SPECIAL CONTRIBUTIONS

Research Fundamentals: Statistical Considerations in Research Design: A Simple Person's Approach

JAMES B. JONES, MD, PHARMD

Abstract. A basic understanding of statistical methodology is essential, both for designing quality research projects and for evaluating the medical literature. Careful statistical planning, including the selection of study endpoints, the determination of the required sample size, and the selection of statistical tests to be used in the data analysis, is important to ensure a successful research project. The purpose of this article is to provide a basic review of statistical terms and methods for both the researcher and the clinician, as well as to clarify questions that need to be answered prior to embarking on an experimental study. The advantages of collaborating with statistical consultants, and some guidelines for such collaborations, are discussed as well. **Key words:** statistical methods; clinical research; hypothesis testing; statistical tests. ACADEMIC EMERGENCY MEDICINE 2000; 7:194–199

THE MOST powerful tool for advancing the knowledge base of a medical specialty is a well-designed clinical research program. Information generated by such a program should directly influence the practice of clinical medicine in the emergency department (ED). A basic understanding of statistical methodology is essential, both for designing quality research projects and for evaluating the medical literature. However, just the thought of statistics intimidates many physician investigators. Statistical planning, prior to the initiation of a study, is essential to ensure a successful research project. Particular attention should be paid to choosing the proper study design,[1] analysis methods,[2] and number of subjects.[3]

The purpose of this article is to provide a basic review of statistical terms and methods for both the researcher and the clinician, as well as to clarify questions that need to be answered prior to embarking on an experimental study.

ESTABLISHING THE RESEARCH QUESTION

One of the most important and difficult steps in clinical research is the selection of the research question.[4,5] A good research question must fulfill several criteria, including originality, feasibility, and clinical relevance. Research questions can have their origins from various sources: one's own clinical experience and observations, gaps in the medical literature, or conversations with your colleagues. Identifying a research question frequently takes a significant amount of time and review of the literature.[5]

HYPOTHESIS TESTING

After a research question has been defined, the endpoint must be determined. For example, an investigator is interested in whether a new analgesic can reduce an ED patient's level of pain. To evaluate this, the researcher could use a 100-mm visual analog scale to allow the patient to assign a number to his or her level of pain at any given point in time.

When both the research question and the endpoint have been agreed upon, the investigator needs to formally define the hypothesis for the study.[6,7] The hypothesis is the statement that the project is designed to prove or to disprove. Statistical methods allow investigators working with a sample of observations (data) derived from a group of subjects to make generalizations regarding the population from which the subjects were obtained. This process is referred to as hypothesis testing.[7] There are certain assumptions regarding the population that must be made in order to allow statistical interpretation of the hypothesis. For example, one must assume that the study group is a random sample from a given population. If the measured

From the Methodist Hospital, Clarian Health Partners, Indiana University School of Medicine, Indianapolis, IN (JBJ).
Received November 4, 1999; accepted November 8, 1999.
Series editor: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor–UCLA Medical Center, Torrance, CA.
Address for correspondence and reprints: Mary Ann Schropp, Executive Director, Society for Academic Emergency Medicine, 901 North Washington Avenue, Lansing, MI 48906-5137. Fax: 517-485-0801; e-mail: saem@saem.org

characteristic of the population being studied follows a normal distribution (i.e., a bell-shaped curve), the data generated from the study will also follow a normal distribution. Such data are called "parametric" data.

In classic hypothesis testing, a calculated p-value is used to decide which of two hypotheses is supported by the data.[7–9] The two hypotheses are the null hypothesis and the alternative hypothesis (Table 1). The null hypothesis states that there is no difference between the groups being compared, with respect to the characteristic being measured. For example, in a study evaluating a new analgesic for pain control, the null hypothesis might state that no difference exists in the amount of pain relief obtained from the new analgesic compared with the control agent. The alternative hypothesis, on the other hand, might state that there is a difference between the two treatments being tested. The difference defined by the alternative hypothesis is the treatment effect the study is designed to detect. For example, the treatment effect may be a difference of 30 mm using a 100-mm visual analog pain scale.

The alternative hypothesis can be stated in different ways, depending on the investigator's expectation about the direction of the treatment effect. It can state that there is a difference of a certain magnitude without indicating a direction, leaving open whether the effect is positive or negative in the test group relative to the control. This is called a two-tailed or nondirectional alternative hypothesis. It can also be given a direction, stating that the effect of the treatment is either better or worse than the control. This is called a one-tailed or unidirectional hypothesis. In general, two-tailed alternative hypotheses are preferred. In either case, the magnitude and direction of the difference, if any, should be clearly stated in the study proposal. Defining the null and alternative hypotheses is a prerequisite to: 1) determining the study design; 2) determining the sample size; and 3) determining the statistical methods to be used in the analysis.

Once the hypotheses have been clearly defined and the study data obtained, a statistical test (which is really a mathematical formula) is used to calculate a p-value. The p-value is used to determine which of the hypotheses will be accepted as true. A great deal of statistical literature revolves around the topic of selecting the correct statistical test.[2] To select the correct test, one must assess the type of data (nominal or categorical, ordinal, continuous) and, for continuous data, decide whether it is safe to assume that they will be normally distributed.[10] In addition, one must define whether the measurements are paired (e.g., before-and-after measurements on each subject) or unpaired. Table 2 lists some common statistical tests.

TABLE 1. Definitions of Commonly Used Statistical Terms

Term	Definition
Null hypothesis	The hypothesis that there is no difference between the groups being compared, with respect to the measured variable.
Alternative hypothesis	The hypothesis that a difference does exist between the groups being compared, with respect to the measured variable. This hypothesis must define the magnitude of the difference.
α	The maximum p-value to be considered statistically significant. The risk of committing a type I error, assuming the null hypothesis is true.
β	The risk of committing a type II error, assuming the alternative hypothesis is true.
Power	The probability that the study will detect the treatment effect defined by the alternative hypothesis, if it truly exists.
Type I error	The error that results if a true null hypothesis is rejected, i.e., if one concludes a difference exists when there really is no difference; a false positive.
Type II error	The error that results if a false null hypothesis is accepted, i.e., if a difference is not detected when a difference exists and it is as big as that defined by the alternative hypothesis; a false negative.
p-value	The probability of observing a difference as large as or larger than the one actually observed, by chance alone, assuming there is no difference between the groups.

Next, one must select the "level of significance," denoted as α. The value of α is the maximum p-value to be considered statistically significant. This value also gives the probability of rejecting the null hypothesis, when the null hypothesis is true and there is no difference between the groups. For example, in the case of the new analgesic, one could conclude that there is a difference between the new analgesic agent compared with the control when, in fact, the two provide the same amount of pain relief. The value of α has traditionally been set at 0.05. The smaller the value of α, the less likely we will incorrectly reject the null hypothesis, but the harder it is to detect a true difference between the groups.

The next step in hypothesis testing is to calculate the p-value, using the appropriate statistical test. The p-value is the probability of obtaining data showing as large a difference between the two groups as that actually observed, or a larger difference, by chance alone under the assumption that there really is no difference between the groups (the null hypothesis is true). The smaller the p-value, the more inconsistency exists between the observed data and the null hypothesis. If the p-value is less than α, then the null hypothesis is rejected and the alternative hypothesis is accepted by default.

Types of Errors

In determining whether the two groups are the same (the null hypothesis is accepted) or they are different (the alternative hypothesis is accepted), two potential errors can be made. These errors are called a type I error and a type II error. A type I error occurs when the null hypothesis is rejected (a p-value less than α is obtained) when it is actually true and there is really no difference between the two groups being compared. When this occurs, the investigators will report that a difference exists between the two groups being studied, when in fact no difference exists. The chance of this occurring, when there is no difference between the populations of subjects being compared, is given by α.

A type II error occurs when the null hypothesis is not rejected (a p-value greater than α is obtained), when in fact a difference does exist between the two groups, *and* the difference is as large as that defined by the alternative hypothesis. The probability of committing this type of error, when the difference defined by the alternative hypothesis does exist, is denoted by β. This type of error commonly occurs when an insufficient number of subjects are studied. In fact, the smaller the size of the treatment effect the study is designed to detect, the larger the sample size must be to reliably reach statistical significance and minimize the chance of committing a type II error.[3]

The power of the study is the probability that the study will detect the predetermined treatment effect between the two study groups, if it truly exists, given the value of α and the sample size.[11] Since β is the chance of *not* finding the true treatment effect, power is given by $(1 - \beta)$. Since the power of the trial is the chance of finding a true treatment effect, the quantity (1 − power) is the chance of committing a β (type II) error. Power can be calculated using a number of software packages, or published tables[12,13]; however, it is often useful to consult with a statistician or mentor with a strong statistical background to ensure the calculation is performed correctly. Power is commonly set at 80% or 95%. A power of 95% is generally preferred, although it requires a larger sample size. A larger power assures the reader that, if the treatment effect defined by the alternative hypothesis was really there, a statistically significant p-value would likely have been obtained. It is essential to assess the power of a study that fails to identify an expected treatment difference. If the power is too low, the negative result is unreliable and does not reliably exclude the existence of the defined difference between the groups.[3]

Sample Size Planning

Determination of the appropriate sample size is sometimes referred to as power analysis, since the sample size is the primary determinant of the power of a study, which is under the control of the investigator. The calculation of a sample size required to find a treatment effect between two study groups is based on several important factors, including: the desired effect size and expected variability in the observed data, the value of α, the desired β or power, and the study design.[3,11–13] Often trade-offs must be made between α, β or desired power, and the minimum effect size sought, to yield a practical sample size. This is an area where a good statistical consultant or mentor can be quite helpful. With currently-available statistical software programs, sample size calculations can be performed for most studies.

Planning the Statistical Analysis

Careful planning of the statistical analysis to be performed at the end of a study is essential to minimize the chance of drawing erroneous conclusions. In general, one should define a single primary comparison that will answer the research question. Secondary comparisons, such as the comparison of secondary endpoints or the comparison of baseline (prior to intervention) characteristics, should be defined as well. These comparisons must be defined prior to data collection. The study should be designed to have an adequate sample size to reliably detect a clinically important treatment effect in the primary outcome, as defined by the alternative hypothesis. Any subgroups of individuals to be considered separately must be defined prospectively.[14,15] Any baseline characteristics that may affect patient outcome of the study must be accounted for in the statistical analysis, either by comparing these baseline characteristics between treatment groups, or by their inclusion in a multivariate model (e.g., logistic regression).

The choice of the appropriate statistical test to use to determine the p-value rests on the choice of

the endpoint, the selection of the most efficient design (e.g., paired vs unpaired measurements), and the type of data to be collected. The most important of these criteria is the type of data collected, whether categorical or nominal, ordinal, or continuous. For continuous data, it is important to assess whether the data are likely to be normally distributed. It is important that researchers understand the type of data they are collecting before they choose a statistical method. The most common statistical tests,[2] and the type of data for which they are suited, are shown in Table 2.

Categorical refers to data without numerical value, which are also referred to as nominal data. Categorical data divide subjects into categories (e.g., gender, ethnicity, survival to hospital discharge). Analysis methods for this type of data include the chi-square test and Fisher's exact test.

Ordinal data are characteristics that have an underlying order to their values, but the particular numbers are arbitrary. For example, the Glasgow coma score is probably best treated as ordinal, since it is unknown whether the difference between 14 and 15 is the same as the difference between 3 and 4. This type of data requires ranking or categorical methods of analysis.[16]

Continuous or interval data are data from a scale that measures numerical characteristic with values that occur on a continuum; for example, age, heart rate, or body temperature. In addition, differences between two values have specific meaning. This type of data are commonly analyzed using Student's t-test or the Wilcoxon rank sum test. These tests compare the means or medians of variables between two groups of subjects. Student's t-test requires that the data from the two groups be normally distributed and have equal variances, whereas the Wilcoxon rank sum test does not have these requirements.[10,16] If three or more groups are being compared, one-way analysis of variance (ANOVA) and the Kruskal-Wallis tests can be used.

Student's t-test and ANOVA are examples of parametric statistical tests. These tests assume that the data follow a normal distribution and all groups yield data with equl variances. If the data are not normally distributed, then nonparametric statistical tests are indicated.[10,16] The Wilcoxon rank sum test can be used for unpaired samples, while the Wilcoxon signed rank test can be used for paired samples. Samples can be paired if the two data points are obtained from an individual subject (e.g., predrug and postdrug pain scores). Unpaired samples occur when single data points from many individuals are being compared.

A potential problem exists if a single patient has multiple observations in the course of a study, or if there are multiple endpoints to be compared.[17–22] For example, consider a study in which a visual analog pain scale is administered every 5 minutes for 30 minutes (a total of six observations) after a baseline measurement. Normally, when two groups of measurements are compared statistically (baseline and 30-minute pain scores), if the two groups are identical there is still a 5% or α chance that a statistically significant p-value will be obtained. If the investigator performs multiple comparisons using the data (baseline vs 5 minutes, baseline vs 10 minutes, etc.), the risk of at least one false-positive p-value is increased, because the risk associated with each test is incurred multiple times.[17] The overall risk of at least one type I or false-positive error is roughly equal to the α value used in each test, multiplied by the total number of tests performed. This is the basis for the Bonferonni correction.[17] The Bonferonni correction is a

TABLE 2. Common Statistical Tests

Statistical Test	Description
Student's t-test	A test for continuous data, to determine whether the means of two groups are equal. Assumes the data follow a normal distribution and have equal variances in the two groups. Both paired and unpaired forms of the test exist.
Wilcoxon rank sum test	A test for ordinal or continuous data, similar to Student's t-test, but does not require that the data be normally distributed or that the variances be equal in the two groups. For paired measurements the Wilcoxon signed rank test is used instead.
Chi-square test	A test for categorical data with two or more treatment and outcome categories, to determine the effect of the treatment on outcome. Assumes five or more subjects expected in each "cell" of the contingency table.
Fisher's exact test	A test similar to the chi-square test, but useful when five or fewer observations are expected in some "cells."
Analysis of variance (ANOVA)	A test for continuous, normally distributed data, to compare the means of three or more groups. Also assumes that data from all groups have equal variances.
Kruskal-Wallis test	A test for ordinal or continuous data, analogous to the Wilcoxon rank sum test, but used when there are three or more groups being compared. Also analogous to one-way ANOVA, but does not require normally distributed data.

TABLE 3. Services a Statistician May Provide

1. Recommend study design
2. Recommend sample size
3. Contribute technical writing/editing of proposal
4. Help design data collection forms
5. Recommend data entry/management systems
6. Oversee data entry/data management
7. Provide data quality checks and cleaning
8. Plan data analysis
9. Recommend software for proposed analysis
10. Implement data analysis
11. Summarize results
12. Prepare tables, figures, grafts
13. Technical writing/editing of manuscript
14. Review a dry run of oral presentations and slides
15. Help respond to referees of a submitted manuscript

method for reducing the overall risk of a type I error for the whole study, by reducing the maximum statistically significant p-value or α value used for each of the individual statistical tests. The Bonferonni correction consists of dividing the overall desired α by the number of tests, and using the smaller value in interpreting the p-value of each individual test. For example, in the study mentioned above, one would take the overall α of 0.05, and divide by 6 (the total number of comparisons desired), yielding a maximal significant p-value for each comparison of 0.008. This latter value is sometimes called an "adjusted" α. The Bonferonni adjustment is generally a conservative adjustment, and it does not take into account any correlations between the different comparisons.

The downside to using the Bonferonni correction is that it controls the overall chance of committing a type I error by increasing the chance of committing a type II error, given a particular sample size. Since each test now uses a smaller "adjusted" α, making it more difficult to achieve statistical significance, the chance that an important difference will be missed increases. One can avoid using the Bonferonni correction, in specific cases, by using statistical tests that have been developed to compare three or more groups, such as ANOVA, the Kruskal-Wallis test, the chi-square test, and Fisher's exact test. These tests are useful for detecting differences in three or more groups, with relatively high power for a given sample size, and simultaneously control the risk of committing a type I error.

Many investigators preferentially use tests with which they are familiar, or that are commonly reported in the literature. However, these tests are not necessarily the best choice for a particular type of data or study design. Choosing the most appropriate statistical test for the data generated can be quite complex, and a statistician or a mentor with a strong statistical background should be consulted in the early stages of all research endeavors. Test selection is particularly difficult when multiple measurements are made on individual subjects (repeated or blocked measurements), data are not normally distributed, or adjustments must be made for baseline characteristics.

THE ROLE OF STATISTICAL CONSULTANTS

Many investigators use statistical consultants or experienced mentors to assist them in the design and prestudy planning of their clinical research projects. Such consultants can help identify potential problems within the proposal, including vagueness in the statement of the research question or hypotheses, aid in the choice of study design, and help determine sample size. Table 3 shows the range of services a good statistical consultant can

TABLE 4. Steps to Facilitate Interaction with a Statistical Consultant

1. Send several articles from the literature search to the consultant, prior to your first meeting, so that he or she may become familiar with the current knowledge base.
2. Supply the consultant with a brief statement regarding the motivation for and importance of the proposed topic, based on the available literature and/or other background material.
3. Send several articles that illustrate the type of study you would like to perform, and that illustrate the type of statistical analysis you would like to have conducted.
4. Have a list of the people involved, their roles, and an idea of the budget or other funding sources available to complete the project. Be prepared to discuss compensation for the consultant. Statisticians typically charge $50 to $100 per hour. Also be prepared to discuss authorship on the final manuscript, depending on the degree of input by the statistician.
5. Be open and honest in communicating your knowledge and areas of weakness. Immediately tell the consultant when you do not understand him or her (it will not become clearer later).
6. When choosing a consultant, ask about his or her familiarity with the subject area, and make sure he or she has the time available to assist you in your project. Failure to confirm the amount of time available to work on your study may result in serious delays in the data evaluation and presentation of the research. You may want to request a résumé or list of publications.
7. Ask for references for the statistical approach the consultant is proposing.
8. Ask for assistance with the creation of data collection tools and databases. Even if the statistician does not help with these steps, make sure he or she reviews all data collection tools and databases prior to use.

provide. Unfortunately, talking to a consultant can sometimes be intimidating, especially for the novice investigator. In order to make the initial meeting with the statistician easier, the steps shown in Table 4 may be helpful.

CONCLUSIONS

The design of high-quality clinical research studies can be difficult. A basic understanding of statistical issues is critical to the design and analysis of all research projects. Proper prestudy planning, and early consultation with either a statistician or experienced research mentor, is critical to the success of any research project.

References

1. Hall KN, Kothari RU. Research fundamentals: IV. Choosing a research design. Acad Emerg Med. 1999; 6:67–74.
2. Menegazzi JJ, Yealy DM, Harris JS. Methods of data analysis in the emergency medicine literature. Am J Emerg Med. 1991; 9:225–7.
3. Brown CG, Kelen GD, Ashton JJ, Werman HA. The beta error and sample size determination in clinical trials in emergency medicine. Ann Emerg Med. 1987; 16:183–7.
4. Lewis LM, Lewis RJ, Younger JG, Callaham M. Research fundamentals: 1. Getting from hypothesis to manuscript: an overview of the skills required for success in research. Acad Emerg Med. 1998; 5:924–9.
5. Kwiatkowski T, Silverman R. Research fundamentals: II. Choosing and defining a research question. Acad Emerg Med. 1998; 5:1114–7.
6. Silverman R, Kwiatkowski T. Research fundamentals: III. Elements of a research protocol for clinical trials. Acad Emerg Med. 1998; 5:1218–23.
7. Kelen GD, Brown CG, Ashton J. Statistical reasoning in clinical trials: hypothesis testing. Am J Emerg Med. 1988; 6: 52–61.
8. Lewis RJ, Bessen HA. Statistical concepts and methods for the reader of clinical studies in emergency medicine. J Emerg Med. 1991; 9:221–32.
9. Gaddis GM, Gaddis ML. Introduction to biostatistics: part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med. 1990; 19:820–5.
10. Lewis RJ. Parametric statistical tests: unnecessary assumptions, computers, and the search for the trustworthy p-value [editorial]. Acad Emerg Med. 1998; 5:1048–50.
11. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994; 272:122–4.
12. Fleiss JL. Statistical Methods for Rates and Proportions. Second Edition. New York: John Wiley & Sons, 1981.
13. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
14. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA. 1991; 266:93–8.
15. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med. 1992; 116:78–84.
16. Gaddis ML. Introduction to biostatistics: part 5, statistical inference techniques for hypothesis testing with nonparametric data. Ann Emerg Med. 1990; 19:1054–9.
17. Smith DG, Clemens J, Crede W, Harvey M, Gracely EJ. Impact of multiple comparisons in randomized clinical trials. Am J Med. 1987; 83:545–50.
18. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. Mayo Clin Proc. 1988; 63:813–5.
19. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment 2. Comparisons among several therapies. Mayo Clin Proc. 1988; 63:816–20.
20. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 3. Repeated measures over time. Mayo Clin Proc. 1988; 63:918–20.
21. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 4. Performing multiple statistical tests on the same data. Mayo Clin Proc. 1988; 63:1043–5.
22. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 5. Comparing two therapies with respect to several endpoints. Mayo Clin Proc. 1988; 63:1140–3.