Selecting an Appropriate Statistical Test for Research Conducted in Engineering/Graphics Education: A Process

Alice Y. Scales, Julie H. Petlick
North Carolina State University

Abstract
Individuals in institutions of higher education who are involved in research on teaching engineering graphics, and other projects, are frequently confounded by the process of selecting the appropriate statistical test to analyze the data they collect. Research studies are usually only a portion of faculty member's work, and they generally only have taken one or two required courses in statistics during their graduate work. For these reasons, they either have to consult with a statistician or may select a test that is inappropriate for the research they are conducting.

This paper will layout a flow chart for selecting appropriate statistical tests based on the nature of the data being collected for a study. Examples used in the paper will use situations that might be encountered by individuals conducting research in teaching engineering/technical graphics.

Introduction
At the heart of statistics is answering questions with data. According to Finnney, statistics is "concerned with finding out about the real world by collecting, and then making sense of data" (p. 164).[1] However, statistical findings are only meaningful if the data are analyzed by using an appropriate statistic test. When the test is wrong, then the data analysis may be meaningless. Individuals in our field involved in research have frequently had a problem when it comes to the selection of a statistical test. There are several reasons for this difficulty.

Issues Underlying Test Selection
The first issue that creates problems in selecting an appropriate test is there is more to it than meets the eye. The form of the data, sample size, sample distribution, test power, and test robustness are all part of the equation of test selection. The second issue is that most researchers usually only have taken the prerequisite one or two required graduate statistics courses, which are designed to make them low-level practitioners of statistics, not experts in the field. These courses primarily teach the technical aspects of using the statistical formulas and often have developed a reputation for being difficult, mechanical, and boring. Where they fail, even in the eyes of many teachers of statistics, is in the larger area of research design. According to Wild, "The process of investigation as a whole should be the heart of any statistics program."[1] Most graduate level courses are about calculations rather than role of statistics in an investigation.[1,2] In the courses attempt to cover the technical side of statistics, a clear pattern of how and why

individual tests and models should be selected are lost. Therefore, Chance and others in the field advocate that statistics courses should be redesigned to teach statistical thinking.[3] Still another reason that statistics educators feel that their courses fail is they generally enroll students from a variety of curricula. Therefore, the examples and content is not specific to individual needs.

The other problem with statistical training received by the average researcher in our field is that most courses focus on classical or parametric statistics, which frequently are not appropriate for research conducted in educational settings.[4] Educational research is not as cut and dry as it is in science and technical fields. Cause and effect is much harder to examine because of the many facets of the human experience.[5] Using a parametric statistical test when it is not appropriate can be problematic for several reasons. "First, the analysis of the data may result in a rejection of the null hypothesis, not because the data indicate that the null hypothesis is false, but because the data indicate that one of the assumptions of the test is invalid. Hypothesis tests in general are sensitive detectors not only of false hypotheses but also of false assumptions in the model. The second danger is that sometimes the data indicate strongly that the null hypothesis is false, and neutralize each other in the test, so that the test reveals nothing and the null hypothesis is accepted" (p 85).[6] Parametric statistical tests are often very sensitive to violations of their assumptions, and the results of the analysis may be due to this violation rather than a true view of the patterns in the data.[7]

The predominance of the use of parametric statistical methods and statistic courses' concentration on teaching them began with the mathematicians, Gauss and Laplace, in the 1700s. Both of these individuals discovered the normal distribution of errors, although, it had been demonstrated by de Moivre some years earlier. Gauss and Laplace were influential men in their time and derived this asymptotic distribution from their work examining errors in astronomical observations. Gauss and Laplace's work was based largely on instinct, rather than mathematical proof. Their influence lead to the application of their work to other fields, which was not always appropriate. Other mathematicians later discovered that this distribution was a fairly good fit for other data. Quetelet, for instance, found that biological and anthropological phenomena also closely conformed to the Gaussian law of error. According to Bradley, Gauss' work on least squares and the seeming elegance and simplicity created by the assumption of a normal distribution, lead to an almost deification of the normal distribution. It was understood that exactly normal distributions were not possible because it was based on errors extending to infinity, but it was found that many populations often were quasi-normal, which, unfortunately, lead to a general belief in a universal quasi-normality for all populations.[8]

When the development of many of the classical parametric tests occurred in the early 20th century, they were based largely on work in agriculture and anthropology. It was found that derivations could be greatly simplified by, not only assuming a normal distribution of errors, but also setting the variance of the various distributions to be equal to each other. This second precondition, that a populations variance be homogenous, along with normalcy and the assumption that populations were identical when sampling from more than one, meant that only if you knew these conditions to be met did the test statistics provide accurate results. This reliance on parametric statistics was also due to the lack of alternative tests. Distribution-free and nonparametric statistical tests were not widely developed until the middle 1930s. Even then, they were looked down on for their simplicity. However, nonparametric and distribution-free tests do not require the same level of assumptions required by classical parametric tests, nor do they require the large N needed to produce a normal distribution of errors. Normal distributions approach normalcy only with large N.[8]

Failure to understand when specific tests are appropriate can lead to inferences from the statistical analyses that are not appropriate. For example, it is often the case that a parametric statistical procedure is chosen even when the data cannot meet the assumptions for the use of that test. This is due to several false ideas about the tests as well as ignorance of alternative tests. First, researchers often feel that parametric tests are more valid and are robust against the failure to meet specific assumptions, particularly normalcy. Robustness is defined as the ability of a statistical procedure to be used when specific violations are not met. Many statisticians feel that even a small violation of assumptions for a test is a problem. In particularly, statistical inferences become increasing less robust and less accurate as distributions depart more significantly from normalcy.[9]

Statistical Selection: A Process

When determining the appropriate statistical analysis, there are several questions one must answer. The first question concerns how many groups/treatments are being compared. The second question concerns whether a between groups or a within subjects (also called repeated measures) research design was utilized. If a between groups design is used, then one is comparing groups, and if a within subjects design is used, one is comparing treatments. For example, if pre and posttests are given to an introductory engineering class, then it is a within subjects design and performance on the two treatments (pre and post) are being compared. On the other hand, if the researcher is examining whether achievement in an engineering graphics course differs for students with and without prior CAD experience, you have a between groups design.

The last question a researcher must answer in determining the appropriate statistical analysis to use concerns the type of data collected. Data can be nominal, ordinal, interval, or ratio. This question is particularly important because it will determine whether a parametric or nonparametric statistic is appropriate. Parametric tests require the data to be either interval or ratio. Nonparametric tests can be used with nominal, ordinal, interval, or ratio data.[10]

Parametric statistics are the most widely used procedures and are well represented throughout the literature. Some examples of parametric tests commonly used in educational and psychological research include t tests and ANOVA's. For many of the parametric statistical tests there is an equivalent nonparametric test. For example, the parametric ANOVA test for between groups designs is analogous to the nonparametric Kruskal-Wallis test. The parametric t test used for a within subjects design is analogous to the nonparametric Wilcoxin matched-pairs signed-ranks test.[10]

If data meet all of the required assumptions, then parametric statistics are more powerful in most cases than nonparametric tests, but they also require more assumptions than their nonparametric counterparts. In order for a statistical test to be valid, the data must meet the assumptions of the test. Some common assumptions of parametric tests, are the data be normally distributed, and be interval or ratio. If the assumptions of a statistical test are not met, or if small sample sizes are used, nonparametric tests are more appropriate. One advantage of nonparametric tests is that they make no assumption about the data being normally distributed. In fact, nonparametric tests are often referred to as distribution free tests. In general, nonparametric tests tend to have fewer and milder assumptions that must be met in order for the test to be appropriate.[11] The following flowchart represents some of the most commonly used parametric statistical tests and their nonparametric counterparts (See Figure 1). When answering the three questions stated earlier, enter the flowchart on the left-hand side, under the parametric column. The first question concerns the number of comparisons being made, and answering that question will determine where on the left-hand side you would actually enter the flowchart. Answering the second question will indicate which line in the flow chart to consider. The answer to the third question will indicate whether or not you should remain in the parametric column or follow the line into the nonparametric column. Note: that for each parametric test listed, the nonparametric counterpart can be determined by moving to the nonparametric column and looking at the same line. Recall that one of the assumptions of parametric tests is that data are interval or ratio. The nonparametric tests can be used with data that are only of nominal or ordinal scale. Nonparametric tests are also appropriate for interval or ratio data when the data fails to meet the other assumptions required by parametric tests.

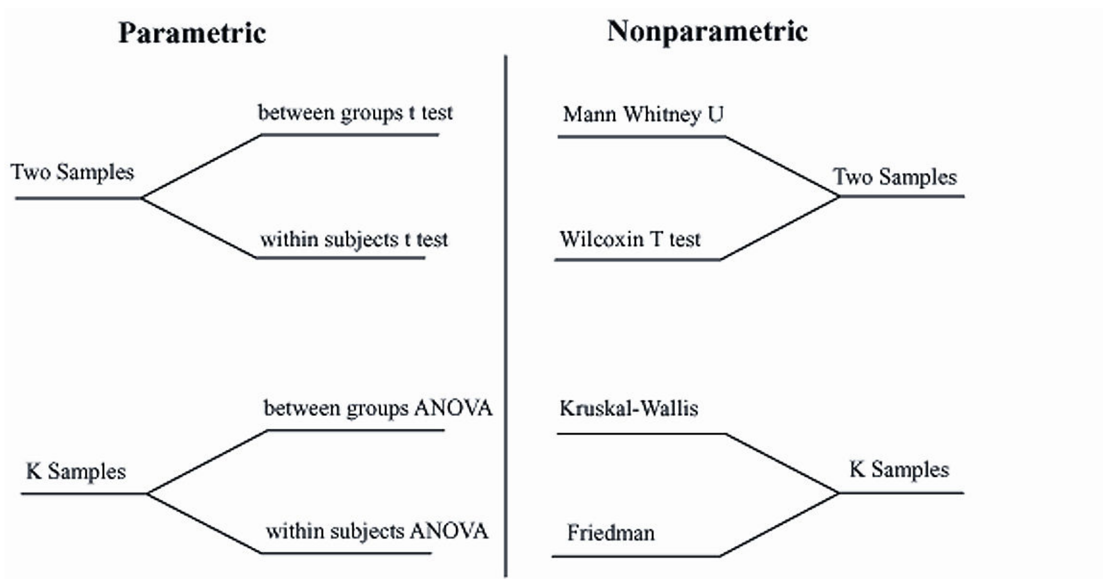

Figure 1: Flow Chart for Selecting Appropriate Statistical Tests.

For example, if one is comparing two groups/treatments, then a t test for between groups design, t test for within subjects design, Mann Whitney U test, or the Wilcoxin matched-pairs signed-ranks T test comprise the pool of tests from which the researcher might select. Answering the second question, whether the research design is between or within, will eliminate some of the alternatives. For between groups research designs, the test will be either a t test for between groups or a Mann Whitney U test. If the data are interval or ratio, then it is very likely that a t test for between groups will be the appropriate test, providing other assumptions of that test are met. Likewise, if the researcher is comparing three or more groups/treatments, then the available tests would be ANOVA for between groups, ANOVA for within subjects, Kruskal-Wallis test, or Friedman test.

One possible example from the engineering literature is one in which a survey concerning the importance of topics taught in an introduction to engineering course was given to instructors, professionals in industry, and to students. The respondents were asked to indicate for each topic the amount of time they felt should be spent on the topic for a standard three credit hour 15-week semester course (i.e. 45 total contact hours). For this example, answering the first question would result in an answer of "three comparisons are being made" (the response of instructors vs. industry vs. students). The answer to the second question would indicate that the design is between groups. The third question asks about the data type. For this example, respondents are giving estimates of time. Time is ratio scale data; therefore, the appropriate test would be an ANOVA for between groups.

Now suppose, instead of having the respondents indicate the amount of time that should be spent on topics, the researchers listed the course topics and asked the instructors, industry professionals and students to rank the topics in order of importance. The design is still a between groups design with three comparisons, however, rank data is ordinal, so the appropriate test would be the Kruskal-Wallis test.

Another example from research in engineering education concerns job satisfaction for individuals in the profession. A survey containing a seven point Likert type scale was given to assess job satisfaction for individuals in the profession for 10 or more years versus those in the profession fewer than 10 years. The design is one in which two comparisons are being made (individuals with 10+ years vs. fewer than 10 years in profession) and is a between groups design. Quite often in the literature, Likert scale data is treated as interval data, and, therefore, parametric statistics are used. However, those taught from a purist perspective have been taught that Likert scale data is ordinal and for that reason nonparametric tests are more appropriate. When a Likert Scale is used, although possible responses are numbered, the differences between the numbers do not represent a constant. For situations like this, utilizing classic parametric procedures are just not appropriate. For the current research example, the test selected should be a Mann Whitney U test.

Conclusions

Selecting an inappropriate analysis undermines the time and effort that go into doing rigorous research. Errors in test selection that leads to incorrect inferences weaken our knowledge base in the field. Furthermore, when new research is based on the inaccurate conclusion from previous work, it undermines the validity of the research process as a whole. The lack of understanding of statistical appropriateness is, in part, due to the over emphasis in statistics courses on learning how to perform statistical calculations rather than on learning how to determine the appropriateness of the statistical test. Confounding the issue further, is a lack of familiarity with nonparametric tests, which likely inhibits their use, even when they are the more appropriate choice. Another reason for the unfamiliarity with nonparametric and distribution-free tests is the historical over reliance on parametric analyses.

References

1. Wild, C.J. (May, 1994). Embracing the "wider view" of Statistics. *The American Statistician*, 48 (2), 163-171.

2. Snee, R.D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician*, 44, 116-121.

3. Chance, B.L. (2002). Components of statistical thinking and implications for instruction and assessment, *Journal of Statistics Education*, 10 (3).

4. Scales, A.Y. (October, 2003). *Non-parametric statistics analyses and their application to educational research studies.* Paper presented at the Engineering Design Graphics Division of the American Society for Engineering Education Midyear Meeting held in Scottsdale, AZ.

5. Kish, L. (March, 1978). Chance, statistics, and statisticians. *Journal of the American Statistical Association*, 73 (361), 1-6.

6. Conover, W. J. (1971). *Practical nonparametric statistics*. John Wiley & Sons, New York, NY.

7. Gibbons, J. D. (1993). *Nonparametric statistics: An introduction.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-090: Newbury Park, CA.

8. Bradley, J.V. (1968). *Distribution-free statistical tests*. Prentice-Hall: Englewood Cliffs, NJ.

9. Leech, N.L. and Onwuegbuzie, A.J. (November, 2002). *A call for greater use of nonparametric statistics*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Chattanooga, NT.

10. Ferguson, G.A. (1981). *Statistical analysis in Psychology and Education, 5rd edition*. McGraw-Hill: New York, NY.

11. Hollander, M. and Wolfe, D.A. (1973). *Nonparametric statistical methods*. John Wiley and Sons: New York, NY.

ALICE Y. SCALES, Ed.D.

Dr. Alice Y. Scales is an Assistant Professor at North Carolina State University, where she has taught since 1988. At NCSU, she serves as the Assistant Department Head of the Department of Mathematics, Science, and Technology Education and Coordinator of the Graphic Communications Program. She received her B.S. in Science Education in 1965, her M.Ed. in Industrial Arts Education in 1983, and her Ed.D. in Occupational Education in 2000.

JULIE H. PETLICK, Ph.D.

Is a Research Assistant with the Department of Mathematics, Science, and Technology Education at NC State University. She has a Ph. D. in Psychology with a focus in the area of learning and cognition. Her research interests include the role of technology in learning, and the use of technology to accommodate perceptual learning style preferences.