Shane Allua, PhD, and Cheryl Bagley Thompson, PhD, RN

Inferential Statistics

This article is the 15th in a multipart series designed to improve the knowledge base of readers, particularly novices, in the area of clinical research. A better understanding of these principles should help in reading and understanding the application of published studies. It should also help those involved in beginning their own research projects.

Descriptive statistics (see part 13 in this series[1]) summarize the data with the purpose of describing what occurred in the sample. In contrast, inferential statistics are calculated with the purpose of generalizing the findings from a sample to the entire population of interest. For instance, an investigator would use inferential statistics to determine whether differences between groups (ie, treatment and control groups) are unique to his or her sample (because of chance) or are a result of real differences between the population represented by group 1 and the population represented by group 2 (or however many groups are involved). Inferential statistics, therefore, rely on appropriate sampling methods (see part 5 of this series[2]) to ensure maximal representation of the population of interest. Inferential statistics are based on probability theory and the process of hypothesis testing (see part 14 in this series[3]).

Inferential statistics can be classified as either parametric or nonparametric. Nonparametric statistics are most commonly used for variables at the nominal or ordinal level of measurement, which basically means that they are used for variables that do not have a normal distribution. Statistical significance is calculated using information contained only in the sample (rather than the population) and may use measures of central tendency appropriate for nominal or ordinal level data (ie, the median rather than the mean). Parametric statistics are the most common approach to inferential statistical analysis. Parametric statistics require that the variables be measured at the interval or ratio level. Use of parametric statistics also relies on other assumptions, such as the expectation that values for a given variable will be normally distributed in the population.

Inferential statistics encompass a variety of statistical significance tests that investigators can use to make inferences about their sample data. These tests can be divided into three basic categories depending on their intended purpose: evaluating differences, examining relationships, and making predictions. The decision of which procedure to use is determined, in part, by the investigator's research question or research design. Level of measurement of the data (see part 13 in this series[1]) is also an important determinant in choice of significance test. Table 1 summarizes commonly used parametric and nonparametric statistical analyses. Research questions addressed by more commonly used parametric inferential statistics are discussed, followed by a discussion of statistical reporting and interpretation. Brief mention is made of common nonparametric equivalent statistical tests. However, the reader may want to consult with a statistician or other resources for further information on these as well as other advanced statistical procedures.[4]

Evaluating Differences

Significant difference tests can be used to evaluate differences on one interval or ratio level dependent variable of interest between two groups and three or more groups.

Two Groups

Research question: Is there is a significant difference in the mean flight time for trauma and nontrauma patients?

The independent samples *t*-test is used to test the statistical significance of the differences in means between two groups (a dichotomous independent variable) on some dependent variable measured at the interval or ratio level. For example, in an investigation of transport times, a *t*-test can be used to determine whether there is a significant difference in mean flight time for trauma and nontrauma patients. The *t* is the actual test statistic that is calculated and compared with the critical values of *t* that mark the critical region(s) indicating the presence of statistical significance. The critical *t* value is determined by the researcher-selected significance level or alpha level (eg, $\alpha = .05$) and the degrees of freedom that represent the conditions under which the *t* is calculated (related to numbers of subjects, number of groups, and the statistic). The *P*-value is the probability of whether the differences seen in the two flight times is present because of a true difference (in population means) or because of chance (seen only in this sample, not the population). If the calculated *t* value falls within the critical region and thus $P < .05$, the null hypothesis is rejected in favor of the research hypothesis.

Three or More Groups

Research question: Is there is a significant difference in mean systolic blood pressure for the control group, the group with drug A, the group with drug B, or the group with both drugs A and B?

Table 1. Common Statistical Techniques

Statistical Test	Independent Variable	Dependent Variable	Comments
Non Parametric			
Chi-Squared	Nominal	Nominal	Need > 5 expected subjects/cell
Fisher's Exact	Nominal	Nominal	
Mann-Whitney *U*	Dichotomous	Ordinal	
Kruskal-Wallis	Nominal	Ordinal	3 or more values for Independent Variable
Spearman Rho	Ordinal	Ordinal	
Parametric			
t-Test	Dichotomous	Interval/Ratio	
ANOVA	Nominal	Interval/Ratio	
MANOVA	Nominal	Interval/Ratio	Multiple Dependent Variables
Pearson's *r* (Correlation)	Interval/Ratio	Interval/Ratio	
Simple Regression	Interval/Ratio	Interval/Ratio	Single Independent Variable
Multiple Regression	Interval/Ratio	Interval/Ratio	Multiple independent variables
Canonical	Interval/Ratio	Interval/Ratio	Multiple Independent and dependent variables

An analysis of variance (ANOVA) is slightly more complex than a *t*-test but is based on the same mathematical principles. In fact, when an ANOVA is calculated for a two-group independent variable, the conclusions (significance vs nonsignificance) are exactly the same as the results obtained with a *t*-test. Although ANOVA can be used with two groups, it is most commonly used for independent variables that have three or more groups (possible values for the independent variable). Again, the dependent is assumed to be measured at the interval or ratio level.

For ANOVA, the statistic calculated is an F, rather than a *t*. The F statistic is the value compared against the critical value of F, which defines the critical region to determine statistical significance. The *P*-value is interpreted exactly the same as for a *t*-test such that if the *P*-value is below the selected alpha and therefore the obtained F value falls within the critical region, the null hypothesis can be rejected. With an ANOVA, a statistically significant *P*-value indicates that there are group differences present in the data but does not indicate which groups are different. Thus, in an analysis of four treatment groups, if group 4 has much higher blood pressure than any of the other three groups, the investigator cannot assume that this specific difference is statistically significant. Further analysis (post hoc analysis) is needed to determine the nature of the differences. Post hoc analysis is beyond the scope of this paper.

Both the *t*-test and ANOVA noted above assume independent samples or groups. For dependent samples where, for example, the same sample is measured at time 1 and time 2, a dependent samples *t*-test or repeated-measures ANOVA (time 1, time 2, and time 3) should be used. Furthermore, evaluation of group means can be done for multiple independent variables (factorial ANOVA) and multiple dependent variables (multivariate analysis of variance, MANOVA). Other nonparametric statistical tests used to evaluate group differences include the Mann-Whitney *U* test, the Wilcoxon T test, and the Kruskal-Wallis test.

Examining Relationships

Statistical tests also evaluate the significance of the relationship between two variables and the strength of the relationship. A correlation, however, cannot be used to infer a causal relationship between two variables. In other words, an investigator should only draw the conclusion that a relationship between two variables does or does not exist, not which variable is the cause of the other variable.

Research question: Is there is a positive relationship between scene time and flight time?

The most common statistic used to describe the relationship (the correlation) between two variables is the Pearson product-moment correlation or Pearson's *r* (r_p). Pearson's *r* is a descriptive statistic when used only to describe a relationship; it is an inferential statistic when used to infer a relationship in the population. Pearson's *r* requires that both variables be measured at least at the interval level of measurement. Consequently, a correlation between role and age is not appropriate, even if role is assigned a numerical value (1 = nurse, 2 = physician).

Pearson's *r* is used to evaluate both the statistical significance of the relationship and the magnitude and direction of the relationship. Pearson's *r* ranges from -1.0 to +1.0: a +1.0 indicates a perfect direct (positive) relationship, and a -1.0 indicates a perfect inverse (negative) relationship. Pearson's *r* of +1.0 or -1.0 are both represented as a straight line when displayed graphically. The further the data are from a perfect relationship, the more the points spread out from a straight line. One assumption of correlation is that a linear relationship exists between variables. To the extent that there is a nonlinear relationship between the two variables being correlated, correlation will understate the relationship.

Similar to both the *t* and F statistic, the obtained r_p is compared against the critical *r* value to define the critical region. If the obtained r_p falls within the critical region and thus the *P*-value is less than the selected alpha level, the investigator can conclude that there is a statistically significant relationship between the variables. The r_p also informs the investigator about the strength and direction of the relationship. For example, if r_p = +0.37, there exists a moderate, positive relationship between scene time and flight time, meaning that as scene time increases, so does flight time. It does not mean that an increase in scene time *causes* an increase in flight time or vice versa.

When data are not measured at the interval or ratio level, other variations of correlations are more appropriate. For example, the correlation between two ordinal level variables should be analyzed using the Spearman rho correlation coefficient, the nonparametric equivalent to the Pearson *r*.

A common nonparametric statistic used to examine the relation between nominal level variables is the chi-squared test of association.

Research question: Is there a relationship between intubation success and use of neuromuscular blockade?

Data analyzed using the chi-squared test statistic, χ^2, are typically organized into a contingency table. The chi-squared procedure calculates the expected number of observations in each cell of the contingency table and compares them with the number of observations actually occurring in each cell (observed frequencies). The greater the deviation of the observed frequencies from the expected frequencies, the greater the chance for statistical significance, providing evidence that the variables are related. In contrast to the correlations described, the interpretation of the χ^2 does not include the direction of the relationship (ie, positive or negative). Rather, it is described by looking at the pattern of observed frequencies in the contingency table and identifying which categories of one variable go with which categories of the other variable. For example, given χ^2 = 5.3, which is determined to be statistically significant at the .05 level, a statistically significant relationship between intubation success and use of neuromuscular blockade exists. Thus, evidence supports the conclusion that the use of neuromuscular blockade increases the intubation success rate.

Making Predictions

Although correlation does not allow the investigator to infer causation, other statistical tests can be used to predict one variable from another.

Research question: Do weight, height, and smoking status influence resting pulse rate?

Simple regression analysis is used for a single independent and dependent variable, whereas multiple regression analysis is used for multiple independent variables. Multiple regression assumes that the dependent variable is measured at the interval or ratio level. Regression analysis is computationally related to ANOVA but is used for independent variables that are measured at the interval or ratio level, rather than nominal or ordinal level. Interval or ratio level independent variables can be broken down into ordinal categories and analyzed using ANOVA (eg, ages 20-39 = 1, ages 40-59 = 2, ages 60-79 = 3). However, because a great deal of information is lost with this approach, regression analysis is preferred for interval or ratio-level independent variables.

Statistical significance obtained using regression is somewhat different from that described for the statistical tests presented thus far. First, to determine whether the overall model or, if together, all independent variables are significant predictors or the dependent variable, the F statistic is evaluated as described previously. In addition, regression provides the investigator with a measure of predictive accuracy that is determined by the strength of the relationship. The predictive accuracy or R^2 is the percentage of variance in the dependent variable explained by the independent variable(s). Another way to say this is that the R^2 is the percentage of overlap between the independent and dependent variables. For example, given a significant F value, the investigator would conclude that, given R^2 = 0.3064, approximately 31% of the variation in pulse rate can be explained by weight, height, and smoking status. Regression is also used to determine the significance and importance of each predictor or each independent variable in the model. The discussion of regression coefficients (eg, beta weights) is beyond the scope of this paper.

By convention, a small r^2 is used when referring to a single independent variable (eg, correlation), and a large R^2 is used for multiple independent variables. Reporting the adjusted R^2 is recommended, especially when the number of independent variables is high. Other variations of regression are available for research questions that involve the prediction of nominal or ordinal level dependent variables or for the prediction of more than one dependent variable.

Interpretation

The preceding discussion focused on the mechanics of statistical analysis. This section focuses more on the fuzzier aspects of statistical analysis, statistical reporting, and interpretation.

Although limited somewhat by journal requirements, authors have some leeway in how they present the results of their statistical analysis. The reader is advised to examine several journals for statistical output formats to find one that is

most appealing and understandable (unless restricted by journal). The author also may elect to place results only within the text of the manuscript. Although tables often are easier to read, simple results may not merit the space required for a table or figure. When reporting findings in text, the author must report the names of the independent and dependent variables, the statistical procedure used, the statistic calculated (eg or F), the value of the statistic, appropriate degrees of freedom (*df*, from the computer printout), and the *P* value (in relation to alpha or actual value). For example, a narrative report of the *t*-test described earlier would be that the mean flight time for nontrauma patients was statistically higher than for trauma patients, $t = 2.82$, $df = 644$, $P < .05$. More commonly, the exact *P*-value provided in the computer printout is reported (eg, $P = .0049$).

A final topic for discussion is the interpretation of statistical output. Knowledge of more than the numbers obtained from the computer is needed to adequately interpret a statistical analysis. The first consideration is clinical versus statistical significance. Not all results that achieve statistical significance have clinical significance. For example, in a study of the effect of drug A on chronic hypertension, a statistically significant difference between the control and experimental group was found. However, the average systolic blood pressure for the control group was 162, and the average systolic blood pressure for the experimental group was 158. Although this drop in blood pressure was statistically significant, such a small drop in blood pressure has no clinical significance to the patient. In addition, a larger drop still might be considered clinically insignificant if the cost of the drug was extreme in comparison with a moderate drop in blood pressure.

Statistically significant results that do not have clinical significance are more common in studies with large sample sizes. For this reason, investigators often decrease alpha in the case of a large sample so as to diminish this problem (and decrease the chance of a type I error).

On a very rare occasion, a study might have clinically significant findings without achieving statistical significance. This most often occurs in the event of small sample sizes. For example, a pilot study of 10 patients—five in the control and five in the experimental group—might show a large difference in the means of the systolic blood pressure for the two groups. However, large variability in systolic blood pressure between subjects in each group (large standard deviation) and a small sample size may preclude finding statistical significance. The researcher still may wish to consider the results clinically significant and worthy of further investigation.

A final caution about the interpretation of statistical analyses is that statistics can be misleading. A reader should look to a report of the study to find evidence that many of the decisions that are important to statistical analysis are made before the study, not after the results are obtained. For example, an investigator should not wait until after obtaining the *P*-value before selecting an alpha.

Other methods for manipulating an analysis of which the reader should be aware include subdividing or combining groups within the sample to maximize the results. At times there are good reasons to alter group designations (too few subjects per cell), but these possibilities should be discussed before the analysis and not after. Finally, statistical significance can be found if the investigator looks hard enough. As discussed previously, the *P*-value relates to the probability of finding results only by chance. If an alpha is set at .05, and 20 analyses are computed, at least one of the analyses is likely to be statistically significant merely by chance (5% chance). Because research journals often are more interested in studies that demonstrate statistical significance, an author may be tempted to do multiple exploratory analyses merely to find something that is publishable. Here, being able to link the review of literature to the research questions and results becomes very important. Such links help to demonstrate that the questions were well thought out in advance and not merely the result of creative data analysis.

Conclusion

We would like to stress that conducting and interpreting a statistical analysis are not impossible. Understanding a few basic concepts will provide a background helpful for interpreting a wide range of statistics, even if the details of the math are not understood. The researcher also is reminded to seek out the assistance of a statistician early in the process of research. A statistician can provide direction, education, and reassurance as you grow in your statistical sophistication.

References

1. Thompson CB. Basics of research part 13: Descriptive data analysis. Air Med J 2009;58:56-9.
2. Panacek EA, Thompson CB. Basics of research part 5: Sampling methods: Selecting your subjects. Air Med J 2007;26:75-8.
3. Allua S, Thompson CB. Basics of research part 14: Hypothesis testing. Air Med J 2009;.
4. Diekhoff GM. Basic statistics for the social and behavioral sciences. Upper Saddle River, NJ: Prentice Hall; 1996.

Shane Allua, PhD, is a senior consultant for a global business consulting company located in Bethesda, MD. Cheryl Bagley Thompson, PhD, RN, is an associate professor and assistant dean of informatics and learning technologies at the University of Nebraska Medical Center College of Nursing in Omaha. She can be reached at cbthompson@unmc.edu.

1067-991X/$36.00

doi:10.1016/j.amj.2009.04.013