

Overview of biostatistics used in clinical research

JAMES E. DE MUTH

The purpose of this article is to provide a brief overview of the types of statistical tests that are available to analyze pharmacy research data. Many of the statistical and mathematical foundations on which these tests are based were presented in a previous article.¹ Descriptive statistics can be used to summarize data collected during research, but this article will focus solely on inferential statistical tests in which statements or decisions are being made about a larger population based on sample information. As described in the previous article, the process of a statistical test can be divided into seven steps: (1) establish the research question, (2) formulate a hypothesis, (3) select an appropriate test, (4) sample correctly, (5) collect data, (6) perform the test, and (7) make a decision.² This article will focus on three of those steps: selecting the appropriate test, performing the statistical test, and making a decision based on the results of the analysis. The most commonly used statistical tests will be presented under the conditions (types of variables) in which these tests are appropriate. The assumptions required in order to use the tests and how to interpret the results of the calculations from the statistical analysis (most commonly

Purpose. A brief overview is given of the types of statistical tests that are available to analyze pharmacy research data.

Summary. The most important aspect of selecting the correct statistical test is defining the types of variables being analyzed. Variables that are controlled or determined by the researcher are referred to as independent variables. Dependent variables are those that are observed and are out of the researcher's control. There are two types of random error that exist with inferential statistics: rejecting a null hypothesis (H_0) when it is true and failing to reject H_0 when it is false. There are two primary ways to interpret the significance of results from an inferential statistical test: (1) creation of a confidence interval and determination of whether a value falls within the interval and (2) calculation of a ratio and determination of whether the resultant value exceeds an established critical value. Student's t test is one of the simplest inferential tests and can be used to illustrate both the confidence interval

and the ratio approaches to evaluating sample data. The p value indicates the amount of error that can exist if the researcher chooses to reject H_0 . Parametric tests require two additional assumptions in order to be applied correctly. Some examples of these include the two-sample t test and the paired t test. Nonparametric tests are designed for small sample sizes and are easy to calculate. These tests use the median as the measure of center. Some examples of nonparametric tests include the chi-square test and the Fisher exact test. Other statistical tests that are available to help the pharmacist researcher include equivalency testing, survival statistics, and noninferiority studies.

Conclusion. Selection of the proper statistical test depends on the type and number of variables and whether parametric conditions are met.

Index terms: Methodology; Pharmacy; Research; Statistics

Am J Health-Syst Pharm. 2009; 66:70-81

seen as a computer output) will also be presented.

Selecting the most appropriate test

In order to select the correct statistical test, the most important aspect is to define the types of variables being analyzed. Each variable (also

referred to as a factor) needs to be defined with respect to both the type of measurement scale involved and whether or not the researcher has control over that specific variable.

Discrete versus ordinal versus continuous variables. There are four different types of measurement scales that can be used, and these

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

The author has declared no potential conflicts of interest.

Copyright © 2009, American Society of Health-System Pharmacists, Inc. All rights reserved. 1079-2082/09/0101-0070\$06.00. DOI 10.2146/ajhp070006

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.

have been discussed in a previous article.¹ A nominal scale represents categories (e.g., males versus females, control group versus experimental group). Each observation is required to fall into one of the mutually exclusive and exhaustive categories. Factors measured on nominal scales are sometimes referred to as discrete variables, and the outcomes are reported as frequency counts or percentages.

In contrast, interval and ratio scales represent quantitative data that can be measured, and there is relative positioning with no gaps or interruptions in the continuum (e.g., height, weight, percents, cholesterol level, blood pressure). The difference between interval and ratio scales is that the former has no true zero value. Factors that are measured using interval and ratio scales are often referred to as continuous variables. For continuous variables, the most commonly used measures of central tendency would be the sample mean and standard deviation.

A fourth scale, ordinal measures, falls between discrete and continuous scales. This type of scale represents information that has ascending or descending order, but the difference between units is not necessarily the same. Examples of ordinal scales would include stages I–IV for tumors and 0–10 Apgar scores for assessing the health of newborns. When there is a large number of divisions on the ordinal scale or multiple subscales are combined, they may be treated as interval or ratio scales. An example would be the Epworth

sleepiness scale.³ The total scores on the Epworth scale can vary from 0 to 24, but results can be classified nominally as normal (0–10), sleeping disorder (11–17), and sleep apnea (18–24). Each of the eight questions on the Epworth scale can be rated on a 0–3-point scale. For example, one question assesses sleepiness in the situation of “sitting and talking to someone” and the only possible ordinal responses are

- 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

Multiple questions can be combined to create an overall score ranging from 0 to 24 that could be treated as ordinal or ratio data. For ordinal variables, the measures of central tendency would be the sample median (50th percentile) and dispersion as the 25th and 75th percentiles.

Independent versus dependent variables. Some variables are controlled or determined by the researcher. These are referred to as independent or predictor variables. For example, if a pharmacist researcher wishes to compare three hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors on a hospital formulary, the three products chosen (different manufacturers) would be the researcher-controlled independent variable. However, individual patient responses to the three different products (e.g., change in total cholesterol levels after six months of therapy) are beyond the researcher’s control. The change in cholesterol level would be considered either a response variable or a dependent variable (dependent on the level of HMG-CoA reductase inhibitors received). The result measured for the dependent variable is associated with the classification (or

level) of the independent variable. In this example, there is a discrete independent variable with three levels or possibilities (drugs A, B, and C) and an associated continuous dependent variable (change in cholesterol). The most appropriate inferential statistic for this scenario would be the one-way analysis of variance, as will be seen below. The types of variables involved will dictate this choice in statistical test to be performed.

Dealing with random error

Systematic error, or bias, is minimized by good study design and care in collecting and recording observations. Uncertainty or random error always exists with inferential statistical tests and exists as two types: (1) rejecting the null hypothesis (H_0) when it is true and (2) failing to reject H_0 when it is false. These are labeled as type I errors (α or p) and type II errors (β), respectively. A complete discussion of hypothesis testing has been presented in a previous article.¹ When performing a statistical test, the researcher can never be 100% certain of the results because results from a small subset (sample) are being used to predict a larger population. The traditional acceptable level for type I error is less than 5%, or a p value of less than 0.05. Therefore, in published reports in which authors reject H_0 , they will usually cite a corresponding p value of less than 0.05. The smaller the p value, the more confident the researcher is in his or her decision to reject H_0 . At the same time, a type II error rate of 20% or less is usually acceptable and is a complement to the statistical power of the test results ($1 - \beta$).

As a consequence of the statistical test, H_0 can be rejected in favor of the alternative (research) hypothesis with less than a certain probability of being wrong in making this decision. However, if the researcher fails to reject H_0 , H_0 is not proven; there is simply not enough evidence to reject it.

Interpretation of statistical results

There are two primary ways to interpret the significance of results from an inferential statistical test: (1) creation of a confidence interval and determination of whether a value falls within the interval and (2) calculation of a ratio and determination of whether the resultant value exceeds an established critical value.

For the confidence interval approach, boundaries are established using the following general equation:

$$\text{population parameter} = \text{sample statistic} \pm (\text{reliability coefficient} \times \text{standard error})$$

These boundaries for the population parameter (i.e., the population mean, μ) will be estimated based on the sample results (i.e., the corresponding descriptive sample statistic, \bar{X}). These borders are calculated by adding and subtracting a correction factor to the observed sample statistic (the observed outcome from the sample). This correction factor is the product of a reliability coefficient (how much confidence the researcher requires [e.g., 5% or less error]) and an error term that accommodates for random variability (this standard error term is defined differently for each specific inferential statistic and usually represents a mathematical manipulation of the sample standard deviations; see the example below using the two-sample t test). The significance is determined by whether or not a predicted value (0 in the case of test of differences or 1 in the case of ratios) falls within the interval. Figure 1 illustrates two confidence intervals in which the predicted population parameter $\mu_C - \mu_D = 0$ represents a significant difference because a difference of 0 does not fall within the interval (when the difference between means for populations C and D could not possibly be 0). In contrast, $\mu_A - \mu_B = 0$ cannot be rejected as a possible outcome because 0

is a possible value inside the interval (where the difference between means for populations A and B could possibly be 0).

With the ratio approach to hypothesis testing, a test statistic is calculated by placing the comparison of interest in the numerator and the error term in the denominator:

$$\text{Test statistic} = \frac{\text{sample comparison of interest}}{\text{standard error}}$$

For example, the sample comparison of interest might be a difference (e.g., the difference between two sample means) and the error term might account for variability within each of the two samples. Figure 2 illustrates the interpretation of the ratio method. From one of many statistical tables, a critical value is selected based on the amount of acceptable error (usually 5% or less) and sample size (expressed as degrees of freedom). The result of the ratio is then compared with the critical value. Anything less than the critical value is considered random error or “noise” within the data. However, if the ratio (in terms of an absolute number) is larger than the critical value, there is a significant result and H_0 is rejected.

Using the Student t test to illustrate interpretation of test results. The Student t test is one of the

simplest inferential tests and can be used to illustrate both the confidence interval and ratio approaches to evaluating sample data. The two-sample t test is employed when there are only two levels or possibilities for a discrete independent variable and there is a continuous dependent variable as the outcome. To illustrate the two-sample t test, data will be used from the fictional TRIAL study in a previous article.¹ The TRIAL study was intended to evaluate the effectiveness of a new sleep aid by comparing the new product against a placebo control. Secondary objectives were to assess any relationship between sleep deprivation and depression (as measured by traditional sleep and depression scores) and determine if selected factors, such

Figure 1. Confidence intervals (horizontal lines) for predicted differences of 0 between the population mean (μ) of populations A and B and of populations C and D. The interval $\mu_A - \mu_B = 0$ cannot be rejected as a possible outcome because it includes the value 0 (the vertical dotted line). The interval $\mu_C - \mu_D = 0$ does not include 0 and thus represents a significant difference.

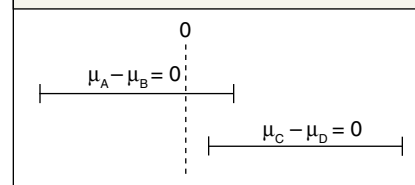
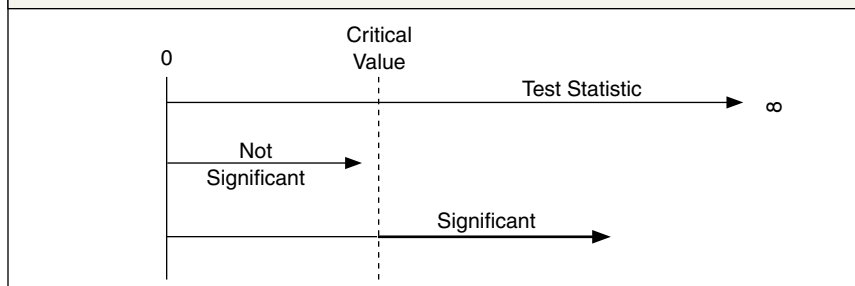


Figure 2. In the ratio method of hypothesis testing, the critical value for the test statistic is determined on the basis of the amount of acceptable error and the sample size. If the actual test statistic, calculated by dividing the sample comparison of interest by an error term, is less than the critical value, it is considered to represent random error. If the actual test statistic exceeds the critical value, the null hypothesis is rejected.



as age or sex, influenced changes in the scores. For the *t* test example, the pharmacist researcher compared Epworth sleepiness scale scores at the beginning of the study to determine if there were any differences between the control group (placebo) and the experimental group (new sleep aid) before exposure to the experimental conditions. At the initiation of the TRIAL study, 100 volunteers were randomly divided into each group. The initial Epworth scores for the experimental group (sample mean $[\bar{X}_E] = 12.36$, S.D. $[S_E] = 3.34$, sample size $[n_E] = 50$) and the control group ($\bar{X}_C = 12.18$, $S_C = 3.38$, $n_C = 50$) were different but represented sample data. There was a difference between the two sample means, but the researcher wished to know if that difference was important for the populations (i.e., all people meeting inclusion or exclusion criteria) or simply due to random variability in the data. To determine if the difference was significant, a two-sample *t* test was used (“two-sample” implies two levels of the discrete independent variable). In this case, the hypotheses (evaluating population means, μ_E for the experimental condition and μ_C for the control condition) could be written two different ways: (1) for the confidence interval:

$$H_0: \mu_E - \mu_C = 0$$

$$H_1: \mu_E - \mu_C \neq 0$$

where H_1 is the research or alternative hypothesis, or (2) for the critical value approach:

$$H_0: \mu_E = \mu_C$$

$$H_1: \mu_E \neq \mu_C$$

When the hypotheses are written in either style, the researcher wishes to reject H_0 (type I error, α) with a probability of less than 5% ($p < 0.05$). Going to a table of Student *t* values with 98 degrees of freedom ($n_E + n_C - 2$), the researcher would find a value of 1.98. This number would be used as

both the reliability coefficient in the confidence interval and the critical value for the ratio method of testing.

For the confidence interval approach, the equation is

$$\mu_E - \mu_C = (\bar{X}_E - \bar{X}_C) \pm t_{n_E+n_C-2}(1-\alpha/2) \sqrt{\frac{S_p^2}{n_E} + \frac{S_p^2}{n_C}}$$

where S_p^2 is the pooled variance (a weighted average of the squared standard deviations):

$$S_p^2 = \frac{(n_E - 1)(S_E)^2 + (n_C - 1)(S_C)^2}{n_E + n_C - 2}$$

In this example, the pooled variance would be 11.29. The calculations for the confidence interval boundaries would be

$$\mu_E - \mu_C = (12.36 - 12.18) \pm 1.98 \sqrt{\frac{11.29}{50} + \frac{11.29}{50}}$$

Thus,

$$-1.51 < \mu_E - \mu_C < 1.15$$

The researcher does not know for sure what the true difference between the two populations is but can estimate with 95% confidence that the true difference between the two populations ($\mu_E - \mu_C$) is somewhere between -1.51 and 1.15 score points on the Epworth scale. Because 0 falls within the interval, the researcher cannot reject the possibility that $\mu_E - \mu_C = 0$ and thus fails to reject H_0 .

Using the ratio approach, the test statistic and calculations would be

$$t = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{12.36 - 12.18}{\sqrt{\frac{11.29}{50} + \frac{11.29}{50}}} = \frac{0.18}{0.67} = 0.27$$

In this case, the *t* statistic (0.27) does not exceed the critical value (1.98), and the result is identical: failure to reject H_0 . In both cases, the researcher failed to find any significant difference between the two groups on the average Epworth scores at the beginning of the study. It is important to note that the researcher did not prove the two population means were equal (H_0), only that no significant difference could be found. As noted in a previous article, H_0 can never be proven.¹

The one-sample *t* test can be used with descriptive statistics (sample mean and standard deviation) to create an interval within which the true population mean is expected to be located with a given degree of certainty. This test will be discussed in greater detail below.

Interpretation of the *p* value.

Most computer programs will produce an output that not only lists the *t* statistic for the ratio method but also a corresponding *p* value. In the previous example, output would look similar to the following:

T-Test of difference = 0 (vs not =):
T-Value = 0.27 P-Value = 0.789

The *p* value indicates the amount of error that can exist if the researcher chooses to reject H_0 . In this case, the researcher could decide to reject H_0 but would need to accept a type I error rate (rejecting H_0 when it is true) of 79%. This far exceeds the normal acceptable standard of 5% or less. To reject H_0 , the computer report would need to present a *p* value of less than 0.05.

One-tailed versus two-tailed test of hypothesis.

Some inferential statistics allow directional testing. For example, if the researcher wishes to determine if one level of the independent variable is significantly larger or smaller, the hypotheses could be written

$$H_0: \mu_E \geq \mu_C$$

$$H_1: \mu_E < \mu_C$$

In this case, the hypotheses are presented such that the researcher wants to determine if the new sleep aid is significantly better (has a lower average Epworth score) than the placebo control. In this case, the research (alternative) hypothesis is directional and, if H_0 is rejected, would show that the new sleep aid is superior. To test these hypotheses, a different critical value needs to be taken from a Student t table where $1 - \alpha = 0.95$. This is referred to as a one-tailed test. In the case of a two-tailed test, the type I error is divided equally ($\alpha/2$) over both tails of the potential distribution of outcomes. If the one-tailed approach were used for the TRIAL study, the computer printout in Figure 3 would be created. In this case, based on the p value presented in Figure 3, the researcher would reject H_0 with less than a 0.1% error rate and conclude that the test sleep aid was superior to the placebo in decreasing the Epworth sleepiness scores.

Paired versus unpaired tests. Although the reduction in the group sleep scores in the previous example is informative, more useful information could be obtained by observing how each individual's sleep score changed during the study. This pairing of information can be more powerful and reduce the amount of random error. For example, when the same volunteers are evaluated twice (i.e., under both the experimental and the control conditions or before and after an intervention), the results have less variability because the volunteers are serving as their own control. Once again, a t test can serve to illustrate this approach to inferential testing.

For the TRIAL data, it is possible to evaluate the change for each individual at the end of the study compared to his or her sleep score at the beginning. This can be accomplished using a paired t test where the null and research hypotheses would be

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

where μ_d would be the difference in a much larger population based on the average sample difference. Each volunteer in the treatment group was measured twice: before the study and six months later at the end of the study. The difference was compared for each of the remaining 49 volunteers (1 person did not complete the study). The average difference (X_d) and standard deviation for the difference were -2.84 and 1.57 points on the Epworth scale, respectively. The critical value and reliability coefficient for $n - 1 = 48$ pairs (pre-scores and post-scores) would be 2.00 . The paired t test can be also evaluated using the confidence interval approach:

$$\mu_d = \bar{X}_d \pm t_{n-1}(1 - \alpha/2) \frac{S_d}{\sqrt{n}} =$$

$$-2.84 \pm 2.00 \frac{1.57}{\sqrt{49}}$$

Thus,

$$-3.29 < \mu_d < -2.39$$

Zero does not fall within the confidence interval, and $\mu_d = 0$ cannot be a possible value; therefore, the null hypothesis is rejected and the decision was a significant decrease in the Epworth sleepiness scores for the treatment group. Identical results are found if the ratio method is used:

$$t = \frac{\bar{X}_d}{\frac{S_d}{\sqrt{n}}} = \frac{-2.85}{\frac{1.57}{\sqrt{48}}} = -12.57$$

The absolute value 12.57 exceeds the critical value of 2.00 , so H_0 would be rejected and the resulting p value on the computer printout would be less than 0.001 . Therefore, the researcher can be 99.9% confident that there is a significant decrease in the sleepiness scores and, if similar patients with the same inclusion and exclusion criteria were administered the new sleep aid, there would be an expected average decrease in sleepiness scores somewhere between 2.39 and 3.29 points.

Survey of other statistical tests

Table 1 summarizes most of the statistical tests commonly seen in the pharmacy and medical literature. The selection of the inferential test is based on the types of variables involved. Is there an independent variable? Is it discrete or continuous? Are the dependent variables discrete or continuous? These are the questions that need to be answered, and this information provides guidance to test selection (Table 1). All of the tests listed in the table will be briefly discussed below.

Tests for a single variable

In some instances, there may be a single variable and the researcher may wish to make a statement about a population parameter based on a test statistic (e.g., the population mean estimated from the sample mean). Given sample data for a continuous variable (e.g., total cho-

Figure 3. Hypothetical printout from a computer's statistics program for a one-tailed t test. If the acceptable rate for a type I error is set at 5%, the p value, which represents an error rate of $<0.1\%$, indicates that the null hypothesis should be rejected. Statistical abbreviations in printouts may differ from standard abbreviations. Only 97 of 100 volunteers completed the study.

Two-sample T for Post-Epworth			
Group	N	Mean	StDev
Control	48	11.90	3.45
Treatment	49	9.57	2.84
Difference = mu (Treatment) - mu (Control)			
T-Test of difference = 0 (vs <): T-Value = 3.62 P-Value < 0.001			

lesterol levels, blood pressures, length of hospitalization), it is possible to estimate what these same averages would be in a given population. A confidence interval is created similarly to the two-sample *t* test:

$$\mu = \bar{X} \pm \left(t_{1-\alpha/2} \times \frac{S}{\sqrt{n}} \right)$$

In this one-sample *t* test, the researcher is using the sample mean, (\bar{X}) to estimate a range within which the true population mean (μ) should fall. To this best estimate, a factor that accounts for acceptable error and variability is added and subtracted to create an interval centered on the same mean. For example, the TRIAL study assessed the extent of volunteer

depression by using the Hamilton⁴ depression (HAM-D) scale. For all volunteers who entered the study, HAM-D depression scores (\bar{X} = 11.49, S = 6.03, n = 100) were used to create a confidence interval:

$$\mu = 11.49 \pm \left(1.98 \times \frac{6.03}{\sqrt{100}} \right)$$

Thus,

$$10.29 < \mu < 12.69$$

In this example, a *t* value of 1.98 was selected from a traditional Student *t* table, which can be found in statistics books. This value can be used with the sample information to create a 95% confidence interval.

Therefore, the researcher can assume that patients seen in the clinic with similar sleep complaints and the same inclusion and exclusion criteria will, on average, have HAM-D scores between 10.29 and 12.69.

With a discrete variable, there may be a preconceived or defined distribution expected across several categories. In this case, a chi-square (χ^2) goodness-of-fit test can be used to determine if the observed frequency for outcomes in the sample data is significantly different than hypothesized outcomes for each category.

In most inferential tests, the researcher will be concerned with at least two variables: the relationship between multiple dependent variables and the significance of one or

Table 1. Selection of Statistical Tests Based on Types of Variables

Independent Variable	Dependent Variable	Condition(s)	Test(s)
None	Continuous	One variable	One-sample <i>t</i> -confidence interval
		Two variables • Parametric requirements ^a • Nonparametric requirements ^b	Correlation Spearman ρ
None	Discrete	One variable	Chi-square goodness of fit
		Two variables	Chi-square test of independence
		Two variables, very small data sets	Fisher exact test
Discrete	Continuous	Independent variable (two levels) ^c • Unpaired, parametric requirements ^a • Unpaired, nonparametric requirements ^b • Paired, parametric requirements ^a • Paired, nonparametric requirements ^b	Two-sample <i>t</i> test Mann-Whitney <i>U</i> Paired <i>t</i> test Wilcoxon matched-pair test
		Independent variable (two or more levels) ^c • Unpaired, parametric requirements ^a • Unpaired, nonparametric requirements ^b • Paired, parametric requirements ^a	One-way analysis of variance Kruskal-Wallis Complete randomized block
		More than one independent variable ^a	<i>n</i> -way analysis of variance
		Two variables • Unpaired • Unpaired, very small data sets • Paired	Chi-square test of independence Fisher exact test McNemar test
		Risk estimates • Retrospective • Prospective	Odds ratio Relative risk ratio
		Two variables ^a	Linear regression
Continuous	Continuous	More than one independent variable	Multiple regression

^aParametric requirements are those in which a population is normally distributed and the variances are approximately equal.

^bNonparametric requirements are those that do not require the data to be equally distributed and are appropriate for ordinal dependent variables.

^cLevels refer to the number of possibilities in a discrete variable (e.g., three treatment options for the independent variable would indicate three treatment levels).

more independent variables on a dependent outcome.

Testing at least one discrete independent variable and one continuous dependent variable

The most commonly encountered statistics likely involve a discrete independent variable with multiple levels or categories (e.g., control versus experimental, drug A versus drug B versus drug C, males versus females) with some type of quantifiable outcome or response (continuous dependent variable). Two of these tests have already been described: the two-sample *t* test and the paired *t* test. These are parametric tests because they require two additional assumptions in order to be applied correctly. First, it is assumed that the samples are taken from populations that are normally distributed (i.e., they have a bell-shaped curve or Gaussian distribution). The sample may not look exactly like a bell-shaped curve, but it should have some of the characteristics of a bell shape, such as values clustered near the center with a few data points falling to the extreme in both directions of the continuum. Second, for each level of the independent discrete variable, the dispersion of the data should be approximately equal. This homogeneity of variance is assessed by observing similarities in the sample variances. A quick rule of thumb is that the largest variance should not be more than twice as large as the smallest variance to assume homogeneity.

As discussed, Student's *t* test can be used when there are one or two levels of the independent discrete variable. This test is not applicable when there are more than two levels or categories (e.g., interventions 1, 2, 3, up to *K* possible interventions). For these situations, a one-way analysis of variance (ANOVA or *F* test) is the method of choice. ANOVA is a parametric procedure and must meet the same assumptions of equally distributed populations and homo-

geneity of variance. The hypotheses are similar to those for the *t* test but expanded:

$$H_0: \mu_1 = \mu_2 = \mu_3 \dots = \mu_K$$

$$H_1: H_0 \text{ is false}$$

Note that the alternative hypothesis is not $\mu_1 \neq \mu_2 \neq \mu_3 \dots \neq \mu_K$. If H_0 is rejected, the result is some existing difference, but ANOVA does not indicate where that difference exists. In the case of a rejected H_0 , a special set of post hoc procedures can be employed to assess where differences exist among the various levels of the discrete independent variable.

The mathematics for calculating the *F* statistic are beyond the scope of this article, and the reader should refer to a basic statistics book for this information. However, the resultant *F* statistic is very similar to the ratio approach for the two-sample *t* test:

$$F = \frac{\text{difference between the means}}{\text{standard error of the difference of the means}} = \frac{MS_B}{MS_W}$$

The *F* value is calculated from a measure in the numerator, mean square between (MS_B), that accounts for differences between the sample centers and the denominator, mean square within (MS_W or mean square error), that accounts for the variability within each level of the independent discrete variable.

The computer output for the analysis of variance will be in the form of a table with a variety of information, the most important of which is the *F* statistic with its associated *p* value. This *p* value is interpreted the same as the two-sample *t* test described previously. For example, the researcher may wish to assess three age categories at the beginning of the TRIAL study to determine if there is any difference in the HAM-D scores. Using information from these three age groups (young adults, <46 years;

mature adults, 46–65 years; geriatrics, >65 years), the resultant ANOVA table is presented in Figure 4. The most important information of the ANOVA table is presented in the last two columns (*F* statistic, *p* value). For this example, assume the researcher wanted to be 95% confident in the decision ($p < 0.05$). Since the resultant *p* value is greater than the acceptable amount of error, the researcher would fail to reject H_0 and assumes there is no difference in depression score by age group.

The advantage of pairing individual outcomes was discussed when the paired *t* test was presented as a traditional before-and-after-type study design. But what if there are more than two time points being assessed, such as a test of volunteers' knowledge before and after an intervention and then again at 6 and 12 months postintervention? A paired *t* test is not the recommended test for this type of situation. A model that can be used to test *K* time periods is the complete randomized block design.

In certain instances, a researcher may wish to evaluate several discrete independent variables at the same time. For such situations, there are *n*-way ANOVAs. The simplest example would be the two-way ANOVA, in which two independent variables are tested at the same time on a single continuous dependent variable. In the TRIAL study, assume the researcher wanted to evaluate entry-level HAM-D scores for both sex and age groups. An advantage of this type of test is that not only can the effects of two independent (or predictor) variables be assessed in the test, but also any potential interaction between the two independent variables can be tested. In this example, there would be three null hypotheses being tested at one time with the creation of an ANOVA table evaluating each hypothesis:

$$H_{01}: \mu_{\text{young adult}} = \mu_{\text{mature adult}} = \mu_{\text{geriatric}}$$

$$H_{02}: \mu_{\text{male}} = \mu_{\text{female}}$$

- H_{03} : no interaction between gender and age groups
- H_{11} : H_{01} is false
- H_{12} : H_{02} is false
- H_{13} : H_{03} is false

Figure 5 illustrates the results of the analysis. The F statistic and p values are interpreted similarly to previous tests (looking at the two columns on the right side of the ANOVA table): (1) there is no significant difference based on gender (would require the acceptance of >9% rate for a type I error), (2) there is a significant difference based on the age group (<1.5% rate for a type I error), and (3) no significant interaction was observed between the two factors (>97% rate for a type I error). Larger n -way designs become more complex with multiple main effects, more two-way interactions, and additional n -way interactions to assess.

Testing two continuous variables

If all the data being evaluated are on continuous scales, there are two primary inferential tests: correlation and regression. The selection of the appropriate test depends on the presence or absence of an independent

variable. If there are no independent variables, the calculation of a correlation coefficient (r_{xy}) can be used to determine a significant relationship between the continuous variables and the strength of that relationship. The correlation coefficient can range from -1.00 (a perfect negative relationship) to 1.00 (a perfect positive correlation), with zero representing absolutely no relationship between the two variables. The hypotheses would be written

$$H_0: r_{xy} = 0$$

$$H_1: r_{xy} \neq 0$$

Rejecting H_0 indicates a significant relationship between the variables. The sign indicates the direction of that relationship.

In the case of two continuous variables, data can be plotted on graph paper with each variable on the x or y axis. For each data point, there will be a corresponding x and y value. These are used to calculate the r statistic. Each x and y value is squared and multiplied together, and all the values are summed to create the values Σx , Σy , Σx^2 , Σy^2 , and Σxy . These are used in the formula to calculate the correlation coefficient:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Most computer programs will produce the correlation coefficient and a corresponding p value that can be used to determine whether or not to reject the null hypothesis. If the researcher wanted to see if there was a correlation between the initial depression scores and sleep scores in the TRIAL study, the resulting computer output would be

$$r = 0.92, p < 0.001$$

There is a significant relationship, and this can be assumed with a less than 0.1% type I error. Does a bad depression score cause a bad sleep score or vice versa? Not necessarily. Correlation does not predict causality; it only confirms that a relationship exists. Using a scatter plot, it is possible to visualize this definite positive correlation (Figure 6).

With regression models, in contrast to correlation, there is at least one continuous independent variable that the researcher controls. For example, potential concentrations for a solution, based on percent, are on a continuum. However, usually a researcher will not just use random concentrations but will carefully prepare specific concentrations of a solution (e.g., 1%, 2%, 4%, 8%, and 16%) and measure some type of response on a second continuous variable (e.g., light absorption). Regression-type models are commonly used to determine if a straight line can be drawn through sample data points and how much of the total variability on the y axis (dependent variable) can be accounted for by the straight line (called the coefficient of determination and represented by the symbol r^2). A line is defined as

$$y = a + bx$$

where b is the slope of the line and

Figure 4. Hypothetical printout from a computer's statistics program for analysis of variance. The larger p value associated with the F test indicates that the null hypothesis cannot be rejected. DF = degrees of freedom, SS = sum of squares, MS = mean square.

Source	DF	SS	MS	F	P
Age Group	2	199.9	100.0	2.85	0.063
Error	97	3405.0	35.1		
Total	99	3605.0			

Figure 5. Hypothetical printout from a computer's statistics program for an n -way analysis of variance. The p value of 0.014 associated with the F statistic of 6.30 indicates a significant difference in the dependent variable on the basis of age group. DF = degrees of freedom, SS = sum of squares, MS = mean square.

Source	DF	SS	MS	F	P
Gender	1	83.2	83.2	2.43	0.094
AgeGroup	2	432.4	216.2	6.30	0.014
Gender*AgeGroup	2	1.4	0.7	0.02	0.979
Error	97	3223.7	34.3		
Total	99	3740.7			

a is the y intercept (where $x = 0$). If a linear relationship exists, then a point on the y axis can be predicted for any point on the x axis by inserting x into the previous equation. However, note that it is impossible to predict beyond the smallest or largest sample collected on the x axis. The straight line is defined between these two extreme points, but no information exists beyond these values.

Calculation of the linear regression model involves numerous equations. The end result is an ANOVA table that tests the hypotheses

$$H_0: x \text{ and } y \text{ are not linearly related}$$

$$H_1: x \text{ and } y \text{ are linearly related}$$

The coefficient of determination defines how much variability is accounted for by this line. It is also possible to calculate a confidence interval at any point on the x axis (Figure 7). If this were a stability study to determine if the researcher could be 95% confident that at least 95% of the active ingredients were available after three months of storage under normal conditions, the answer would be yes. The dots represent the actual sample results, and the straight line is the line of best fit calculated to have the least amount of variability between all the dots and the line on the y axis. The curved lines are the confidence bands (boundaries at any given point on the x axis), and the horizontal line represents the 95% threshold above which the product should be maintained for use.

Multiple independent (or predictor) variables can be tested for their effect on a continuous dependent variable, similar to the n -way ANOVA. This is referred to as a multiple regression. The linear regression model for a straight line can be written as follows:

$$y_i = a + \beta x_i + \varepsilon_i$$

where β is the true population slope

and ε is a measure of the random error. This formula can be expanded for a multiple regression model:

$$y_i = a + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_j x_{ij} + e_j$$

Here, the β s are referred to as beta coefficients, and they measure the influence of each predictor variable on the regression model. Computer output will present a table

Figure 6. Scatter plot for the hypothetical TRIAL study. Preintervention scores for the Epworth sleepiness scale (x axis) are plotted against the corresponding preintervention scores for the Hamilton depression (HAM-D) scale (y axis). r = correlation coefficient.

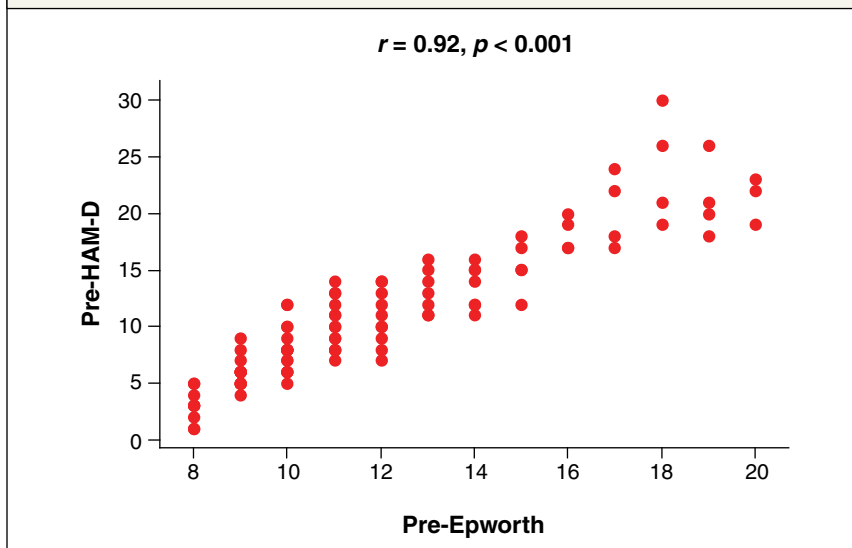
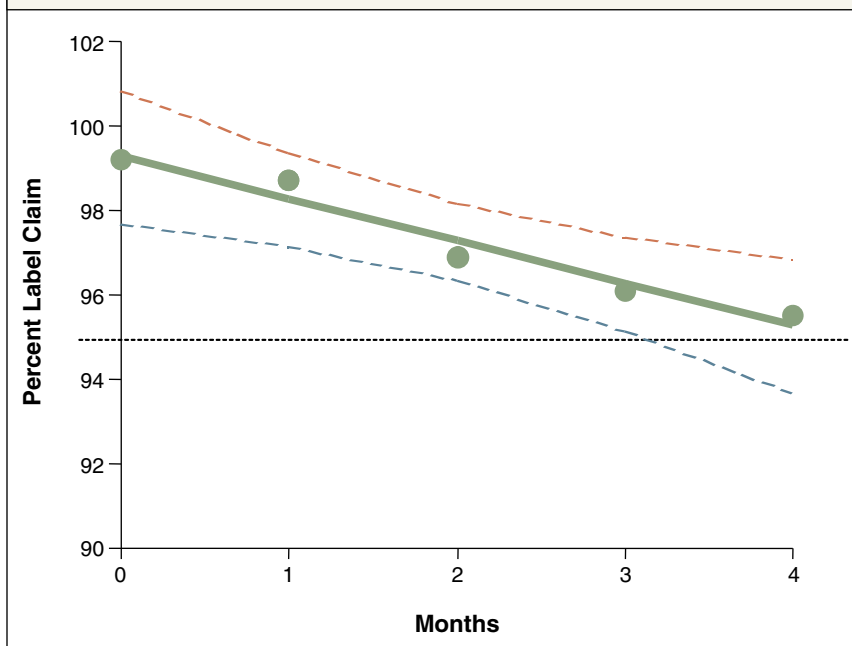


Figure 7. Linear regression plot for a hypothetical study of the stability of a drug. The solid line is the line of best fit. The curved broken lines show the boundaries of confidence limits for any given point on the x axis. The straight dotted line shows the study's arbitrary definition of stability (i.e., retention of 95% of drug claimed on the label).



that includes the beta coefficients and p values for each predictor variable. These are interpreted in a way similar to the results on an n -way ANOVA table.

Nonparametric alternatives

All the tests discussed to this point are described as parametric procedures. In order to use these tests, there are several required assumptions.⁵ These tests are considered robust tests if samples are drawn from populations that are equally distributed and if dispersions are similar (homogeneity of variance).^{6,7} Nonparametric tests offer a convenient alternative when the researcher is concerned that the population distribution may not be symmetrically bell shaped or when there is a large difference in the sample variances, which are the best approximations for the population variances. These nonparametric tests are also referred to as distribution-free statistics and use the median as the measure of center instead of the mean; therefore, they are not affected by outliers. They are designed for small sample sizes and are easy to calculate.

There continues to be spirited debate on the role of nonparametric tests for ordinal data.⁸ It has been recommended that parametric tests be limited to continuous outcomes measured on either interval or ratio scales.⁵ By default, nonparametric tests appear to be most appropriate for ordinal data.

There are two major problems with nonparametric tests. First, continuous data (ordinal, index, or ratio) are converted from the original measurement scale to a ranking from smallest to largest. These ranks are used for the actual statistical test. Information about the population is lost when the data are converted. Second, nonparametric tests have less statistical power than their parametric counterparts and offer a greater chance of causing a type II error (rejecting a true H_0).⁹⁻¹³

Table 2 lists the most commonly used nonparametric alternatives. The parametric test is more powerful and should be the test of choice unless visual inspection of the data shows a distribution greatly differing from normality or if there is a large disparity in the variances. Nonparametric statistics should be used when dealing with dependent variables that are measured on ordinal scales.

Testing two discrete variables

Evaluating two discrete variables will require a χ^2 test of independence and numerous related tests. At least one variable will be dependent, but the other variables could be independent or dependent based on the study design. Therefore, there may or may not be an independent variable involved in the analysis.

The test statistics involve creation of a contingency table, with one variable represented by the columns and another represented in the rows. A third possible confounder variable could be evaluated going into a third dimension using the Mantel-Haenszel χ^2 , but it will not be discussed in this article. The determination of significance is to compare the differences between observed sample outcomes and what would be expected under complete independence (no relationship) between the row and column variables:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency (count) of events for each cell in the contingency table and E is the corresponding expected value under independence. The expected value for any given cell in the table can be calculated by multiplying the sums of the observation for the cell's respective column ($\sum C_k$) and row ($\sum R_j$) and dividing that product by the total number of observations (n):

$$E = \frac{\sum C_k \times \sum R_j}{n}$$

The degrees of freedom equal the number of rows minus one times the number of columns minus one ($df = [R - 1][C - 1]$). If every observed value (the outcome) is equal to the expected value, the resultant χ^2 will be zero. As resulting deviations from the expected values increase, χ^2 will increase and eventually exceed the critical value, as illustrated in Figure 2. H_0 states that independence exists between the row and column variables and that no relationship exists between the two variables. If H_0 is rejected, the researcher can conclude that the row variable and column variable are not independent of each other.

In the TRIAL study, the researcher wanted to determine if the age group was independent of achieving the goal of normal sleep (Epworth score of less than 11). Both the observed and expected scores are presented

Table 2.

Parametric Tests and Some of the More Common Nonparametric Alternatives

Parametric Test	Corresponding Nonparametric Test(s)
Two-sample t test	Mann-Whitney U , median test
Paired t test	Wilcoxon signed rank test, sign test
One-way analysis of variance	Kruskal-Wallis test
Two-way analysis of variance	Friedman two-way analysis of variance
Correlation	Spearman ρ , Kendall τ
Linear regression	Theil's incomplete method

in Figure 8. The resultant χ^2 statistic would be

$$\chi^2 = \frac{(18 - 20.68)^2}{20.68} + \frac{(19 - 21.29)^2}{21.29} + \dots + \frac{(6 - 10.97)^2}{10.97} = 5.22$$

To be significant, the result would need to exceed a critical value of 5.99 from a χ^2 distribution table. As seen in Figure 8, the result represented a *p* value greater than 0.05; therefore, the researcher failed to reject H_0 and his or her best guess was that age grouping and meeting the goal of normal sleep were not related to each other. In other words, normal sleep was independent of age grouping.

Similar to other tests discussed in this article, certain conditions are required for the χ^2 test of independence. For example, each cell has to have at least one observation (frequency greater or equal to one) and the expected values for each cell have to be equal to or greater than five. If these criteria are not met, the χ^2 test should not be performed. The size of the matrix can be reduced by combining contiguous columns or rows to meet these criteria. For example, in the previous test, two of the adjacent age groups could have been combined if the criteria were not met. A reduction in columns and rows could continue for any size table until the smallest matrix is obtained (Figure 9). For small data sets in which the criteria cannot be met with a two-by-two table, the Fisher exact test is an acceptable method for analyzing the data.

The two-by-two matrix shown in Figure 9 is used for a variety of other tests of association, including the McNemar test, which is equivalent to the paired *t* test for dichotomous, discrete outcomes. As mentioned earlier, a third-dimensional confounding factor can be assessed for a two-by-two contingency table using the Mantel-Haenszel χ^2 (in survival statistics, this test is referred to as the Cochran-Mantel-Haenszel).

The most common use of the matrix illustrated in Figure 9 is for odds ratio (OR) and relative risk ratio (RRR) evaluations. One simple way to differentiate these two tests is that ORs are generally used for retrospective studies and RRRs are generally used for prospective studies. There are exceptions to this broad, simplified division. The odds for a given condition or exposure of interest is calculated as *a/c* whereas the relative risk of an experimental condition is calculated as *a/(a + c)*. The ratio is the likelihood of the outcome occurring divided by the likelihood of the outcome not occurring. For an OR, the formula is the experimental-event odds (EEO) divided by the control-event odds (CEO):

$$OR = \frac{EEO}{CEO} = \frac{a/c}{b/d}$$

The RRR is calculated by dividing the experimental-event rate (EER) by the control-event rate (CER):

$$RRR = \frac{EER}{CER} = \frac{a/(a + c)}{b/(b + d)}$$

Both ratios can be used in formulas to calculate a confidence interval similar to the two-sample *t* test illustrated previously; however, the interpretation is different. If the experimental group and the control group both had a 0.7 rate of risk, the RRR would be 0.7/0.7 = 1. Therefore, a significant confidence interval for

a ratio would be determined by the absence of the value of one within the interval and not the absence of zero as used in confidence intervals that test for differences:

$$H_0: RRR = 1$$

$$H_1: RRR \neq 1$$

Summary

Many other statistical tests are available to help the pharmacist researcher, including equivalency testing, survival statistics, and non-inferiority studies. All of these tests are beyond the scope of this article. More in-depth discussions on any of these tests and those presented in this article can be found in biostatistical textbooks.¹⁴⁻¹⁸ The purpose of this article is to provide examples of the wide variety of tests available and how tests are chosen based on the type of variables being analyzed in a study. This article is a brief overview of how tests are performed and how results are interpreted.

Unfortunately, mistakes in research design and statistical errors are still seen in the literature.¹⁹⁻²¹ Carefully planned studies that are evaluated with appropriate statistical tests can help to eliminate these mistakes. Employing the skills and knowledge of a professional statistician can assist in this effort and should be considered by any pharmacist who engages in research activities.

Figure 8. Contingency tables for chi-square analysis of one aspect of the hypothetical TRIAL study. The tables show the observed and expected numbers of subjects who met a goal of normal sleep. Numbers outside of the boxes are totals. df = degrees of freedom. Only 97 of 100 volunteers completed the study.

	Observed (Age Groups)				Expected (Age Groups)			
	Young Adult	Mature Adult	Geriatric		Young Adult	Mature Adult	Geriatric	
Met Goal?								
Yes	18	19	22	59	20.68	21.29	17.03	59
No	16	16	6	38	13.32	13.71	10.97	38
	34	35	28	97	34	35	28	97

Chi-Square = 5.22, df = 2, P-Value = 0.074

Figure 9. Classic two-by-two contingency table for chi-square analysis. "Exposure" would be used for retrospective odds ratios. Experimental and control conditions would be used for prospective relative risk ratios.

		Exposure		
		Yes	No	
Outcome	Yes	a	b	a + b
	No	c	d	c + d
		a + c	b + d	n

Conclusion

Selection of the proper statistical test depends on the type and number of variables and whether parametric conditions are met.

References

- De Muth JE. Preparing for your first meeting with a statistician. *Am J Health-Syst Pharm.* 2008; 65:2358-66.
- De Muth JE. Basic statistics and pharmaceutical statistical applications. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2006.
- Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep.* 1991; 14:540-5.
- Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960; 23:56-62.
- Siegel S. Non-parametric statistics for the behavioural sciences. New York: McGraw-Hill; 1956:19.
- Gaito J. Non-parametric methods in psychological research. *Psychol Rep.* 1959; 5:115-25.
- McNemar Q. Psychological statistics. New York: Wiley; 1969:431.
- Gardner PL. Scales and statistics. *Rev Educ Res.* 1975; 45:43-57.
- Dixon WJ. Power under normality of several nonparametric tests. *Ann Math Stat.* 1954; 25:610-4.
- Boneau C. A comparison of the power of the U and t tests. *Psychol Rev.* 1962; 69:246-56.
- Kerlinger FN. Foundations of behavioral research. New York: Holt, Rinehart and Winston; 1964:259.
- Siegel S, Castellan NJ Jr. Nonparametric statistics for the behavioural sciences. Singapore: McGraw-Hill; 1988:34.
- Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
- Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York: Wiley; 1999.
- Dawson B, Trapp RG. Basic and clinical biostatistics. 3rd ed. New York: Lange; 2001.
- Forthofer RN, Lee ES. Introduction to biostatistics: a guide to design, analysis and discovery. San Diego: Academic Press; 1995.
- Glantz SA. Primer of biostatistics. New York: McGraw-Hill; 1987.
- Zar JH. Biostatistical analysis. 4th ed. Englewood Cliffs, NJ: Prentice-Hall; 1999.
- Murphy JR. Statistical errors in immunologic research. *J Allergy Clin Immunol.* 2004; 114:1259-63.
- Schatz P, Jay KA, McComb J et al. Misuse of statistical tests in Archives of Clinical Neuropsychology publications. *Arch Clin Neuropsychol.* 2005; 20:1053-9.
- Neville JA, Lang W, Fleischer AB Jr. Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January through December 2003. *Arch Dermatol.* 2006; 142:737-40.