pharmacypractice

FEATURE

JOURNAL SERIES:

Statistics Review Part 5: Basic Statistical Tests

by Kevin Look, PharmD, PhD and Amanda Margolis, PharmD, MS, BCACP

This is the fifth article in a series of articles designed to review the basic, fundamental concepts of biostatistics. This article describes an approach to picking the correct statistical test, and outlines the correct situation for each statistical test.

Objectives:

1. Describe an approach to choosing a statistical test.
2. Define the assumptions and situation when each statistical test is appropriate.

Being able to determine if a research article has chosen the correct statistical test is an invaluable skill for a competent reader of biomedical literature. This article will review several commonly used statistical tests and discuss the appropriate situation in which each test is used.

Approach for choosing the correct statistical test

There are three major groups of basic statistical tests, which are grouped based upon the type of variable of dependent variable collected (discussed in part 1 of this series): continuous tests, ordinal tests, and categorical tests. Once the correct dependent variable is chosen the number of independent groups needs to be considered. A single-group pre-post test will use a different test than a calculation comparing the results between two independent groups. Lastly, normality must be considered to determine if all assumptions for each test are met. Once the type of dependent variable is determined, the number of independent groups are known, and normality assumptions are considered, the statistical test decision tree can be used to verify the appropriate statistical test (Figure 1).

Rationale of Statistical Testing

Both continuous and ordinal tests of statistical significance follow the same general approach. The investigator is trying to determine if the efficacy demonstrated (or the difference between group means) is large enough to convince a reader a true difference exists when the variation within groups is taken into account (standard deviation or background noise). Once all aspects of the specific equation have been entered into the formula, the test statistic is calculated. (The actual formulas and calculations are beyond the scope of this paper.) The larger the effect size and the smaller the noise, the larger the test statistic will be. A larger test statistic is more likely to indicate a statistically significant difference between the means. To determine if a test statistic is large enough to indicate statistical significance it is compared to a critical value. Critical values are determined using probability tables that account for the number of subjects in the sample. A test statistic must be larger than the critical value in order to be considered statistically significant. If a test statistic is larger than a critical value the null hypothesis can be rejected and the difference between the means is considered statistically significant (see paper 4 in this series for a review of hypothesis testing). If the test statistic is smaller than the critical value we fail to reject the null hypothesis due to insufficient evidence. When we do not find a statistically significant difference it does not mean that the two groups are equally efficacious; a different type of study (i.e., a non-inferiority study) would need to be conducted to determine if two therapies are similar.

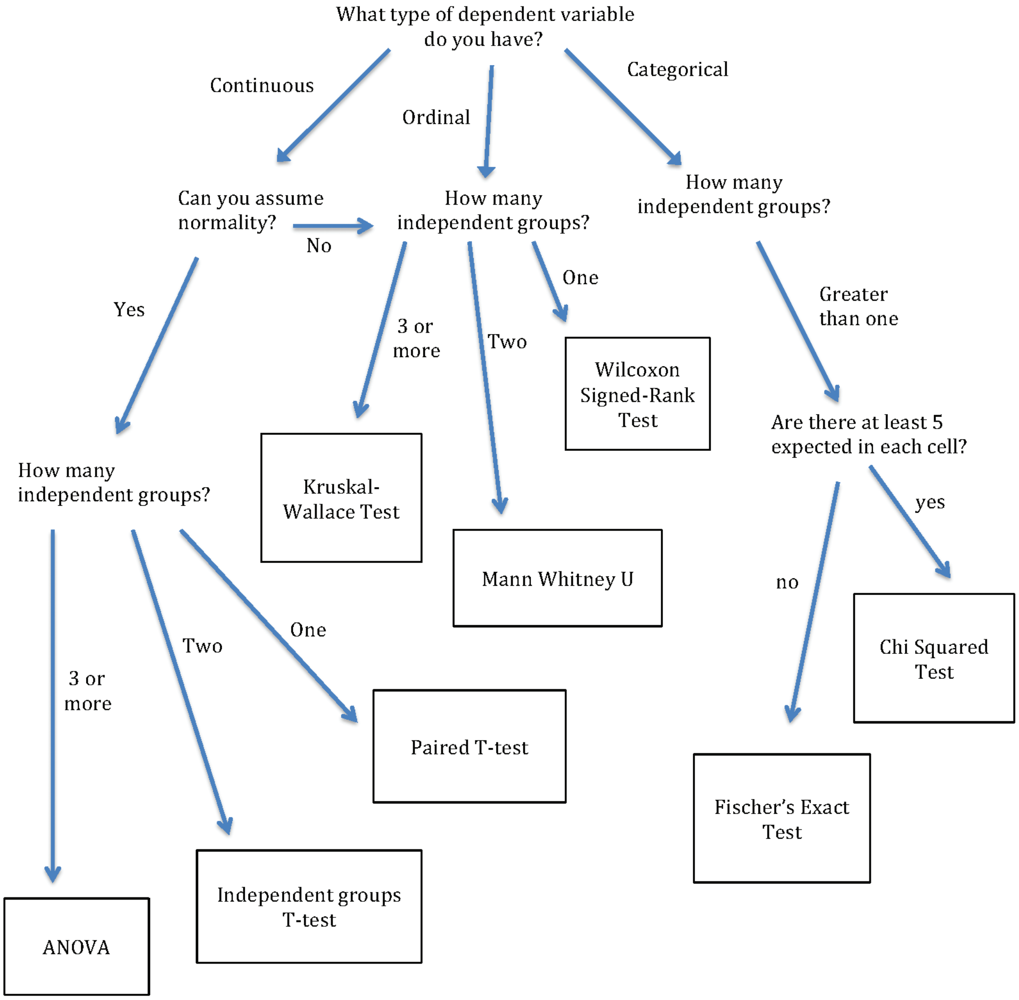
What type of dependent variable do you have?
Continuous
Ordinal
Categorical
Can you assume normality?
No
How many independent groups?
How many independent groups?
Yes
3 or more
Two
One
Greater than one
Wilcoxon Signed-Rank Test
Kruskal-Wallace Test
Are there at least 5 expected in each cell?
How many independent groups?
yes
no
Mann Whitney U
3 or more
Two
One
Chi Squared Test
Paired T-test
Fischer's Exact Test
Independent groups T-test
ANOVA

FIGURE 1. Statistical Test Decision Tree

Continuous Dependent Variables

The student's t-test or the two-sample t-test is one of the most commonly used statistical tests. It is used to see if the means between two groups for a continuous variable is statistically significant. However, there are several assumptions (or criteria) that must be met in order to appropriately use a student's t-test. The first assumption is that the groups must be independent from each other. This means that the individuals in the study were not able to influence each other's results. The second assumption is that the data are normally distributed (see paper 3 in this series for a description of normality). A Kolomogorov-Smirnov or goodness-of-fit test can be used to determine if the sample is normally distributed. A very loose rule of thumb is that at least 30 subjects are needed in each sample group to use a t-test; if the sample is less than 30 in at least one group a non-parametric test should be used instead (see nonparametric tests section). A t-test is also inappropriate in situations where the data are skewed (i.e., not normally distributed). However, if independence and normality are present in the sample the student's t-test can be used.

When a pre-post study is conducted in the same group of sample members the assumption of independent groups is no longer met (also known as paired data) and a one-sample test must be used. This would occur if a study was conducted in a single group of subjects and a continuous dependent variable was measured prior to the intervention and measured again post-intervention. In this case, those individuals with lower pre-intervention values will likely have lower post-intervention values and those with higher pre-intervention values will likely have higher post-intervention values. By pairing the data, or only looking at the pre-post difference for each individual, a large amount of the variability or background noise can be minimized. The decreased noise increases the power of the statistical test or increases the likelihood of being able to find a statistically significant difference when compared to using a two-sample test. When the dependent variable is continuous and normality can be assumed a paired t-test can be used.

Analysis of Variance or ANOVA is used to test for statistical significance of a continuous dependent variable between three or more independent groups assuming the data are normally distributed. ANOVA involves a complicated series of calculations to determine statistical significance, but in essence it is still comparing whether the observed differences between the groups is greater than the variability within the groups (similar to a t-test). When a statistically significant p-value is found using ANOVA, it means that there is a statistically significant difference between two or more of the groups. A student's t-test can then be used to tease out exactly where the differences are. ANOVA can be used for a comparison between two groups, but the student's t-test will give the same answer and is much easier to calculate.

Nonparametric Tests

In the case where normality cannot be assumed or the dependent variable has an ordinal measurement, a Mann Whitney U Test (also called the Wilcoxon Rank Sum Test) can be used instead. The Mann Whitney U Test is a non-parametric test, meaning it does not require a normally distributed sample. The Mann Whitney U Test is a ranked test, meaning the lowest value in the sample gets the lowest rank, and the second lowest sample value gets the second lowest rank, etc. Therefore, the Mann Whiney U test can be used for ordinal dependent variables due to the data ranking; it no longer matters if the difference between variables is linear, as only a general magnitude is preserved in the test. A test statistic is calculated from the ranks that is compared to a critical value to determine if the difference between the two groups is statistically significant. However, unlike the t-test, the Mann Whitney U Test is not testing for the difference between two means; it is actually testing for a statistically significant difference in the median, or for a shift in the overall sample distribution between the two groups. Samples tested for statistical significance using the Mann

Whitney U test do not have a minimum sample size but still need to meet the assumption of independence between the two groups.

Just like the Mann Whitney U is the nonparametric equivalent test for the Student's t-test, both the paired t-test and ANOVA have equivalent non-parametric tests. These tests are also used then the sample data do not meet the assumptions for normality or the dependent variable is ordinal. Analogous to the paired t-test, the Wilcoxon Signed-Rank Test should be used when a nonparametric one-sample paired test is needed. In cases where three or more independent groups are being compared with an ordinal dependent variable or for data that is not normally distributed, the Kruskal-Wallace Test should be used instead of ANOVA. The Kruskal-Wallace is performed using the same series of calculations as ANOVA, except it uses the data ranks instead of the actual results.

Categorical Dependent Variables

The Chi-Squared Test is used to test for statistical significance of a dependent categorical variable between two or more independent groups. Chi-Squared tests are usually set up as grids for calculations purposes with the different study arms as the rows and the outcomes as the columns. The only assumption for the Chi-Squared Test is that all cells must have an expected value of at least five. To calculate an expected cell value, the row total is multiplied by the column total for that cell, which is then divided by the total number of subjects in the study (Figure 2). The actual cell values are compared to the expected cell values to determine the Chi-Squared test statistic, which is then compared to the critical value to determine if the groups have a statistically significant difference in outcomes. In this case a difference between the two groups is a shift in the proportion of those with the outcome from one group to the next; a larger change in the proportion with the outcome will be seen as a larger Chi-Squared test statistic. In cases where the expected value is not at least five for all cells a Fischer's Exact Test should be used. The Chi-Squared Test is actually an approximation of the Fischer's Exact Test, but the Chi-Squared Test is more commonly used, as it is easier to calculate and is more intuitive to understand.

$$\text{Expected Cell} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total Number Subjects}} = \frac{\text{Total with Outcome} \times \text{Total in Group}}{\text{Total Number Subjects}}$$

	Total with outcome	Total without outcome	
Group A	16 17.7	35 33.3	51 (Row total Group A)
Group B	20 18.3	33 34.7	53 (Row total Group B)
	36 (Column total with outcome)	68 (Column total without outcome)	104 Total sample size (n)

Blue = actual numbers of subjects seen in study
Green = expected calculated numbers for chi-squared test
n= number of subjects

FIGURE 2. Equation to determine a Chi-Squared expected cell value

Summary

This article explained various assumptions of several commonly used statistical tests. Readers should be able to use the statistical test decision tree in Figure 1 to choose when each test is appropriate to use. The next article in this series will begin to shift the focus of this series to research methods by describing key points in randomized clinical trials. ●

Kevin Look is an Assistant Professor in the Social & Administrative Sciences Division at the University of Wisconsin School of Pharmacy, Madison, WI. Amanda Margolis is a Lecturer at the UW-Madison School of Pharmacy and a Clinical Pharmacist at the William S. Middleton Memorial Veterans Hospital, Madison, WI.

Practice Questions

1. You are reading a paper that presents the number of patients who had a stroke in a randomized controlled trial. 1 out of 52 patients in the intervention group had a stroke and 3 out of 47 patients in the control group that had a stroke. Which test would be most appropriate to determine significance in this case?
   a. Student's T-test
   b. Mann Whitney U Test
   c. Fischer's Exact Test
   d. Chi Squared Test

2. A study of a new asthma medication randomized patients to a new inhaler versus standard of care. There are 82 patients in the new inhaler group and 78 patients in the control group. Which test is the most appropriate to test for statistical significance of FEV_1 between the two groups?
   a. T-test
   b. ANOVA
   c. Chi Squared Test
   d. Mann-Whitney U Test

3. A study collects pre-post depression scores (range 1-5 with 1 being less depressed) for 130 participants before and after starting Cymbalta. Which is the best test to check for statistical significance?
   a. Student's T-test
   b. Paired T-test
   c. Mann Whitney U Test
   d. Wilcoxon Signed-Rank Test

Answers:

1. **c** Stroke is a categorical dependent variable and this data does not meet the chi-squared assumption for at least five expected in each cell. Therefore, the Fischer's Exact Test is the best option.

2. **a** FEV1 is a continuous dependent variable. As we are testing two independent groups and it meets the general rule of thumb having at least 30 subjects in each group, the t-test is the best option. A Kolomogorow-Smirnov or goodness-of-fit test could also be performed to determine if the sample is normally distributed.

3. **d** The depression scores are an ordinal variable so a non-parametric test must be used. As the data is a pre/post comparison of paired data the Wilcoxon Signed-Rank Test is the best option.

References and suggestions for further review:

1. Overholser BR, Sowinski KM. Biostatistics primer: part 2. Nutr Clin Pract. 2008;23(1):76-84.
2. Plichta SB, Kelvin E. Statistical Methods for Health Care Research. Sixth Edition. Philadelphia, PA: Lippencott Williams and Wilkins; 2013.