REVIEW

A Clinician-Educator's Roadmap to Choosing and Interpreting Statistical Tests

Donna M. Windish, MD, MPH,[1] *Marie Diener-West, PhD*[2]

[1]Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA; [2]Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

As educators seek confirmation of successful trainee achievement, medical education must move toward a more evidence-based approach to teaching and evaluation. Although medical training often provides physicians with a general background in biostatistics, many are not prepared to apply these skills. This can hinder clinician educators as they wish to develop, analyze and disseminate their scholarly work. This paper is intended to be a concise educational tool and guide for choosing and interpreting statistical tests aimed toward medical education assessment. It includes guidelines and examples that clinician-educators can use when analyzing and interpreting studies and when writing methods and results sections of reports.

KEY WORDS: medical education; educational research; statistics; faculty development.
DOI: 10.1111/j.1525-1497.2006.00390.x
J GEN INTERN MED 2006; 21:656–660.

As accreditation bodies seek confirmation of successful trainee achievement,[1,2] medical education must move toward a more evidence-based approach to teaching and evaluation.[3,4] To meet these challenges, educators must have knowledge and skills in developing, analyzing and disseminating educational interventions as part of their scholarly work. Effective development and evaluation require a fundamental knowledge of study design and statistical methods. Although medical training often provides physicians with a general background in epidemiology and biostatistics, many physicians are not prepared to apply these skills.[5–7]

While an effort has been made to help educators apply epidemiology to educational research,[8] we found no references that help educators understand how to use statistical tests to evaluate educational interventions. This paper is intended to be a concise educational tool and guide for choosing a statistical test during medical education assessment and for interpreting and analyzing educational studies without relying on mathematical theory. To provide a framework for understanding statistical concepts and to illustrate the decision-making process needed to choose a statistical test, we present an educational intervention detailing the hypothesis testing, data analysis, and interpretation of the results. Examples of statistical tests recently used in the educational literature are provided in Appendix 1, and statistical terms appearing in boldface are defined in Appendix 2.

BACKGROUND CONSIDERATIONS

Before determining which statistical test to use, one must consider study hypotheses, study design, number of study groups, whether groups are matched or paired for certain characteristics, type of outcome data, and how data are distributed in the sample. A checklist of questions addressing these areas is provided in Table 1. First, we present a sample educational intervention to illustrate the statistical concepts presented later in the text.

Hypothetical Example: Study Design and Methods

We developed a 1-month curriculum to improve second-year medical students' physical examination skills, interpersonal skills and confidence level. We conducted a randomized controlled trial in which half of the class received the new curriculum and the other half served as controls. We collected information regarding student age, gender, and college major. We evaluated all students' physical examination and interpersonal skills using a standardized patient exam 1 week after the intervention (Note: for simplicity, we will consider only one station of a standardized patient exam). We assessed the number of relevant physical examination maneuvers performed correctly by each student (total of 6 manuevers), a 20-item interpersonal score rated by the standardized patient, and whether the patient would recommend the student to a friend. Each interpersonal item was rated on 5-point Likert scale (1 = poor, 5 = excellent). We assessed each intervention student's confidence level in performing physical examination techniques before and after the curriculum using a 4-point Likert scale (1 = not very confident, 4 = very confident).

We used a Student's *t*-test to compare the mean number of physical examination maneuvers performed correctly and the Wilcoxon rank-sum test to compare overall interpersonal scores between groups. We used the Wilcoxon signed-rank test to compare intervention students' confidence level before and after the curriculum. To assess the relationship between student characteristics and the likelihood of being recommended to a friend, we performed simple logistic regression.

With a sample size of 60 students in each group, the study had 80% power to detect a difference of 1.2 maneuvers between the intervention and control groups in the mean number of relevant physical examination maneuvers performed correctly.

The authors have no conflicts of interest to report.

The approach to choosing and interpreting statistical tests described in this paper was presented at a workshop during the Society of General Internal Medicine's National Meeting in New Orleans, LA, May 11–14, 2005.

Address correspondence and requests for reprints to Dr. Windish: Yale Primary Care Residency Program, 64 Robbins Street, Waterbury, CT 06721 (e-mail: donna.windish@yale.edu).

Manuscript received September 22, 2005
Initial editorial decision November 2, 2005
Final acceptance December 12, 2005

Table 1. Questions to Consider When Selecting the Appropriate Statistical Test

1. What is the study design and study question?
   Are you interested in describing the data or testing a hypothesis?
   (a) Describing the data
   (i) Describe a group
   → *Exploratory Data Analysis (Descriptive Statistics)/ Summary Measures* (see Appendix 1a)
   (b) Testing a hypothesis
   (i) Compare 2 or more groups
   → *Inferential Statistics/Hypothesis Testing* (see Appendix 1b – e)
   How many groups are involved?
   Are they paired (matched) in some way?
   (ii)Quantify the association between variables/predict outcomes
   → *Inferential Statistics* (see Appendix 1f –g)
   (iii) Assess time to an event
   → *Inferential Statistics* (see Appendix 1g)
2. What type of outcome variable is being assessed?
   (a) Continuous
   (b) Dichotomous
   (c) Ordinal
   (d) Nominal
3. What type of distribution does the outcome variable have?
   (a) Normal or binomial
   → *Parametric test*
   (b) Skewed
   → *Nonparametric test*

Statistical Overview

Statistics is the scientific use of data to describe and draw inferences about true associations or phenomena by assessing the strength of the evidence for or against a hypothesis. It is used to make predictions and comparisons about a larger population based on data collected from a smaller sample. Since we usually cannot test an entire population (e.g., all second-year medical students), we must rely on sample data to guide our understanding of the truth. How well our sample represents the larger population determines how **generalizable** our findings are.

Data collected in any study are subject to variation. Some variation comes from random error and some from statistical error (measurement variation). **Bias** can be introduced in any stage of the study from its development to reporting of the results.[9] The goals of any study should include decreasing bias and minimizing error.

Variable Types

Studies generally have 2 variable types: the response variable (also called the outcome or **dependent variable**) and the explanatory variable (also called a **covariate** or **independent variable**). These variables can be quantitative or qualitative in nature. **Quantitative variables** are numerical and can be continuous or discrete. **Continuous variables** have no gaps in the values (e.g., age), whereas **discrete variables** have gaps (e.g., the number of study participants). **Qualitative variables** describe certain attributes and are either ordinal or nominal. **Ordinal variables** have an implicit ranking associated with them (e.g., Likert scales), whereas **nominal variables** are descriptive and cannot be ordered (e.g., college major). The types of dependent and independent variables used to make comparisons influence what statistical tests are needed.

Study Design

The appropriate use of statistics depends upon the research question(s) being asked. These questions and study hypotheses influence the study design and should be determined before conducting a study. Two types of study designs are commonly used in research: observational and experimental. **Observational studies** examine groups at one or more points in time (e.g., case-control, cross-sectional, and cohort studies). **Experimental studies**, or controlled trials, allocate participants to one or more groups and make comparisons across groups to assess differences in outcomes. Our study was a randomized controlled trial. Random allocation involves chance in the assignment of participants to intervention and control groups. This avoids a potential bias called selection bias that may be present if group assignment is known, as is often the case in observational studies. Selection bias can produce comparison groups that are different from each other from the study onset. This can limit the interpretation and generalizability of the study results.

The study design and the type of comparison group influences the statistical analyses performed. If the study uses a pre-post design, each participant is assessed by the same instrument at different points in time. The results obtained for each individual during different measurements are more likely to be highly correlated than the results of 2 randomly selected participants. Statistical analyses in this case should be performed using **paired methods** such that each participant serves as his/her own comparison. Our study requires the use of paired methods to assess differences in student confidence level before and after the intervention.

Exploratory Data Analysis (Descriptive Statistics)

The first step in any analysis is to explore the data collected to ensure that they are reasonable, accurate and not affected by measurement or recording errors. Exploratory data analysis, or **descriptive statistics**, is a method of organizing, summarizing and displaying data. It includes calculating measures of central tendency (e.g., **mean** and **median**) along with measures of dispersion (e.g., **standard deviation** and **interquartile range**). Graphically displaying the data in histograms, **stem-and-leaf plots** or **box-and-whisker plots** will also aid in assessing patterns of dispersion and can identify potential outlying values that may influence study results. Understanding the type of data collected and how it is dispersed helps determine which types of statistical analyses can be performed.

Confirmatory Data Analysis (Inferential Statistics)

Confirmatory data analysis, or **inferential statistics**, uses estimation and hypothesis testing to assess the strength of the evidence, make comparisons, make predictions and draw conclusions about a population based on the sample data. Types of inferential statistics include **bivariate analyses** that investigate relationships between 1 dependent and 1 independent variable, and **multivariable analyses** that investigate relationships between 1 dependent and multiple independent variables while controlling for the possible **confounding** influence of several independent variables on the dependent variable. In our example, we use bivariate analyses to compare differences in interpersonal scores between groups and multivariable

analyses to quantify the association of student characteristics with the interpersonal score.

The results of inferential statistics are reported according to the type of data collected and the statistical test or method used to determine the result (e.g., mean number of physical examination maneuvers performed correctly in each group using a Student's t-test). Results are also described by a level of **statistical significance** expressed as a **P-value** or estimated with a confidence interval (CI).

Hypothesis Testing

In hypothesis testing, the **null hypothesis** is a statement of no effect or no association. The null hypothesis regarding our main study goal would be: Participants and controls do not differ in the mean number of relevant physical examination maneuvers performed correctly at the end of the curriculum. The alternative hypothesis is that there is a difference.

Two types of errors can occur when making conclusions regarding the null hypothesis: **Type I error** and **Type II error**. A Type I error refers to rejecting the null hypothesis when the null hypothesis is true (false positive). A Type II error refers to accepting the null hypothesis when it is false (false negative). The goal is to minimize the probability of making a Type I error. Most studies set this probability, known as the significance level, at .05. In statistical tests, *P*-values are calculated as the probability of obtaining an outcome as extreme or more extreme than the observed study result under the assumption that the null hypothesis is true. If the *P*-value is less than the significance level, the result is considered statistically significant (e.g., $P<.05$). When statistical significance is not observed, either the null hypothesis is true (i.e., no difference really exists) or the sample size was not large enough to detect a difference (i.e., insufficient statistical **power**). The relationship between sample size, **effect size**, and statistical power is important to consider and is described elsewhere.[10,11]

Although *P*-values are used ubiquitously in the literature, they have several limitations. *P*-values do not indicate the strength or direction of the association, nor do they provide a direct interpretation of the results. For this reason, a 95% confidence interval (CI) associated with the result should be used when possible. A 95% CI indicates 95% certainty that the interval contains the true value. The true value refers to the outcome that we would expect if we could test the entire population. In our example, we wanted to determine whether there was a difference in the mean number of relevant physical examination maneuvers performed correctly between groups. The 95% CI for the true difference in mean scores was 0.85 to 1.7 suggesting that the true difference lies approximately in the range of 1 to 2 maneuvers. Studies with larger sample sizes and less variation will have narrower CIs indicating more precision in the results. Those with smaller sample sizes and higher variation will have larger CIs indicating less precision.

Before conducting a study, determination of statistical significance and clinical (practical) significance should be made. To do this, one needs to define the magnitude of detectable difference that would provide a meaningful change in outcome. In some studies, statistical significance may be reached due to large sample size, but the practical significance of the outcome may not be noteworthy. On the contrary, statistical significance may not be reached due to low sample size, but the outcome may be clinically relevant. In our example, we wished to see if the intervention improved the average number of physical exam maneuvers performed correctly by students. We needed to ascertain in advance, either from other research or practical experience, the increase in average number of exam maneuvers that would constitute a meaningful change in results, and establish a sample size that would allow statistical detection of this change.

Data Distribution

The distribution of data assessed during exploratory data analysis helps determine whether **parametric** or **nonparametric tests** should be used to make comparisons. Parametric tests are based upon the assumption that the data are sampled from a known population distribution (Note: we will consider only the **normal (bell-shaped) distribution** for continuous outcome data and the **binomial distribution** for dichotomous outcomes). If continuous outcome data in a sample are **skewed** toward either higher or lower values, or if the sample size is small, nonparametric tests should be used. Ordinal variables are usually analyzed using nonparametric tests; however, parametric tests can be used when values of separate variables are summed together to produce a total score which follows a normal distribution (e.g., summing each student's 20-item interpersonal ratings to obtain an overall score). Nonparametric tests use ranked observations rather than the actual values and do not assume that the shape of the distribution is known.[12] These tests are more conservative, but are important to use when parametric considerations do not hold.

SELECTING THE APPROPRIATE STATISTICAL TEST

We will use the steps outlined in Table 1 and the diagrams in Appendix 1 to illustrate how to select the appropriate statistical test for each of the 4 study hypotheses.

Hypothesis 1: Participants and controls do not differ in the mean number of relevant physical examination maneuvers performed correctly at the end of the curriculum.

1. *Study design and study question*: Randomized controlled trial comparing 2 unpaired groups (intervention and control students) (Appendix 1b).
2. *Outcome variable*: The number of relevant physical exam maneuvers performed correctly is handled as a continuous variable for analysis purposes.
3. *Distribution of the outcome variable*: The distribution of the number of physical exam maneuvers for each group plotted on a histogram appeared normally distributed, suggesting a parametric test should be used.
4. *Statistical test*: Student's *t*-test.
5. *Results*: The mean number (standard deviation) of relevant physical examination maneuvers performed correctly by the intervention group was 4.4 (1.1) compared with 3.1 (1.1) for the control group, $P<.0001$, 95% CI for the true difference in means (0.85 to 1.7).

6. *Interpretation*: Our *P*-value suggests a highly statistically significant difference, a difference that is unlikely due to chance alone, in mean number of physical examination maneuvers performed between groups. The 95% CI for the true difference in means also indicates a significant difference as it does not include the value of 0 (which would suggest that each group performed similarly). Thus, we reject the null hypothesis and conclude that the intervention students scored higher than the controls.

Hypothesis 2: Participants and controls do not differ in their overall interpersonal scores at the end of the curriculum.

1. *Study design and study question*: Randomized controlled trial comparing 2 unpaired groups (intervention and control students) (Appendix 1b).
2. *Outcome variable*: The overall interpersonal score is the sum of the 20-item interpersonal scores rated on a 5-point Likert scale. This score is continuous ranging from 20 to 100.
3. *Distribution of the outcome variable*: Although the outcome is continuous, the distribution of the scores plotted on a histogram appeared skewed toward higher values, suggesting a nonparametric test should be used and the median rather than the mean for the summary measure.
4. *Statistical test*: Wilcoxon rank-sum test.
5. *Results*: The median number (interquartile range, IQR) of the interpersonal score for the study students was 78 (IQR 66 to 94) compared with 73 (IQR 66 to 84) for the control students, $P=.07$. (The *P*-value in this case refers to the test of the difference in the distribution of ranked scores as assessed by the Wilcoxon rank-sum test and not the direct comparison of median scores. There is no analog of the 95% CI for this test).
6. *Interpretation*: The *P*-value is not statistically significant and the interquartile ranges overlap. Thus, we cannot reject the null hypothesis and conclude that our curriculum did not improve interpersonal skills.

Hypothesis 3: Participants' confidence level in performing physical examination maneuvers does not differ before and after the curriculum.

1. *Study design and study question*: Pre-post design comparing 1 paired group (intervention students before and after the curriculum) (Appendix 1c).
2. *Outcome variable*: The confidence level is measured on a 4-point Likert scale and is an ordinal variable.
3. *Distribution of the outcome variable*: The distribution of the confidence level plotted on a histogram is nonnormally distributed suggesting a nonparametric test should be used.
4. *Statistical test*: Wilcoxon signed-rank test.
5. *Results*: The median number (IQR) for the confidence level of students before the intervention was 2 (2 to 3) compared with 3.5 (3 to 4) after the course, $P<.0001$.
6. *Interpretation*: The *P*-value suggests a statistically significant difference between pre and postintervention ratings. The IQRs show minimal overlap between the two scores which also supports a statistically significant difference. Thus, we reject the null hypothesis and conclude that the intervention was successful at improving students' confidence.

Hypothesis 4: No association exists between a student's age, gender, and college major with the patient's recommendation of the student to a friend.

1. *Study design and study question*: Randomized controlled trial quantifying the association between 3 independent variables with the outcome variable (patient's recommendation) (Appendix 1g).
2. *Outcome variable*: The recommendation is dichotomous (yes or no).
3. *Distribution of the outcome variable*: The distribution of the outcome variable is binomial.
4. *Statistical test*: A simple logistic regression was used to test the hypothesis of no association between each individual covariate with recommendation. A more advanced analysis would extend this to a multiple logistic regression where potential **confounding variables** could be controlled for in the analysis.
5. *Results*: For each increase in 1 year of age, the odds are reduced by 1% that the student will be recommended to a friend (odds ratios [OR]=0.99; 95% CI, 0.83 to 1.19), $P=.99$. Compared with males, females have a 25% decrease in the odds of being recommended, (OR=0.75; 95% CI, 0.33 to 1.69), $P=.49$. Compared with science majors, nonscience majors have a 23% decrease in the odds of being recommended (OR=0.77; 95% CI, 0.28 to 2.11), $P=.61$.
6. *Interpretation*: For each of the hypotheses, there was no statistically significant association between the covariate and the outcome as observed by the large *P*-values and 95% CIs overlapping the value one. Thus, we cannot reject each null hypothesis of no association between each student characteristic and the likelihood of recommendation by the standardized patient. This may be due to insufficient statistical power in our study.

FINAL CONSIDERATIONS

This paper illustrates the decision-making processes clinician-educators can use to select statistical tests for interventions with 2-group comparisons. Examples of comparisons between 3 or more groups, correlations, and different regression analyses can be found in Appendix 1. Other tests or analyses may be needed depending on the research question of interest. Studies using observer ratings should be analyzed for **interrater** and/or **intrarater reliability** to assess consistency of results. When multiple comparisons will be performed, researchers may need to adjust the significance level to a smaller value (e.g., $P=.001$) to decrease the probability of finding a statistically significant result by chance alone. When performing **regression analyses**, certain assumptions must be checked to assess whether a specific regression model is appropriate and whether the potential for **confounding** and **effect modification** by certain covariates should be considered.[13]

With this guide, we hope to provide educators with a tool for improving the quality of medical education research conducted and presented in the literature. To obtain appropriate advice for both statistical design and analyses, we suggest the consultation of a statistician early in a study. Other resources such as textbooks and references for clinical research [10,11] may be needed to address areas not covered in this paper.

REFERENCES

1. **Liaison Committee on Medical Education.** Available at http://www.lcme.org. Accessed December 1, 2005.

2. **Accreditation Council for Graduate Medical Education Outcome Project.** General Competencies: minimum program requirements language. Available at http://www.acgme.org/outcome/comp/compMin.asp. Accessed December 1, 2005.
3. **Chen FM, Bauchner H, Burstin H.** A call for outcomes research in medical education. Acad Med. 2004;79:955–60.
4. **Dauphinee WD, Wood-Dauphinee S.** The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. Acad Med. 2004;79:925–30.
5. **Berwick DM, Fineberg HV, Weinsteing MC.** When doctors meet numbers. Am J Med. 1981;71:991–8.
6. **Wulff HR, Andersen B, Brandenhoff P, Guttler F.** What do doctors know about statistics? Stat Med. 1987;6:3–10.
7. **Lurie SJ.** Raising the passing grade for studies of medical education. JAMA. 2003;290:1210–2.
8. **Carney PA, Nierenberg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM.** Educational epidemiology. Applying population-based design and analytic approaches to study medical education. JAMA. 2004;292:1044–50.
9. **Hartman JM, Forsen JW, Wallace MS, Neely JG.** Tutorials in clinical research: part VI. Recognizing and controlling bias. Laryngoscope. 2002;112:23–31.
10. **Riegelman RK.** Studying a Study and Testing a Test. How to Read the Medical Evidence. 4. Philadelphia: Lippincott Williams & Wilkins; 2000.
11. **Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB.** Designing Clinical Research. 2. Philadelphia: Lippincott Williams & Wilkins; 2001.
12. **Siegel S, Castellan Jr. NJ.** Nonparametric Statistics for the Behavioral Sciences. 2. Boston: McGraw-Hill; 1988.
13. **Katz MH.** Multivariable analysis: a primer for readers of medical research. Ann Intern Med. 2003;138:644–50.