Understanding Statistical Tests

Todd Neideen, MD, and Karen Brasel, MD, MPH

Division of Trauma and Critical Care, Department of Surgery, Medical College of Wisconsin, Milwaukee, Wisconsin

INTRODUCTION

Critical reading of the literature requires the capability to determine whether the conclusions are supported by the data presented. Part of this determination involves deciding whether the results are statistically valid. Although much statistical analysis may be beyond those without advanced statistical training, basic knowledge will significantly enhance the ability to both read and interpret medical literature. Part of this knowledge is determining which statistical test is appropriate for a given data set.

KEY WORDS: parametric, nonparametric, statistical analysis

COMPETENCY: Medical Knowledge

PARAMETRIC AND NONPARAMETRIC TESTS

Data can either be continuous, discrete, binary, or categorical. Continuous, or interval, data have units that can be measured with a value anywhere between the lowest and the highest value. An example is platelet count. Discrete, or ordinal, data have a rank order, but the scale is not necessarily linear. A pain scale from 1 to 10 is a good example; a pain score of 8 is not necessarily twice as bad as 4. Binary data are simply yes/no data: alive or dead. Examples of categorical, or nominal, data are color or shape. The data are different, but no rank order exists. The test chosen to analyze the data is based on the type of data collected and some key properties of that data.

PARAMETRIC TESTS

Parametric tests are more robust and for the most part require less data to make a stronger conclusion than nonparametric tests. However, to use a parametric test, 3 parameters of the data must be true or are assumed. First, the data need to be normally distributed, which means all data points must follow a bell-shaped curve without any data skewed above or below the mean. Ca-125 levels are an example of non-normally distributed data. In the general population, normal Ca-125 values range from 0 to 40. The median is 15, which leads to a skewed rather than a normal distribution. The data also need to have equal variance and have the same standard deviation. Finally, the data need to be continuous. Commonly used parametric tests are described below.

Pearson Product Correlation Coefficient

The correlation coefficient (r) is a value that tells us how well 2 continuous variables from the same subject correlate to each other. An r value of 1.0 means the data are completely positively correlated and 1 variable can be used to compute the other. An r of zero means that the 2 variables are completely random. An r of −1.0 is completely negatively correlated. As an example, consider the correlation between the systolic and the diastolic blood pressures from the Framingham study data. Figure 1 shows the plot of the data.

It seems that the diastolic and systolic blood pressures are highly correlated. This r value corresponds to an 88% association. The important thing to remember is that this is only an association and does not imply a cause-and-effect relationship.

Student t-Test

The Student t-test is probably the most widely used parametric test. It was developed by a statistician working at the Guinness brewery and is called the Student t-test because of proprietary rights. A single sample t-test is used to determine whether the mean of a sample is different from a known average. A 2-sample t-test is used to establish whether a difference occurs between the means of 2 similar data sets. The t-test uses the mean, standard deviation, and number of samples to calculate the test statistic. In a data set with a large number of samples, the critical value for the Student t-test is 1.96 for an alpha of 0.05, obtained from a t-test table. The calculation to determine the t-value is relatively simple, but it can be found easily on-line or in any elementary statistics book.

As an example, given 1000 men measured for height in China and Japan, are the mean heights different? China's mean is 169.1 cm with a standard deviation of 6.21 cm, and Japan's

Correspondence: Inquiries to Karen Brasel, MD, MPH, Division of Trauma and Critical Care, Department of Surgery, Medical College of Wisconsin, 9200 W. Wisconsin Avenue, Milwaukee, WI 53226; fax: (414) 805-8641; email: Kbrasel@mcw.edu

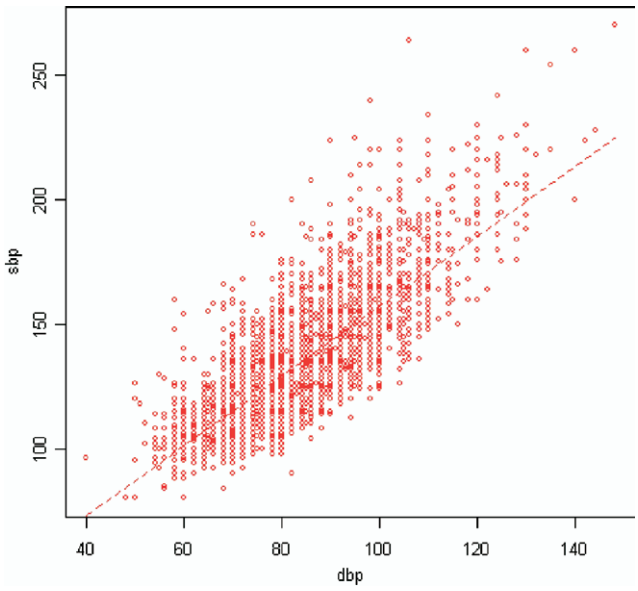

FIGURE 1. Framingham study data—diastolic versus systolic blood pressure of individual patients.

mean height is 168.6 cm with a standard deviation of 5.7 6cm. The t-value is 1.88; therefore, the mean heights are not statistically different.

The z-Test

The next test, which is very similar to the Student t-test, is the z-test. However, with the z-test, the variance of the standard population, rather than the standard deviation of the study groups, is used to obtain the z-test statistic.

Using the z-chart, like the t-table, we see what percentage of the standard population is outside the mean of the sample population. If, like the t-test, greater than 95% of the standard population is on one side of the mean, the p-value is less than 0.05 and statistical significance is achieved.

As some assumption of sample size exists in the calculation of the z-test, it should not be used if sample size is less than 30. If both the n and the standard deviation of both groups are known, a two sample t-test is best.

ANOVA

Analysis of variance (ANOVA) is a test that incorporates means and variances to determine the test statistic. The test statistic is then used to determine whether groups of data are the same or different. When hypothesis testing is being performed with ANOVA, the null hypothesis is stated such that all groups are the same. The test statistic for ANOVA is called the F-ratio.

As an example, ANOVA is used to compare values for pulse rate in 3 groups of trauma patients over 65, one group without comorbidities (control), a second group status post-heart attack without beta blocker treatment, and the third group status post-heart attack on beta blocker therapy. H_0: Myocardial infarction and beta blocker therapy do not affect pulse rate in trauma patients greater than 65 years of age. H_1: One of these groups is different than the other groups. Table 1 shows the data from such a study.

The F-statistic for this data set is 2.34. As with the t- and z-statistics, the F-statistic is compared with a table to determine whether it is greater than the critical value. In interpreting the F-statistic, the degrees of freedom for both the numerator and the denominator are required. The degrees of freedom in the numerator is the number of groups minus 1 (2 in this example), and the degrees of freedom in the denominator is the number of data points minus the number of groups (197 in this example). The critical value for the F-statistic in this case is 3.04; therefore, the null hypothesis cannot be rejected for this data set. The groups are not statistically different.

This comparison is a basic ANOVA; more and more variables can be added, which increases the mathematical complexity, although the concepts remain the same.

NONPARAMETRIC TESTS

If the data do not meet the criteria for a parametric test (normally distributed, equal variance, and continuous), it must be analyzed with a nonparametric test. If a nonparametric test is required, more data will be needed to make the same conclusion. For this reason, categorical data are often converted to continuous data before analysis. Many nonparametric tests and multiple variations of each of those specific tests exist. Discussion will be limited to the few that are most used and correspond to the previously discussed parametric tests.

Chi-Squared

The chi-squared test is usually used to compare multiple groups where the input variable and the output variable are binary. This test is very well illustrated with an example of a 2×2 table of data comparing 2 groups, one receiving a treatment and the other not, and 2 outcomes of cure versus continued disease (Table 2).

The expected data table was constructed by taking the totals of the rows, multiplying them by the column totals, and then dividing by the grand total of subjects. Next a critical value is determined. The degrees of freedom for a chi-square table are the number of rows minus one times the number of columns minus one. For a 2×2 table, the degrees of freedom is always 1. For this example, the calculated chi-squared statistic is 27.5, and the

TABLE 1. Example ANOVA. Pulse rate in elderly trauma patients.

Pulse	MI + βØ	MI − βØ	Control	All groups
Mean	80.25	85.77	88.93	87.11
Standard deviation	16.81	16.59	19.96	18.97
Patients	25	48	127	200

TABLE 2. Example chi-square. Generic data.

	Outcome		
Treatment	+	–	**Total**
Observed data			
+	45	20	65
–	5	30	35
total	50	50	100
Expected Data			
+	32.5	32.5	65
–	17.5	17.5	35
total	50	50	100

critical value is 3.841, which allows us to reject the null hypothesis; the treatment positively affects outcome.

Along the same lines as the chi-squared test is Fisher exact test. This test is complicated to calculate but is used if the table of expected outcomes has any single value of less than 2 or if more than 20% of the values are less than 5. The test then compares the probability that the study data could have randomly been acquired and compares it with the probabilities of all expected data tables that could be generated with n subjects.

Spearman Rank Coefficient

Like the Pearson product correlation coefficient, the Spearman rank coefficient is calculated to determine how well 2 variables for individual data points can predict each other. The difference is that the data need not be linear. To start, it is easiest to graph all the data points and find the x and y values. Then rank each x and y value in order of occurrence. Similar to the Pearson correlation coefficient, the test statistic is from −1 to 1, with −1 being a perfect negative correlation and 1 a perfect positive correlation.

Table 3 and Fig. 2 show spontaneous tidal volume from a random sample of patients on any given day at a family practice clinic.

Mann-Whitney U Test

This test, sometimes referred to as Wilcoxon rank test, uses rank just as the previous test did. It is analogous to the t-test for continuous variable but can be used for ordinal data. This test compares 2 independent populations to determine whether they are different. The sample values from both sets

TABLE 3. Example Spearman Rank Coefficient. Spontaneous tidal volume and age.

Rank (age)	Age	Tidal volume (ml)	Rank (TV)
1	12	954	8
2	15	785	5
3	24	806	7
4	45	458	3
5	54	785	6
6	60	325	2
7	78	154	1
8	86	621	4

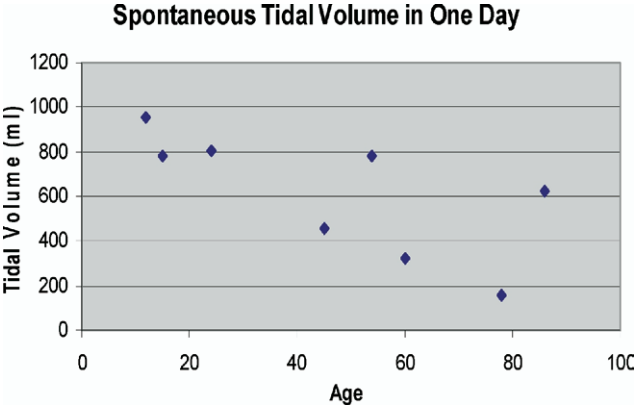

FIGURE 2. The test statistic for this data set is −0.71. It seems as though age negatively correlates with spontaneous tidal volume.

of data are ranked together. Once the 2 test statistics are calculated, the smaller one is used to determine significance. Unlike the previous tests, the null hypothesis is rejected if the test statistic is less than the critical value. The U-value table is not as widely available as the previous tables, but most statistic software will give a p-value and state whether statistical difference exists. Two valuable websites that provide this information are (http://fsweb.berry.edu/academic/education/vbissonnette/tables/mwu.pdf) and (http://www.physicalgeography.net/fundamentals/3g.html).

The following example will compare different pain medications and their effectiveness in controlling pain after inguinal hernia repair. There are 24 patients, and each report their pain score on postoperative day 1 before going to bed. The first group receives oxycodone/acetaminophen, and the other group receives ibuprofen/acetaminophen, each with a dosage of 1 to 2 tablets every 4 hours. Table 4 illustrates their pain scores and relative ranks.

The test statistic is less than the critical value. The null hypothesis that these 2 pain control regimens are equivalent should be rejected.

TABLE 4. Example Mann-Whitney U test (Wilcoxon Rank test). Effectiveness of pain medication regimens.

Oxycodone APAP score	Ranks	Ibuprofen APAP score	Ranks
1	1	4	6
2	2.5	5	9
2	2.5	6	13
3	4	6	13
4	6	6	13
4	6	7	17.5
5	9	7	17.5
5	9	7	17.5
6	13	7	17.5
6	13	8	21.5
8	21.5	8	21.5
8	21.5	9	24
Sum ranks	109	Sum ranks	191

Kruskal-Wallis Test

The Kruskal-Wallis test uses ranks of ordinal data to perform an analysis of variance to determine whether multiple groups are similar to each other. This test, like the previous example, ranks all data from the groups into 1 rank order and individually sums the different ranks from the individual groups. These values are then placed into a larger formula that computes an H-value for the test statistic. The degrees of freedom used to find the critical value is the number of groups minus 1. If in the previous example an additional group takes another pain regimen, such as tramodol, the Kruskal-Wallis test would be an appropriate test.

This compendium of tests is by no means exhaustive. Even with the examples presented, many different tests could be used to analyze the data. The tests outlined here are commonly used in clinical studies. Understanding these tests will provide some framework for analyzing test results when critically reading journal articles.