.e., when adhered to). In addition, intention to treat does not alter the final power of the trial with regard to the original sample size calculations. This stands in contrast to per-protocol analysis.

Per Protocol

As an alternative to intention-to-treat analysis, a per-protocol approach provides a superior estimate of how well an intervention performs when executed as designed (i.e., method effectiveness). Those who do not adhere to their assigned treatment are dropped from the final data analysis. Although this approach is sometimes judged a more forthright accounting of what has occurred during a trial, readers should be mindful of several caveats to interpret per-protocol analyses correctly.

First, for trials of drug interventions is the question of fully assessing compliance. If this assessment is indeed possible, a second consideration in using per protocol is

Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. Clin Pharmacol Ther 1995;57:6–15.

establishing explicit criteria for patients considered compliant with therapy. This may seem straightforward on the surface, but consider the patient who misses only one or two doses. Researchers may decide to count as compliant only patients who have taken at least 75% of their prescribed doses. Other researchers may establish a threshold well above or below this level. Regardless of the methodology used, justifications for such demarcations often introduce a subjectivity to the decision of which data are included for analysis. An even more threatening problem when using a per-protocol approach is the danger of having to disregard so much data that the power to detect important differences is jeopardized. Because of this danger, per-protocol analysis should always be specified during the planning stages so that the planned sample size accounts for such potential losses of data. Differences suggested by post hoc per-protocol analysis, as may occur in some subgroup reviews, must be viewed as simply suggesting hypotheses for future study.

As Treated

An as-treated analysis offers a solution to some of the vulnerabilities of the per-protocol approach. With the as-treated method, patient data are analyzed based on the actual drug-taking behavior. If a patient was assigned to an active drug but never complied, his or her data would be retained but reassigned to the placebo group. Alternately, patients allocated to placebo who receive active drug as part of their care outside of the study protocol would be included in the active drug group. Although requiring many of the same considerations about what constitutes adherence as per protocol, this method does not throw out any data and, hence, preserves a study's power. Unfortunately, this occurs at the expense of jeopardizing the initial randomization. Such a breech can eliminate the ability to conclude that the intervention being studied caused any observed differences. That potential outcome calls into question the original justification for the study and introduces ethical concerns beyond the scope of this chapter. As-treated analysis needs to be considered carefully in both the design and interpretation of a study to avoid significant undesirable consequences.

Confidence Intervals

Confidence intervals remind readers that reported differences between interventions are not hard and fast values, but rather estimates based on a sample. As estimates, these reported values (e.g., mean) may or may not portray the true underlying dissimilarity of the groups being compared. They are, in effect, a "best guess" of any true differences based on the information provided from the sample that was used. Constructed using the summary estimate of difference or change and some approximation of the sampling variation, CIs represent the possible values for the true difference between interventions that are supported by the data. They can be constructed for most types of variables, including averages, proportions, odds ratio (ORs), and relative risks. As an example, the formula mean ± 1.9 (SEM) provides the 95% CI for the mean of a continuous variable.

A 95% CI means that if a study were repeated many times, then 95 of the 100 CIs constructed using this method would contain the true mean difference. As with a p value less than 0.05, this meaning also prompts readers of statistical estimates to recall the inherent uncertainty of these figures. The 95% CI is sometimes interpreted as being "95% certain that the true difference between groups lies within the reported interval." Although scoffed at by statistical purists, such a definition portrays a practical, if technically imprecise, definition of what a CI represents.

Take for example a hypothetical study of antipyretic drugs. If one drug lowered temperature by an average of 4°F and another drug by an average of 2°F this difference might be reported as a difference and CI of 2°F (95% CI = 0–4). The CI would have been constructed by means of a formula that, using an approach analogous to the example for the mean of continuous variable, calculated the uncertainty around any estimated difference in efficacy by adding and subtracting some multiple of the SEM to and from the mean difference. In this case, the results of this trial can be interpreted as a mean difference in antipyretic drug effect of 2°F, and a CI showing that the true difference between these drugs could be as small as 0° (i.e., no difference) or as large as 4°.

A CI allows conclusions about the significance of results. A 95% CI is the customarily accepted level to determine statistical significance, but higher levels may be chosen (e.g., 99%). Because the numbers encompassed by the CI are all possible values for the real difference between groups, a CI that includes a value consistent with no difference cannot represent a statistically significant finding. When reviewing results from trials where a value of 0 would represent no difference between groups, a 95% CI that contains 0 within its range corresponds to a p value greater than 0.05. Using the previous example of antipyretic drugs, the reported difference in effectiveness of 2° (95% CI = 0–4) would not be statistically significant. Again, this is because the data as reflected in the CI cannot exclude the possibility that a disparity of 0° represents the true difference between these drugs.

For observational or other types of studies where comparisons are reported as ratios (e.g., OR and relative risk) a value of 1 would signify the lack of a difference between comparators. For example, in cohort trials, relative risk is calculated by dividing the chance of developing a disease in people with a certain exposure by the chance of developing that same disease in people without such an exposure. Relative risks of less than 1 imply less risk of a disease given a particular exposure, and relative risks greater than 1 imply an increased risk. The OR, reported in case-control trials, divides the odds of having a particular risk factor given for people with a specific disease by the odds of having that same risk factor in patients without the disease in question. An OR is interpreted in the same way as a relative risk, by considering the direction and magnitude of any deviations from the value of 1. For instance, a case-control study examining a link between Bell's palsy and a previously (but no longer) marketed intranasal flu vaccine in Switzerland reported an OR of 84.0

(95% CI = 20.1–351.9). Such a finding can be interpreted to mean that the chances of having Bell's palsy symptoms are increased 84 times in people who receive this particular intranasal vaccine. For both of these types of studies, a value of 1 in the CI would mean the results are not statistically significant. In the end, it should seem sensible that if values representing no difference cannot be excluded as the real difference between groups, then such findings could not have a p value less than the significance level (i.e., could not be statistically different).

In addition, CIs can help in determining the clinical significance of results. Because the CI reports the range of possible true differences that are consistent with the data, a reader of such results can take into consideration the full spectrum of possible changes that might be expected when using an intervention. This property, combined with the previously discussed properties, has led to more frequent reporting of CIs in published studies. Still, this practice has not been adopted universally.

Common Statistical Tests and Their Interpretation

Deciding which statistical test is best suited to a particular analysis scenario often may be best left in the hands of statisticians. Just the same, pharmacists must have an understanding of the appropriate application of a core set of statistical tests. A functional "statistical formulary" should minimally include a command of the specific tests discussed in this section. An appreciation for the underlying assumptions of these methodologies also will serve to ensure their correct interpretation.

Parametric Versus Nonparametric Tests

Many familiar statistical tests (e.g., t test and analysis of variance) belong to a group of tests termed parametric tests. Tests belonging to this group all assume certain characteristics about the nature of the data being analyzed. To varying degrees, violations of these assumptions result in inaccurate conclusions about the statistical differences between groups. For example, three assumptions associated with the use of analysis of variance are: 1) the data have a normal distribution; 2) each observation is independent of the others; and 3) the variances within the groups being compared are equal (homoscedasticity). Sample data that do not meet the conditions of these assumptions are sometimes still analyzed using parametric tests if their departure from these assumptions is not extreme. In addition, some data are subject to mathematical transformations (such as a logarithmic transformation) that result in meeting the assumptions (e.g., normality). Nonparametric tests are statistical tests that can be used in cases where the assumptions of parametric tests cannot be guaranteed. Statisticians can use separate methods to assess whether a sample exhibits properties such as normality or homoscedasticity and, thus, help to ensure the use of the best type of test. Prior work with comparable data also can provide information about whether the application of a parametric test is appropriate. It is important to use statistical tests whose assumptions match the parameters of the data under examination. When this is not done, statistical accuracy and precision (as well as the validity of any resulting inferences) are endangered.

Along with these considerations, the type of variables that are being measured also helps to determine which statistical test should be used. Inferential tests assess whether the data drawn from the samples differ by an extent greater than that expected due to chance. Each type of test completes this evaluation mathematically by comparing the probability of the observed results with those that would be expected due to the underlying distribution of the data. Different tests are needed, in part, because each type of variable (e.g., categorical and continuous) has a different distribution (e.g., binomial and normal) of expected values.

A third consideration for some testing scenarios addresses the need to adjust the statistical calculations to consider influences on the variables that are not otherwise accounted for in the design of the trial. Referred to as confounders, such influences can bias the data to favor a particular outcome. For example, if studying the effects of a recently marketed oral contraceptive on the incidence of myocardial infarction, smoking status should be considered a potential confounder because it has an independent effect on the incidence of myocardial infarction. Certain tests adjust for such biases, or try to, to allow valid inferences about causality in circumstances that might otherwise be difficult to judge. Such tests are not a substitute for a well-designed investigation but aid in minimizing the risk of a decision error that can occur due to an uncontrolled or unrecognized bias.

Categorical Variables

Categorical or nominal variables measured from independent samples often are analyzed using a chi-square test. For nominal, ordinal, and continous variables, independent observations occur when the results of one measurement are not influenced by or dependent on the value observed at some prior time. There exist several variations of this test for particular circumstances; but, in general, this test compares the expected frequency of events to that actually observed during an investigation. Baseline characteristics of patients in randomized trials (e.g., sex, smoking status, and β-blocker use/nonuse) may be analyzed using a chi-square test to demonstrate no significant differences between comparison groups. Viewed a different way, this process can detect any important dissimilarity between the groups. Often associated with a 2-row-by-2-column table, the chi-square test is used in the analysis of rates, percentages, and proportions. For categorical variables using a smaller sample size, the Fisher exact test accomplishes the same goal. The Fisher exact test usually is recommended when any one cell in a 2-by-2 table has an

Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med 1991;114:515–17.

Gaddis ML, Gaddis GM. Introduction to biostatistics: part 1, basic concepts. Ann Emerg Med 1990;19:86–9.

expected value of less than 5. Because of the underlying distribution used to calculate the expected values and the chi-square test statistic, other corrections are sometimes used for samples of intermediate size (e.g., Yates correction). Yates correction is most likely to be mentioned by researchers when n=25–40. As with a 2-by-2 table, a contingency table (a table with more than two columns and two rows) is analyzed using a chi-square test but one designed for more than two comparisons. The effects of confounders, if not adequately considered during the design of a trial, may be accounted for at least partially during the statistical analysis. The Mantel-Haenszel test can be used to adjust for confounding variables when comparing two independent groups. This test also often is associated with the determination of statistical significance of an OR in certain types of investigations.

When comparing categorical data between groups of observations that are not independent (e.g., crossover studies and paired/matched observations), another test must be used. In this case, the McNemar test can be chosen to analyze results from studies with related or dependent measures. For example, studies of treatments for rare neurological disorders may match study patients based on certain underlying characteristics to assemble patients into treatment groups that are as comparable as possible. If the outcome of interest is categorical (e.g., mortality and hospitalization), then such a trial could use McNemar to analyze this matched data. For comparisons involving more than two matched groups, a chi-square test with a correction for multiple comparisons should be selected.

Ordinal Variables

Ordinal data are analyzed based on an evaluation of the respective ranking of values in the data set without measuring the magnitude of the difference between values. Such an approach considers the non-normal distribution of possible values of data for these noncontinuous variables.

As with the tests for nominal variables, the tests for ordinal variables depend on the number of groups being compared and whether the measurements are independent. For two groups derived from independent observations, the Mann-Whitney *U* and the Wilcoxon rank sum tests commonly are suggested. These tests would be useful for comparing drug interventions in two groups when the outcome measure uses an ordinal rating scale. Ordinal rating scales are common in trials of psychiatric interventions, pain management, and patient satisfaction. For a comparison of more than two groups, the Kruskal-Wallis test is appropriate. Regardless of the number of comparisons, the analysis of variance ranks test helps to adjust for confounders. Nonindependent samples are compared by a test such as the Wilcoxon signed rank test (two groups) or the Friedman test (more than two groups).

Continuous Variables

Continuous variables are candidates for analysis using parametric tests. If the data do not meet the assumptions of a parametric test, they may be analyzed using a nonparametric test. The classic Student t test, so named because of the pseudonym used by the person who first described the technique, compares the means of two groups. Forms of this test exist to accommodate the analysis of groups with either equal or unequal variances. T tests that compare the means of two independent samples typically are known as two-sample t tests. Consider a trial in which patients are randomized to two antihyperlipidemic drugs. If the outcome of interest were the cholesterol concentrations at the end of the trial, then a two-sample t test appropriately would be applied to the data. A one-sample t test is selected for comparisons that are made from two sets of observations taken from a single sample. A one-sample t test also is used to compare the mean of a sample to a known population mean, or to a predetermined target value. For paired (dependent) measures of two groups, a paired Student t test is recommended.

Analysis of variance is used when more than two groups are being compared. The repeated measures variation of analysis of variance analyzes data that are paired. Analysis of covariance can be used for independent samples when the effects of confounders, or covariates need to be considered. This technique might be used in trials of weight-loss drugs because it is recognized that patients with more excess weight will lose weight more quickly, on average, than patients with less excess weight. Baseline weight represents a confounder that must be accounted for in the statistical analysis.

Analysis of variance and multiple comparison procedures represent a family of special techniques used to avoid the dangers of performing multiple t tests to assess differences among three or more groups. The risk in using multiple t tests is that a type I error becomes more likely with each subsequent comparison. For example, if researchers are making three comparisons (i.e., Group A to Group B, Group B to Group C, Group A to Group C), each calculation brings with it a risk of making a decision error. The probabilities of such errors must be combined for researchers to understand the chance of reaching an unsound conclusion. Given that each comparison has a 5% (0.05) chance of leading to a type I error, making these three comparisons would have about a 15% chance of permitting such an error. The formula $1-(1-\alpha)^k$, where α represents the significance level and k the number of comparisons, will give the precise cumulative probability of a type I error in such situations. Analysis of variance in conjunction with multiple comparison procedures allows for multiple comparisons without an increase in spurious significant findings. A trade-off is that analysis of variance, when indicating a statistical difference between groups, does not reveal which specific groups differ. Further testing of individual H_0s is conducted using other appropriate statistical tests, some of which are discussed in the following paragraph.

One method that can be used to account for such probabilistic shifting is a technique known as the Bonferroni correction. In this approach, the p value that is considered statistically significant results from dividing the normally accepted significance level (0.05) by the number of comparisons. For example, if three t tests were performed when analyzing data from a study with three groups, then the minimally significant p value would be 0.017. There are many more types of multiple comparison procedures (e.g., Tukey, Student-Newman-Keuls, and Sheffé tests) to handle this

common problem. Although the preceding examples have focused on continuous data, multiple comparisons can threaten statistical inference for nominal and ordinal data as well.

Figure 1-1 shows how an appropriate statistical test can be determined. This figure allows for analogies that can aid in recalling which statistical test is used for which type of data to be easily discerned. For example, looking down the right-hand side of the figure, it can be seen that the chi-square test, Mann-Whitney *U*, and Student t test all analyze differences between two independent groups for nominal, ordinal, and continuous variables, respectively.

Subgroup Analysis

Subgroup analysis, the statistical evaluation of smaller groupings from a data set, often can be the source of much consternation for readers of the medical literature. First, there is no empirical limit on how often a group of study patients might be further divided (in ways that make clinical sense or not). Because of this, subgroup analyses are subject to the previously discussed concerns related to multiple comparisons. Subgroup analyses are categorized into two types: *a priori* (conceived beforehand) or post hoc (formulated after the fact) subgroups.

Subgroups of samples specified before the start of a trial can be analyzed in the same way as any H_0. Whether the subgroups make sense clinically is not guaranteed and must be considered when drawing conclusions. However, if the divisions suggested provide useful information, then their statistical analysis would proceed along the same lines discussed previously in the Common Statistical Test and

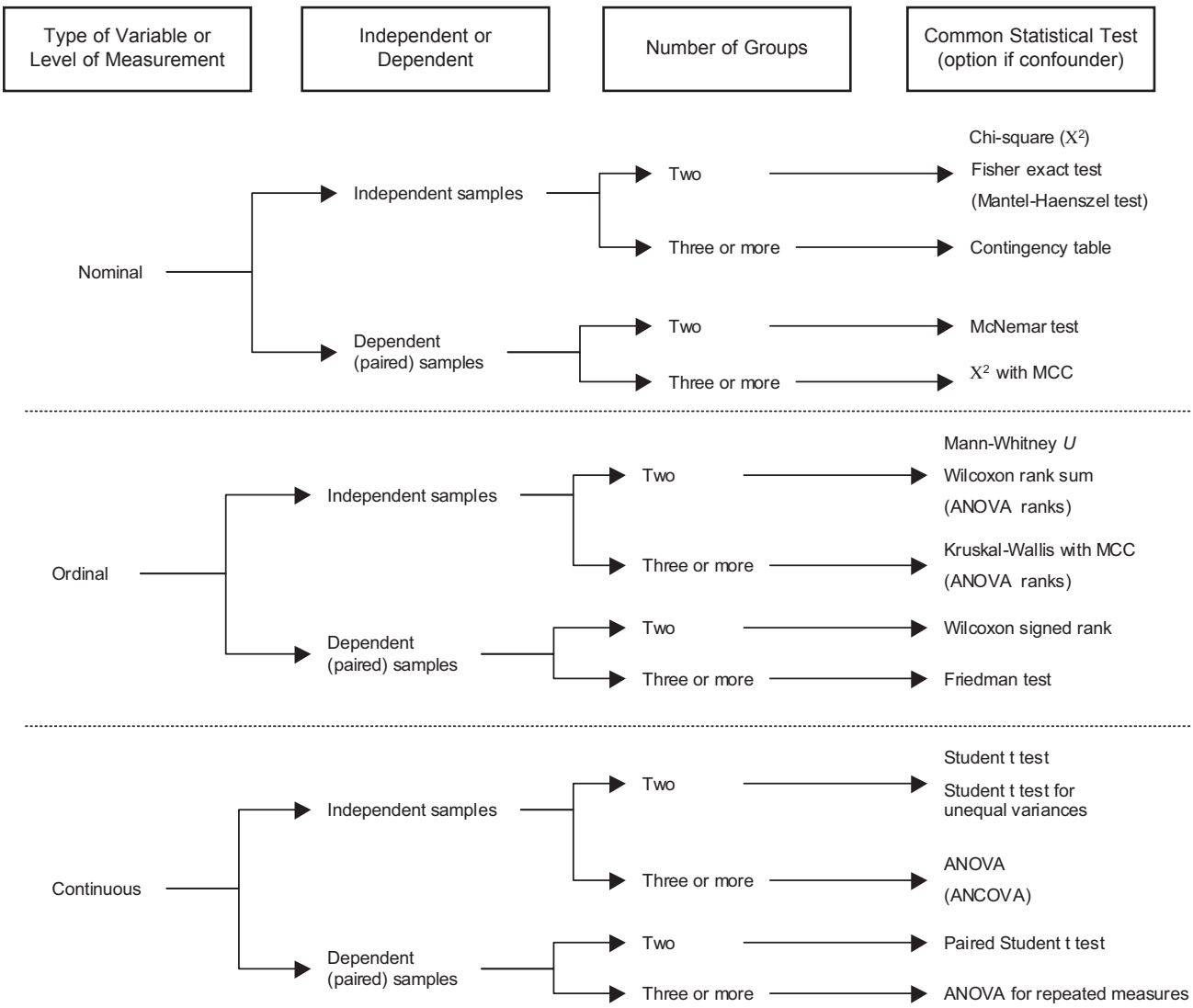


Figure 1-1. Statistical tests for common comparisons.[a]
[a]Many other tests exist, these being some of the most frequently encountered.
ANCOVA = analysis of covariance; ANOVA = analysis of variance; MCC = multiple comparison correction.

Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ 2001;322:989–91.

Their Interpretation section. Deliberations about power and multiple comparisons take on a heightened importance when deciding whether suggested differences should be trusted. Often, a trial will be powered to detect meaningful differences only for the main outcome using a predetermined sample size. When further dividing that sample into subgroups, the power to detect differences diminishes quickly so that the decision to not reject the H_0 cannot exclude the possibility that differences between subgroups do exist. On the other side of the coin, the multiple comparisons that are made between subgroups must be considered in the statistical analysis so as to avoid a type I error. However, when these difficulties have been thought about and sufficiently addressed, subgroup differences associated with preestablished subsets of patients can provide a more refined understanding about the effects of a drug therapy. For example, it may be the case that a drug provides a benefit for a sample of patients and that it provides even more benefit for a subgroup of those same patients.

Subgroups that are defined at the end of a trial should be considered hypothesis-generating as they typically have insufficient power to detect differences, are subject to biases that are not accounted for in the initial randomization, and present more opportunities for spurious results as a consequence of multiple comparisons. Significant differences that are suggested by post hoc subgroups should be further confirmed in subsequent controlled trials.

Composite End Points

Primary end points are those outcomes around which trials are designed. The primary end point serves as the basis for considerations of power, sample size, duration, and other aspects of a trial. Many investigations also will have several secondary end points to derive the most useful information from the study. Primary or secondary end points can be composite end points. Composite end points combine the data of more than one outcome into a single analysis. A composite end point of hospitalization, worsening symptoms, or death might be used as the primary end point in a trial of drugs for congestive heart failure. Any significant differences between groups on this outcome would imply that patients who receive the study therapy are less likely to experience worsening symptoms, a hospital admission, or death. A criticism of this approach is that it may be that only one or two of the outcomes measured in the composite end point are truly different between groups. Differences observed in the remaining two or three outcomes factored into the collective end point may not be of statistical significance. In particular, the combination of subjective (e.g., worsening symptoms) and objective (e.g., death) outcomes into a single end point contributes to unreliable conclusions. Pharmacotherapy studies often include such composite end points and should be evaluated with considerations of these limitations. There are statistical techniques to help deal with some of this confusion. The tests used are unusual and typically not well-known to a clinician reader, but include certain ranking tests for each component of the composite end point and subsequent specialized tests to ascertain which parts of the end point are significantly different from the comparison group. A more optimal approach would be to design the trial to assess individual outcomes, but this may not be feasible, affordable, or of interest.

Statistical Techniques in Systematic Reviews (Meta-analysis)

Meta-analysis is a method of using mathematical techniques to evaluate the numerical results of past studies. A systematic review is an explicit process by which trials dealing with a particular therapeutic question are brought together for a collective weighing of their results. A discussion of the steps necessary to conduct a high-quality systematic review is not addressed in this chapter. However, at least three topics related to meta-analysis are germane to this chapter: interpreting results, sensitivity analysis, and heterogeneity. These are addressed in the following sections.

Interpreting Results

Figure 1-2 illustrates a common graphic format for presenting the results of a meta-analysis. Referred to as a Forest plot, it depicts the results of all of the individual trials included in the systematic review by plotting each trial's estimate of the difference between groups with a dot, and the variability around that estimate with a line through the dot. This line represents the values comprising the CI for

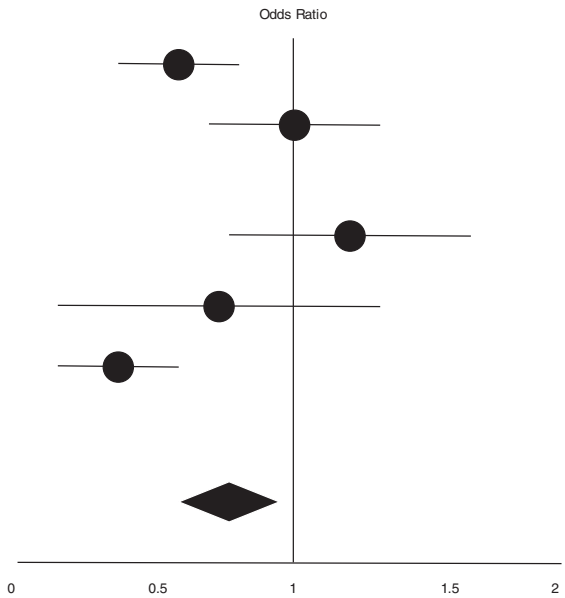


Figure 1-2. A forest plot of a hypothetical meta-analysis.
The results of individual trials are plotted as a difference between study groups (dots) along with their attendant confidence intervals (lines). The summary estimate of treatment effect from all of the trials is shown as a diamond where the vertical center represents the estimate of effect overall and the width of diamond represents the confidence interval around this estimate. The results of many such analyses are reported as an odds ratio. If this plot represented some drug effect on mortality then the overall result would favor the drug because an odds ratio less than 1 implies less risk of death.

Abbreviations

each estimate of difference found in the respective trials. A final assessment of efficacy based on the collective results is depicted using a square or diamond shape. The vertical center of this shape represents the calculated effect of treatment based on the results of all the individual trials. A horizontal line through the square or the width of the diamond denotes the CI associated with this effect. The results can be plotted as percentage change, percentage difference, or in any other units. If the results being analyzed are nominal data, it is common for results to be conveyed as an OR. Regardless, the center vertical line of the plot represents the finding of no difference. This means that estimates of effect, whether the summary estimate or its CI, which include this line, signify nonsignificant differences.

Sensitivity Analysis and Heterogeneity

One of the weaknesses of combining results of trials to reach conclusions is that the results are susceptible to influences other than the effects of treatment. As such, systematic reviews may undergo one or more statistical tests to detect the presence or extent of such influences. Sensitivity analysis seeks to ascertain if the final results are particularly influenced by the inclusion of one or more individual studies. Examination of the different effects of including or excluding certain studies from the meta-analysis is an essential exercise of any good meta-analysis. These analyses may include not only consideration of the trials themselves, but also any subgroups within the trials. A discussion of these factors adds to an understanding of the final results by alerting the reader to possible groups to whom the results do not apply.

Meta-analysis combines the results from different trials to reach a conclusion and so it should be expected that differences between studies, broadly termed heterogeneity, need to be addressed in interpreting any final results. The processes used to test for heterogeneity assume that the effects evaluated in each study are identical. A test statistic named the Cochran Q often is cited as the method used to test this H_0. A shortcoming of this test is that it only tests for the presence of heterogeneity and does not quantify its extent. A value known as the I^2 recently has been proposed as a way to indicate the degree of heterogeneity in or between studies in a meta-analysis. I^2 also can be used to compare heterogeneity between meta-analyses done on the same topic. This statistic expresses the amount of variation between studies as resulting either from heterogeneity or from chance. The I^2 takes on values ranging from 0% to 100%: the higher the value, the greater the heterogeneity. A value of 0% represents the absence of heterogeneity.

Characterizing Relationships Among Variables

Regression and Correlation

Regression techniques describe or summarize the relationship between two or more variables. Linear regression examines straight-line relationships among variables. This examination is accomplished by using methods that minimize the distance of all individual observations from a best-fit line drawn through the corresponding values of x and y (when studying only two variables) from each observation. Any such line can be represented algebraically as:

$$y = mx + b$$

where m represents the slope of the line and b represents the intercept on the y-axis of a graph of the data. Statistical tests investigate whether the slope of the line representing the observed relationship between these variables is different from a line with a slope of 0 (i.e., a horizontal line). Regression techniques also can consider relationships between variables other than linear relationships (e.g., quadratic relationships). Regression methods are used to develop models that allow clinicians to predict a variable of interest (e.g., mortality) by assessing one or more predictor variables (e.g., age, oxygen saturation, and heart rate).

Correlation answers the question, "How much of the variation in the value of x is associated with changes in y?" Correlation estimates the strength of the relationship between the two variables. Statistical tests to assess correlation depend on the type of underlying data. A well-known test of correlation, the Pearson product moment correlation coefficient, is a parametric test that can be used to examine the relationship between two continuous variables. A nonparametric test, Spearman rank correlation, is used for ordinal data or for continuous data that are not normally distributed. For example, if researchers are wishing to discern a correlation between metered-dose inhaler technique ranked on a 4-point scale (i.e., 0 = poor to 3 = perfect) and an estimate of effectiveness (i.e., 0 = no effect to 3 = complete relief of symptoms) the Spearman rank correlation method could be used. There are many more tests that can be applied for this purpose for different combinations of variable types.

Reported with these tests for correlation is an r value. In simple linear regression, this value is squared (r^2) and represents the percentage of variation in x that is accounted for by y. A hypothetical set of data from a study of the dose of a drug and its effect on heart rate could be subject to such an analysis. If at the end of the study a correlation analysis results in an r of 0.82, the r^2 equals 0.67 and indicates that 67% of the variation in heart rate can be associated with the change in dose of the drug. Correlation values can range from -1 to 1 where 0 would be no correlation, -1 is a perfect inverse correlation, and 1 represents a perfect positive

Etminan M, Levine M. Interpreting meta-analyses of pharmacologic interventions: the pitfalls and how to identify them. Pharmacotherapy 1999;19:741–5.

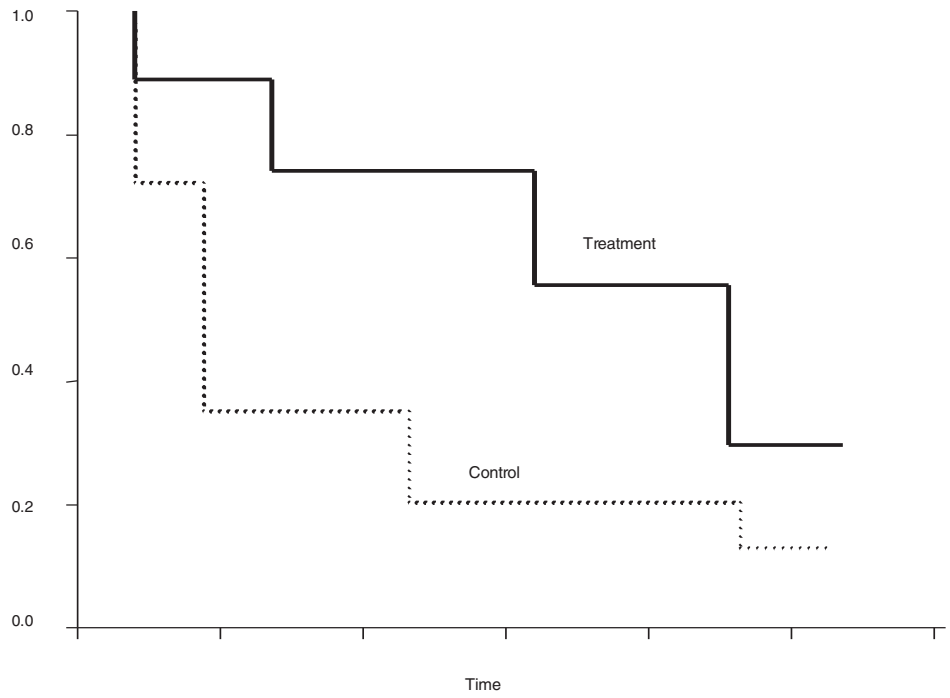


Figure 1-3. Example plot of survival analysis.
An example of a plot of a survival analysis showing the differences in proportions surviving during a hypothetical drug trial. The proportion surviving is plotted against time, with each time increment associated with a percentage of the initial sample still alive. The values used to construct the control and treatment lines are tested to determine if the differences between them are significant.

correlation. Of importance, correlation should not be assumed to imply that one change causes the other change.

Multivariable Analysis/Regression

Multivariable analysis takes the place of simple correlation for questions that involve more than two variables. It often is used to develop predictive models of risk, such as the Framingham model that predicts cardiovascular risk based on the values of several other variables (e.g., blood pressure, sex, and serum lipid values). Multiple linear regression accomplishes this task for continuous outcome variables. For dichotomous outcome variables, a technique named "logistic regression" is used. Regardless of the statistical tests used in performing the regression analysis, the general method involves adding or deleting variables in a model to see if an association between the variables in the model and the outcome variable exists. This process results in a model that should account for the maximum amount of variation in the outcome explained by the model. When a new variable is added or subtracted from the model to predict or explain variations in outcome, the results are compared to the previous model to check for statistically significant changes. This type of modeling becomes mathematically complex and is performed using statistical computer software. Models that use variables with dichotomous outcomes will have results reported as an OR. It is from these types of models that some statements of the form "having disease x increases the risk of outcome y (blank) times" can be derived.

Survival Analysis

Survival analysis uses the technique of proportional hazards regression (Cox) to assess the effects of two or more variables on the time to an event. A simpler test, the log-rank test, is used for the same purpose when examining only two variables. The log-rank test can be used to construct Kaplan-Meier curves, which estimate the proportion of people who would survive a certain length of time under the same conditions as the study. Here the term "survival" also can represent discrete outcomes other than death (i.e., the analysis examines time to an event, where the event may be something other than death). Common outcomes of interest in pharmacotherapy trials, including time to relapse, time to hospitalization, or time to death, are all analyzed using survival analysis. Survival analysis uses censored survival times in its methods. This term merely recognizes the fact that not all patients in a study will reach an end point of interest during the study period. Censored data are simply those data included in the analysis for which the time to event was not observed for that patient. Data may be censored for patients who withdraw from the study, are lost to follow-up, or for whom the event exceeded the study period. The results of survival analysis, regardless of the specific statistical techniques used, are reported using a format similar to that shown in Figure 1-3. Another term sometimes associated with these types of analyses is hazard, or hazard function. Hazard represents the chance that a patient will survive through a certain time interval in the study, not just the probability of surviving until the end of the study period.

Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med 2003;138:644–50.

Abbreviations

Interpreting Summary Reports of Effect

Absolute Versus Relative Changes

Simply put, absolute changes are more important than relative changes. The baseline rate of an outcome needs to be understood to make clinical sense of any changes in that outcome. Increasing the rate of a rare outcome (i.e., one in 100,000) by a factor of three (i.e., a 300% increase) has little clinical impact. The absolute change in risk in this scenario is 0.003% minus 0.001%, or two additional events for each 100,000 patients treated or exposed. The relative change, as previously discussed, can be described as a 3-fold increase in the number of events. Clearly, this latter number seems like a more notable result. The nature of relative changes in the reporting of pharmacotherapy trials can be just as misleading. Because relative changes do not give any indication of the usual rate of an event, they provide no useful information about whether an intervention should be used or avoided. The absolute changes represent more clearly the clinical importance of such changes. These distinctions are of enormous importance not just in clinical trials, but also in epidemiological (observational) trials where changes are always reported as a relative change.

Number Needed to Treat

The number needed to treat represents another way to characterize changes in absolute risk. The number needed to treat can be calculated easily by subtracting the absolute difference between groups and then taking the reciprocal of this difference. The resulting number represents the average number of patients who would need to be treated to have (or to prevent) one additional outcome of interest. For example, if a study found that 13% of patients randomized to receive active drug had an emergency department visit and that 36% of the placebo group had an emergency department visit, then the number needed to treat to prevent one emergency department visit would be calculated as:

$$1/(0.36 - 0.13)$$

This computation yields a number needed to treat of about four patients. Because it is an estimate, a CI can be calculated for the number needed to treat. The number needed to harm uses this same calculation to ascertain, for example, the number of patients who would need to receive a drug therapy to have one adverse event (e.g., rash and edema).

Examples of how absolute and relative differences result in much different numbers, along with number needed to treat calculations for these differences, are shown in Table 1-4. When the number needed to treat or number needed to harm is being used to summarize differences between groups, it should be remembered that the values only apply to the time intervals from which the data come. If a trial lasted 5 years, the number needed to treat would apply only to the outcome at 5 years. It is inappropriate to double or halve the 5-year number needed to treat to estimate a number needed to treat that would occur after 10 years. The use of the number needed to treat and number needed to harm is limited to nominal types of data as number needed to treat values are only useful in summarizing differences involving dichotomous outcomes.

Putting Results into Perspective

The distinctions between absolute and relative changes are of more than statistical importance. It is known that relative changes, because of their association with bigger numbers, often are viewed as more convincing by both

Table 1-4. Relative Changes Versus Absolute Changes Versus Numbers Needed to Treat—Primary End Points From the Heart Outcomes and Prevention Evaluation (HOPE) Study

	Primary end points after 4.5 years (all differences in outcomes, p<0.001)				
	Ramipril	Placebo	Relative Reduction[a]	Absolute Reduction	NNT[b]
Patients suffering a myocardial infarction	9.9%	12.3%	20%	2.4%	1/(0.123 - 0.099) = 42
Patients suffering a stroke	3.4%	4.9%	31%	1.4%	1/(0.049 - 0.034) = 67
Patients with death from a cardiovascular cause	6.1%	8.1%	25%	2.0%	1/(0.081 - 0.061) = 50
Patients suffering a myocardial infarction, stroke, or death from a cardiovascular cause	14.0%	17.8%	21%	3.8%	1/(0.178 - 0.140) = 26

[a]Calculated as (percentage in placebo group - percentage in ramipril group)/percentage in placebo group.
[b]Number needed to treat data from The Heart Outcomes Prevention Study Investigators. Effects of an angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients.
NNT = number needed to treat.
The Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. N Engl J Med 2000;342:145–53.

Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452–4.

clinicians and patients. Studies demonstrate that when the results of pharmacotherapy trials are reported as relative changes, patients and practitioners are more likely to be convinced of their importance. This phenomenon helps to explain why drug advertisements report relative changes almost exclusively. It also serves to remind clinicians that when discussing the risks and benefits of drug therapy with patients that presentations of both types of data may be necessary to fully convey the impact of a therapy in a way that is understandable. Keeping such numbers from becoming confusing may be a challenge, but one that can be overcome by avoiding technical terms or jargon. Using regular number descriptions, such as x out of 100 instead of percentages, and conveying changes in terms of absolute risk of an event have been suggested as ways to ensure a better understanding of the effects of pharmacotherapy.

Summary

Clearly, pharmacists must understand the nature and interpretation of statistical analysis to function effectively as drug experts. Statistics is a powerful tool to assist decision-making about the optimal use of drug therapies. Proper application of statistical results to pharmacotherapy decision-making is as essential as a knowledge and understanding of pharmacology for the pharmacist of today.

Annotated Bibliography

1. Riegelman RK. Studying a Study and Testing a Test—How to Read the Medical Evidence, 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2000.

   This book, now in its fourth edition, has served as a handy reference regarding the application of statistical tests to clinical questions for many years. Not quite a statistics text, yet more than an overview, this resource is divided into sections including study design and selecting statistics. Specific areas covered include uni-, bi-, and multivariable analyses. A sizable section on the basic principles of statistical analysis provides a great introduction to the assumptions of inferential statistics. The portion of the book dealing with meta-analysis is clear and easy to understand. This book provides some detail about the calculation methodology of some tests and techniques but is intended for the nonmathematician. The book also includes separate sections on determining the utility of diagnostic tests and understanding rates in health care. This is a highly accessible resource for clinicians who do not deal with these topics frequently.

2. STATS—Steve's Attempt to Teach Statistics. Available at *http://childrens-mercy.org/stats*. Accessed October 5, 2004.

   There are many Web sites that deal with the topic of statistics. This site, developed by a health care statistician, is one of the most straightforward. The "Ask Professor Mean" page presents, in question-and-answer format, many topics that a pharmacist might seek a clearly worded explanation (or refresher). Within this page are answers to common questions about confidence intervals (CIs), specific statistical tests, descriptive statistics, decision errors, parametric versus nonparametric tests, and many other topics. There also is even a link that explains "obscure statistical jargon". Other parts of the Web site include explanations that may be informative for those who are actively involved in research. Some of the details are intended for people seeking information about the underlying or necessary mathematics associated with particular statistical tests. As a resource currently available at no cost, STATS provides an easy-to-understand, practical, and sometimes humorous approach to the topics of biostatistics.

3. Greenhalgh T. How to read a paper: statistics for the non-statistician. I. Different types of data need different statistical tests. BMJ 1997;315:364–6.

4. Greenhalgh T. How to read a paper: statistics for the non-statistician. II. "Significant" relations and their pitfalls. BMJ 1997;315:422–5.

   These two references are a series by a well-known author in the field and address the basics of data analysis. The focus of these articles is on the nature of different variable types and common problems in interpreting p values, respectively. As the titles imply, the articles do not assume a background in calculus or much previous exposure to statistics. If readers' recollections of introductory statistics are more than a little fuzzy, these two short, easy-to-understand articles are a great way to fill in the blanks.

5. BMJ Publishing Group. Statistics at Square One. Available at *http://bmj.bmjjournals.com/collections/statsbk/index.shtml*. Accessed October 5, 2004.

   Another resource available without charge, Statistics at Square One is an online version of the often-recommended book by the same title. The table of contents lists 13 chapters, each dealing with a separate area of statistics: data display and summary; mean and standard deviation; populations and samples; statements of probability and CIs; differences between means: type I and type II errors and power; differences between percentages and paired alternatives; the t tests; the chi-square tests; exact probability test; rank score tests; correlation and regression; and survival analysis. Health care providers are the intended audience for this text and most of the examples use scenarios familiar to pharmacists. Chapters provide ample demonstration of the techniques discussed and include examples of calculations for many statistical tests.

6. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM, 2nd ed. Edinburgh, Scotland: Churchill Livingstone, 2000.

   This book encompasses more than just statistics. As the title implies, the broader question of how to use the techniques of evidence-based medicine is addressed throughout each chapter. However, within this context, many useful points are made about the use of statistical results in arriving at an evidence-based recommendation. The chapters on therapy and harm are of particular interest to pharmacists. In addition, an appendix presenting methods to calculate standard errors and CIs for several common types of results can be handy when assessing trials where such calculations are not reported. The glossary at the back of the book also is

Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003;327:741–4.

useful. Because of its broader focus, this text provides a further understanding of biostatistics by placing them into the more comprehensive framework of clinical decision-making.

7. Rosner B. Fundamentals of Biostatistics, 3rd ed. Boston, MA: PWS-Kent Publishing Company, 1990.

This textbook often is used in college biostatistics courses. It includes a large amount of detail about the mathematical calculations necessary to derive a value for a given test statistic. Because of the detail, this book allows the reader to better understand biostatistical concepts and applications in a clear and logical fashion. All examples used in this book are related to health care, so the inquiring pharmacist should feel at home with the exercises that illustrate the use of these methods. The book also contains many statistics problems, with answers to some, which can aid in reinforcing the understanding of biostatistics.

Abbreviations

SELF-ASSESSMENT QUESTIONS

1. You wish to describe the types of patients who use the anticoagulation monitoring services that your department provides. Specifically, the pharmacy and therapeutics committee is interested in the age of the population that is served. Which one of the following pieces of information is most useful to the committee members?

   A. Standard error of the mean (SEM).
   B. T test.
   C. Mean.
   D. Chi-square test.

As part of a process improvement committee, you are responsible for determining the impact of a recent educational campaign to improve the recording of patient allergies in the medical record. Before the education efforts, you record the allergy status of 100 patients on one unit using their admission orders. After the education, you assess the allergy status of another 100 patients from another unit using the same method. Your results are as follows:

	Allergy recorded	Allergy not recorded
Before education	78	22
After education	90	10

2. In preparing to analyze the results of this intervention statistically, you consider the data in the table as which one of the following types?

   A. Nominal.
   B. Ordinal.
   C. Interval.
   D. Ratio.

3. A study is conducted to test if the use of a new drug adherence aid has an effect on congestive heart failure exacerbations. Researchers want to evaluate the proportions of patients with an exacerbation of their congestive heart failure symptoms in a group of people using the adherence aid compared to a group who did not. To assess if the changes observed are statistically significant, which one of the following tests is best?

   A. McNemar.
   B. Chi-square.
   C. Two-sample t test.
   D. Mann-Whitney *U* test.

4. In a cohort study designed to determine an association between measles, mumps, and rubella vaccination and autism, investigators report the relative risk of autistic disorder in the vaccinated group compared to the unvaccinated group as 0.92 (95% confidence interval [CI] = 0.65–1.07). Which one of the following p values is consistent with these reported findings?

   A. A p value of less than 0.05.
   B. A p value of less than 0.01.
   C. A p value of greater than 0.05.
   D. A p value of greater than 0.10.

5. A prospective, randomized, placebo-controlled trial of a new antidepressant drug reports that for the primary outcome of response rate (50% decrease in Hamilton Rating Scale for Depression) there is no difference between the drug and placebo ($p>0.05$). The researchers also report that they decided to do an additional previously unplanned analysis of the data after the conclusion of the trial; and that they were able to demonstrate a better response rate for the new drug versus placebo in the women in the trial ($p=0.04$). Which one of the following is the most valid conclusion from this trial?

   A. The new drug works in women but not in men.
   B. The trial should have listed two primary outcomes.

C. The response rate reported for the entire group of participants should be analyzed as a secondary outcome.
D. A prospective trial designed to test the drug in men compared to women should be considered.

6. A systematic review evaluated the effect of bisphosphonates on skeletal metastasis in patients with cancer. A meta-analysis that included placebo-controlled trials of at least 6 months duration showed a combined odds ratio (OR) for the use of radiation therapy of 0.67 (95% CI = 0.57–0.79). For spinal cord compression, bisphosphonates compared to placebo resulted in a combined OR of 0.71 (95% CI = 0.47–1.08). Which one of the following is the best interpretation of these results?

A. Bisphosphonates favorably impact the use of radiation therapy and the occurrence of spinal cord compression in patients with cancer with skeletal metastasis.
B. Bisphosphonates affect neither the use of radiation therapy nor the occurrence of spinal cord compression in patients with cancer with skeletal metastasis.
C. Bisphosphonates favorably impact the use of radiation therapy but not the occurrence of spinal cord compression in patients with cancer with skeletal metastasis.
D. Bisphosphonates do not affect the use of radiation therapy but do favorably impact the occurrence of spinal cord compression in patients with cancer with skeletal metastasis.

7. The Women's Health Initiative Study compared the use of conjugated equine estrogens plus medroxyprogesterone to placebo in healthy postmenopausal women and reported a higher number of cardiovascular events in women receiving this hormone replacement therapy regimen. Which one of the following correct descriptions of the results about cardiovascular disease reported in the trial would be best to use in discussions with patients?

A. Cardiovascular events increased in women taking the active drug.
B. Patients taking the drug had an increased risk of cardiovascular events that was statistically significant.
C. For every 10,000 women who take the drug for 1 year, there will be seven extra cardiovascular events.
D. The rate of cardiovascular events was increased by 29% in women taking hormone replacement therapy.

8. A trial (n=48) reports that the average dose of an intravenous analgesic drug needed to keep postoperative pain below a rating of 2 on a 10-point scale is 67 mg with a standard deviation of ± 17 mg. You wish to calculate a CI for this mean dose. Which one of the following is the SEM associated with this result?

A. 0.35.
B. 1.39.
C. 2.45.
D. 6.93.

Questions 9 and 10 pertain to the following case.
A trial compares drug X and drug Y for treating nausea and vomiting associated with pregnancy because clinicians believe that there may be differences in their efficacy in preventing nausea and vomiting in pregnant women. Drug X has been used for many years and has a large evidence base demonstrating efficacy and safety. Drug Y recently has been introduced to treat nausea and vomiting associated with chemotherapy, but has not been well studied in patients with nausea and vomiting due to pregnancy. Patients will be randomized to one of these two drugs.

9. Which one of the following represents the correct statement of the null hypothesis (H_0) for this trial?

A. Efficacy of drug X equals the efficacy of drug Y.
B. Efficacy of drug X is not equal to the efficacy of drug Y.
C. Efficacy of drug X is greater than the efficacy of drug Y.
D. Efficacy of drug X is less than the efficacy of drug Y.

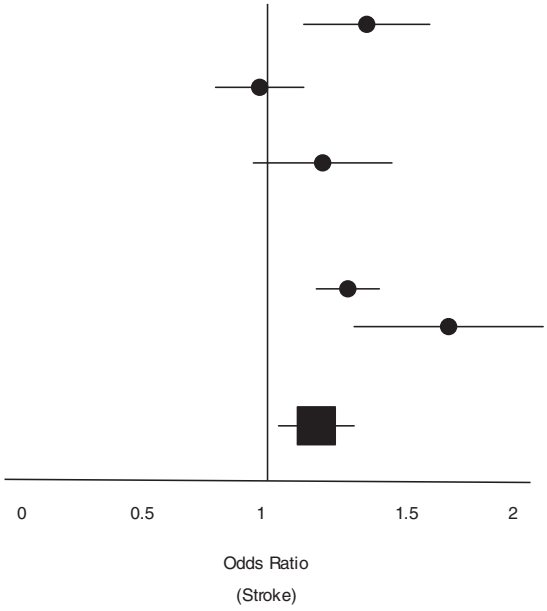
10. The end point for this trial will be based on patients' ranking of their nausea and vomiting 3 hours after taking the drug to which they have been assigned. Nausea and vomiting will be graded using the following scale: 0 = no nausea, 1 = mild nausea, 2 = moderate nausea, 3 = severe nausea, and 4 = vomiting. Which one of the following types of statistical test is best suited to test for differences between these drugs?

A. Wilcoxon rank sum test.
B. Wilcoxon signed rank test.
C. Student t test.
D. Paired Student t test.

11. A prospective, randomized, double-blind study (n=2200) finds that when comparing two oral drugs for treating type 2 diabetes, the final mean percent hemoglobin A1C value for patients in group 1 is 8.21 and for patients in group 2 is 8.27. This difference between the groups is stated as having a calculated p value of less than 0.05. Assuming similar baseline characteristics and appropriate final statistical analysis, which one of the following statements best characterizes these findings?

A. The difference between the drugs is not statistically significant, but is clinically significant.
B. The difference between drugs is both statistically and clinically significant.
C. The difference between the drugs is statistically significant, but not clinically significant.
D. The difference between the drugs is neither statistically significant nor clinically significant.

Question 12 pertains to the following figure.



12. The results of a meta-analysis investigating the effects of a drug on stroke are shown in the figure above. Which one of the following is the best interpretation of these results as portrayed in this figure?
    A. The drug being studied does not have an effect on the frequency of stroke.
    B. The drug being studied decreases the frequency of stroke.
    C. The drug being studied increases the frequency of stroke.
    D. The drug's effect on stroke cannot be determined from the figure.

13. A new antipsychotic drug has been compared to a previously available drug in a prospective, randomized, and blinded trial. This 12-month trial measured the frequency of inpatient psychiatric admissions in both treatment groups. At the end of the trial, 6% of the patients taking the new drug had an inpatient psychiatric admission compared to 11% of patients taking the older drug (p=0.03). In presenting the results of this trial to members of the formulary committee at the health maintenance organization where you work, which one of the following statements would provide the committee with the best information to use in deciding the formulary status of the new drug?
    A. The new drug decreased psychiatric admissions by 45%.
    B. The new drug decreased psychiatric admissions by 83%.
    C. You would need to treat two patients with the new drug to avoid one admission.
    D. You would need to treat 20 patients with the new drug to avoid one admission.

14. A trial studied an antihypertensive drug to assess its effects on blood pressure. Researchers compared the blood pressure of 150 patients at baseline and then again after having taken the drug for 2 weeks. The results show that this antihypertensive drug lowers systolic blood pressure by an average of 9 mm Hg (p=0.04; two-sample t test) and diastolic blood pressure by an average of 7 mm Hg (p=0.03; two-sample t test). Which one of the following statements is consistent with these reported results?
    A. This drug is not effective at lowering blood pressure.
    B. This drug is effective at lowering only diastolic blood pressure.
    C. This drug is effective at lowering both diastolic and systolic blood pressure.
    D. This drug may or may not lower blood pressure; the results are unreliable.

15. A case-control study is performed to judge whether a drug is associated with an increased incidence of early miscarriage. The final analysis shows that the OR for miscarriage with drug exposure is 1.3 (95% CI = 0.9–1.7). Which one of the following provides a correct description of these results?
    A. This drug increases the risk of miscarriage by 70%.
    B. This drug increases the risk of miscarriage by 30%.
    C. This drug decreases the risk of miscarriage by 10%.
    D. This drug is not associated with an increased risk of miscarriage.

Questions 16 and 17 pertain to the following case.
A linear relationship between the dosage of a new chemotherapeutic drug and pulmonary function is being investigated. Measurements of the forced expiratory volume in 1 second are collected as a measure of lung function and plotted against the corresponding dosage of drug that each patient received.

16. Which one of the following is an appropriate statistical approach to assess any correlation between drug dose and forced expiratory volume in 1 second?
    A. Pearson product moment coefficient.
    B. Analysis of variance.
    C. Spearman rank correlation.
    D. Analysis of covariance.

17. The statistical test applied to the data studying the relationship between this drug and lung function reports an r = -0.46 ($p<0.05$). Which one of the following represents the best interpretation of these results?
    A. Seven percent of the variation in forced expiratory volume in 1 second is associated with the dose of this drug.
    B. Twenty-one percent of the variation in forced expiratory volume in 1 second is associated with the dose of this drug.

C. Forty-six percent of the variation in forced expiratory volume in 1 second is associated with the dose of this drug.
D. Ninety-two percent of the variation in forced expiratory volume in 1 second is associated with the dose of this drug.

18. The manager of an obesity clinic in your health care system approaches you about selling a recently marketed herbal weight-loss supplement in the clinic. She tells you that unlike other products making claims about weight loss, this product has been described to her as containing no ephedra (ma huang) or other stimulants, and no dangerous herbal derivatives. She goes on to show you a copy of a trial proving that this supplement works. You review the study that claims this product increases metabolism. The study shows that patients taking the supplement burned an average of 20 calories more during a 700-calorie workout than those who were not taking the supplement ($p<0.05$). Which one of the following is an appropriate response to the manager based on the information provided?
    A. The statistical differences show that the product is worth using.
    B. The differences demonstrated do not appear to be clinically significant.
    C. The study shows the product is effective regardless of the p value.
    D. The study demonstrates neither statistical nor clinical differences.

19. A new laxative has been compared to psyllium in adults and children between the ages of 2 and 65 years. The study assesses the length of time to the first bowel movement after taking one of the two study drugs. At the end of the trial, a statistical analysis of the outcomes resulted in a p value of 0.30. Based on this result, the researchers report that they then decided to look for differences in effect between men and women and among different age groups (i.e., 2–5 years old, 6–12 years old, 13–18 years old, 19–55 years old, and older than 55 years). At the end of these analyses, the new laxative was found to provide superior relief of constipation in women older than 55 years of age ($p<0.05$). Which one of the following is the best interpretation of these results?
    A. This drug only works better than psyllium in women older than 55 years of age.
    B. Until more trials are conducted, it should be concluded that this drug works no better than psyllium.
    C. The researchers found no overall difference, so it cannot work better in certain subgroups.
    D. Because it works better in women older than 55 years of age, it should work in all women.

20. A systematic review reports that its meta-analysis includes studies of the following sizes: two studies with sample sizes between 200 and 400 patients; four studies with sample sizes between 401 and 1000 patients; one study with a sample size of 1100 patients; and one study with a sample size of 5200 patients. In reviewing the results for use in practice, which one of the following is the most important type of analysis to seek out?
    A. A regression analysis of variables that might increase the risk for the outcome of interest.
    B. A sensitivity analysis and tests for heterogeneity.
    C. A calculation of the hazard function for the total number of patients.
    D. A Cox regression analysis.

21. A randomized, double-blind trial is conducted to test the hypothesis that a new vasodilator improves the symptoms of congestive heart failure. Patients are randomized to either the new drug or to placebo and continue their existing congestive heart failure pharmacotherapy. At the conclusion of the trial, the intention-to-treat analysis shows no statistically significant difference between the groups. The authors then decide to perform an analysis based on the actual therapies that patients received. Data for patients who did not finish at least 60% of their assigned drug were evaluated as if those patients were in the placebo group. This analysis showed that the new drug was more effective than placebo ($p<0.05$) in lessening the symptoms of congestive heart failure that were measured. Which one of the following is the best course of action based on these results?
    A. Recommend the new drug for all patients with congestive heart failure.
    B. Recommend the new drug for patients who adhere to their current therapies.
    C. Only recommend the new drug for early-stage congestive heart failure.
    D. Do not recommend the new drug; wait for further studies.