Abbreviations

BIOSTATISTICS FOR THE CLINICIAN

Mary Lea Harper, Pharm.D.

Learning Objectives

1. Understand when to use and how to calculate and interpret different measures of central tendency (mean, median, and mode) and dispersion (range, interquartile range [IR], and standard deviation [SD]).
2. Identify the types of error encountered in statistical analysis, the role of sample size, effect size, variability, and power in their occurrence, and implications for decision making.
3. Describe basic assumptions required for utilization of common statistical tests including the Student's t-test, paired t-test, Chi square analysis, Fisher's exact test, Mann-Whitney U test, Wilcoxon signed rank, log rank test, and Cox Proportional Hazards model.
4. Understand the common statistical tests used for analyzing multiple variables, including one-way and two-way (repeated measures) analysis of variance.
5. Interpret a confidence interval (CI) with inferential functions.
6. Differentiate between association and correlation analysis, interpret correlation coefficients and regression coefficients, and describe the application of single and multivariable statistical models.
7. Describe and interpret common statistical techniques used in performing meta-analysis.

Introduction

Pharmacists need to have a basic understanding of statistical concepts to critically evaluate the literature and determine if and how information from a study can be applied to an individual patient. By using data from well-designed studies that have been appropriately analyzed and interpreted, clinicians are able to determine how the "average patient" might respond to a new medication. The problem of course, is how many times do clinicians care for "average patients"? Statistics would be much less sophisticated if every individual were alike; response to therapy would be highly predictable. The science of statistics allows clinicians to describe the profile of the average person, then estimate how well that profile matches that of others in the population under consideration. Statistics do not tell the reader of the clinical relevance of data. When a difference is considered to be statistically significant, the reader rejects the null hypothesis, and state that there is a low probability of getting a result as extreme as the one observed with the data. Clinicians need to interpret this data and to guide the clinical decision-making process (i.e., the decision to use or not use a medication for a specific patient). By staying current with the literature and critically evaluating new scientific advancements, clinicians are able to make educated decisions regarding how to care for patients.

Several published studies have critiqued the selection and application of statistics in the medical literature. These articles, many in well-respected medical journals, have consistently found high rates of inappropriate application, reporting, and interpretation of statistical information. Although it is critical to collaborate with a biostatistician in all phases of research, it is not necessary to be a statistician to present or interpret basic statistics appropriately. This chapter provides an overview of basic concepts related to

Abbreviations in this Chapter

NSAID	Nonsteroidal anti-inflammatory drug
SEM	Standard error of the mean
IR	Interquartile range
SD	Standard deviation
df	Degrees of freedom
CI	Confidence interval
ANOVA	Analysis of variance
H_0	Null hypothesis

descriptive and inferential statistics. To begin, one needs to have a basic understanding of terminology commonly used in statistics, including variables and types of data.

Variables

A variable is any characteristic that is observed or measured (e.g., sex, baseline fasting blood sugar, or weight). Variables may be described as either independent (predictor) or dependent (response). The independent variable is the intervention, or what is being manipulated in a study. Most statistical tests require at least one independent variable that is established in advance and controlled by the researcher. The dependent variable is the outcome of interest, which should change in response to some intervention. At least one or more dependent variables are then measured against their independent counterparts. For example, in a study that compares a new nonsteroidal anti-inflammatory drug (NSAID) to standard therapy for the treatment of pain, the degree of symptom relief (dependent variable) depends on whether or not the patients received the new NSAID (independent variable). Independent variables that can affect the dependent variable are called confounding variables. These variables must be controlled through the design of the study or analysis. An example of a confounding variable is severity of disease. If there are more patients with severe disease in one group than in the other, this may ultimately affect the outcome of the study.

Types of Data

Variables are frequently classified as nominal, ordinal, interval, or ratio. Nominal data can be sorted into one of a limited number of categories, but the categories cannot be ordered. An individual may belong to one and only one group. Examples of nominal data include demographic data such as sex (i.e., male or female) and risk factors (e.g., smoker or nonsmoker), or may include a clinical outcome (e.g., presence or absence of disease). Nominal variables do not have to be dichotomous, they can have any number of categories. For instance, a patient may have blood type A, B, AB, or 0. The important concept is that any one blood type is not better or worse than another. When entering these data into a spreadsheet, researchers commonly code nominal data by assigning a number to each value. For example, female patients may be coded as "0" and male patients as "1". The ordering is arbitrary and no information is gained or lost because of the order. The conclusions that are drawn will be identical regardless of the ordering of the samples. Nominal data are described by using frequencies (e.g., percent of females).

Ordinal variables are also sorted into mutually exclusive groups based on some common characteristic that all members of the group possess. However, unlike nominal data, ordinal data are sorted by categories often with numbers denoting a rank order. Again, whether numbers are assigned to these values is irrelevant. Ordinal data are used when the relative degree of presence or absence of a certain characteristic can be measured qualitatively, not quantitatively. The magnitude of difference is either unequal or unknown. The type of data is often collected when the evaluation is subjective, such as when assessing patient attitudes or clinical symptoms. For example, when patients complain of a headache during a study, they may be asked to describe the severity of that headache on a scale of 1–4 (1 = mild, 2 = moderate, 3 = moderately severe, and 4 = severe). Although a rating of 4 is worse than 2, clinicians do not really know by how much. When describing the severity of headache, a number 4 is not necessarily twice as severe as a number 2. Because these are not real numbers, arithmetic means are not generally calculated with ordinal data. Median and frequency are used to describe this type of data.

Interval and ratio are the most powerful and specific type of data. Unlike ordinal data, the distance between the consecutive numbers on the scale is constant, and therefore, one can appropriately perform arithmetic (e.g., sum, difference, multiplication, or division). Interval and ratio data are equivalent in all characteristics except that ratio data have a true zero. Examples of interval and ratio data are body temperature and body weight, respectively. Interval and ratio data may be described using the mean or median.

The terms discrete or continuous are also used to describe interval and ratio data. Data are considered to be discrete if the observations are integers that correspond with a count of some kind. These variables can take on a limited number of values. For example, if a patient was asked to rate his or her pain on a 5-point scale where only the values 1, 2, 3, 4, and 5 were allowed, only five possible values could occur. Such variables are referred to as "discrete" variables. In contrast, data are considered to be continuous if each observation theoretically falls on a continuum. Continuous data may take any value within a defined range; the limiting factor is the accuracy of the measuring instrument. Examples of continuous data include uric acid or glucose blood levels, body temperature, and body weight. Although it is feasible to consider a body temperature of 98.6, one does not discuss the concept of counting white blood cells as a percent of a cell. The concept of data being described as discrete or continuous will be important when examining the assumptions for statistical testing later in the chapter.

Investigators may transform the data collected to a lower type of data. In other words, interval and ratio data may be reported as ordinal or nominal, and ordinal data may be reported as nominal. Nominal data may not be reported as ordinal, interval, or ratio data. For example, measurement

of blood pressure is normally collected using an interval scale. It could also be reported as follows:

≤ 90 mm Hg	10 patients
> 90 mm Hg	16 patients

In this example, data from patients are inserted into one of two mutually exclusive categories. It is not possible for patients to fit into more than one category. The occurrence of one event excludes the possibility of the other event. This is an example of how nominal data can be presented. It is generally undesirable to transform data to a lower level because information is lost when individuals are collectively included in more general categories. If data are transformed, it should be presented in both ways. For instance, in the presentation of blood pressure measurements, it may be worthwhile to present the mean change in blood pressure (ratio data), as well as the numbers of patients that achieve goal blood pressure values (nominal data).

Descriptive Versus Inferential Statistics

Descriptive statistics are concerned with the presentation, organization, and summarization of data. In contrast, inferential statistics are used to generalize data from our sample to a larger group of patients. A population is defined as the total group of individuals having a particular characteristic (e.g., all children in the world with asthma). A population is rarely available to study as it is usually impractical to test every member of the population. Therefore, inferences are made from taking a randomized sample of the population. A sample is a subset of the population. Inferential statistics require that the sampling be random; that is, each patient has an equal chance of receiving either treatment. Some types of sampling seek to make the sample as representative of the population as possible by choosing the sample to resemble the population on the most important characteristics (surveys for assessing medication histories relating to risk of side effects). For instance, when investigators study a new NSAID in 50 geriatric patients with osteoarthritis, they are not just interested in how these patients in the study respond, but rather they are interested in how to treat all geriatric individuals with osteoarthritis. Thus, they are trying to make inferences from a small group of patients (sample) to a larger group (population).

Frequency Distribution

Data can be organized and presented in such a way that allows an investigator to get a visual perspective of the data. This is accomplished by constructing a frequency distribution. These distributions may be described by the coefficient of skewness or kurtosis. Skewness is a measure of symmetry of a curve. A distribution is skewed to the right (positive skew) when the mode and median are less than the mean. A distribution that is skewed to the left (negative skew) is one in which the mode and median are greater than the mean. The direction of the skew refers to the direction of the longer tail. Kurtosis refers to how flat (platykurtic) or peaked (leptokurtic) the curve appears. The frequency distribution histogram (i.e., a type of bar graph representing an entire set of data) is often a symmetric, bell-shaped curve referred to as a normal distribution (also referred to as Gaussian curve, curve of error, and normal probability curve). Under these circumstances (i.e., normal distribution), the mean, median, and mode are all similar, and the kurtosis is zero. The mean plus one standard deviation (SD) includes approximately 68% of the data. The assumption that there is normal distribution of variables in the population is important because the data are easy to manipulate. Several powerful statistical tests (e.g. Student's t-test, as well as other parametric tests) require a normal distribution of data. However, the Student's t test and other parametric tests assume rather than require normal distributions. The central limit theorem states that given a distribution with a mean (m) and variance (s^2) the sampling distribution of the mean approaches a normal distribution with a mean (m) and a variance (s^2)/N as N (sample size) increases. This assumption is based on the premise that when equally sized samples are drawn from a non-normal distribution from the same population, the mean values from the samples will form a normal distribution, regardless of the shape of the original distribution. For most distributions, a normal distribution is approached very quickly as the sample increases (e.g., N>30).

Mean, Median, and Mode

There are three generally accepted measures of central tendency (also referred to as location): the mean, median, and mode. The mean (denoted by $\bar{x}$) is one acceptable measure of central tendency for interval and ratio data (Table 1-1). It is defined by the summation of all values (denoted by X for each data point) divided by the number of subjects in the sample (n) and can be described by the equation

$$\bar{x} = \sum x/n.$$

For instance, the mean number of seizures during a 24-hour period in seven patients with the following values 9, 3, 9, 7, 8, 2, 5, is calculated by dividing the sum of 43 by 7, which is equal to 6.14, or an average of approximately six seizures per patient.

The median is the value where half of the data points fall above and half below it. It is also referred to as the 50th percentile. It is an appropriate measure of central tendency for interval, ratio, and ordinal data. When calculating the median, the first step is to put the values in rank order. For example, the median number of seizures during a 24-hour period in seven patients with the following values 2, 3, 5, 7, 8, 9, 9 is 7. There are three values below and three values above the number 7. If we added one more value (e.g., 11), the median would be calculated by taking the two middle numbers and dividing by two. Under these circumstances, the calculation would change to (7 + 8)/2 to get a median of 7.5. Half of the numbers are below 7.5 and half are above. The median has an advantage over mean in that it is affected less by outliers in the data. An outlier is a data point that is an extreme value either much lower or higher than the rest of the values in the data set. Mathematically, outliers can be determined by using the following formulas: values greater than 1.5 times the interquartile range (IR) plus the upper

Table 1-1. Common Statistical Applications

Type of Data	Measures of Location	Measures of Variability	Common Statistical Tests for Independent Groups of Data	Common Statistical Tests for Paired Groups of Data
Nominal	Mode	None	**2 groups of data:** Chi square **3 or more groups of data:** Chi square	**2 groups of data:** McNemar's Test **3 or more groups of data:** Cochrane Q
Ordinal	Median and mode	Range and interquartile range	**2 groups of data:** Wilcoxon rank sum or Mann-Whitney U **3 or more groups of data:** Kruskal-Wallis test	**2 groups of data:** Wilcoxon signed rank test **3 or more groups of data:** Friedman two-way analysis of variance
Interval/Ratio	Mean, median, and mode	Range, interquartile range, and standard deviation	**2 groups of data:** Student's t-test **3 or more groups of data:** One-way analysis of variance	**2 groups of data:** Paired t-test **3 or more groups of data:** Two-way (repeated measures) analysis of variance

quartile or values less than the lower quartile minus 1.5 times the IR are often considered outliers. A more extensive description of IR is described below.

The measure of central tendency for nominal data is the mode. The mode is the most frequently occurring number in a dataset. The mode of the above series of numbers is 9. The mode is not frequently presented in clinical trials unless a large data set is described.

Measures of Dispersion—Range, Interquartile Range, and Standard Deviation

The measure of dispersion (also referred to as measures of variability or spread) describes how closely the data cluster around the measure of central tendency. Data points that are scattered close to the measure of central tendency give a different perspective than those not as close to the value. For instance, the mean may be seemingly very different between two groups, but when examining data with a large amount of variability around the mean, they may begin to look similar. There are three common measures of dispersion: the range, IR, and SD. The range is defined as the difference between the highest and lowest values. If we had the same numbers as above (i.e., 2, 3, 5, 7, 8, 9, 9) to describe the total number of seizures at baseline, the range of values is 9-2 or 7. Frequently, when presenting data in clinical trials, investigators will describe this data as "9-2", and not use the single number. This provides more information about the sample. There is an advantage in providing the mean plus the range over providing the mean alone. For instance, a mean age of 50 in a study without a measure of dispersion gives a different sense of the data than when you tell individuals that the range included individuals from 12 to 78 years of age. This tells the reader that data were obtained from both the adolescent and geriatric populations. The disadvantage of the range over other measures of dispersion is that it is greatly affected by outliers. In the above example, if there was only one 12 year old in the study, and the next youngest individual in the study was 47, the range was greatly affected by this one outlier.

The IR is another measure of spread or dispersion. The lower quartile is also referred to as the 25th percentile and the upper quartile is the 75th percentile. The IR is defined as the difference between the lower quartile (often referred to as Q_1) and the upper quartile (often referred to as Q_3) and comprises the middle 50% of the data. The formula for the IR is $Q_3 - Q_1$. To determine the IR, the numbers are again sorted in rank order. Consider the following example of number of seizures:

2	3	5	7	8	9	9
	Q_1		Q_2		Q_3	

In this example, the first quartile is 3, the second quartile (which is the median) is 7, and the third quartile is 9. The IR is 3 to 9. Although the IR is not used extensively, it is considered to be underutilized because it is considered a stable measure of spread.

The SD is the most widely used measure of dispersion. It is defined as an index of the degree of variability of study data about the mean. For example, assume you need to determine what the mean and SD for the following data set of scores on a test are: 56, 62, 52, 50, and 45. The sample mean is 53 mm Hg (56 + 62 + 52 + 50 + 45)/5. The deviations are calculated in Table 1-2.

The sum of the deviations will always be zero. When the sum of the squared differences between the individual observations and the mean is computed, and this value is divided by the degrees of freedom (df), it produces an intermediate measure known as the sample variance (s^2). The degrees of freedom (n-1) are used in this equation to correct for bias in the results that would occur if just the

Table 1-2. Calculation of Standard Deviation for a Group of Test Scores

Observation X	Deviation $X-\bar{X}$	Squared Deviation $(X-\bar{X})^2$
56	3	9
62	9	81
52	-1	1
50	-3	9
45	-8	64
365	**0**	**164**

number of observations (n) was used. In general, the degrees of freedom of an estimate are equal to the number of independent scores that go into the estimate minus the number of parameters estimated. If the average squared deviation was divided by n observations, the variance would be underestimated. As the size of the sample of data increases, the effect of dividing by n or n-1 is negligible. The sample SD, equal to the square root of the variance, is denoted by the letter *s* as defined by the following formula:

$$s = \frac{\sqrt{\sum(X-\bar{x})^2}}{n-1}$$

Using this formula, the SD is the square root of 164/4 or 6.4. From this example, one can see that each deviation contributes to the SD. Thus, a sample of the same size with less dispersion will have a smaller SD. For example, if the data were changed to: 55, 52, 53, 55, and 50, the mean is the same, but the SD is smaller because the observations lie closer to the mean.

The usefulness of the SD is related only to normally distributed data. If one assumes that data are normally distributed, then one can say that one SD below and above the mean includes approximately 68% of the observations, two SDs above and below the mean include approximately 95% of the observations, and three SDs in either direction include approximately 99% of the observations. The histogram in Figure 1-1 describes the distribution of test scores for a larger sample.

In Figure 1-1, the mean was calculated to be 37.78 and SD is 13.15. Therefore, approximately 68% of the values will be between 24.63 and 50.93 (mean ± 1 SD), approximately 95% of individuals will have scores between 11.48 and 64.08 (mean ± 2 SD), and approximately 99% of the sample will be between 0 and 87.23 (mean ± 3 SD).

The SD and standard error of the mean (SEM) are frequently confused terms. The SD indicates the variability of the data around the mean for a sample, whereas the SEM is a measure of precision for the estimated population mean. This estimate is most commonly expressed using a confidence interval (CI) and is related to SD by the equation:

$$SEM = SD/\sqrt{n}.$$

The use of CIs is important in hypothesis testing and is described later in the chapter.

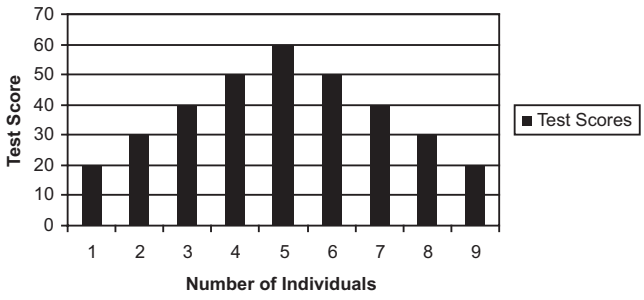

Figure 1-1. Test scores.

Hypothesis Testing and Meaning of P

A hypothesis is an unproved theory. The null hypothesis is defined as the theory that no difference exists between study groups. If a study were to compare two means, the null hypothesis (H_0) is $\mu_A = \mu_B$ (i.e., the population mean of group A is equal to the population mean of group B). The alternate (or research) hypothesis is the theory that a difference does exist between groups. This may be that the mean of group A is greater or less than the mean of group B ($\mu_A > \mu_B$ or $\mu_A < \mu_B$). If the change can be in either direction (i.e., μ_A is not equal to μ_B), this is a two-tailed test of significance. If a change is in only one direction (e.g., $\mu\mu_A > \mu_B$), then a one-tailed test of significance is used. This has implications for type I error rate (also referred to as alpha or α), or p value.

One needs to have a basic understanding of probability to appreciate the meaning of p. Probability deals with the relative likelihood that a certain event will or will not occur, relative to some other events. The probability is always a number between 0 and 1. The concept of probability is discussed only in the context of a chance operation; that is, an operation whose outcome is determined at least partially by chance. This can be illustrated with a coin toss. In this case, the chance operation is a toss of the coin. The event is heads. Each time the coined is tossed, it either falls heads or it does not. If the coin is equally likely to fall heads or not, then the probability is 0.5. The p value in clinical trials is the probability that chance alone would yield a difference among the groups as large or larger then the observed if the null hypothesis is really true. In other words, it is the probability that a type I error was committed. In general, the p value should not exceed 0.05 to reject the null hypothesis. In other words, there is a one in 20 (5%) chance that the investigator will be wrong in concluding that a difference exists between the study groups. The 0.05 threshold is an entirely arbitrary level and has been a subject of much debate in the literature. Once the alpha level has been set, the researcher collects the data and is interested in determining if there is a statistically significant difference. Once a statistical test is selected, a "t statistic" is calculated and a p value is determined. The p value is the probability of obtaining a result as extreme or more extreme than the actual sample value obtained given that the null hypothesis is true. If the p value is less than or equal to the alpha level established (typically set at a threshold of 0.05), then the null hypothesis is rejected and the difference between the groups is considered to be statistically significant. If the p value is greater than the alpha established (typically set at a threshold of 0.05), the null hypothesis is accepted and the difference between the groups is not considered to be statistically significant. Any measurement based on a sample of individuals will differ from the true value by some amount as a result of the random process. Whenever two treatments are compared, some differences will be present purely by chance. This is referred to as random error. As a result, unless one takes the role of chance into account, every experiment will conclude that one treatment is better than another.

The use of a one-tailed or two-tailed test can have implications on the risk of making a type I error. A two-tailed test is preferred in hypothesis testing; if

investigators use a one-tailed test, they need to justify its use. There are two ways in which the type I error can be distributed. In a two-tailed test, the rejection region is equally divided between the two ends of the sampling distribution. A sampling distribution can be defined as the relative frequency distribution that would be obtained if all possible samples of a particular sample size were taken. A two-tailed test divides the alpha level of 0.05 into both tails. In contrast, a one-tailed test is a test of hypothesis in which the rejection region is placed entirely at one end of the sampling distribution. A one-tailed test puts the 5% in only one tail. A two-tailed test requires a greater difference to produce the same level of statistical significance as a one-tailed test. The two-tailed test is more conservative and thus preferred in most circumstances.

The Significance of No Significant Difference

The failure to find a difference between (among) a set of data does not necessarily mean that a difference does not exist. Differences may not be detected because of issues with power. Power is the ability of a statistical test to reject the null hypothesis when it is truly false and therefore should be rejected. Type II error (also referred to as beta or β) is defined as not rejecting the null hypothesis when in actuality it is false; that is, to falsely consider that no difference exists between study groups. Power and type II error are related in the equation

$$1 - \text{type II error} = \text{power}.$$

Statistical power is not an arbitrary number of a study, but rather it is controlled by the design of the study. In studies, a desirable power is at least 80%. This means that there is an 80% chance of detecting a difference between two groups if a difference of a given size really exists.

Sample size is related to power; the higher the power that is desired by the investigator, the larger the sample size required. If there are insufficient numbers of patients enrolled in a study, a statistically significant difference will not occur. The sample size is the one element that can easily be manipulated to increase the power. When calculating the sample size for a study that compares two means, several elements are used: desired detectable difference, variability of the samples, and the level of statistical significance (α). The type I error is typically set at 0.05. There is an inverse relationship between type I and type II errors. If investigators choose to lower the risk of type I error in a study, they increase the risk of type II error. Therefore, the sample size needs to be increased to compensate for this change.

Likewise, effect size (minimum clinically relevant difference) is also determined *a priori* (*a priori* is a term used to identify a type of knowledge that agrees with reason and is frequently obtained independent of experience), and is selected based on clinical judgment and previous literature. There are times when a 1% difference is irrelevant, as in the case of a 70% success rate compared to 71% rate for a new antibiotic compared to standard. In contrast, investigators may be able to defend a difference of 2% in the rate of a fatal myocardial infarction after receiving a new medication compared to standard therapy. A sufficient number of patients need to be recruited so that any clinically meaningful differences are also statistically significant. Given enough study subjects, any true difference among study groups can be detected at a chosen p value, even if the effect size is clinically unimportant. The smaller the effect size that is clinically important, the greater the number of subjects needed to find a difference if one truly exists. For fixed sample sizes, as the effect size increases, the p value decreases. The clinical question is if it would be worthwhile to enroll these additional subjects to attain statistical significance if the difference between the two groups is not clinically important. Therefore, it is important for investigators to stipulate the minimum effects when planning a study. The variance is also set at the beginning of the study and is generally based on previous literature. If the variance is low, a given sample of a group is more likely to be representative of the population. Therefore, with lower variance, fewer subjects are needed to reflect the underlying population accurately and thus fewer patients are needed to demonstrate a significant difference if one exists. The best way to prevent a type II error from occurring is to perform a sample size calculation before initiation of the study.

Selection of Statistical Test

If the incorrect statistical test is used, a misleading or inaccurate result may occur. There are many statistical tests, and several may be appropriate to use for a given set of data. The test that investigators use needs to be identified in the statistical methods section of the published report and in the footnotes of tables. Several commonly used statistical tests are described in Table 1-1. Among key considerations for choice of an appropriate test is the type of data, whether the data are paired (dependent) or unpaired (independent), and number of groups of data being compared. Statistical tests are also categorized into parametric or nonparametric tests. If appropriate criteria are met, a parametric test is preferred. Parametric tests are used to test differences using interval and ratio data. Samples must be randomly selected from the population and they must be independently measured. In other words, the data should not be paired, matched, correlated, or interdependent in any way. Two variables are independent if knowledge of the value of one variable provides no information about the value of another variable. For example, if you measured blood glucose level and age in a diabetic population, these two variables would in all likelihood be independent. If one knew an individual's blood glucose, this would not provide insight into a person's age. However, the variables were blood glucose and hemoglobin A_{1c}, then there would be a high degree of dependence. When two variables are independent, then the Pearson's correlation (further information on Pearson's correlation is provided in the Regression and Correlation Analysis section) between them is 0. When the phrase "independence of observations" is used, reference is being made to the concept that if two observations independent of the sampling of one observation do not affect the choice of the second observation. Consider a case in which the observations are not independent. A researcher wants to estimate how productive a person with osteoarthritis is at work compared to others without the disease. The researcher randomly chooses one person who has the

condition from an osteoarthritis disease registry and interviews that person. The researcher asks the person who was just interviewed for the name of a friend who can be interviewed next as a control (person without osteoarthritis working the same job). In this scenario, there is likely to be a strong relationship between the levels of productivity of the two individuals. Thus, a sample of people chosen in this way would consist of dependent pieces of information. In other words, the selection of the first person would have an influence on the selection of other subjects in the sample. In short, the observations would not be considered to be independent. The data also need to be normally distributed or the sample must be large enough to make that assumption (central limit theorem) and sample variances must be approximately equal (homogeneity of variance). The assumption of homogeneity of variance is that the variance within each of the populations is equal. As a rule of thumb, if the largest variance divided by the smaller variance is less than two, then homogeneity may be assumed. This is an assumption of analysis of variance (ANOVA), which works well even though this assumption is violated except in the case where there are unequal numbers of subjects in various groups. If the variances are not homogeneous, they are heterogeneous. If these characteristics are met, a parametric test may be used. The parametric procedures include tests such as the t-tests, ANOVA, correlation and regression. The list of tests in Table 1-1 is not all-inclusive of tests used in clinical trials, but it represents the most common analyses. Complex or uncommon statistical tests may be appropriate, but they should be adequately referenced in the publication of a clinical trial.

Comparing Two or More Means

The Student's t-test is a parametric statistical test used to test for differences between means of two independent samples. This test was first described by William Gosset in 1908, and was published under the pseudonym "student". Because the t-test is an example of a parametric test, the criteria for such a test needs to be met before use. The measured variable is approximately normally distributed and continuous. The variances of the two groups are similar. The Student's t-test can be used in cases where there is either an equal or unequal sample size between the two groups. Once the data are collected, and the t value is computed, the researcher consults a table of critical values for t with the appropriate alpha level and degrees of freedom. If the calculated t value is greater than the critical t value, the null hypothesis is rejected and it is concluded that there is a difference between the two groups.

In contrast to the Student's t-test, the paired t-test is used in cases in which the same patients are used to collect data for both groups. For example, in a pharmacokinetic study where a group of patients have their drug serum concentration measured while taking brand name medication A, and the same group of patients have their drug serum concentration measured while taking medication B, the differences between these two means will be determined using a paired t-test. In this case, patients serve as their own control. With the paired t-test, the t-statistic is not describing differences between the groups, but actual individual patient differences.

When the criteria for a parametric test are unable to be met, a nonparametric test can be used. These tests are traditionally less powerful. Nonparametric tests do not make any assumptions about the population distribution. The requirements of normality or homogeneity of variance associated with the parametric tests do not need to be met. These tests usually involve ranking or categorizing the data and in doing so may decrease the accuracy of the data. It may be more difficult to identify differences that are actually there. The investigator needs to evaluate the risk of type II error. The Mann-Whitney U test is one of the most powerful nonparametric tests, and tests a hypothesis that the medians of two groups are significantly different. The Mann-Whitney U test is the nonparametric equivalent to the Student's t-test. The test is based on ranks of the observations. Data are ranked and a formula is applied. As with all statistical tests, there are certain assumptions that need to be met. Both samples need to be randomly selected from their respective populations, the data need to be at least ordinal, and there needs to be independence between the two samples. The Wilcoxon rank sum test has similar assumptions that need to be met and when used, will give similar results to the Mann-Whitney U test.

The Wilcoxon signed ranks test is a nonparametric equivalent of the paired t-test for comparing two agents. The test is based on the ranks of the differences in paired observations. To appropriately use this analysis, the differences are mutually independent, and they all have the same median.

The one-way ANOVA (also referred to as the F-test) is an expansion of the t-test to include more than two levels of discrete independent variables.

The same assumptions for parametric tests need to be met with this procedure, including the need for the measured variable to be continuous from populations that are approximately normally distributed and have equal variances. The null hypothesis states that there are no differences among the population means, and any differences identified in the sample means are due to chance error alone. The alternate hypothesis states that the null hypothesis is false; that there is not a difference among the groups. This is because the test statistic identifies that a difference does occur somewhere among the population means. If the null hypothesis is rejected, then an *a posteriori* test must be done to determine where the differences lie. These post hoc procedures can evaluate where the differences exist while maintaining the overall type I error rate at a level similar to that used to test the original null hypothesis (e.g., 0.05). Examples of these tests include the Tukey Honestly Significant Difference (HSD) test, Student-Newman-Keuls test, Dunnett test, Scheffe Procedure, Least Significant Difference (LSD) test, and Bonferroni Method. The Bonferroni Method is the simplest and is best suited for a small number of preplanned comparisons. Just as the t-test involves calculation of a t-statistic, which is compared with the critical t, ANOVA involves calculation of an F-ratio, which is compared with a critical F-ratio.

The ANOVA is preferred over using multiple t-tests because when more than one hypothesis is tested on the same data, the risk is greater of making a type I error. If three groups of data were being compared (i.e., $\mu_A = \mu_B = \mu_C$), and a Student's t-test was used to compare the means of A versus B, A versus C, and B versus C, then the type I error rate would be three comparisons times 0.05 or 0.15. If multiple testing did occur, the investigator needs to either use a stricter criterion for significance or would need to apply the Bonferroni's correction. This factor reduces the threshold p value by the number of comparisons made. For example, if there were six comparisons using multiple t-tests, the results would only be accepted as being statistically significant if the new p value was less than 0.008 rather than 0.05.

Two-way (repeated measures) ANOVA is an expansion of the paired t-test and is used when there are more than two groups of data and the same group of subjects is studied using various treatments or time periods. Several assumptions need to be met to use the two-way ANOVA, including independent groups, normally distributed data, similar variance within the groups, and continuous data. The difference between a one-way and two-way ANOVA is that when using a one-way ANOVA there is a single explanatory variable, and a two-way analysis is applied to 2 (two) explanatory variables. The Kruskal-Wallis (one-way) ANOVA is a nonparametric alternative to the one-way ANOVA. The Friedman two-way ANOVA is used as a nonparametric alternative to the two-way ANOVA. For both of these tests, data need to be measured on at least an ordinal scale.

Finding a Difference with Proportions

When a researcher has nominal data and want to determine if frequencies are significantly different from each other for two or more groups, this can be determined by calculating a chi square statistic (X^2). The chi square analysis is one of the most frequently used statistical tests, and compares what is observed with the data with what one would expect to observe if the two variables were independent. If the difference is large enough, researchers conclude that it is statistically significant.

To perform a chi square analysis, one must be sure that the data in the contingency table meet several requirements. When using a 2X2 contingency table, if n is greater than 20, the chi square analysis may be used if all expected frequencies are five or more. If greater than 2 (two) groups are compared, the chi square may be used if no more than 20% of the cells may have expected frequencies less than 5 and none may have expected frequencies less than 1. An example of how to set up a contingency table is as presented in Figure 1-2. A contingency table has two variables. The categories (or levels) of the intervention, the fictitious medication magnadrug or no magnadrug, are represented in k rows in the table and the category of the outcome, gastrointestinal upset, are represented by the m columns in the table. This is a 2X2 contingency table and has 4 (four) cells. A 2X2 contingency table is called this because it has two rows and two columns and "contingency" because the values in the cells are contingent on what is happening at the margins.

By inspecting the observed frequencies (cells A to D), or those found as a result of the experimental program, there appears to be differences in the numbers of patients who had gastrointestinal upset in each group. The cell frequencies are added to obtain totals for each row and column. An expected frequency is the number of patients one would expect to find in a given cell if there were no group differences. The formula to calculate the expected frequency of a cell is as follows:

$$\text{Expected frequency of cell} = \frac{(\text{cell's row total})(\text{cell's column total})}{(\text{total number of patients in study})}$$

$$\text{Expected frequency of cell A} = \frac{(76)(80)}{(280)} = \frac{6080}{280} = 21.7$$

In the example, all of the expected frequencies for cells A through D were greater than 5. The next step is to determine if the frequencies observed in the experiment are significantly different from the frequencies that would be expected if there were no group differences. The chi square statistic is calculated.

If the chi square statistic is equal to or greater than the critical value, the difference is considered to be statistically significant. Chi square analysis does not tell which of the observed differences is statistically significant from the others, unless there are only two categories of the variable being compared. Further statistical analysis is required to single out specific differences.

In this case, n is greater than 20. If the n was less than 20 and if each cell had an expected frequency of at least 5, a Fisher's exact test could have also been used. If more than 2 (two) groups are being compared, a Fisher's exact test may be used if the sample size is at least 20 and any cell has an expected frequency of less than 5.

The McNemar's test and Cochran's Q test are tests of proportions based on samples that are related. McNemar's test involves dichotomous measurements (e.g., present or absent) that are paired. Cochran's Q test can be thought of as an extension of the McNemar's test concerned with three or more levels of data.

Regression and Correlation Analysis

Both regression and correlation analysis are used to evaluate interval or ratio data. Correlation analysis is concerned with determining if a relationship exists between

		Gastrointestinal Upset				
		Yes		No		Total
Magnadrug	Yes	42	A	34	B	76
	No	38	C	166	D	204
	Total	80		200		280

Figure 1-2. Example of a contingency table.

two or more variables and describes the strength of that relationship. Regression on the other hand describes the magnitude of the change between the two variables. In other words, regression is both descriptive and predictive, whereas correlation is only descriptive.

Regression analysis provides a mathematical equation that can be used to estimate or predict values of one variable based on the known values of another variable. Regression analysis is used when there is a functional relationship that allows investigators to predict the value of a dependent (or outcome; y) variable from the known values of one or more independent (predictor; x) variable(s). When there is one explanatory variable, it is referred to as simple regression. When two or more explanatory variables are tested, it is referred to as a multiple regression analysis. When the response variable is a binary categorical variable (e.g., dead or alive), the procedure is called logistic regression. Logistic regression may be either simple or multiple logistic regression. In a study, an investigator may collect data on several explanatory variables, determine which variables are more strongly associated with the response variable, and then incorporate these variables into a regression equation. Cox proportional hazards regression is used to assess the relationship between two or more continuous or categorical explanatory variables and a single response variable (time to the event). Typically, the event (e.g., death) has not yet occurred for all participants in the sample, which creates censored observations. Elements that need to be presented when describing the results of a study include the methods for selection of independent variables, threshold of significance, and overall, how well the model worked. In many cases, this information is underreported.

In the case of simple linear regression, the formula for the model is:

Dependent variable =
intercept + (slope × independent variable)

The regression line is the straight line that passes through the data that minimizes the sum of the squared differences between the original data and the line, and is referred to as the least squares regression line. Once the linear relationship has been determined, the next step is to determine if there is a statistically significant relationship present. The coefficient of determination, r^2, describes the proportion of the variation of the data presented by the dependent variable that is explained by the independent variable. An r^2 of 1.0 is a perfect relationship between the two variables. If the r^2 value were 0.5, this is interpreted as 50% of the variation in the data presented by the dependent variable can be described by the independent variable. An ANOVA is used to determine if the differences identified are due to chance. If a p value was found to be less than 0.05, one would conclude that there is a significant relationship between the two variables of interest. The closer the coefficient of determination is to "0", the less likely it would be to find a difference. A significant value indicates that there is an association, and typically not a cause and effect relationship. This is true in most cases. An exception to this rule is in the case of stability studies in which the independent variable is controllable. In this case, a cause and effect relationship can be claimed.

Correlation is used to determine if two independent variables are related in a linear manner. For instance, an investigator wants to determine if there is a relationship between bone mineral density and the number of fractures in postmenopausal women. A unitless number, called the correlation coefficient "r", summarizes the strength of the linear relationship between the two variables. The r value varies from "-1 to +1". A "-1" indicates a perfect linear relationship in which one variable changes while the other changes in an inverse fashion. The closer the calculated value is to this number, the stronger the negative relationship. A "0" indicates no relationship exists between the two variables. A "+1" indicates that there is a perfect positive linear relationship with one variable changing as the other changes in the same direction.

Although there are several formulas used to determine the correlation coefficient, the most common method is the Pearson's product-moment correlation. This formula assumes normal distribution. The Spearman correlation coefficient is the comparable nonparametric statistic if the data are not normally distributed. When interpreting the r value, a p value needs to be considered to help assess how likely the correlation is due to chance. If in the above example with bone mineral density and risk of fractures the relationship between these two variables was determined to have an r value of 0.97 and a p of 0.01, then the r value is significantly different from 0 (no correlation) and that the finding is probably not due to chance. If investigators conclude that there is a relationship between the two variables, this does not imply that there is a cause and effect relationship. Unlike the regression analysis, the correlation analysis does not describe the magnitude of the change between the two variables. In other words, regression is both descriptive and predictive, whereas correlation is only descriptive.

Confidence Intervals

A CI is the range of values consistent with the data, which is believed to contain the actual or true mean of the population. The estimate of the population mean from the sample is referred to as the point estimate. The range of possible means is defined as the confidence limits. The CIs can be used to estimate mean differences between groups or estimate the true mean for the population from which the sample was drawn. When considering the CI of the difference between two groups, the 95% CI is related to statistical significance at the 0.05 level. When the 95% CI for the estimated difference between groups or in the same group over time does not include zero, the results are significant at the 0.05 level. For example, if the difference in mean blood pressure measurements between two groups was 10 mm Hg (95% CI = 6–10 mm Hg), the difference between the groups in mean blood pressure would be considered to be statistically significant at the 0.05 level. Zero is not included in the area for 95% of the values over which the observed difference is likely to range; therefore, it must be in the remaining 5%. The likelihood of obtaining a difference of 0 mm Hg is less than 5 times in 100.

In contrast, the mean difference in blood pressure measurement between two groups was 2 mm Hg (-1–5 mm Hg). Here, the CI includes zero, so the difference is not statistically significant at the 0.05 level. The likelihood of obtaining a difference of 0 mm Hg is greater than 5 times in 100. The CI can be used as an alternative to conventional statistical tests of significance in hypothesis testing, and is most often preferred because of the information that can be obtained from these values. The width of the CI is an indicator of the precision of the estimate; the level of significance is an indicator of the accuracy. For a given level of confidence, the narrower the CI, the greater the precision of the sample mean as an estimate of the population means. There are three factors that can influence the width of the CI. First, the variance of the sample scores on which the CI is calculated can affect the width of the CI, with a smaller variance resulting in a narrower CI. Although efforts could be made to obtain a more homogenous sample from the population to help decrease the width of the CI, in general, the investigator typically has little control over this variable. Secondly, sampling precision can also influence the size of the interval. Because sample precision is related to the square root of the sample size, doubling the sample size will decrease the width by 25%. And the last factor is the level of confidence that is used. If an investigator wants to be 99% confident that the true mean of the population is included in the range of values, the CI will be wider than if the investigator sets the level of confidence at 95%.

In addition to the CI being used to describe differences between group means or mean changes within the same group, the CI can also be used for proportions, odds ratios, and risk ratios. Over time, the CI can also be used for proportions, odds ratio, and risk ratios. Other common estimates that may be accompanied by a CI include survival rates, slopes of regression lines, effort to yield measures, and coefficients in a statistical model. The CI can also be used for a single clinical trial, but are also routinely used to describe aggregate data in a meta-analysis.

Measures of Association with Categorical Data

Relative risk and odds ratio are two measures of disease frequency. The relative risk is the ratio of the incidence rate of an outcome in the exposed group to the incidence rate of the outcome in the unexposed group. The incidence rate of a disease is a measurement of how frequently the disease occurs. It is the number of new cases of the disease (in a defined time period) divided by the number of individuals in that population at risk.

If the relative risk is 1, the risk of unintended drug effect for an exposed person is the same as the risk for the nonexposed person. If the relative risk is greater than 1, the risk of unintended drug effect for an exposed person is X times greater than that for a nonexposed person. If the relative risk is less than 1, the risk of unintended drug effect for an exposed person is X times less than that of a nonexposed person. Frequently, we would like to not only give an indication of risk or benefit in relative terms, but one would like to examine the actual risk. One way to describe this is by presenting the attributable risk. Attributable risk is defined as follows:

Attributable risk =
incidence of gastrointestinal disease in exposed group –
incidence of gastrointestinal disease in unexposed group

Another important method to describe risk is as an odds ratio. The odds ratio is an estimate of the relative risk when the disease under study is relatively rare. When using the odds ratio as an estimator of risk, one must assume that the control group is representative of the general population, the cases are representative of the populations with the disease, and the frequency of the disease in the population is small. The odds ratio is mathematically obtained by multiplying the number of cases with the disease and exposed to the factor by the number of cases without the disease and not exposed to the factor and dividing this number by the number of cases with the disease without exposure to the factor multiplied by those cases without the disease but exposed to the factor. It is defined by the following equation:

$$\text{Relative odds} = \frac{\text{odds of exposure for cases A/C}}{\text{odds of exposure for controls B/D}}$$

The odds ratio of 1 is interpreted as the number of cases that are just as likely to have been exposed as the controls. An odds ratio of greater than 1 is interpreted as the number of cases that are X times more likely to have been exposed than are the controls. An odds ratio of less than 1 is interpreted as the number of cases that are X times less likely to have been exposed than the control.

Confidence intervals are frequently used as a measure of testing the significance. When a 95% CI contains 1, there is no difference between exposed and nonexposed groups.

A more detailed description of statistics commonly used with the pharmacoepidemiology literature is described in the Pharmacoepidemiology chapter.

Survival Analysis

Data in clinical trials may be presented using survival curves with time-to-the-event as the dependent variable. The event or outcome may be treatment response. Patients are followed until either they experience a predefined event or follow-up is terminated without an end point event. Two common ways to calculate a life table are the actuarial approach and the Kaplan-Meier approach. There are several assumptions that need to be met in order to use this analysis. There needs to be an identifiable starting point. With the use of medications, the identifiable starting point is immediately after the medication is given. There needs to be a well-defined outcome that is dichotomous, such as death or hospitalization. The Kaplan-Meier approach should not be used if any patients are lost to follow-up because this event may be related to the outcome of interest. Under these circumstances, the survival function will be biased with an

underestimation of the risk of death. Lastly, there should not be significant differences in how patients are handled. Secular changes are changes (diagnostic practices and treatment regimens) that occur over time, and as such, patients who were enrolled in the trial early may differ from those who were enrolled later. The investigator, under these circumstances, could not assume that he or she is dealing with a homogeneous group, and therefore, the data should not be combined.

The log-rank statistic is commonly used to compare two survival distributions. This test compares the observed number of events with the numbers expected. This test works under the assumption that if there is no difference between the groups, then at any interval the total number of events should be divided between the groups approximately to the number of subjects at risk. A log rank test assigns scores to each uncensored and censored observation based on the logarithm for the estimated probability at each observation. If two curves are being compared and there are an equal number of patients in both groups, then each group should have about the same number of events.

Another test commonly used to assess survival curves is the Cox proportional hazards model. This test has been compared to the analysis of covariance for handling survival data in that it can handle any number of covariates in the equation quantifying the effect of covariates on the survival time. The Cox proportional hazards model is used when the researcher is concerned about group differences at baseline and related to a covariate that is measured on a continuous scale. This allows the investigator to evaluate survival data and adjust for confounding variables such as severity of disease or age.

Meta-Analysis

Meta-analysis is a discipline that provides methods for finding, appraising, and combining data from a range of studies to answer important questions in ways not possible with the results of individual studies. Meta-analysis can be used if 1) definitive clinical trials are impossible, unethical, or impractical, 2) randomized trials have been performed, but the results are conflicting, 3) results from definitive trials are being awaited, 4) new questions not posed at the beginning of the trial need to be answered, and 5) sample sizes are too small. From a logistical standpoint, a written protocol needs to be strictly followed consistent with a good research design, including a clearly defined research question, search strategy, abstraction of data, and statistical analysis. Data are typically inspected using an L'Abbe plot. This technique is used to inspect data for heterogeneity. The outcome rates in treatment and control groups are plotted on the vertical and horizontal axes. The graphical display reveals heterogeneity of both size and direction of effect, and indicates which studies contribute most to it.

Pooling refers to methods of combining the results of different studies. This is not simply combining the data from all trials into one very large trial, but rather statistically combining results in a systematic way. Pooling must maintain the integrity of individual studies. In general, the contribution of each study to the overall result is determined by its study weight, usually the reciprocal of its variance. Details regarding the statistical methods for pooling are beyond the scope of this chapter.

Summary

A basic understanding of statistical concepts and application is important when assessing data as the foundation of the clinical decision-making process regarding the use of medications in patients. This chapter provides a general overview of statistical concepts including both descriptive and inferential statistics. One of the more common mistakes made by readers of the scientific literature is the failure to distinguish between the clinical and statistical significance of the data. In general, data that are clinically significant are relevant to patient care. When claiming that data are statistically different, this refers to a mathematical term to express a conclusion that there is evidence against the null hypothesis. The probability is low of getting a result as extreme or more extreme than the one observed in the data if the null hypothesis is accepted. Merely achieving statistical significance does not characterize the author's data as clinically important. Having a sound understanding of statistical concepts allows readers to make good judgments regarding the validity and reliability of the data, and to assess the value of the data to an individual patient or patient group. Readers are referred to the Annotated Bibliography for more detailed information regarding the topics described.

Annotated Bibliography

1. Campbell MJ, Gardner MJ. Calculating intervals for some nonparametric analyses. BMJ 1988;296:1454–6.

   This article reviews the methodology to calculate confidence intervals (CIs) for some nonparametric analyses, including a population median and mean (when the criteria for normal distribution are not met). It is a well-organized document that nicely outlines these two approaches. The authors also describe the limitations of providing CIs with nonparametric analysis. It is not always possible to calculate CIs with exactly the same level of confidence. The authors recommend that the 95% CI be routinely calculated. The paper is concisely written, easy to read, and provides some nice examples to reinforce concepts.

2. Fitzgerald SM, Flinn S. Evaluating research studies using the analysis of variance (ANOVA): issues and interpretation. J Hand Ther 2000;13:56–60.

   This article provides a critical evaluation of the single factor analysis of variance (ANOVA) test by providing a question and answer format addressing the issues that are pertinent to the evaluation of a test in a clinical trial. It describes whether the ANOVA is appropriate to use, discusses its interpretation, use of post hoc test of analysis, examination of error, and clinical interpretation. The authors do not discuss which post hoc test is best based for a particular data set, but rather they provide a general review of important concepts to consider. The authors discuss the elements that affect type I and type II error, including all elements that affect power of a study.

3. Fleming TR, Lin DY. Survival analysis in clinical trials: past developments and future directions. Biometrics 2000;56:971–83.

This article is a nice review of standard statistical procedures for assessing survival in clinical trials. The article discusses the more conventional methods such as Kaplan-Meier approach to estimating the survival function, the log-rank statistic for comparing two survival curves, and proportional hazards model for quantifying the effects of covariants on survival time. The authors provide an overview of the direction anticipated for future research activities. This type of analysis has gained a significant popularity in this era of outcome management, with the science of survival analysis changing and evolving in several directions. For example, the authors discuss the concept of integrating a Bayesian approach in survival analysis, especially in the areas of noncompliance and multivariate failure time.

4. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg 1999;81:1454–60.

This paper provides a nice overview of the relationship between sample size and statistical power. The authors discuss the concept of hypothesis testing and using *a priori* selection of both types I and type II error rates. The implications of not having enough power relating to the inability to find a difference when a difference truly exists are nicely described. The authors also nicely describe the relationship between the elements that affect power, including an extensive discussion on effect size, variance, error rate, and the role of sample size. The authors also reinforce the need to do a post hoc analysis of power after completion of the study.

5. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978;299:690–4.

This article describes an analysis where the authors reviewed 71 trials in which the investigators did not find a difference among the patient groups. The authors were interested in assessing the frequency by which these 71 trials lacked sufficient sample size to find a difference if one may have actually existed. Sixty-seven of the trials had a greater than 10% risk of missing a 25% difference between the groups, and 50 of the trials missed a 50% improvement. These authors describe the implications of low power and reinforce the occurrence of this problem in the literature with the results of their analysis.

6. Gardner MJ, Altman DG. Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J 1986;292:746–50.

This paper provides a nice review of CIs and compares their value to conventional hypothesis testing. It also describes how to calculate CIs for means and proportions. Simple and practical examples are provided to reinforce the concepts. The paper also describes how CIs should be presented in the literature, including suggestions for graphical display.

7. Greenfield ML, Kuhn JE, Wojtys EM. A statistical primer. Correlation and regression analysis. Am J Sports Med 1998;26:338–43.

This review provides a nice overview of correlation and regression analysis. The information is an overview for individuals who are unfamiliar with the concepts, interpretation, and presentation of information. The authors also provide some examples throughout the document to highlight and reinforce basic concepts. The authors reinforce the differences between the two analyses and emphasize areas that are commonly confused between the two functions. To emphasize the value of regression analysis, they discuss only the concept of simple linear regression. This paper does not discuss differences with other regression analyses such as multiple regression or logistic regression. Readers are referred to other publications for these discussions.

8. Guyatt G, Walker S, Shannon H, Cook D, Jaeschke R, Heddle N. Correlation and regression. CMAJ 1995;152:497–504.

This article provides a nice overview of correlation and regression. With the increasing emphasis on outcome studies using large databases, using these statistical tests will continue to increase. This article is a nice primer of when these tests are used and how to interpret the data. The authors use several real examples to help describe and differentiate these concepts. Sets of data are provided that help readers use and interpret the information. The difficulty with this article is that the authors assume that readers have a basic understanding of concepts, including calculation of values.

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

This article provides a nice review of the statistical concepts of hypothesis testing and p values. The role of chance and its relationship to probability and p value are discussed. Simple examples, such as the toss of the coin, help explain these concepts to clearly reinforce the basic concepts. The authors discuss in brief the concept of type II error, type I error related to the multiple testing problem, and limitations of hypothesis testing. The concept of successful randomization and the potential need to adjust baseline values to improve the validity of the results also are discussed. This article reviews the basics of hypothesis testing and describes core concepts relating to the testing process.

10. Hartzema AG. Guide to interpreting and evaluating the pharmacoepidemiologic literature. Ann Pharmacother 1992;26:96–7.

The article provides an overview of criteria for evaluating the pharmacoepidemiologic literature. The author discusses the elements needed to evaluate research design, including the case-control and cohort studies. He discusses how data can be interpreted, including the use of the odds ratio, relative risk, measures of association (p value), and CIs. This is not a complete review, but rather a simple, concise article that addresses what a reader of pharmacoepidemiologic literature may need to consider in interpretation. The author does not provide calculations or common errors in their use or interpretation. This would not be a good article for an individual well versed in basic concepts, but rather someone who is new to pharmacoepidemiologic concepts.

11. Henry DA, Wilson A. Meta-analysis. Part 1. An assessment of its aims, validity and reliability. Med J Aust 1992;156:31–8.

This review is part one of a two-part series that addresses the issue of meta-analysis, including the purpose, controversies, and the reliability and validity of the technique. This article is basic in its approach, yet it addresses many issues pertinent to review of the data produced by this technique. The authors describe the literature that addresses the reliability and validity of meta-analysis. The authors provide a balanced review, and

use data to reinforce some of the areas of controversy or value of combining data. The article also reinforces the need to provide a systematic approach to performing a meta-analysis, as the results that are obtained can be misleading without a logical and statistically sound approach.

12. Khan KS, Chien PW, Dwarakanath LS. Logistic regression models in obstetrics and gynecology literature. Obstet Gynecol 1999;93:1014–20.

    This study examines the variations in quality of the reporting of logistic regression in the scientific literature. Criteria were described and used to assess the accuracy, precision, and interpretation of logistic regression in 193 articles from four generic obstetrics and gynecology journals in 1985, 1990, and 1995. The authors found that the proportion of articles that used logistic regression increased over this time period. There were several violations in quality and presentation, including the lack of clear reporting of dependent and independent variables (32.1%), the selection of variables being inadequately described (51.8%), and 85.1% did not report assessment of conformity to linear gradient. This article is a nice review for how logistic regression should be presented in the literature. It also reinforces the need to critically evaluate its presentation because of the misinformation that is frequently described.

13. Kuhn JE, Greenfield ML, Wojtys EM. A statistics primer. Hypothesis testing. Am J Sports Med 1996;24:702–3.

    This article focuses on the concept of hypothesis testing and compares and contrasts the null and research hypothesis. The authors reinforce the need to clearly state the hypothesis within a publication, and for the reader of such an article to identify the elements important to these concepts. The readers should identify the research and null hypothesis of the article that they are reviewing, and the meaning of the type I and type II error of a trial. This is an important article that addresses the development of the null hypothesis and appropriate selection of statistical analysis.

14. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2:93–113.

    This article summarizes sample size determination and its relationship to power when planning a research project. Frequently, articles will review only one or two testing procedures. The advantage of this article is that the authors discuss methods for sample size determination for the t-test, tests for proportions, tests for survival time, and tests for correlations. For each example, sample values are given and calculated. There is also a detailed discussion of power and the elements that affect power. The article is written in a concise, practical, and easy-to-read manner.

15. Lee ET, Go OT. Survival analysis in public health research. Annu Rev Public Health 1997;18:105–34.

    Common statistical techniques for assessing survival data in public health research are reviewed. The authors discuss both nonparametric and semi-parametric approaches, including the Kaplan-Meier Product Limit Method, methods of testing the equality of survival distributions, and Cox's regression model. The authors also discuss parametric models that are commonly used, such as the accelerated failure time model. Hazard functions for the exponential, Weibull, gamma, Gompertz, lognormal, and log-logistic distributions are described. Examples from the literature help reinforce principles. There is a nice overview of commercially available software packages that can help perform these analyses, including SAS, BMDP, SPLUS, SPSS, EGRET, and GLIM. The first two are discussed more extensively than the others in this article.

16. Levine MA. A guide for assessing pharmacoepidemiologic studies. Pharmacotherapy 1992;12:232–7.

    The article is organized into eight primary questions that need to be considered when reviewing the pharmacoepidemiology literature, such as elements of study design, association, temporal relationship, evaluation of a dose-response relationship, and other practical points relating both to the logistics, as well as interpretation of the data. The authors address many of the basic elements in evaluating this type of literature, but that would be considered to be too basic for a researcher in the area. This area is frequently overlooked in published reviews or general tests on statistics. For more extensive, detailed information, the reader is referred to other publications from this author and others.

17. Loney PL, Chambers LW, Bennett KJ, Roberts JG, Stratford PW. Critical appraisal of the health research literature: prevalence of incidence of a health problem. Chronic Dis Can 1998;19:170–6.

    This article provides an overview of how to evaluate an article that estimates the prevalence or incidence of a disease or health problem. These two terms are different, but terminology is frequently misused in the literature. This article is a primer for health professionals who have an interest in either performing this type of research or who review these publications to make changes in their practice. The concepts of design, sampling frame, sample size, outcome measures, measurements, and response rates are discussed. Examples are provided that help reinforce how data need to be presented, interpreted, and applied to practice.

18. Mathew A, Pandey M, Murthy NS. Survival analysis: caveats and pitfalls. Eur J Surg Oncol 1999;25:321–9.

    This article discusses the concept of survival analysis, its purpose, and appropriate use. It also discusses many methods used to estimate the survival rate and its standard error. The authors discuss the concept of misusing these types of tests and guide the reader on how to properly consider issues with data. The authors make some general recommendations, including the support of the Kaplan-Meier approach, suggesting that the median (instead of the mean) survival time be provided whenever possible, and those confidence limits be used as a measure of variability. The information that is shared is practical and frequently overlooked by the investigators publishing studies that use survival analyses.

19. Pathic DS, Meinhold JM, Fisher DJ. Research design: sampling techniques. Am J Hosp Pharm 1980;37:998–1005.

    Several statistical procedures require that one have a basic understanding of what constitutes a population versus a sample, and whether a trial used appropriate randomization techniques. This article describes different types of sampling procedures, including nonprobability samples such as convenience samples, judgment samples, and quota sampling, and compares them to probability samples such as simple random sampling, stratified samples, and cluster samples. The required formulas and examples for how to calculate sample size for both estimating the population mean and establishing population proportions are provided. Overall, it is a concise, easy-to-read article that addresses a topic that is so frequently a source of error in clinical trials.

20. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–8.

This article nicely describes the concept of multiple testing and the increased risk of type I error. Although textbooks frequently describe this issue, they also describe using Bonferroni adjustments to resolve some of these problems. This article discusses the mechanism to adjust for multiple testing and describes the problems with this technique, including testing irrelevant null hypothesis, the increased risk of type II error, and difficulties with interpretation. The authors also discuss circumstances when the Bonferroni adjustment may be implemented, although they suggest that few situations actually exist. This is a nice document to have in your files when reviewing the concept of multiple testing.

21. Porter AM. Misuse of correlation and regression in three medical journals. J R Soc Med 1999;92:123–8.

This review discusses some commonly found errors relating to the use, presentation, and interpretation of both correlation and regression tests, which are frequently used in the published literature. Several clinical trials found in the British Medical Journal, the New England Journal of Medicine, and The Lancet were used. The authors identified 15 different errors in the publication process: eight were considered to be common. This paper is important for individuals who are either using regression or correlation in clinical research, or for individuals who are routinely integrating these types of papers into their clinical practice.

22. Rigby AS. Statistical methods in epidemiology: statistical errors in hypothesis testing. Disabil Rehabil 1998;20:121–6.

This article on hypothesis testing describes the concept of statistical testing and the need to provide a systematic approach to testing a hypothesis. The authors discuss in detail the concept of type I and type II error, and the implications of their occurrence. The problem of multiple testing and its relationship to type I error are discussed. The concept of one- and two-tailed tests are discussed, and how p values should be presented in the literature and what they mean. The authors also introduce the concept of using CIs and briefly describe the value of their use.

23. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987;316:450–5.

This study provides an overview of elements that should be routinely included in a meta-analysis. The authors reviewed 86 meta-analyses in the published literature and assessed each related to six major areas: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and application of results. Only 28% of the papers reviewed addressed all six areas. In the six major areas, a total of 23 separate items were examined. Of the reviewed articles, up to 14 items were satisfactorily addressed in any of the studies. This analysis of the literature reinforces the need to educate researchers on the appropriate planning, implementing, and presenting meta-analyses in the literature.

24. Sim J, Reid N. Statistical inference by CIs: issues of interpretation and utilization. Phys Ther 1999;79:186–95.

This article provides a nice review of the value of CIs, including their advantages over conventional hypothesis testing. There is a well-organized section on basic principles, interpretation, and calculation. The authors describe two basic advantages to using CIs: 1) to attach a measure of accuracy to a sample statistic, and 2) to interpret the questions of clinical importance of the data. The role of CIs in meta-analysis is also described. The authors strongly reinforce the need to include a CI (in addition to the results of hypothesis tests) with the level of statistical inference consistent with the level of statistical significance for the hypothesis test (95% CI for $p<0.05$).

25. Wilson A, Henry DA. Meta-analysis. Part 2: assessing the quality of published meta-analysis. Med J Aust 1992;156:173–4, 177–80, 184.

This review is part two of a two-part series (part one is described in Reference 11) that reviews the elements that need to be considered in a well-designed meta-analysis. The article is organized by using 10 questions that every investigator needs to address. Although they do not provide details regarding calculations, several issues relating to statistical presentation, including the need to plot results, test for heterogeneity of outcome, and to calculate a summary estimate are addressed. The authors also discuss the concept of publication bias, with a description of a funnel plot. As the technique of using meta-analysis to answer important questions in the literature continues to grow, this article can serve as an important checklist to use when assessing the quality of the design and validity of the results.

26. Gaddis GM, Gaddis ML. Introduction to biostatistics: Part 5, statistical inference techniques for hypothesis testing with nonparametric data. Ann Emerg Med 1990;19:153–8.

This article provides several nonparametric statistical tests used when analyzing nominal and ordinal data. It is part five of a six-part series that discusses basic concepts of statistics. This article is written for the individual who has little background in statistics. An individual with a foundation in statistics would find this article elementary. The authors have taken an approach of presenting information to clinicians who will primarily be in the role of evaluating the published literature instead of actually performing mathematical calculation of a set of data. There are few mathematical formulas or problem-solving examples in which the reader needs to manipulate data.

27. De Muth JE. Basic Statistics and Pharmaceutical Statistical Applications. New York: Marcel Dekker Inc., 1999:115–48.

This chapter provides a nice overview of the concept of normal distribution and CIs. It reviews how to determine if the distribution is normal and defines and describes the central limit theorem. It also has a nice overview of CIs and their application. There are practice problems at the end of the chapter with answers to help reinforce concepts.

Please start new PSAP-IV answer sheet before continuing.

SELF-ASSESSMENT QUESTIONS

Questions 1 and 2 pertain to the following case.
In a study that compared two different medications for the treatment of allergic rhinitis, each patient was asked to rank his or her severity of symptoms on a scale of 0 to 4 (0 = none, 1= mild, 2 = moderate, 3 = moderately severe, and 4 = severe) 2 hours after the medication was given.

1. What type of data will be obtained from this measurement?
   A. Interval.
   B. Ratio.
   C. Nominal.
   D. Ordinal.

2. Which one of the following is an appropriate method of central tendency and dispersion for the data that will be obtained above?
   A. Mean and standard deviation.
   B. Median and interquartile range.
   C. Median and standard deviation.
   D. Median, no measure of dispersion is acceptable.

3. Examine the data set of eight students with the following grades on a possible 100-point examination: 85, 79, 30, 94, 97, 35, 87, and 88. Which one of the following measures of central tendency is least affected by outliers?
   A. Median.
   B. Mean.
   C. Standard deviation.
   D. Standard error of the mean.

4. The following data are the number of seizures for five patients in a study: 3, 4, 4, 6, and 8. The mean is 5. Calculate the standard deviation?
   A. 1.
   B. 2.
   C. 3.
   D. 4.

5. In a study that compares the cholesterol-lowering effect of a new medication to standard therapy, the mean age of 100 patients receiving the new therapy is 40 years old, and the standard deviation is 4 years. Assuming normal distribution, about what percentage of patients is expected to be between 32 and 48 years of age?
   A. 20%.
   B. 68%.
   C. 95%.
   D. 99%.

6. Normal distribution is an important assumption to make when using several statistical analyses. Which one of the following is an indicator of normal distribution?
   A. The kurtosis is 1.
   B. The frequency curve is skewed to the right.
   C. The mean, median, and mode have similar values.
   D. The distribution is only modestly defined by the mean and standard deviation.

7. The mean difference in diastolic blood pressure between patients taking either a new or traditional antihypertensive is 10 mm Hg (95% confidence interval = -15 mm Hg to 51 mm Hg). The difference in diastolic blood pressure was not statistically significant at the 0.05 level. What number is included in the range of values that allows you to make this decision?
   A. 0.
   B. 1.
   C. 2.
   D. 20.

8. A p value of less than 0.05 implies which one of the following?
   A. Less than 5% of the time the conclusion from the statistical test will be due to chance alone.
   B. The power of the study is at an acceptable level.
   C. There is only a 5% probability that if a true difference existed that it would be found.
   D. A one-tailed test was used.

9. A study that compares the cholesterol-lowering effect of a new medication to traditional therapy found no difference in total cholesterol between the two groups after 1 year of treatment. What type of error would you be making if you wrongly concluded that there is no difference between the two different treatment arms in total cholesterol?
   A. Type I error.
   B. Type II error.
   C. Power.
   D. α Error.

10. When calculating the sample size to determine the difference between two means, which one of the following are elements to consider in the equation?
   A. Effect size.
   B. Sensitivity.
   C. The statistical test that will be used.
   D. Measurement process.

Questions 11–16 pertain to the following case.
A bioequivalence study was performed that compared a generic form of isosorbide dinitrate with a brand name formulation Isobide. Fifty-five subjects were randomly assigned to receive either a single dose of Isobide or a new generic product. After a 2-week washout period, subjects received the alternative agent. A difference of 2% was considered to be clinically important between the two groups, with a power of 60%, and an α value of 0.05. A one-tailed test was performed.

11. Why is using a one-tailed test instead of a two-tailed test a problem in this study?
   A. There is a greater chance of making a type II error.
   B. There is a greater chance of making a type I error.
   C. There is a greater sample size required to compensate for evaluating the difference in only one direction.
   D. The α is not actually 0.05 but rather 0.025, which is more stringent than required for a study.

12. Which one of the following is the probability of not finding a difference if there is indeed a difference?
   A. 2%.
   B. 5%.
   C. 60%.
   D. 40%.

13. Which one of the following elements could have affected the low level of power in this study?
   A. The difference that the investigators are looking for is small.
   B. Sample size is large (greater than 30).
   C. The α is 0.05.
   D. The variability is low.

14. Which one of the following is the appropriate statistical test to use?
   A. Chi square analysis.
   B. Fisher's exact test.
   C. Paired t-test.
   D. Student's t-test.

15. Which one of the following criteria need to be met to use a parametric test?
   A. Sample sizes need to be equal.
   B. Data need to be at least ordinal.
   C. Data need to be normally distributed in both sample groups.
   D. The power needs to be at least 90%.

16. If it were found that a parametric test could not be used, which one of the following is the nonparametric alternative?
   A. One-way analysis of variance.
   B. Student's t-test.
   C. Wilcoxon signed rank test.
   D. Mann-Whitney U test.

17. In a study that compared 16 patients who received either a standard pain medication or a new medication, the primary end point was documenting whether patients had complete pain relief. In this study, five-eighths (62%) and six-eighths (75%), respectively, had complete pain relief. Which one of the following is the best statistical test to use to determine if this difference is important?
   A. Fisher's exact test.
   B. t-test.
   C. Mann-Whitney U test.
   D. Paired t-test.

18. A study that compares three different antihypertensive therapies in 500 young females used the mean decrease in diastolic blood pressure as the primary end point for success. Investigators chose to use the t-test to determine if the differences between the three groups were significant. The authors found a statistically significant difference among groups. Why would the t-test be inappropriate?
   A. When using the t-test for three sample groups randomly drawn from independent populations, the risk of type I error increases.
   B. A Student's t-test should not be performed when the variable is measured on an interval or ratio scale.
   C. The sample size was not large enough to use a t-test.
   D. Investigators used young women in the study who frequently do not have hypertension.

19. A study of 40 patients was performed to examine the association of weight and serum drug levels of a new antibiotic. Linear regression was used to assess the association. The slope of the regression line (slope = 30) was significantly greater than 0, indicating that the serum drug level increases as weight increases. The r^2 value was calculated as 0.75 (r=0.86). Which statement provides the most accurate interpretation of these data?
    A. Seventy-five percent of the variance in serum levels is likely to be explained by its relationship with weight.
    B. Twenty-five percent of the variance in serum levels is likely to be explained by its relationship with weight.
    C. Thirty percent of the variance in serum levels is likely to be explained by its relationship with weight.
    D. Seventy percent of the variance in serum levels is likely to be explained by its relationship with weight.

20. A meta-analysis was performed with the purpose of assessing the risk of hypertension following the use of a newly approved medication for weight loss to patients not taking the medication. The odds ratio was calculated to be 5.6. This is important because an odds ratio of greater than 1 indicates that there is which one of the following?
    A. A decreased risk in the treatment group.
    B. An increased risk in the treatment group.
    C. There is no statistically significant difference between the groups at the 0.05 level.
    D. The difference in risk is statistically significant at the 0.05 level.

21. A meta-analysis was performed with the purpose of assessing whether using a new medication for hypertension will increase mortality in patients with congestive heart failure. Investigators were interested in determining if the differences in the individual trial outcomes were greater than any one could reasonably expect by chance alone. Which one of the following is used to test for heterogeneity?
    A. L'Abbe plot.
    B. There is no need to test for heterogeneity.
    C. t-test.
    D. Paired t-test.

22. A study that examined the relationship between birth weight and salary at age 50 found an r-value of 0.8. This value can be interpreted as which one of the following?
    A. As birth weight increases, the salary increases.
    B. The low birth weight caused a high salary at age 50.
    C. There is a statistically significant relationship between these two values.
    D. There is a good to excellent relationship between these two values.

23. In a study that examines a group of patients with cancer, the Kaplan-Meier approach of the 5-year survival rate after treatment was 25% (95% confidence interval [CI] = 20–35%) for the patients receiving drug A (n=100) and 65% (95% CI = 60–75%) for patients receiving drug B (n=100). The log-rank test revealed a statistically significant difference between the survival rates over time (p<0.01). What test can be used to assess the association between explanatory variables (age, family history) and survival rate?
    A. Wilcoxon test.
    B. Life table method.
    C. Cox proportional hazards regression.
    D. Student's t-test.

24. An investigator wanted to test the risk of getting thrombocytopenia with a new low-molecular-weight heparin compared to patients not receiving heparin, and found the odds ratio of 5.2 (95% CI = 1.5–10.5). Which one of the following answers provides the most accurate interpretation of this information?
    A. Because the CI does not include 0, patients who are taking the new low-molecular-weight heparin are 5.2 times more likely to have thrombocytopenia than individuals on heparin and this is considered to be statistically significant.
    B. Because the CI does not include 1, patients who are taking the new low-molecular-weight heparin are 5.2 times more likely to have thrombocytopenia than individuals on heparin and this is considered to be statistically significant.
    C. Because the CI does not include 0, patients who are taking the new low-molecular-weight heparin are 5.5 times more likely to have thrombocytopenia than individuals on heparin, but this is not considered to be statistically significant.
    D. Because the CI does not include 0, patients who are taking the new low-molecular-weight heparin are 5.2 times more likely to have thrombocytopenia than individuals on heparin, but this is not considered to be statistically significant.

25. Which one of the following statements is most accurate regarding the appropriate selection and use of simple linear regression?
    A. It is used when there is only one dependent variable with only one independent variable being analyzed.
    B. It routinely uses Pearson's product-moment correlation coefficient, "r".
    C. It assesses the linear relationship between two or more continuous or categorical variables and a single continuous response variable.
    D. It is an aspect of time-to-event (survival) analysis.

26. Below are data from a study that was performed to determine if there was a difference in the length of recovery room stay for patients undergoing surgery of the spine who received a new nonsteroidal anti-inflammatory drug (NSAID), morphine, or the combination of the two. Ten records were randomly

selected from patients receiving each of the three combinations during a 12-month period. What test was used to determine if there was a difference in length of recovery room stay for these patients?

Recovery room time (hours)

	NSAID	Morphine	Combination
	2	3	2
	1	2	3
	1	2	1
	2	2	2
	2	4	2
	3	1	2
	1	1	1
	5	2	2
	1	2	2
	2	2	2
Mean =	2	2.1	1.9

A. Repeated measures analysis of variance.
B. One-way analysis of variance.
C. Three-way analysis of variance.
D. Bonferroni adjustment.