

Preparing for the first meeting with a statistician

JAMES E. DE MUTH

The first visit with a statistician can be an intimidating experience. In order to take full advantage of this professional's expertise, this meeting should take place early in the research development process before any study data are collected. This can avoid many problems associated with both ill-conceived research designs and the collection of data that cannot be interpreted through traditional statistical analysis.

The purpose of this article is to review practical statistical issues that should be considered when performing data collection and analysis. It is intended to assist pharmacists in communicating more effectively with statisticians, with the ultimate goal being better quality research. This article focuses on questions to consider before the first meeting with a statistician. The questions relate to the intent of the study, whether the results will be representative of only those individuals involved in the study or projected to a large group of individuals, the type of data that will be collected, what measurements represent predictor and response variables in the study, and the role of hypothesis testing in providing interpretable results from the data analyzed. Closely related to the intent of the research is the study design used to meet the objectives of the

Purpose. Practical statistical issues that should be considered when performing data collection and analysis are reviewed.

Summary. The meeting with a statistician should take place early in the research development before any study data are collected. The process of statistical analysis involves establishing the research question, formulating a hypothesis, selecting an appropriate test, sampling correctly, collecting data, performing tests, and making decisions. Once the objectives are established, the researcher can determine the characteristics or demographics of the individuals required for the study, how to recruit volunteers, what type of data are needed to answer the research question(s), and the best methods for collecting the required information. There are two general types of statistics: descriptive and inferential. Presenting data in a more palatable format for the reader is called descriptive statistics. Inferential statistics involve making an inference or decision about a population based on results obtained from a sample of that population. In order for the results of a statistical test to be valid, the sample should be representative of the

population from which it is drawn. When collecting information about volunteers, researchers should only collect information that is directly related to the study objectives. Important information that a statistician will require first is an understanding of the type of variables involved in the study and which variables can be controlled by researchers and which are beyond their control. Data can be presented in one of four different measurement scales: nominal, ordinal, interval, or ratio. Hypothesis testing involves two mutually exclusive and exhaustive statements related to the research question. Statisticians should not be replaced by computer software, and they should be consulted before any research data are collected.

Conclusion. When preparing to meet with a statistician, the pharmacist researcher should be familiar with the steps of statistical analysis and consider several questions related to the study to be conducted.

Index terms: Data collection; Methodology; Pharmacists; Research; Statistics

Am J Health-Syst Pharm. 2008; 65:2358-66

study. Specifics of study design go beyond the scope of this article and will be addressed in another article in this series.

The process of statistical analysis involves the following seven steps: establishing the research question, formulating a hypothesis, selecting an

appropriate test, sampling correctly, collecting data, performing tests, and making decisions.¹ The actual mathematical manipulation of research data only constitutes one step in the process (performing tests). The other steps are of equal importance, and most are nonmathematical and

JAMES E. DE MUTH, B.S.PHARM, M.S., PH.D., is Professor, School of Pharmacy, University of Wisconsin, 777 Highland Avenue, Madison, WI 53705 (jedemuth@pharmacy.wisc.edu).

The author has declared no conflict of interest.

Copyright © 2008, American Society of Health-System Pharmacists, Inc. All rights reserved. 1079-2082/08/1202-2358\$06.00.
DOI 10.2146/ajhp070007

The Research Fundamentals section comprises a series of articles on important topics in pharmacy research. These include valid research design, appropriate data collection and analysis, application of research findings in practice, and publication of research results. Articles in this series have been solicited and reviewed by guest editors Lee Vermeulen, M.S., and Almut Winterstein, Ph.D.

should be considered before ever meeting with a statistician. In fact, with easy access to computer software for statistical analysis, performing tests may be the least important component. This article will focus on the first five steps of statistical analysis. The pharmacist seeking help from a statistician should consider several questions about the study and what they are trying to evaluate. A subsequent article in this series will address the actual statistical tests and the most important step: making a decision by a correct interpretation of the results.

What are the objectives of my study?

It is critical to thoroughly understand the intent of any research and to list the primary and any secondary objectives of a study. You should know your specific research questions. Why waste your time and associated expenses, not to mention potential risk to patients in the case of therapeutic intervention, without a clear understanding of what you are trying to accomplish as a result of the study? Once the objectives are established, the researcher can determine the characteristics or demographics of the individuals required for the study, how to recruit volunteers, what type of data are needed to answer the research questions, and the best methods for collecting the required information.

In order to receive institutional review board (IRB) approval to conduct any type of research study involving human subjects, the researchers are

required to provide clearly stated objectives as part of the purpose of the study. In addition, the variables to be collected and proposed methods for data entry and statistical analysis should be specified in the protocol presented to the IRB. If a description of the proposed methods of analysis is not included with the study protocol submitted, the final study report should describe how the methods used were selected.²

To illustrate many of the statistical and research design issues presented in this article, a fictitious study (the TRIAL study) is described. In this simple clinical trial, the effectiveness of a new sleep aid was compared with that of a placebo. In the protocol submitted to the IRB, the investigators listed their primary objective as “to evaluate changes in sleep status for patients administered a new sleep aid and compare the results to a control group.” The secondary interest was to make an assessment of any relationship between sleep deprivation and depression (as measured by traditional sleep and depression scores) and determine if factors such as age and gender influence changes in these scores. To accomplish the objectives of the study, a parallel design with two groups was used.³ The benefit of the new drug was assessed by measuring changes in sleep scores for the experimental group who received the drug in the recommended evening dosage while also observing any changes in sleep scores in the control group who received a placebo (identical in appearance to the sleep aid). Both groups were evaluated at the beginning of the study and again six months into the study. Patients returned to the clinic every two months for a progress evaluation until the final assessment at six months. To determine the primary measurement of quality of sleep, the Epworth sleepiness scale was administered.⁴ The secondary measure of depression was assessed using the Hamilton depression (HAM-D) scale.⁵ Only the

first 17 questions of the HAM-D scale were used to calculate a total score.

Do I plan to simply report results or make a statement about a population?

There are two general types of statistics: descriptive and inferential. In most studies, data are collected and the researcher is faced with the challenge of taking a large quantity of observations or data points and trying to present them in some type of condensed, logical arrangement. Presenting research data in a palatable format for the reader is called descriptive statistics. Results could be presented visually using images such as pie charts, bar graphs, histograms, and scatter diagrams. Results could also be presented numerically in tables or expressed as a center for a distribution with an associated dispersion of values around the center of that distribution. The presentation of descriptive statistics may be all that is required for a relatively simple study, and the pharmacist researcher may not require the assistance of a statistician. For example, outcomes from a Phase I clinical trial may require only that descriptive summaries include the number (n) and percentage for categorical variables; that n , mean, geometric mean, 95% confidence interval (CI), standard deviation, standard error, median, minimum, and maximum values for continuous variables are calculated; and that summary tables and listings are presented by dosage received. All of these factors, except the 95% CI, are descriptive statistics.

Inferential statistics represent what most people think of as statistics. Inferential statistics involve making an inference or decision about a population based on results obtained from a sample of that population. For example, taking information about a small subset of patients (e.g., patients given a new diuretic and classified as having congestive heart failure) and making a statement about a larger group of

patients (e.g., how all patients with congestive heart failure might react if given this diuretic) is an inferential statistic. Inferential statistical tests involve hypothesis testing. Common examples of inferential tests include the Student *t* test, analysis of variance models, correlation, linear regression, chi-square tests of independence, and nonparametric procedures. Based on the hypothesis being tested and the types of variables involved with data, the most appropriate test can be selected. The first three steps of statistical analysis should be performed before any data are collected.

A sample is used to make a statement about a population with an inferential statistic; therefore, a measurement of uncertainty (or random error) is required because error can be associated with the decision-making process. Based on the complexity of the analysis, a statistician can help choose the most appropriate test or tests, help determine sample size, and assist with the interpretation of the results of the statistical analysis or analyses.

In order to perform inferential statistics, the pharmacist researcher must first create descriptive statistics to summarize the collected data. These descriptive statistics are then incorporated into the equations used for inferential purposes.

How do I obtain a representative sample if I am trying to make a statement about a population?

As noted, inferential statistics make statements about a population based on a small subset of the population (referred to as a sample). Therefore, in order for the results of a statistical test to be valid, the sample should be representative of the population from which it is drawn and for which a determination is being made. If the sample is not representative of the population, there will be bias in the statistical results and this could lead to erroneous or misleading results. Even though the test

results may have internal validity for the data being analyzed, there could be a threat to the external validity of making generalizations to a larger population. It is important to know what population you are trying to make a determination for. Examples of populations include volunteers with normal health between the ages of 18 and 45, all patients seen in an emergency room over a 12-month period, all patients in a clinic with the chief complaint of a sleep disorder, all subscribers to a particular health maintenance organization (HMO), all patients in a specific geographic area diagnosed with Lyme disease, and all individuals in the world with chronic renal failure. Theoretically, by using a random numbers table, each individual in a population would have an equal chance of being selected for study inclusion in an attempt to avoid any bias in the results. Inclusion and exclusion criteria must be established and carefully observed when recruiting volunteers for a research project.

The first population example represents a typical Phase I clinical trial where healthy volunteers are recruited and anyone willing to take part in the study, and who also meets the inclusion and exclusion criteria, is eligible for the study. The second example represents a random selection of individual charts during the time period defined in the study. In a similar manner, patients in the HMO can be randomly selected. If the study involves patients in the HMO who receive care in multiple clinics in different cities and states, a sampling plan would be required to allow all patients at all clinics to have an equal chance of being selected in order to have results representative of all HMO subscribers. A design involving nested sampling or cluster sampling would be used. The Lyme disease population example offers similar problems. There are several different sampling plans that would be applicable given the objectives of the study and the population being sampled.⁶ The population example

that is representative of patients with a certain complaint is eligible for inclusion. The results of the study can be interpreted as representative of all patients seen in this particular clinic with the chief complaint of a sleeping disorder, but this sample might not be representative of patients worldwide who present with sleeping problems. The results of this study could possibly be extrapolated to larger populations but would require a similar extrapolation as the last population example of patients with chronic renal failure. More than likely, multiple sites over a wide geographic area would be required to be able to approximate all patients worldwide. The characteristic “chronic renal failure” needs to be carefully defined through specific inclusion and exclusion criteria to establish certain physiological values necessary to qualify a volunteer to be included in the study.

For clinical trials in which a pool of qualified volunteers has been established and multiple treatment options (including an active control or a placebo) are available, patients need to be assigned to these various treatments. In this scenario, random assignment to the various treatment subgroups is critical. Methods for randomization of patients are beyond the scope of this article, but information is readily available describing the process.⁷ In the case of a randomized clinical trial, every patient has an equal probability of being assigned to any one of the various therapy subgroups. This can be accomplished by using a random assignment list created by either a statistician or the pharmacist researcher. In the TRIAL study, the researchers recruited 100 volunteers who met the inclusion and exclusion criteria. These patients were then randomly divided into two equal-sized groups; one subset of 50 patients were the experimental group and the second subset of 50 were the control group.

The TRIAL study was designed to evaluate changes from the beginning to the end of the study. The

study design assumes that any variable that can influence both groups will be distributed equally between the two groups; therefore, any difference in changes in sleep scores between the two groups is attributable to the intervention received. In most journal articles that describe studies that compare multiple treatment approaches or drugs, the authors present a table that compares patient demographics and other clinical variables that may affect the outcome between the treatment and control groups. The purpose of this comparison is to assess whether the randomization process worked or whether significant differences existed between the study groups. If no differences are identified before treatment, then subsequent differences observed later in the study are assumed to be caused by the different treatments received during the study.

In the TRIAL study, the volunteers were divided into two equal-sized groups using a random numbers table. For this fictitious study of a new sleep aid, the researchers established the inclusion criteria as otherwise healthy patients (no known comorbidities), at least 18 years of age, and a chief complaint of difficulty sleeping. Exclusion criteria included any previous treatment for sleep disorders with any product in the same class as the experimental agent. Based on the data from TRIAL study, the demographic information at the beginning of the study is provided in Table 1. The randomization of patients in the study was successful, and there were no significant differences between the two subgroups based on the chosen variables established on the assumption that any *p* value greater than 0.05 is not statistically significant. A detailed discussion of *p* values appears later in this article.

All inferential statistics have two basic requirements. First, data must come from a representative sample, ideally a random sample. Second, observations should be measured

independently of each other, in that no member of the sample affects the outcomes for any other member in the sample. These requirements can be accomplished through careful study design and data collection. In the case of the TRIAL study (if measured independently), it was assumed that the extent of depression (measured by the HAM-D score) for any given patient would not influence the depression score for any other volunteer in the study and that volunteers were randomly assigned to the two experimental conditions.

In Table 1, the information from the TRIAL study represents sample data, and the measures of central tendency are presented as the mean and the standard deviation. The mean (sometime the median but rarely the mode) is the measure of the center of the data when the variable is measured on a continuum or is quantifiable. In general, researchers will not know the true population mean (μ), but the sample mean (expressed as \bar{X}) will represent the best estimate. The sample mean (\bar{X}) is the weighted center or average, the sum (Σ) of all the observed values (x_i) divided by the total number of observations (n):

$$\bar{X} = \frac{\sum x_i}{n}$$

For dispersion of data around the

center, the common measure is the sample standard deviation (*S* or S.D.; *S* is typically used in statistical equations, while S.D. is used when study results are reported) as the best estimate of the population standard deviation (σ). To determine the sample standard deviation, an intermediate measure, the variance, is calculated. The variance (S^2) is the average of the squared differences for the observations from the sample mean:

$$S^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}$$

This average is calculated by dividing by the degrees of freedom ($n - 1$) rather than the actual number of observations (n). This creates a slightly larger value for the dispersion since the researchers are naive about the actual population variability. To get this measure of dispersion back into the same units used to express the sample mean, the square root of the variance is calculated to produce the standard deviation:

$$S = \sqrt{S^2}$$

Table 2 lists the most commonly used measures of center and dispersion and the appropriate symbols. Once again, it is important to emphasize that a sample needs to be representative of the population in order to

Variable	Experimental Group (n = 50)	Control Group (n = 50)	<i>p</i>
Mean ± S.D. age (yr)	54.2 ± 17.4	54.8 ± 16.7	0.87
Female (%)	62	62	1.00
Mean ± S.D. pretreatment HAM-D score ^a	11.7 ± 6.3	11.2 ± 5.8	0.68
Mean ± S.D. pretreatment Epworth sleepiness score	12.2 ± 3.4	12.4 ± 3.3	0.79
Diagnosis of insomnia (Epworth score ≥10) (%)	80	76	0.63
Diagnosis of sleep apnea (Epworth score ≥18) (%)	12	10	0.75

^aHAM-D = Hamilton Depression Scale.

have sample descriptive statistics that are appropriate estimates (sample descriptive statistics are sometimes referred to as point estimates).

In the TRIAL study, the calculated sample means and standard deviations on the Epworth sleepiness scale for the experimental and control groups at the beginning of the study were 12.2 ± 3.4 and 12.4 ± 3.3 , respectively. They are not identical, but inferential tests can determine if the difference between the groups is statistically significant.

When collecting information about volunteers, researchers should only collect information that is directly related to the study objectives. The researcher should avoid “nice to know” data that are not critical to the study. For example, in the TRIAL study, the researchers could have collected information about height, weight, ethnicity, or religious orientation, but these factors probably would not have greatly impacted the volunteers’ sleep habits. Age and gender are also questionable factors for the purpose of this study. Adding unimportant variables to a study can increase the time for collecting the data (and subsequently the cost) and decrease the response rate (especially with written surveys).

What types of data are involved in my study?

Important pieces of information that a statistician will require first are an understanding of the type of variables (or factors) involved in the study and which variables can

be controlled by researchers (e.g., measurement scales) and which are beyond their control (e.g., predictors versus response variables). All this information is critical in choosing the most appropriate inferential statistical tests for evaluating research data.

Data can be presented in one of four different measurement scales: nominal, ordinal, interval, or ratio.⁸ A nominal scale involves categorical or group information. Such variables represent qualitative data. Examples of nominal data (sometimes referred to as discrete variables) include distinct dosage forms (tablets versus capsules), therapeutic subgroups (treatment A, treatment B, placebo), and outcomes that pass or fail a specific criterion. Levels for a nominal variable must be exhaustive (the categories account for all possible outcomes) and mutually exclusive (observations cannot fall in more than one category). For example, in the TRIAL study, the two treatment groups represented a nominal variable. All volunteers were classified as either experimental or control and could not fall into more than one group. Also, the control and experimental groups exhausted all the possibilities as interventions for this study. A nominal variable can represent predetermined blocks of data, such as above and below a certain point in a distribution. The total scores on the Epworth scale can vary from 0 to 24 (discussed below), but results can be classified as normal (0–10), sleeping disorder (11–17), and sleep apnea (18–24). In

the TRIAL study, the distribution of possible scores was broken into three discrete categories. Based on sleep score at the beginning of the study, each volunteer fell into only one possible category, and the three categories exhausted all possible outcomes. Descriptive statistics for nominal variables are limited to reporting frequency counts and percentages (or proportions).

Two closely related continuous scales are the interval and ratio measurement scales. For both scales, the difference between each measurement unit is equally distant (e.g., the distance between 1 and 2 in is the same as between 6 and 7 in). The scales represent a quantitative measure with an equal difference between units. However, for an interval scale, the ratio between the scale values have no meaning because of an arbitrary zero (e.g., temperature that has a ratio between 40 and 20 °F does not imply that 40 °F is twice as hot as 20 °F). With the ratio scale, there is a genuine zero point (e.g., length, weight, percent). Using weight as an example, an object weighing 500 mg is twice as heavy as an object weighing 250 mg. The precision of the measuring instrument is the limitation to how fine a measurement can be made on an interval or ratio scale. For example, a digital bathroom scale may report weight to the whole kilogram (assuming the use of a metric scale). The same weight could be taken with more precise balances and reported to the gram, milligram, or even microgram. Factors that are measured using interval and ratio scales are often referred to as continuous variables. For continuous variables, the measures of central tendency would be the sample mean and standard deviation.

Ordinal measures fall between nominal and interval or ratio scales. An ordinal scale involves relative positioning. A five-point Likert scale, with “5, strongly agree” being the most positive response and “1,

Table 2. Symbols Used for Sample Statistic versus Population Parameter

Description		Symbol		
Measure	Name	Sample		Population
Center	Mean	\bar{X}	≈	μ
Distribution	Variance	S^2	≈	σ^2
Distribution	Standard deviation	S	≈	σ
Volume	Sample size	n	≈	N

strongly disagree” being the most negative response, is an example of an ordinal scale. The TRIAL study used an ordinal scale for each of the eight questions on the Epworth scale. For example, one question assessed sleepiness in the situation of sitting and talking to someone, and the only possible responses were

- 0 = would never doze or sleep
- 1 = slight chance of dozing or sleeping
- 2 = moderate chance of dozing or sleeping
- 3 = high chance of dozing or sleeping

As the number increases, there is an increase in sleepiness, but the magnitude of change might not be the same between “never” and “slight” as between “moderate” and “high.” The numbers represent subjective responses on a limited scale. Ordinal scales are often considered continuous, but special caution should be used in determining which statistical test is most appropriate for these data. For ordinal variables, the measures of central tendency would be the sample median (50th percentile), with the corresponding 25th and 75th percentiles as measures of dispersion.

Problems with the ordinal scale arise when calculating a single total sleepiness score by summing up the eight questions and when a resultant score ranges from 0 to 24. Is the result still an ordinal scale, or can it be considered a ratio because there are now 24 possible values with relative positions? It is not a perfect continuum, because a value such as 7.68 is not a possible outcome; only whole numbers can be reported and evaluated. In situations with large numbers of ordinal values, statistical tests that are appropriate for index or ratio scales can be used.⁹ However, some statisticians would argue that even an ordinal scale with only two possible outcomes could be treated

with tests appropriate for ratio data,¹⁰ while others would argue that such a test should not be used no matter how many possible ordinal levels, even if it represents the sum of multiple smaller ordinal scales (e.g., the Epworth total score).¹¹ Other statisticians would argue that, even if a logical order exists, data should not be considered ordinal unless there are at least five levels.¹² Subscribing to this philosophy, individual questions on the Epworth instrument could not be evaluated as ordinal data. It is up to the individual researcher, possibly in consultation with a statistician, to determine whether to treat such data as ratio or ordinal data. Different inferential statistical tests should be used when dealing with ordinal data as an outcome.

Once individual variables are defined by the type of measurement scale involved (ratio, nominal, or ordinal), the second thing to determine is whether each variable can be controlled by the researcher. In the example of the TRIAL study, the researchers determined that there would only be two types of intervention: (1) the experimental group with the new sleep aid and (2) the control group receiving a placebo. The intervention group is referred to as the independent variable, or predictor variable (in the case of observational studies). For this study, the intervention represented an independent discrete variable with two levels. The dependent variable is something that is beyond the researcher’s control: a response variable or an outcome. In the TRIAL study, the researcher could not control how sleep or depression scores changed over the six-month study period. Those responses were dependent on the group of volunteer patients. Thus, the total Epworth score at the end of the study was a continuous dependent variable. For this example, a 24-level ordinal scale was considered a ratio scale for statistical evaluation.

Selecting an appropriate test should be revisited at this point of statistical analysis. If the statistician or pharmacist researcher can define the independent and dependent variables for a specific research question and can determine whether the variables are discrete, ordinal, or continuous, the most appropriate statistical test can be selected. Examples will be presented in a subsequent article; for the purposes of this article, it is sufficient to say that the type of variable and the corresponding scales will dictate which inferential statistic to use in the analysis.

What are the hypotheses that I am trying to test in my study?

As mentioned previously, the second step of statistical analysis is to formulate a hypothesis, which is a simple statement that is important for interpreting the result of an inferential test. The decision to accept or reject the hypothesis is based on its associated *p* value.

Hypothesis testing involves two mutually exclusive and exhaustive statements related to the research question. In other words, the results need to be either A or B. There is no possibility for an outcome to be interpreted as C, D, etc. For the time being, A will be called the hypothesis being tested or, more commonly, the null hypothesis. If A is proven false, the only possible alternative is B, which is called the research hypothesis or alternative hypothesis. Traditionally, the symbols H_0 and H_1 are used for the null hypothesis and alternative hypothesis, respectively. Examples of these two statements are

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

The null hypothesis states that the two population means are equal. Notice that the two hypotheses are mutually exclusive (cannot happen both ways) and exhaustive (there is no possible third alternative). In the

example hypotheses presented, the two comparison groups are either equal to each other or statistically different. Note that the interpretation is expressed in terms of the two population means (μ_1 and μ_2), but data to make the decision will be derived from two sample means (\bar{X}_1 and \bar{X}_2).

When preparing hypotheses for testing, the null hypothesis is typically written in terms of zero: no difference or no relationship. Table 3 provides some commonly used null hypotheses associated with different inferential statistical tests. The way these tests are designed, it is not possible to prove that the null hypothesis is true; one simply fails to reject it. However, based on the test results, it is possible to reject the null hypothesis and prove the alternative hypothesis with a certain amount of confidence (e.g., 95% certainty or 5% possibility of being wrong). This creates a problem if the researcher wants to prove that there is no difference. In the previous example, the null hypothesis is $\mu_1 = \mu_2$. To test for equivalency, special tests are required and go beyond the scope of this paper. Bioequivalence testing has been reviewed by Schuirmann¹³ and Chow and Liu.¹⁴

Inferential tests and the associated hypotheses make statements about a population based on sample data. The entire population is usually unknown; therefore, such statements could be erroneous and the researcher and statistician must deal with this

uncertainty. There are two types of error that can occur (Type I and Type II), and these are illustrated by the two-by-two contingency table presented in Figure 1. In the real world, the null hypothesis is either true or false. The researcher must reach a decision based on sample data or a small subset of that population. In an ideal situation, the researcher will fail to reject the null hypothesis when it is true or will reject the null hypothesis if it is false. Unfortunately, since the researcher is making a decision based on sample information, errors can occur. The first error, called Type I error or alpha (α), is rejecting a null hypothesis when it is true. Typically, the researcher would like to have a less than 5% chance of being wrong when rejecting the null hypothesis or be 95% confident of that decision ($1 - \alpha$). The Type I error rate is easily controlled by the researcher through tables of critical values and the decision rule established for rejecting the null hypothesis.

The p value is sometimes mistakenly used synonymously with Type I error. The α , or Type I error, is defined by the researcher in advance as an acceptable level of error in the decision to reject the null hypothesis as false (usually expressed as a proportion, either 0.05 or 0.01). For example, a predetermined decision rule for an inferential statistical test might be as follows: with α equal to 0.05, reject the null hypothesis if the test statistic or resultant value is greater than the value from the

probability table of critical values based on the number of degrees of freedom based on sample size. In this case, the researcher would expect to have a 5% or less chance that the sample results are not true for the population.

The p value is determined after the test is completed and usually reported as part of the output in computer-generated results from an inferential statistic. It indicates how confident the researcher is that the observed results from the sample are also true for the population ($1 - p$). For example, a resultant p value of 0.016 would indicate that the researcher could expect to have a 1.6% chance that the sample results are not true for the population or 98.4% confidence that the sample results are true for the population. Although α and the p value are essentially the same types of errors, α is an established criterion before the test is performed and the p value is probability from the inferential test results.

Type I error can be established by the researcher by setting a predetermined acceptable level from a table of values based on the sample size, but Type II error is more difficult to determine. As seen in Figure 1, in the ideal situation, the researcher will reject the null hypothesis when the null hypothesis is truly false. This is referred to as statistical power ($1 - \beta$). Failure to reject a false null hypothesis is called a Type II error or β . Both Type II error and power involve the

Table 3. Examples of Null Hypotheses

Test	Null Hypothesis	Meaning
Two-tailed Student's t test	$H_0: \mu_1 = \mu_2$	No difference between population means for two groups
One-tailed Student's t test	$H_0: \mu_d = 0$	No mean change between two measures
Analysis of variance	$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$	Alternative hypothesis; H_1 : not all population means are the same
Correlation	$H_0: r_{xy} = 0$	No correlation; H_1 : there is a correlation
Linear regression	H_0 : no linear relationship	H_1 : there is a linear relationship
Chi-square	H_0 : no association (independent)	H_1 : there is a relationship (association)
Many nonparametric tests	H_0 : samples from the same population	H_1 : two or more subgroups are different (i.e., different populations)

Figure 1. Hypothesis testing. The null hypothesis (H_0) is either true or false, and the researcher reaches a decision based on sample data (represented by rows). The four potential outcomes include $1 - \alpha$ (do not reject H_0 when in fact it is true), α (reject H_0 when in fact it is true), $1 - \beta$ (reject H_0 when in fact it is false), and β (do not reject H_0 when in fact it is false).¹

		H_0 Is True	H_0 Is False
		Results of Statistical Test	Fail to reject H_0
Reject H_0	Type I error α or p		$1 - \beta$ power

interrelationship of several determinants including sample size; amount of uncertainty, expressed as variance (the square of the standard deviation); what amount of difference is important (i.e., clinically significant); and the acceptable level of Type I error. This interrelationship can be seen in the example of one type of power determination:

$$t_{\beta} \geq \frac{\delta}{\sqrt{\frac{2S_p^2}{n}}} - t_{\alpha/2}$$

In this example, t_{β} is an estimate of Type II error based on a Student t distribution (to be discussed in a subsequent article), S_p^2 is an estimate of the population variance (σ^2) based on sample data, delta (δ) is the magnitude of an important difference between the comparison groups, and $t_{\alpha/2}$ represents the amount of acceptable Type I error. Power determination would be calculated as $1 - p(t_{\beta})$. One controllable factor for determining statistical power is sample size. If the sample size increases, there will be a corresponding increase in statistical power. Whereas $\alpha = 0.05$ is the most commonly used level for Type I error, based on the type of study, other levels may be employed (e.g., 0.10, 0.01). Power of $1 - \beta = 0.80$ is usually acceptable. Once the statistical test has been performed, the resultant p value will be influenced by the sample size and variability in the data.

The problem that arises with the approximation of the population variability is that the researcher must have some estimate of how data will vary in order to make this determination. Such an estimate could come from previous experience with a drug (in the case of clinical trials) or data from trials with similar agents in the same drug class. The pharmacist researcher can provide the statistician with relevant background information (preferably published reports) in order to estimate expected values (means and variance) in the groups to be compared in a population with characteristics similar to the planned study. If data from a similar population are not available, the researcher should give some thought as to how these values can be conservatively estimated for the planned population. Also, the delta difference must be defined by the pharmacist researcher, not the statistician. What is a clinically important difference (in the case of the TRIAL study, differences between the two levels of the independent variable [new sleep aid versus placebo])? Is a 5%, 10%, or 20% difference in the Epworth sleepiness score important? This type of question cannot be answered or provided by the statistician; it must be determined by the pharmacist based on his or her understanding of the drugs and pharmacologic outcomes. The analysis should be a cooperative effort between the statistician and

the pharmacist involved with the research. A continuing dialogue and relationship between these professionals are important.

What should be the sample size for my study?

The sample size is directly related to the power analysis. If the previous equation is algebraically manipulated, sample size can be estimated using

$$n \geq \frac{2S_p^2}{\delta^2} (t_{\beta} + t_{\alpha/2})^2$$

Here, the researcher is once again faced with the issue of estimating a variance term. If no similar research exists, then no estimate of the variance, S_p^2 , can be made. It becomes a situation in which the variance is required to determine sample size, but the study must be completed before the variance can be calculated. A statistician with experience with similar situations can be extremely valuable in determining not only power but sample size for a desired power in a study. Zar¹⁵ offers an excellent reference to various formulas for determining power and sample size. Unfortunately, almost all of the power formulas require some estimate of the population variance. The sample size can be approximated, but when doing so, the researcher should be conservative and increase the number slightly to ensure that there are enough observations to reach the desired power and adjust for potential dropouts (patients lost on follow-up) or missing data points. Every effort should be made to minimize missing data through the study design and during the information collection phase.

Why should I consult with a statistician?

Many commercial software packages are available for statistical analysis. They offer easier, quicker, and more accurate ways to do statistical analyses compared with hand cal-

culations. In many cases, they have eliminated the need to meet with a statistician. However, the software can give the user a false sense of security. It is important to understand the software and how to enter and query the data. Even more important, the pharmacist researcher must have a good understanding of basic statistics and which test is appropriate for the given situation. Computers cannot make these judgments. Even when using sophisticated packages, the researcher still needs to interpret the output and determine what the results mean with respect to the model used for the analysis. This lack of understanding can result in erroneous interpretations, and the consequence might be the rejection of a poster or manuscript.

Altman¹⁶ showed a dramatic increase in the use of statistics in medical literature during the 1980s. With the increase in access to personal computers and statistical software, there is no reason to believe that this trend has not continued. However, at the same time, results of numerous studies have revealed that statistical and research design errors are far too common in medical and related literature.¹⁷⁻²⁰

Sophisticated computer software packages do not replace the need to work with a professional statistician to determine the most appropriate tests for a given set of data and the correct interpretation of the output. A statistician's input early in the process can aid in choosing the correct experimental design, developing appropriate measurement instruments, and selecting acceptable sampling and randomization strategies. This type of consultation can increase the efficiency of the study design and minimize costs and challenges to the validity of the findings. In some cases, the statistician's input

may result in the decision not to continue with the proposed study. However, as McGuigan²¹ noted in 1995, only a small portion of the articles he reviewed (24–30%) employed a statistician as a coauthor or acknowledged a statistician's help. In fact, in the peer review process, colleagues who review articles submitted to journals probably have about the same statistical expertise as the authors submitting the manuscript. Altman²² also noted that statisticians are not often used and are usually not recognized for their contributions to the published literature.

Statisticians are invaluable resources. In many instances, they will have had previous experience with similar research and can advise the researcher how to best approach and address the research questions and avoid inherent pitfalls. They understand what statistical tests can and cannot prove and their proper selection and use for different research designs. It cannot be overemphasized that statisticians should be consulted before collecting any research data. Finally, communication is essential, and an understanding of some of the basic principles of statistics can help the pharmacist researcher better use the expertise of a statistician.

Conclusion

When preparing to meet with a statistician, the pharmacist researcher should be familiar with the steps of statistical analysis and consider several questions related to the study to be conducted.

References

1. De Muth JE. Basic statistics and pharmaceutical statistical applications, second edition. Boca Raton, FL: Chapman and Hall/CRC; 2006:10-1.
2. Food and Drug Administration. 21 CFR Part 314.126. Adequate and well-controlled studies. www.access-data.fda.gov/scripts/cdrh/cfdocs/cfcfr/

- CFRSearch.cfm?fr=314.126 (accessed 2008 Sep 16).
3. Ascione FJ. Principles of scientific literature evaluation: critiquing clinical drug trials. Washington: American Pharmaceutical Association; 2001:62-3.
4. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*. 1991; 14:540-5.
5. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960; 23:56-62.
6. Bolton S. Pharmaceutical statistics. New York: Marcel Dekker; 1984:93-101.
7. De Muth JE. Basic statistics and pharmaceutical statistical applications. Boca Raton, FL: Chapman and Hall/CRC; 2006:42-4.
8. Stevens SS. On the theory of scales of measurement. *Science*. 1946; 103:677-80.
9. Dawson B, Trapp RG. Basic and clinical biostatistics. New York: Lange; 2001:27.
10. Van Belle G. Statistical rules of thumb. New York: Wiley; 2002:23-4.
11. Hailman JP. The continuing problem of fat classes and a rule of thumb for identifying interval and ratio data. *Bird-Banding*. 1969; 40:321-2.
12. Berry WD. Understanding regression assumptions. Thousand Oaks, CA: Sage Publications; 1993:47.
13. Schuirman DJ. A comparison of the two one-sided test procedure and the power approach for assessing the equivalency of average bioavailability. *J Pharmacokinetics Biopharmaceutics*. 1987; 15:657-80.
14. Chow SC, Liu JP. Design and analysis of bioavailability and bioequivalence studies. New York: Marcel Dekker; 2000.
15. Zar JH. Biostatistical analysis. Englewood Cliffs, NJ: Prentice Hall; 1999.
16. Altman DG. Statistics in medical journals: developments in the 1980s. *Stat Med*. 1991; 10:1897-1913.
17. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation*. 1980; 61:1-7.
18. Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in *Arthritis and Rheumatism*. 1882 versus 1967-68. *Arthritis Rheum*. 1984; 27:1018-22.
19. Vrbos LA, Lorenz MA, Peabody EH et al. Clinical methodologies and incidence of appropriate statistic testing in orthopaedic spine literature: are statistics misleading? *Spine*. 1993; 18:1021-9.
20. Kanter MH, Taylor JR. Accuracy of statistical methods in *Transfusion*: a review of articles from July/August 1992 through June 1993. *Transfusion*. 1994; 34:697-701.
21. McGuigan SM. The use of statistics in *British Journal of Psychiatry*. *Br J Psychiatry*. 1995; 167:683-8.
22. Altman DG. How statistical expertise is used in medical research. *JAMA*. 2002; 287:2817-20.