Invited Review

Biostatistics Primer: Part I

Brian R. Overholser, PharmD; and Kevin M. Sowinski, PharmD, BCPS, FCCP
Department of Pharmacy Practice, Purdue University, School of Pharmacy and Pharmaceutical Sciences, West Lafayette and Indianapolis, Indiana; and the Department of Medicine, Indiana University, School of Medicine, Indianapolis, Indiana

ABSTRACT: Biostatistics is the application of statistics to biologic data. The field of statistics can be broken down into 2 fundamental parts: descriptive and inferential. Descriptive statistics are commonly used to categorize, display, and summarize data. Inferential statistics can be used to make predictions based on a sample obtained from a population or some large body of information. It is these inferences that are used to test specific research hypotheses. This 2-part review will outline important features of descriptive and inferential statistics as they apply to commonly conducted research studies in the biomedical literature. Part 1 in this issue will discuss fundamental topics of statistics and data analysis. Additionally, some of the most commonly used statistical tests found in the biomedical literature will be reviewed in Part 2 in the February 2008 issue.

The Basics

Sampling

Sampling is the most fundamental concept in both descriptive and inferential statistics. Sampling is the process of randomly obtaining information from larger bodies of information called populations. It is these samples that are used to describe or make inferences about the entire population. A *sample* is obtained from a larger population because in most instances, especially in the medical field, it is impossible to study the entire population. Therefore, samples are obtained from the selected population according to the specific research question and are used to predict valuable information about the entire population. Specific sampling procedures and methods for assurance of randomization are beyond the scope of this review, but the importance of the chosen sampling procedure should not be overlooked.[1]

In clinical studies, the specific individuals are commonly patients or healthy research subjects. The information collected from the subjects is referred to as variables. Variables are measurable characteristics or attributes of these research subjects (eg, weight, age, blood pressure). The collected variables are used as estimates of the actual population characteristics. The specific type of the variable collected is important to determine how to properly summarize the data and to determine what type of statistical test should be used to test a specific hypothesis.

Types of Variables

Variables can be classified as qualitative and quantitative. Qualitative variables can be further classified as nominal or ordinal. *Nominal* variables, also referred to as categorical variables, are descriptive for a name or category. For example, the sex of a research subject is a commonly collected nominal variable (ie, male or female). Sex is an unordered categorical variable. Categorical variables that have a specific order associated with them are termed *ordinal*. For example, nutrition studies in patients with chronic liver disease often assess the baseline severity of liver function using the Child-Pugh score (class A, B, or C).[2] These variables are categorical and more specifically ordinal because a class A score has a better prognosis associated with it than class B, which is better than class C.

Quantitative variables can be continuous or discrete. A variable is by definition a *continuous variable* if it can take on any value within a given range. By this convention, a continuous variable could take on an infinite number of possibilities for a given range. For example, age is an example of a continuous variable. Even if the protocol of a research study only recruits subjects between 20 and 30 years old, age remains a continuous variable in that study. There are still an infinite number of values that age could take, even though there is a predefined range for this study. Age can be reported in years, months, days, hours, seconds, and so on. Therefore, there is always a more accurate way to represent a continuous measure, such as age, and it is dependent on the

Correspondence: Kevin M. Sowinski, PharmD, BCPS, FCCP, Purdue University, Department of Pharmacy Practice, W7555 Myers Building, WHS, 1001 West Tenth Street, Indianapolis, IN 46202. Electronic mail may be sent to ksowinsk@purdue.edu.

0884-5336/07/2206-0629$03.00/0
Nutrition in Clinical Practice 22:629–635, December 2007

methods in the study. As an example, a 20-year-old research subject could be classified as 19.8 years old or as 19.76 years old, etc. There are an infinite number of ways to classify the age of this subject, and hence this variable fits the definition of a continuous variable.

Unlike continuous variables, *discrete variables* can only take on a limited number of values in any given range. For example, the Clinical Risk Index for Babies (CRIB) is a scoring system that takes into consideration several continuous and ordinal variables to provide an index of initial neonatal risk. The scoring system generates a whole number. For instance, the CRIB index cannot have a value of 1.4.[3] Therefore, the magnitude of the difference between a score of 2 and a 1 may not be equivalent to the difference between a score of 3 and that of a 2. In some cases, discrete variables may be grouped to make them easier to handle. Ordinal variables, which are categorical, are commonly assigned numeric values, which transform them into discrete variables.

Section 1: Descriptive Statistics

Descriptive statistics are used to summarize and display raw data that are collected or generated in research studies. This can be accomplished by both visual and numerical methods.

Visual: The Histogram

Trends and patterns can be uncovered by the visual display of raw data. This provides a structure that can be used to choose the appropriate methods to summarize the data and choose the most appropriate statistical analysis. There are countless approaches to visually represent data, and it is beyond the scope of this review to give specific examples of graphic representation found in the biomedical literature.[4] However, this section will briefly discuss a simple way to visually inspect raw data that helps determine its underlying distribution and, hence, select the proper statistical approach. Subsequently, this section will introduce the most commonly encountered distribution of continuous data (ie, the normal or Gaussian distribution) and set the foundation for the most commonly used summarization methods and statistical tests found in the literature.

The histogram is a commonly used and relatively simple method to quickly assess the underlying distribution of variables collected in research studies. A *histogram* is a graph used to display the frequency distribution of data. The frequency distribution is an ordered list of possible values that a variable can assume in a research study, along with the frequency that the value occurred in the study. Because continuous variables can take on an infinite number of possibilities for any given study, the frequencies are generally grouped into class intervals.

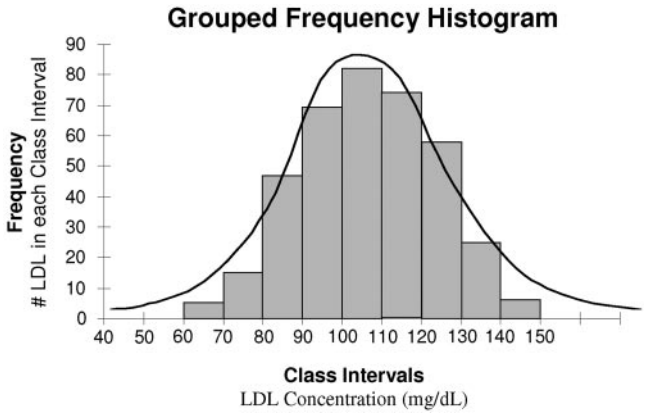

Figure 1. Grouped frequency histogram generated from a simulation of 1000 low-density lipoprotein (LDL) concentrations. The simulated data were broken down into 10 equal class intervals. The smooth line was generated as a symmetrical bell-shaped curve overlying the histogram, representing a normal population distribution.

For example, low-density lipoprotein (LDL) cholesterol concentrations were randomly generated for 1000 hypothetical subjects. LDL is a continuous variable because it can take on an infinite number of values in a given range. The 1000 hypothetical LDL concentrations ranged from 50 to 150 mg/dL. The raw data were grouped into 10 equal class intervals, and observed LDL concentrations were counted in each class interval as the frequency on the *y* axis of the histogram. Figure 1 displays the grouped frequency histogram developed from the generated data in this hypothetical study.

Histograms, such as the one in Figure 1, provide a starting point for researchers to classify data for further analyses. For example, the frequency distribution in Figure 1 displays a trend in the data that is observed for many biologic and physiologic variables. The distribution is approximately bell shaped, as demonstrated by the smooth line overlying the histogram. This smooth line is symmetrical, with either half being a mirror image of the other. These are the characteristics of a normal distribution (also called Gaussian distribution). As the sample size in this example is increased from n = 1000 to the size of the entire population, assuming the data are from a normal distribution, the histogram will more closely approach the smooth line in Figure 1. Data that follow a normal distribution can be appropriately summarized and analyzed by powerful statistical methods. A recurring error in the medical literature is reporting results for data that are clearly skewed by using statistical analyses that are only valid for normally distributed data or for which the variable is not continuous. These data should be analyzed by an alternative statistical method or transformed to approximate a normal distribution.[1] Although data transformation is beyond the scope of this review, alternative statistical methods for nonnormally distributed data will be discussed under

the descriptions of specific statistical tests in Part 2 of this review, to be published in February 2008.

Numerical: Measures of Central Tendency and Variability

The histogram is a powerful tool to sort and organize data, but it does not provide a simple summary indicating where the data are centered or the variability in the dataset. This information is reported using a measure of central tendency that describes the center of the distribution of the observed values and a corresponding measure to represent the variability or degree of dispersion in the dataset. The population measures of central tendency and variability are referred to as population parameters, whereas in a sample they are commonly referred to as statistics. The following notations for size, measures of central tendency, and variability will be used in this review:

Sample size = n
Sample mean = $\overline{X}$
Sample standard deviation = s
Population size = N
Population mean = μ
Population standard deviation = Φ

It is important to note that the population parameters, population mean (μ), and population standard deviation (Φ) will not be known in almost all instances. Therefore, the sample mean ($\overline{X}$) and sample standard deviation (s) are used to estimate the population parameters and are the basis for commonly used methods of statistical estimation and inference.

The 2 most frequently used measures of central tendency are the mean and median. The *mean* is simply the average of the data, whereas the *median* is the midpoint of the variables when they are placed in order of value. Although the calculations of these are fairly intuitive, there are certain types of data in which one is preferred over the other. Choosing the correct measure of central tendency depends on several factors, most importantly the distribution of the data. The most accurate measure of central tendency for data that do not follow a normal distribution is generally the median (eg, data with outliers). Unlike the median, the mean is affected by extreme outliers and will trend toward the tails of skewed distributions (ie, the end of the dataset that has extreme outliers). Biologic and physiologic data are generally skewed in the positive direction, which means that the extreme values are in the positive direction. In these cases, the mean would overestimate the central tendency of the data.

The mean is useful to indicate the center of the distribution for a given dataset. The median describes the middle value of a set of data. However, measures of central tendency alone do not provide any indication about the variability of the dataset. The variability associated with the median is generally reported by the absolute or interquartile range. The *absolute range* of any dataset is simply the maximum value minus the minimum value. The *interquartile range* is the difference in the value at the 75th percentile from the value at the 25th percentile. The value located at the 50th percentile of any dataset is, by definition, the median. This is generally more useful than the absolute range because extreme outliers do not influence the interquartile range.

The variance associated with a mean can be described as a measure of dispersion using the standard deviation, which is the square root of the variance. The absolute and interquartile ranges are limited because they are calculated from only 2 values in any given dataset. On the other hand, the standard deviation is calculated using all of the data in a sample and provides a more complete picture of variability. The standard deviation is not appropriate, however, to describe the variability of a nonnormal distribution. Furthermore, the standard deviation (SD) provides the most valuable information for data that follow a normal distribution, as stated by the empirical rule.[5] The empirical rule states that 68% of all values will be ± 1 SD away from the mean in a given dataset that is normally distributed. Furthermore, 95% of all values will be ± 2 SD away from the mean.

The data presented in Table 1 have been reproduced from a clinical study[3] to provide examples of measures of central tendency and variability among other examples that will be discussed in Part 1 of this review. The investigators were assessing potential mechanisms for a lower infection rate in very-low-birth-weight (VLBW) infants receiving glutamine-enriched enteral nutrition. Table 1 displays the baseline characteristics of infants assigned the glutamine-enriched enteral nutrition and those assigned the control diet. It is important to note that the CRIB is a discrete variable and has been appropriately presented as the median and absolute range in this table. The measure of central tendency and variability for the continuous variable (birth weight) is reported using the mean and standard deviation. The baseline birth weight in the glutamine-enriched enteral nutrition group is reported as 1.18 ± 0.4 kg in Table 1. By applying the empirical rule and assuming a normal distribution, approximately 95% of all babies in this study weighed between 1.10 and 1.26 kg (ie, ± 2 SD from the mean) in the glutamine-enriched enteral nutrition group.

The *standard error of the mean* (SEM) is also commonly reported in the literature. The SEM is used to construct confidence intervals for the population mean and perform hypothesis testing. Although a detailed description of the SEM is beyond the scope of this review, it is important to mention here because it has been used incorrectly in the literature.[5] SEM is calculated as the sample SD divided by the square root of the sample size. The SEM will therefore always be smaller than the

Table 1
Baseline and nutrition characteristics (modified with permission from van den Berg A et al[3])

	Glutamine (n = 52)	Control (n = 50)	p*
Antenatal corticosteroids	39/52 (75%)	39/50 (78%)	.72
Vaginal delivery	23/52 (44%)	24/50 (48%)	.70
Gestational age (wk)	29.3 ± 1.7	28.7 ± 1.8	.07
Birth weight (kg)	1.18 ± 0.4	1.16 ± 0.3	.79
Birth weight <10th percentile	17/52 (33%)	12/50 (24%)	.33
Sex (% male)	28/52 (54%)	27/50 (54%)	.99
Clinical Risk Index for Babies (CRIB)	2.5 (0–12)	3 (0–13)	.45
Age at start of study supplementation, d	2.6 (1.4–4.6)	2.5 (1.8–3.8)	.53
Time to full supplementation dose, d	10 (4–17)	9 (4–23)	.94
Age at increasing enteral nutrition, d	3.6 (0.2–11.8)	3.4 (0.7–10.1)	.92

Values are mean ± SD, median (range), or number (%).
*Student's *t*-test, Mann-Whitney *U* test, χ^2 test, and log rank test for continuous normally distributed data, nonparametric continuous data, dichotomous data, and time-dependent data, respectively.

sample SD and can make sample data seem to have less variability. It is frequently used in figures to increase the clarity of the figure by providing error bars that are shorter than they would be by using the SD. The SEM does not illustrate the variability of the actual population and should be interpreted cautiously.

Section 2: Inferential Statistics

An educated statement about an unknown population is commonly referred to in statistics as an inference. A statistical inference can be made by (1) estimation or (2) hypothesis testing. This section will provide a brief description of these fundamental statistical inferences. The following sections will provide examples of common statistical applications found in the biomedical literature.

Confidence Intervals

Estimation is a method that can be used to make an inference about a population parameter. *Confidence intervals* are commonly reported as a way to estimate a continuous population parameter. Confidence intervals are developed by first obtaining a random sample from the population of interest and then calculating the sample statistics (ie, mean and SD). Of note, in almost all instances the sample mean will not be identical to the true population mean. This phenomenon is due to sampling error and will be described in detail later in this review. Therefore, confidence intervals provide a range of values that are likely to encompass the true population mean with a certain level of confidence.

The first step in estimating a population parameter is to obtain a point estimate from the sample, as displayed in the Figure 2 schematic. The point estimate should be unbiased and the best available estimate of the population parameter of interest. For continuous, normally distributed data, the sample mean ($\overline{X}$) is commonly used as the point estimate for the population mean (μ). As displayed in Figure 2, the confidence interval is constructed from the product of the SEM and the predetermined level of confidence chosen to estimate the population parameter.

In the medical literature, 95% confidence intervals (95% CIs) are the most commonly reported. Although not entirely technically correct, this implies that 95% of the time the true population mean will fall within the given range in the CI. In some cases, 90% or 99% CIs are reported. As an example, refer to the baseline birth weight in the glutamine-enriched enteral nutrition group with a mean ± SD of 1.18 ± 0.4 kg, as reported in Table 1. Using the sample size, mean, and SD and assuming a normal distribution, the 95% CI is calculated to be 1.07–1.29. Essentially, this states that there is 95% certainty that the true mean of the entire population studied will have a mean weight between 1.07 and 1.29 kg. The 90% CI for this sample has been calculated to be 1.09–1.27. It is important to note that the 95% CI will always be wider (have a larger range) than the 90% CI for any given sample. Therefore, the wider the CI, the more likely it is to encompass the true population mean.

As noted above, CIs can be constructed for a single continuous variable with a normal distribution, but they can also be used to estimate the difference between an intervention or proportions such as odds ratios and relative risks. The differ-

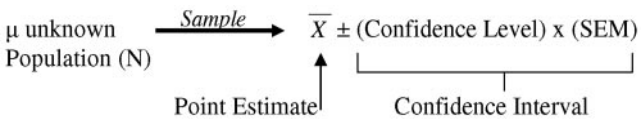

Figure 2. Schematic representing the fundamental elements needed to construct confidence intervals for estimation of the unknown population parameter (μ) from a random sample.
SEM, standard error of the mean.

ences between the SD, SEM, and CIs should be noted when interpreting the literature because they are often used interchangeably. Although it is a common misconception for CIs to be confused with SDs, the information that each provides is quite different and needs to be assessed correctly.

Hypothesis Testing

Hypothesis testing is used to answer specific research questions by making inferences about 1 or more populations. More specifically, hypothesis testing is used to make a prediction or inference about an observed difference in the measure of interest between 1, 2, or more experimental groups. In almost all situations, it is expected that a difference will be observed between the sample means of 2 groups due to random sampling. For example, if 2 random samples of n = 25 are obtained from the same population ($N \sim \infty$), the sample means and SDs may be quite different, and neither may be a good representation of the unknown population parameters. This is referred to as sampling error and is the basis of hypothesis testing. Sampling error is the difference between the parameter estimate based on the sample and the actual population parameter. Therefore, regardless of the scrutiny put into the design and implementation of a clinical trial, there will always be a certain amount of chance to make an incorrect inference due to sampling error.

Hypothesis testing involves 4 sequential steps. **(Step 1) Set up the hypothesis to be tested.** The primary hypothesis to be tested should always be defined *a priori*. If this is not defined before the study initiation, the inferences and study conclusions cannot be properly evaluated. The hypothesis to be tested should initially be set up in the form of a null hypothesis (H_0). The null hypothesis states that there is no difference in the outcomes tested. If the null hypothesis is rejected by hypothesis testing, then the conclusion will be based on the alternative hypothesis. The alternative hypothesis (H_a) is usually the opposite of the null hypothesis and states that there is a difference in the outcomes. Null and alternative hypotheses will be written differently, depending on the study design and the type of variable of interest.

An example of a null hypothesis can be easily imagined using the continuous variable of weight of VLBW infants receiving a glutamine-enriched diet *vs* those receiving a control diet. The null hypothesis could be stated as follows: the mean weight ($\mu_{glutamine}$) of VLBW infants receiving glutamine-enriched diets is equal to the weight ($\mu_{control}$) of VLBW infants receiving control diet. Note that the population mean (μ) is used to state the null hypothesis. Additional examples of null hypotheses for the commonly used statistical tests are provided throughout Parts 1 and 2 of this review.

(Step 2) Set the significance level and generate a decision rule. A decision rule needs to be developed after the research question has been stated in the form of a null hypothesis. The decision rule is used to determine the level of acceptable sampling error, more commonly referred to as the level of significance. Therefore, the decision rule is generated according to an acceptable error rate (α). The types of error associated with statistical tests are discussed in detail in the sections "Power and Statistical Error" and "Interpreting the *p* Value." In most instances, the acceptable α error rate is set at 5% or $\alpha = .05$ in the medical literature.

The 5% error rate ($\alpha = .05$), can be converted to a *critical value* that is specific for any given statistical test. Once the data are collected, a test statistic is calculated using the chosen statistical test. The *test statistic* can be directly compared with the critical value to determine if statistical significance was achieved. *Statistical significance* is achieved when a likely difference exists in the populations and the differences in sample means were likely not due to chance or sampling error alone.

The calculated test statistics and *a priori* critical values are rarely reported in clinical studies. This is due to the fact that each individual statistical test will have a different critical value associated with an $\alpha = .05$, and most statistical software packages will convert the test statistic directly to a *p* value. Therefore, in the medical literature the test statistic is reported as a *p* value and compared directly to the predetermined α. The *p* value is the probability that you obtain a result at least as extreme as you observed if the null hypothesis were true. A detailed discussion of the *p* value and its meaning, including common misconceptions, can be found in the "Interpreting the *p* Value" section.

(Step 3) Perform the experiment and compute the test statistic. This step is individualized, depending on the design of the study and chosen statistical analysis. Experimental design is beyond the scope of this review.[6] The methods for computing test statistics for individual statistical tests are described in "Section 3: Commonly Used Statistical Tests" in Part 2 of this series.

(Step 4) Make an inference. Once the experiment has been completed, and data have been collected and analyzed, an inference will be made. The inference is a prediction based on the sample obtained from the large body of information, the population. It is on this inference that the conclusions of the study will be based. The inference is based on the predetermined critical value and calculated test statistic or, more commonly, the predetermined acceptable error rate and the calculated *p* value. The inference is made by rejecting or failing to reject the null hypothesis. If the *p* value is calculated to be less than the predetermined α, the null hypothesis will be rejected. If the *p* value is calculated to be greater than the predetermined α, there will be a failure to reject the null hypothesis. A

failure to reject the null hypothesis is not the same as accepting the null hypothesis as true. It simply indicates that there was not enough evidence to support the rejection of the null hypothesis.

As an example, refer to the previously stated null hypothesis: the mean weight ($\mu_{glutamine}$) of VLBW infants receiving glutamine-enriched diets is equal to the weight ($\mu_{control}$) of VLBW infants receiving the control diet. Following this study, if a p value were calculated to be less than .05, the null hypothesis would be rejected and the conclusion would be that the mean weight of VLBW infants receiving glutamine-enriched diets is *not* equal to the weight of VLBW infants receiving control diet. Of course when evaluating this conclusion, the reader will have to ensure the study was designed appropriately to minimize bias, that the study was designed for this specific hypothesis, and that the correct statistical test was chosen, given the variable of interest, the distribution, and other factors that are discussed in this review.

Power and Statistical Error

It has become a convention to set the α of a study at .05, and therefore if the calculated p value is less than .05, statistical significance is said to be achieved. However, just because a p value is reported to be less than .05, it does not definitively tell us that there is an actual difference between the populations sampled. By definition, this states that, assuming proper study design and analysis, there was less than a 5% chance to observe the difference in the sample means if they came from the same population. In other words, 5% of the time a researcher will conclude there is a statistically significant difference when one does not exist. This is one form of statistical error and is referred to as type I or α-error; α is the probability of a type I error. On the other hand, it is possible a conclusion could be made that there is not a statistically significant difference when one does exist. This is referred to as type II or β-error; β is the probability of a type II error. Type I (α) error will be described in detail in the section "Interpreting the p Value" of this review.

The *power* of a study is the ability to detect a difference between study groups if one actually exists. Study power is indirectly related to the likelihood of making a β-error; power is $= 1 - \beta$. Therefore, as study power increases, the likelihood of concluding that there is not a difference when there is one will decrease. The power of a study is dependent on (1) sample size, (2) the actual difference between the outcomes of interest (eg, difference between the actual population means μ_1 and μ_2), (3) the variability around each outcome, and (4) the predetermined significance level (α). Because the differences between the population means and the population variance cannot be influenced by the investigator, the only way to increase study power (decrease type II error) without increasing the type I error rate is to increase the sample size.

A statistical power analysis should be performed for every study *a priori* to determine the appropriate sample size in order to decrease the potential for a type II error. The acceptable type II error rate is generally 0.10 or 0.20, depending on the study, and corresponds to 0.90 and 0.80 study power, respectively. Given the acceptable type II error rate, a difference in the outcomes of interest that would be considered clinically significant, the expected variability in the measure, and the type I error rate, an appropriate sample size can be calculated. The sample size calculation is an important step to properly conducting clinical research. If the power of a study is not indicated for an investigation that failed to reject null hypothesis, the occurrence of a type II error should be considered. Furthermore, for studies in which the null hypothesis is not rejected, a power calculation can be recalculated using the actual observed difference in the sample means and the observed variability in that study. This information can then be used to determine the number of subjects needed to detect a difference in the populations of interest if the study were to be repeated or continued.

Interpreting the p Value

An inference is made according to obtaining 1 or more samples and the calculation of the p value. The conclusion of most research reports will rely heavily on the fact that statistical significance has or has not been achieved. In several cases, this statement may come down to the calculation of a single p value. It is therefore important that the calculation of this p value be done correctly and that the study be properly designed for that specific research question. It is also important that the reader have knowledge of the meaning of the p value and thus how to accurately interpret it.

As previously stated, the p value is the probability of obtaining results at least as extreme as observed if the null hypothesis were true. In other words, if 2 independent samples were randomly obtained from the same population, the p value is the probability of the magnitude of the observed difference in the 2 sample means. Therefore, 5% of the time, 2 sample means from the same population will be different enough that one would incorrectly conclude that they were different or from different populations with an $\alpha = .05$. In almost all cases, it will never be known if the null hypothesis is actually true because the entire population cannot be studied. Therefore, an erroneous conclusion suggesting that differences exist will occur 5 times out of 100.

An example illustrating the concept of sampling error and associated p values can be described by evaluating the reported p values in Table 1. This table was originally intended to demonstrate the similarities in the baseline characteristics of the

study participants before the nutrition intervention. Therefore, these subjects were theoretically sampled from the same population (ie, infants with a gestational age <32 weeks or birth weight <1.5 kg admitted to a neonatal intensive care unit). Although this table is reporting baseline characteristics, a *p* value has been reported to indicate whether there were statistically significant differences between the 2 study groups. Statistical tests were performed on these selected variables to indicate that the sampling error did not alter the conclusions after the assigned interventions. The *p* values are all reported to be greater than .05 and, therefore, it is concluded that the 2 study groups had similar baseline characteristics.

As a hypothetical example after the intervention in the study on glutamine-enriched enteral nutrition *vs* control, imagine that glutamine had absolutely no physiologic effect (in reality this would be unknown). Therefore, when the primary outcome is analyzed, (ie, intestinal permeability in this study), the same population would be assessed because no physiologic difference would have occurred. Therefore, if the study were repeated 100 times, 5 of them would incorrectly conclude that glutamine-enriched enteral nutrition altered the measure of intestinal permeability due to sampling error alone.

Statistical Significance vs *Clinical Significance*

As discussed, several important issues should be taken into consideration when evaluating *p* values and hence conclusions of research reports. One recurring issue is that statistical significance does not always relate to clinical significance. When assessing the clinical significance of an observed outcome, considerations should be assessed such as the study design and variable chosen as the outcome. Studies that assess a true clinical outcome, such as mortality, may have more clinical significance than one assessing the change in a clinical or surrogate marker, such as blood pressure. Furthermore, the general acceptance of the marker relating to true clinical events should be evaluated. That is, there are substantial data to suggest that lowering blood pressure below a certain cutoff will decrease mortality; however, such a cutoff may not exist for certain inflammatory markers.

Even using an accepted marker to assess a clinical outcome should be interpreted cautiously. An example is the case with hormone replacement therapy and its effect on LDL cholesterol. Investigations of hormone replacement therapy have demonstrated an LDL-lowering ability, but when clinical outcomes such as mortality for cardiovascular disease were evaluated, hormone replacement therapy was not effective and actually may have been deleterious.[7,8] The problems encountered with hormone replacement therapy are probably not due to the fact that LDL is a poor clinical marker for cardiovascular disease. More likely, the negative clinical outcomes were due to the fact that the negative actions of hormone replacement therapy outweighed the benefits of lowering LDL. Therefore, additional considerations to assess clinical significance include the risks *vs* benefits of the treatments being evaluated, which are often not assessed in clinical studies by statistical methods. If clinical effectiveness is demonstrated for any given intervention, the *p* value alone will not give guidance into the risks, discomfort, time consumption, or economic burdens of the intervention. These are all issues that must be considered when evaluating the biomedical literature for clinical significance, even when statistical significance is achieved.

Nutrition practitioners will benefit from understanding the basics of statistics. Part 2 of this article appearing in the February 2008 issue of *Nutrition in Clinical Practice* will expand on this topic and further address inferential statistics.

References

1. DeMuth JE. *t*-Tests. In: DeMuth JE, ed. *Basic Statistics and Pharmaceutical Statistical Applications.* Boca Raton, FL: Chapman and Hall/CRC; 2006.
2. Albers I, Hartmann H, Bircher J, Creuztfeld W. Superiority of the Child-Pugh classification to quantitative liver function tests for assessing prognosis of liver cirrhosis. *Scand J Gastroenterol.* 1989;24:269–276.
3. van den Berg A, Fetter WP, Westerbeek EA, van der Vegt IM, van der Molen HR, van Elburg RM. The effect of glutamine-enriched enteral nutrition on intestinal permeability in very-low-birthweight infants: a randomized controlled trial. *JPEN J Parenter Enteral Nutr.* 2006;30:408–414.
4. Larson MG. Descriptive statistics and graphical displays. *Circulation.* 2006;114:76–81.
5. D'Agostino RB, Sullivan LM, Beiser AS. *Introductory Applied Biostatistics.* Belmont, CA: Thomson Higher Education; 2006.
6. Stanley K. Design of randomized controlled trials. *Circulation.* 2007;115:1164–1169.
7. Anderson GL, Limacher M, Assaf AR, et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA.* 2004;291:1701–1712.
8. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA.* 2002;288:321–333.