Biostatistics Primer: Part 2

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

ABSTRACT: Biostatistics is the application of statistics to biologic data. This article is the second part of a 2-part series on the application of statistics in nutrition science. The first article, published in the December 2007 issue, reviewed descriptive statistics. Inferential statistics, to be discussed in this article, can be used to make predictions based on a sample obtained from a population or some large body of information. It is these inferences that are used to test specific research hypotheses. This article focuses on inferential statistics and their application in the nutrition and biomedical literature. Additionally, this review will outline some of the most commonly used statistical tests found in the biomedical literature.

In the December 2007 issue of *NCP*, the first part of a 2-part series, "Biostatistics Primer," we reviewed basic concepts of statistics and expounded on descriptive statistics. The concluding part of this series will focus on inferential statistics and some of the most commonly used statistical tests in biomedical literature.

Commonly Used Statistical Tests

The appropriate statistical test for any given dataset should be chosen according to certain characteristics of the collected data. Most of the statistical tests that are described in this review are termed *parametric tests*. Although these tests are the most powerful (ie, more likely to detect a difference if one exists), there are certain assumptions that must be met to appropriately use a parametric test. First of all, the collected data should be on an interval or ratio scale and be describable using the mean and standard deviation (SD). If the correct variable is chosen, the 2 primary assumptions of parametric tests are that samples are obtained from a population that is normally distributed and that the sample variances are essentially equal. In many cases, studies with small samples sizes may not meet these criteria. The Kolmogorov-Smirnov and goodness-of-fit tests are 2 statistical tests that can be used to test the normality of data (ie, the underlying assumption of normally distributed data). If this assumption is not met, the data should be transformed to a normal distribution (eg, log transformation for positively skewed data) or a different statistical test should be used, such as a nonparametric test. Statistical tests can also be used to test the homogeneity of variance to ensure that the second assumption of parametric tests is met.

For each parametric statistical test described in this section, a nonparametric equivalent will be briefly described for data that do not meet the assumptions for a parametric test. Nonparametric tests are more robust and do not make any assumptions about the underlying distribution of the data. However, if the assumptions for a parametric test are met, these should be used because they are more powerful. Finally, data that cannot be placed on an interval or ratio scale are generally described as categorical data. Categorical data (nominal or ordinal) can have different underlying distributions than that of continuous data and therefore a different statistical analysis is required. The most common categorical test in the medical literature is the χ^2 test.

t-Tests

One of the most common questions raised in biostatistics is whether 2 groups differ from one another, making *t*-tests very common applications in biostatistics. These tests are used to compare the average response to a drug between 2 distinct groups, to compare the average value in a group to a known standard, or to compare the baseline drug response with the response after receiving a drug in the same individual. The purpose of this section will be to describe 3 statistical tests that generally are referred to as *t*-tests. In essence, these tests use similar calculations for making the comparisons but

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

0884-5336/08/2301-0076$03.00/0
Nutrition in Clinical Practice 23:76–84, February 2008

are, in fact, quite different in their underlying assumptions and requirements for study design. Each *t*-test uses the ratio of the difference between the group means in the numerator divided by the variability within the groups to calculate the statistical ratio.[1] In the case of a *t*-test, the statistical ratio is the *t*-statistic. This ratio is shown in the following equation:

$$\text{Statistical Ratio} = \frac{\text{difference between group means}}{\text{variability within groups}}$$

The larger the difference between the group means compared with the variability within groups, the greater the absolute value of the ratio. The larger the ratio, the more likely it is to demonstrate a statistical difference between the 2 groups. This ratio is the general form of each of the *t*-test equations, in addition to equations used in the analysis of variance (ANOVA) F-test. *t*-Tests are among the most commonly used,[2] and sometimes misused, statistical tests in the biomedical literature

All *t*-tests are parametric tests, and as such, the user assumes an underlying distribution or structure when using the tests. The derivation of the *t*-test calculation assumes that the outcome response being measured is normally distributed, although it has been suggested[1] that the *t*-test performs reasonably well if the underlying distribution deviates "moderately" from normality. The 3 types of *t*-tests most commonly used are the 1-sample *t*-test, 2-sample *t*-test, and paired *t*-test. The following discussion will address each one.

The 1-sample *t*-test is used to answer the question: is the average of a set of observations equal to the standard value in the population? The study design for this type of *t*-test includes only 1 sample and has a goal of determining if the observed outcome has the same mean of some standard or reference population. The structure of the *t*-statistic is as follows:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample or observed mean, s is the sample or observed SD, n is the sample size, and μ is the population mean. Remember that the statistical ratio described earlier, in this case the numerator, is the difference between the observed mean and the reference mean and the denominator is the variability within the observed group. An example of the use of this type of *t*-test may be used to determine if the mean of 100 blood glucose samples (mean ± SD; 125 ± 15 mg/dL) obtained in the emergency department is different from the established hospital control (115 mg/dL).

A second type of *t*-test, a 2-sample *t*-test, is very commonly used in the biomedical literature. This test is also known as an independent-samples, independent-groups, or unpaired *t*-test. This type of *t*-test is used to compare the responses or outcome measures in 2 independent groups, that is, 2 groups of different subjects. Examples of this in the literature include responses to active drug compared with control, placebo, or standard therapy, or the response to 2 treatments given to different persons. Other examples include comparison of pharmacokinetic parameters between 2 different groups (men *vs* women, young *vs* old, heart failure *vs* control, etc) or comparison of demographics (age, weight, etc) between 2 groups. The question we ask for a 2-sample *t*-test would be: is the mean of one group equal to the mean of the other group (a 2-tailed *t*-test)? Alternatively we may ask, is the mean of one group less than (or greater than) the mean of the other group (a 1-tailed test)? There are 2 types of 2-sample *t*-tests used, one that assumes the variances of the 2 groups are equal (variance homoscedasticity) and one in which the variances are not equal (variance heteroscedasticity). Most statistical software packages perform both of the tests and the analyzer chooses the most appropriate test. The assumption of equal variances can be tested formally or informally. In the latter case, if the ratio of the samples variances (ratio of larger variance to smaller variance) is >2, it is generally assumed that the variances are unequal. This can be formally tested with the F-test for variances, which will not be discussed in this review. If the variances are not equal, the unequal variance *t*-test should be used; otherwise, the equal variance *t*-test is appropriate. The equations below describe the case when the variances are assumed to be equal; interested readers are directed to other reviews[3] for discussion and calculations involved for the unequal variance test and the associated F-test. The structure for the equal variance *t*-test is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the first and second groups, respectively, s_p is the pooled SD, and n_1 and n_2 are the number of observations in the first and second groups, respectively. The pooled SD is the weighted average of the SD in the 2 groups as described in the following equation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2) - 2}}$$

Remember that the numerator of the statistical ratio is the difference between the 2 means and the denominator is the measure of variability within these 2 groups. Using the example from Table 1, we wish to know if the gestational age is statistically different between the glutamine group and the control group. In this example, because these groups are separate individuals, we use the unpaired or independent samples *t*-test rather than the paired *t*-test. The assumptions for this test are that the

Table 1 *Baseline and nutrition characteristics (modified with permission from van den Berg et al[4])*

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

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

variances are equal and that the data are normally distributed. The first assumption appears to be met because the SDs (and variances) are nearly equal. The second assumption is more difficult to verify without access to the actual data, a common problem when evaluating statistical procedures used in published studies. What we do know about the data are that they do not seem to be highly variable (according to the SD) and thus not likely to be skewed or affected by extreme outliers. Thus the use of a *t*-test seems to be appropriate in this case. If you were the researcher with access to the actual data, with the observations from the study, you would be able to evaluate the normality of the data. As an initial step, many investigators would compare the mean and median value and, if similar, assume that the data are normally distributed. Finally, because the number of subjects studied in this case is relatively large, approximately 50 in each group, most would agree that the use of a parametric test is reasonable in this situation.[5]

The third type of *t*-test, a paired *t*-test, also known as a matched-pairs *t*-test, is used to compare 2 responses or outcome measures in the same person. The alternative to this design is the test described previously, the unpaired *t*-test. Instead of 1 treatment administered to 1 group of subjects and the other treatment to another group of subjects (as in a parallel-groups design), in a paired design, 1 subject receives 2 treatments. A commonly used type of study design using a paired *t*-test is a crossover study where 1 subject has 2 measurements obtained after 2 treatments. The benefit of this type of design is the between-subject variability is reduced as each subject serves as his or her own control. The use of a paired study design and paired *t*-test increases the power of this test as compared with a 2-sample *t*-test or a parallel-group study design. The determination of the test statistic for the paired *t*-test is illustrated in the following equation:

$$t = \frac{\bar{x}_d - \mu_d}{s_d / \sqrt{n}}$$

where $\bar{x}_{\mathrm{d}}$ is the mean of the differences in the 2 treatments, s_d is the SD of the difference between the 2, μ_d is the population mean of the difference in the 2 treatments, and n is the number of pairs observations in the first and second group, respectively.

The previous several paragraphs discussed using *t*-test and hypothesis testing to compare means or to determine if a statistical difference exists. A related alternative approach is the use of confidence intervals. Recall from the previous discussions that the confidence interval represents the interval in which the true population mean or difference between means may exist. In most cases we use the 95% confidence interval, which corresponds to the $p < .05$ level of significance used in hypothesis testing.[6]

ANOVA

ANOVA is a commonly used statistical analysis in the biomedical literature. In this paper, we are discussing introductory statistical concepts, so our discussion will focus mainly on 1-way ANOVA, a brief discussion of repeated measures ANOVA, and 2-way ANOVA. ANOVA is one of several multifactorial analyses; factors are groups or treatments. ANOVA can be applied to independent (1-way ANOVA) and related (repeated-measures ANOVA) groups design. The simplest type of ANOVA is a 1-way ANOVA, in which 3 or more groups are being compared. Therefore, it is an extension of an independent-samples *t*-test, in which only 2 groups are compared. There is 1 factor, and the analysis is performed to test whether any of the means differ. ANOVA is a parametric analysis and thus subject to the same assumptions we discussed for *t*-tests and other parametric tests. The assumptions underlying ANOVA are that the dependent variable is continuous or at least interval scaled, the response data are drawn from a normally distributed population, there is a random unbiased selection of cases and assignment to groups, the intersubject variance (or SD) in all groups or factors is equal, and the factors (grouping or independent variables) are categorical or transformed to be categorical.

One-way ANOVA, sometimes referred to as a single-factor experiment, as described above, is applied to an experiment in which 3 or more independent group means are compared. One-way ANOVA implies 1 factor with 3 or more levels. Table 2 illustrates this concept for a hypothetical data set. In this example, 60 subjects were randomly assigned to treatments 1 through 4, which is a parallel-groups design study. The table depicts the individual data and mean and SD in each group. The 1-way or factor, in this example, is the independent variable treatment and the levels are 1–4. The null and alternative hypotheses for this statistical analysis are:

$$H_0 : \mu_{\text{Treatment 1}} = \mu_{\text{Treatment 2}} = \mu_{\text{Treatment 3}} = \mu_{\text{Treatment 4}}$$

H_A: At least 2 means differ significantly

At first look, one may come to the conclusion that multiple independent *t*-tests could be used to perform this analysis. This is a commonly used and incorrect statistical approach for this type of data. If we were to compare each possible permutation in this example, we would be conducting 6 *t*-tests. Assuming an *a priori* type I error rate of 5%, multiple tests would increase that risk of a type I error to >5% and increase the error rate. Using ANOVA keeps the type I error rate at an acceptable level. The statistical test for ANOVA uses the F-test to perform the analysis. The F-test uses the ratio of the sum of squares between groups to the sum of squares within groups to determine the critical value of the F-test statistic. The analysis is similar to the earlier description of an independent *t*-test, except that we must account for the variation between the groups and within the groups for all of the groups rather than just the 2 groups. Just like a *t*-test, the concept of the statistical ratio can be used to understand the components of the statistical analysis used to perform the analysis. Recall that the statistical ratio relates the difference between the means (intergroup variation) and the variability within groups (intragroup variation). The key concept is that ANOVA separates the error into components. In 1-way ANOVA, the calculations for the analysis, which are usually made using statistical software packages, determine the sum of squares total, sum of squares between groups, and the sum of squares within groups. According to how the hypothesis test is written for ANOVA, it should be apparent that the conclusion from the ANOVA will be that either there is no difference between the means or at least 2 of the means differ. In the example in Table 2, the overall ANOVA was statistically significant, as evidenced by the *p* value of < .001, which is below the $\alpha = .05$ set before the study. The overall ANOVA, using the F-test, will not determine which of the 2 means are different. Instead, we would conclude, according to the alternative hypothesis, that at least 2 of the means differ significantly.

Table 2 *Individual and mean and SD data from a hypothetical study in which 4 treatments were administered to 4 different groups*

	Treatment			
	1 (n = 15)	2 (n = 15)	3 (n = 15)	4 (n = 15)
	277	272	255	277
	270	275	253	267
	279	259	268	273
	272	278	264	275
	274	275	279	273
	289	276	253	279
	282	276	278	278
	278	285	275	264
	285	274	265	259
	275	275	251	275
	277	259	258	273
	269	283	255	274
	276	281	264	282
	285	275	268	272
	255	276	265	292
Mean*	276.2	274.6	263.4†	274.2
SD	8.1	7.2	9.2	7.7

*Overall ANOVA, $p < .001$.
†Treatment C different from treatments 1, 2 and 4; $p < .01$, by Tukey HSD.
SD, standard deviation.

Intuitively when one looks at Table 2, it seems as if treatment 3 differs from the others. The use of post hoc testing is the most common approach to statistically evaluate this situation. This approach controls the type I error rate.[6] Although the specific discussion of the calculation of the various post hoc tests will not be done in this article, the various post hoc tests that may be encountered are Tukey's HSD, Scheffe, Bonferroni, Dunnett, Dunn, Newman-Keuls, and least significant difference tests. Each of these procedures has advantages and disadvantages discussed elsewhere.[5]

A more complex ANOVA technique using an independent groups design is 2-way or 2-factor ANOVA. Two-way ANOVA allows the study of 2 independent variables at the same time. An example would be the extension of our previous 1-way ANOVA shown in Table 3. Instead of 1 independent variable (treat-

Table 3 *Mean and SD data from a hypothetical study in which 4 treatments were administered to 4 different groups in 2 different age groups*

		Treatment			
		1	2	3	4
Age	Mean-young (n = 15)	276.2	274.6	263.4	274.2
	SD-young	8.1	7.2	9.2	7.7
	Mean-elderly (n = 15)	262.4	258.5	235.1	225.3
	SD-elderly	4.6	5.2	4.7	5.9

SD, standard deviation.

ment) now we have 2 independent variables, (treatment 1–4) and age (young and elderly). It should be noted that this is not a repeated-measures design, which is the young and elderly receiving 4 treatments, but instead, 4 independent groups receiving the treatments. Using ANOVA we now ask 3 questions: (1) What is the impact of age independent of treatment? (2) What is the impact of treatment independent of age? and (3) What is the joint effect of age and treatment? Instead of 3 sums of squares terms, now there are 4: sum of squares total, sum of squares due to treatment, sum of squares due to age, and sum of squares due to the interaction between treatment and age.

Repeated-measures ANOVA is an extension of a paired *t*-test, which is a related-samples design rather than an independent-groups design. Instead of the example that we have used for 1-way ANOVA (Table 2), in which 4 different groups (n = 15) received different treatments, now consider the example in which 1 group (n = 15) receives 1 treatment, with 4 samples over time. An example of this approach would be the effect of the impact of a natural herbal remedy on total cholesterol, illustrated in Table 4, over the course of a 6-month time period. The study design is an example of a repeated-measures design, or in other words, a paired-study design. We now ask the questions, (1) what is the impact of treatment (ie, time)? (2) what is the impact of subject? and (3) what is the joint effect of subject and treatment? The 4 sums of squares terms are sum of squares total, sum of squares due to treatment, sum of squares due to subjects, and sum of squares error. This could be expanded to another level by adding another group of subjects (like elderly subjects) and become a 2-way repeated-measures ANOVA. There are numerous other types of ANOVA that are more complex and beyond the intent of this article.[1,5,6]

Correlation

In many applications in the biomedical literature, clinicians and researchers wish to examine the relationship between 2 variables. Instead of asking the question that we previously asked from Table 1—are the gestational ages of subjects in the glutamine group different from the gestational ages of the subjects in the control group?—now we ask the question: what is the relationship between birth weight and gestational age? In correlation analysis, we examine whether pairs of data are associated. In this example, intuitively we may think that as gestational age increases, so does birth weight. This is likely to be true up to a certain gestational age. In correlation analysis, this is not to say that increased gestational age causes increased body weight; this would be evaluated by regression analysis. Correlation analysis is often referred to as the degree of association between the 2 variables. In other words, we are investigating the relationship between 2 variables, not suggesting or studying whether or not one causes the other. The 2 variables investigated are both treated equally, and neither is assumed to be the predictor or the outcome.[7] The null hypothesis for a correlations analysis is that the correlation coefficient is equal to zero (no relationship) or that variable 1 and variable 2 are not related.

The statistical parameter quantifying the degree of association between the 2 variables is referred to as the correlation coefficient (r). The correlation coefficient is a unitless measure of the association between the 2 variables. The size and sign of the correlation coefficient communicate important information about the relationship. The correlation coefficient ranges from −1 to +1; a value of −1 indicates a perfect negative relationship (that is, as variable 1 increases, variable 2 decreases). A perfect positive relationship has a correlation coefficient of +1, whereas an r of zero indicates no linear relationship. There is no agreed-upon or consistent interpretation of the value of the correlation coefficient. Table 5 depicts several arbitrary published criteria for defining the strength of the relationship.

The value of r is strongly affected by sample size, measurement error, and the variables being explored. Equally important, the interpretation of the value of the correlation coefficient depends on the type of research being conducted (ie, clinical research, basic research, social science research). The value of r alone also does not communicate any information about whether the relationship between the 2 variables is statistically significant. The formal statistical test for correlation can be performed and is highly dependent on sample size. The larger the sample size, the more likely a statistically significant association will be found. Recall that a statistically significant finding does not necessarily mean that there is a strong relationship, just one that is unlikely to have occurred by chance. Consider Table 6, which illustrates data from 2 correlation analyses. The importance of sample size should be apparent in this example. In study 1, the r value can be described as poor, yet the relationship is statistically significant due to the large sample size. Contrast that with study 2, in which the r value would most

Table 4 *Mean and SD cholesterol data from a study in which 15 subjects received an herbal therapy thought to reduce cholesterol*

	Time			
	Baseline	6 wk	3 mo	6 mo
Mean (n = 15)*	275.8	265.5†	273.7	274.4
SD	3.20	3.12	2.79	3.69

Cholesterol was measured at the intervals indicated over the 6-month period.
*Overall repeated-measures ANOVA, $p < .001$.
†Six weeks different from baseline, 3 months, and 6 months, $p < .001$, by Tukey HSD.
ANOVA, analysis of variance; SD, standard deviation.

Table 5 *Correlation coefficient descriptions*

r Value	Guilford[8]	Rowntree[9]	DeMuth[6]	Portney[1]
<0.10	Slight correlation, negligible relationship	Very weak, negligible	"Weak"	Little or none
0.10–0.20				
0.20–0.30	Low correlation, definite but small relationship	Weak, low	"Fair"	
0.30–0.40				Fair
0.40–0.50	Moderate correlation, substantial relationship	Moderate	"Good"	
0.50–0.60				Moderate to good
0.60–0.70				
0.70–0.80	High correlation, marked relationship	Strong, high, marked		
0.80–0.90				Good to excellent
0.90–1.00	Very high correlation, very dependable relationship	Very strong, Very high		

likely be described as good, yet there is no statistically significant relationship, due to a small sample size. When evaluating correlation coefficients and associated statistical tests, the reader should pay close attention to these issues and appreciate the factors involved in these analyses.

There are several different types of correlations utilized for statistical analysis. The 2 most commonly employed are the Pearson product-moment correlation and the Spearman rank correlation. The Pearson product-moment correlation is the most commonly used correlation analysis, the statistic generated by this analysis is r, or the correlation coefficient, for the sample and for the population, ρ. This is a parametric statistic, and its use assumes that the 2 variables have underlying normal distributions, are continuous or ratio scaled data, and the variables are linearly related. The second type of correlation, the Spearman rank correlation coefficient, is the nonparametric analog of the Pearson product-moment correlation. This correlation is not subject to the same parametric assumptions as Pearson's correlation, and can be used with ordinal data. Instead of using the observed data, the calculation uses the difference in the ranks of the observations to determine r_s (Spearman's ρ).

Crucial to the proper use of correlation analysis is the interpretation of the graphical representation of the 2 variables. Before using correlation analysis, it is essential to generate a scatter plot of the 2 variables to visually examine the relationship. Because standard correlation analyses are only useful for linear relationships, graphical displays are necessary to determine if the relationship is something other than linear, such as curvilinear, or to determine if 1 data point is driving the relationship.[1,10]

Table 6 *Correlation coefficients from 2 different study designs and associated* p *values*

	Study 1	Study 2
r	0.20	0.70
n	102	8
p Value (2-tailed)	<.05	>.05

Regression

A statistical technique related to correlation is regression analysis. Like correlation, there are many different types of regression analysis, including linear, multiple, weighted, and logistic. These techniques are widely used in all disciplines of the biomedical literature, from clinical to basic research.[5,7] It is beyond the scope of this paper to discuss each type of regression analysis; instead, the

focus will be on simple linear regression, the most commonly used technique in basic biostatistics applications. In correlation analysis, we discussed that the assumption was not for one variable to predict another. Instead, we used correlation analysis to determine if an association or relationship existed between 2 variables. In regression analysis, the intent of the analysis is to make predictions based on the observed relationship. That is, given a value of X, what will the observed regression analysis predict to be the most likely value of Y? Also, regression analysis provides us with some degree of certainty relative to the observed relationship. In regression analysis, we refer to studied variables as either the dependent (outcome) and the independent (causative) variable. In addition to making predictions of the dependent variable from the independent variable, regression analysis also quantifies how well the independent variable predicts the dependent variable. An example of regression analysis is shown in Figure 1. In this figure, the outcome variable is resting energy expenditure (Y-axis) and is dependent on the causative variable, weight (X-axis). These data were extracted from a published study.[11] Naturally, body weight does not totally explain resting energy expenditure but only some percentage of the variability. In this example, we are only considering 2 variables and refer to the example as simple linear regression.

The linear regression model, written in a way that is on one hand familiar and easy to understand and then, also, using statistical nomenclature is:

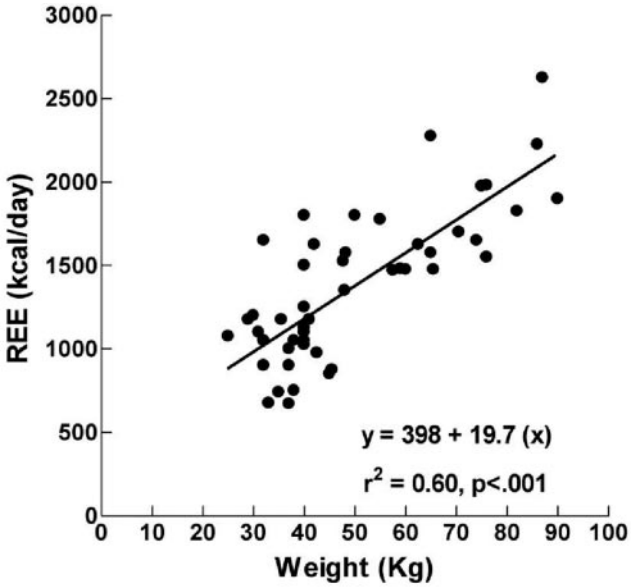

Figure 1. Relationship between resting energy expenditure and body weight. Data were extracted from a previously published study.[11] The coefficient of determination (r^2), *p* value, and regression equation are shown in the figure.

$$Y = mX + b$$

where Y is the dependent variable, X is the independent variable, m is the slope of the relationship, and b is the y-intercept, the location where the regression line crosses the Y-axis. In the sample, in statistical terms:

$$Y = b_0 + b_1X$$

where b_0 is the Y-intercept, the location where the regression line crosses the y-axis, and b_1 is the slope of the relationship. There are numerous ways in which to estimate the linear regression parameters, although the most commonly used in basic statistical analysis and by basic software applications is the method of least squares (ie, least squares linear regression).

The method of least squares regression seeks to minimize the sum of squared differences between the actual values of the dependent variable Y and the predicted values of Y (ie, minimize the sum of squares differences) and provides an estimate of the best-fit regression line, slope, and intercept and regression coefficients. Crucial to the proper use of regression analysis is the interpretation of the graphical representation of the 2 variables. As with use of correlation analysis, it is critical to generate a scatter plot of the 2 variables to visually examine the relationship. Visualization of the data allows us to draw some initial conclusions about the data and helps us to determine what the underlying model for the regression might be. As seen in Figure 1, the relationship between weight and resting energy expenditure appears to be linearly related. The relationship does not seem to be curvilinear or some other relationship.

The assumptions for linear regression are that the dependent variable (Y) is adequately modeled as being linearly related to a single independent variable X. Values of X are "known" or at least observed with negligible error. For each value of X, there is a population of Y values that are normally distributed. The observed values of Y are independent (ie, the value of an observation is not affected by the values of another observation).

The second intent of regression analysis is to determine the extent of variability in the dependent variable that can be explained by the independent variable. The coefficient of determination (r^2) is the ratio of the sum of squares explained by the regression and the total sum of squares. Values of r^2 can range between 0 and 1. The higher the r^2, the stronger the relationship. Like the correlation coefficient used in correlation analysis, the coefficient of determination is a means, albeit with some problems, to quantify the strength of the relationship between the dependent and independent variable. Mathematically, the coefficient of determination is the square of the correlation coefficient. In the example shown in Figure 1, the r^2 is 0.61, suggesting that the body weight explains 61% of the variability in resting energy expenditure.

Like interpretation of r, the interpretation of r^2 is dependent on the scientific arena (ie, clinical, research, and social science research) to which it is applied. In other words, a value of r^2 may have a very different meaning depending on the situation being studied. An r^2 of 0.80 obtained from an analytical standard curve would likely be considered a very poor relationship, whereas the same r^2 in a clinical pharmacokinetic study relating creatinine clearance and drug clearance would indicate a good relationship between the 2 variables.

The biomedical literature is replete with examples of regression analysis being used to make predictions. For example, aminoglycoside elimination rate constants are often estimated from estimated creatinine clearance by the equation:

$$\textit{gentamicin elimination rate} = (0.0024 \times CrCl) + 0.01$$

to make initial dosing predictions. The development of this equation utilized linear regression techniques from a study relating the 2 variables.[12] Usually, equations derived from these studies are only used to make future predictions if several "criteria" are met. In general, we would not use the regression results unless that statistical analysis indicated a statistically significant relationship between the 2 variables. Unfortunately, an error that is frequently made in the interpretation of linear regression analysis is using the regression parameters to predict outcomes without examining all of the pieces of the regression statistics. In evaluating regression statistics, one should evaluate the value of r^2, as well as whether or not a statistically significant relationship between 2 variables was observed. There may be some situations in which a significant relationship between the 2 variables is observed but the value of r^2 is too low to meaningfully predict Y from X. Ideally, we would also investigate the standard error of the regression parameters, the distribution of the residuals, although these parameters are seldom presented in published biomedical manuscripts. The most frequent error that is made in regression analysis is to predict values that were not included in the studied relationship. Extrapolation is discouraged because the nature of the relationship outside of the observed range is unknown. An example of this practice would be to use an assay to determine the concentration of drug below the limit of quantification for a standard curve. Extrapolation outside of the observed values in most cases is discouraged.

Regression analyses are important techniques in the biomedical literature. In this paper, linear regression was discussed. Numerous other types of regression analyses are available, including nonlinear regression, multiple regression, and logistic regression.[1,5,6]

Nonparametric Statistics

Thus far, we have discussed the use of parametric statistics, that is, statistical tests that assume the data have an underlying normal distribution and some degree of variance homogeneity. Nonparametric statistics are not as restricted by underlying distributional requirements required for parametric statistics. Due to this fact, nonparametric statistics are often referred to as distribution-free statistics. In general, nonparametric statistics are less powerful than their parametric counterparts. They are generally useful in several situations when the assumptions of a parametric test are violated: when the outcome variable is an ordered (ordinal, nominal) and not continuous variable, or the outcome variable is continuous but not normally distributed, highly variable, or subject to outliers. The goal of this section is to present 4 commonly used nonparametric tests, to present the situations in which these tests may be used and their parametric counterparts. The nonparametric analog of the paired *t*-test is the Wilcoxon signed-rank test. It is useful in situations described above for the paired *t*-test. In this situation, the null hypothesis is that the median difference between the pair of observations is zero. Consider a situation in which we wish to determine if a patient's Acute Physiologic and Chronic Health Evaluation (APACHE) II score[13] improves from admission to day 3 of a hospitalization. The pair of observations would be the APACHE II score at admission and at 3 days. Because APACHE II scores can only take integer values (that is 1, 2, 3, 4, etc) the use of mean and SDs to describe the central tendency of this type of data does not make sense. Therefore, it should then be intuitive that a parametric test should not be used to perform the statistical analysis. The Wilcoxon signed-rank test is a nonparametric test that uses the ranks of the difference between the paired data, rather than the actual data in the analysis. Another nonparametric test that can be used with this type of data is the sign test.

The Wilcoxon rank sum test is the nonparametric analog of the 2-sample or unpaired *t*-test. The test is also often referred to as the Mann-Whitney *U* test. It is applied in situations described above for the unpaired *t*-test, in which 2 independent groups are being compared, and in which the outcome data do not meet the assumptions of the parametric test. The null hypothesis in this case is that the median is not different between the 2 groups. There are 2 examples of the use of this type of statistic in Table 1. First, the test is used to compare the Clinical Risk Index for Babies (CRIB) between the glutamine and control groups. Recall from earlier in this paper that CRIB is an index that can only have values that are whole numbers. Thus, the CRIB index is not a continuous variable, and comparisons between the 2 groups with an independent groups *t*-test would not be appropriate; instead, the Wilcoxon rank sum test should be used. Another example of use of this test is in the comparison of age at study start, time to full dose, and age at increasing dose. One could certainly make a case that days are continuous data, although we seldom record days as 1.1 days or 1.6 days, etc.

On inspection of the data, which are presented as the median and range in the table, it is evident that these data are skewed to the right and likely not normally distributed. In this case, a *t*-test is also not appropriate, and the Wilcoxon rank sum test would be appropriate. The extension of the Wilcoxon rank sum test for application to >2 groups is the Kruskal-Wallis 1-way ANOVA by ranks. This test is the nonparametric analog of 1-way ANOVA.

Finally, as discussed briefly, the Spearman's rank correlation coefficient is the nonparametric analog of Pearson's correlation. Unlike Pearson's correlation, Spearman's correlation does not require continuous data. Instead, it can be applied to continuous (regardless of the underlying distribution) or ordered data to test the null hypothesis that there is no association between the 2 variables.

χ^2 Test

The χ^2 test is one of the most frequently reported statistical tests in the medical literature. This test can be used on nominal or ordinal (categorical) variables that can take on >2 possible outcomes. Ordinal data can be transformed to discrete data by assigning numeric values and are often assessed using nonparametric tests as opposed to the χ^2 test.

Data that can be described by a contingency table are oftentimes nominal and by definition categorical data. Table 7 provides an example of a 2 × 2 contingency table (pronounced 2 by 2) that has been created using the reported sexes of the very-low-birth-weight (VLBW) infants in Table 1. As previously stated, sex is a nominal variable and can be placed in a contingency table such as the one in Table 7. The inner boxes in this table represent the observed number (percentage) of men and women of the participants in each study group, as reported in Table 1. The outermost boxes are the totals for the rows and columns, and the lower right box is the total of the rows, which is equal to the total of the columns. This is an important attribute of a contingency table.

The χ^2 test is commonly used to test baseline characteristics of treatment groups to ensure that there are no underlying compounding factors between the experimental groups. This is accomplished by testing observed *vs* expected values. The goal of the χ^2 test in the example provided in Table 7 is to test the number of men and women in each study group to ensure there were similar genders in each study group. The null hypothesis for the χ^2 test in this example could state that the sex of the infants in this study is independent of the assigned therapy. Therefore, this is commonly referred to as the χ^2 test for independence. The data presented in Table 7 are the observed values collected from a clinical study. Therefore, expected values are still needed to perform this statistical test. In this example, the expected values can be calculated directly from the observed values in the contingency table.

Table 7 *2 × 2 Contingency table of the sex of research subjects in van den Berg et al*[4]

	Glutamine	Control	Totals
Males	28 (54%)	27 (54%)	55
Females	24 (46%)	23 (46%)	47
Totals	52	50	102

The calculation of the test statistic for the χ^2 test is beyond the scope of this review but it is based solely on observed and expected values. For tests of independence, such as the previous example, the expected values can be calculated from the observed values to test for underlying diseases, sex, or other nominal variables in clinical studies. For studies that have small expected values (<5), the Fisher exact test should be used in place of the χ^2 test.

Summary

Biostatistics are essential tools for all individuals involved in clinical, basic, and translational research. A basic understanding of the underpinnings of biostatistics is necessary for the proper conduct and evaluation of biomedical research papers. This outlines important features of descriptive and inferential statistics as it applies to commonly conducted research studies in the biomedical literature. Additionally, commonly used statistical tests found in the biomedical literature were reviewed.

References

1. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. Upper Saddle River, NJ: Prentice Hall Health; 2000.
2. Lee CM, Soin HK, Einarson TR. Statistics in the pharmacy literature. *Ann Pharmacother.* 2004;38:1412–1418.
3. Davis RB, Mukamal KJ. Hypothesis testing: means. *Circulation.* 2006;114:1078–1082.
4. van den Berg A, Fetter WP, Westerbeek EA, van der Vegt IM, van der Molen HR, van Elburg RM. The effect of glutamine-enriched enteral nutrition on intestinal permeability in very-low-birth-weight infants: a randomized controlled trial. *JPEN J Parenter Enteral Nutr.* 2006;30:408–414.
5. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Hoboken, NJ: Wiley; 2005.
6. DeMuth JE, ed. *Basic Statistics and Pharmaceutical Statistical Applications*. Boca Raton, FL: Chapman and Hall/CRC; 2006.
7. Crawford SL. Correlation and regression. *Circulation.* 2006;114: 2083–2088.
8. Guilford JP. *Fundamental Statistics in Psychology and Education.* New York, NY: McGraw-Hill; 1956.
9. Rowntree D. *Statistics Without Tears: A Primer for Non-Mathematicians*. New York, NY: Charles Scribner's Sons; 1981.
10. Larson MG. Descriptive statistics and graphical displays. *Circulation.* 2006;114:76–81.
11. Winter TA, O'Keefe SJ, Callanan M, Marks, T. The effect of severe undernutrition and subsequent refeeding on whole-body metabolism and protein synthesis in human subjects. *JPEN J Parenter Enteral Nutr.* 2005;29:221–228.
12. Dettli LC. Drug dosage in patients with renal disease. *Clin Pharmacol Ther.* 1974;16:274–280.
13. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* 1985; 13:818–829.