

Statistical Methodology Guidelines for the *Journal of The First Year Experience and Students in Transition*

*Dr. Brian Habing - Department of Statistics, University of South Carolina
August 4, 2003*

It is impossible to deal with all of the complexities that can arise in a statistical analysis in an article in advance. The basics of reporting the results can be found in the *APA Publication Manual, 5th edition*. This document discusses five of the most common reasons that a submission will be returned for revision, resubmission, or even rejection due to its statistical methodology review. These five reasons are:

1. Use of an Incorrect Statistical Method
2. Failure to Check Assumptions
3. Failure of Experimental Design
4. Reliability and Validity of Instruments
5. Multiple Comparisons Problems
6. Failure to Report Effect Sizes and/or Confidence Intervals

Many of these errors can usually be avoided simply by carefully consulting your favorite statistical text. For more complicated analyses, please have the results of your study verified by a statistical consultant or a colleague who has quantitative expertise. If you are not experienced in conducting quantitative research, it is recommended that you have your first consultation before you even gather the data.

Use of an Incorrect Statistical Method

Make sure that the statistical method used is appropriate. That is, make sure that the procedure is appropriate for your type of data and that it tests the appropriate hypotheses. Some common mistakes include:

- Using a two-sample t -test instead of a paired t -test when testing for improvement in a test/re-test situation.
- Using a chi-square or Fisher exact test instead of McNemar's test when testing for a change in proportions.
- Using a regression or ANOVA model instead of logistic regression or log-linear models when the responses fall into only a few possible categories instead of being continuous.

Failure to Check Assumptions

Every statistical procedure has certain assumptions that must be met in order for the resulting analysis to be valid (e.g. normality, constant variance). The plots used to check these assumptions should generally **not** be included in the paper but the fact that they were checked must be stated. Further, if the assumptions were not met corrective measures must be taken and described (e.g. transformations of variables, use of a different test). It might be that no statistical test will have its assumptions met (even approximately), in which case the argument must be based solely on descriptive statistics.

Failure of Experimental Design

A claim that a program results in some improvement (i.e., in GPA, retention) must be supported by appropriate statistical evidence. For example, if it is desired to meaningfully show that a program produces some effect, the program must be compared to something else (either another intervention or lack of intervention).

The preferred method of doing this is to perform an actual experiment where the students are randomly assigned to either be in the program or to not be in the program. If this is done, then any statistically significant differences observed between the two groups will be strong evidence that the program *caused* the difference. Of course, this does not guarantee that the observed result will be of interest. If the program involves the students taking a three-hour study habits/tutoring course tailored for their major instead of a three-hour course in an academic subject unrelated to their major then obviously the group receiving the intervention should be expected to have higher GPAs!

It is usually impossible to randomly assign students to a program/intervention or not. If this cannot be done, the next choice is to compare the students in the program to those outside the program (or perhaps from previous years if necessary) while taking into account variables on which they may differ (i.e., high school GPA, ACT/SAT score, major, course load, motivation) and the issue of self-selection into the program. Ideally, as many of these variables as possible would be included as covariates in the analysis, and those not included would be mentioned in the discussion sections as weakness of the study. At the least, summary statistics describing how the two groups compare on these variables need to be given. In any case, the results will not be able to definitively show the program *caused* the observed differences but will provide evidence linking the two and suggesting a relationship.

When attempting to show that a program/intervention is efficacious, if the students in the program/intervention are not compared to students in another program/intervention, or if possible differences between the two groups are not discussed, then the results of the study are scientifically meaningless and unacceptable for publication.

Reliability and Validity of Instruments

When a psychological instrument is used, a summary of the evidence as to why it should be reliable and valid must be given. In particular the paper must include:

- Basic references to the instrument and studies of its reliability and validity (preferably from peer reviewed sources).
- A brief summary of the evidence of the instrument's validity and reliability.
- A brief overview of the possible differences between the populations that were used in the validity and reliability studies and the population examined in the current study.
- A brief overview of any differences between the intended uses of the instrument in the validity and reliability studies and the intended use in the current study.

The definitive source for the basic definitions of reliability and validity is the current version of the AERA/APA/NCME *Standards for Educational and Psychological Testing*. Many textbooks for evaluation and measurement courses also provide useful information, in a possibly more readable format.

Multiple Comparisons Problems

Whenever several tests of hypotheses are performed at the same time, there is a risk of having an unacceptably large probability of Type I error (false positive). Say you are performing 10 *t*-tests to see how two groups compare on ten different variables of interest. If you use a $\alpha = 0.05$ level on each of the 10 tests, you could actually have up to a 50% chance of having at least one Type I error! Similar difficulties occur if you are trying to see which of the 10 groups have statistically significantly different means following an analysis of variance. There are a variety of ways to address this issue (e.g., MANOVA and Tukey's HSD) depending on the hypotheses being tested, and any appropriate statistical method/adjustment is acceptable.

One easy, powerful, and almost universally applicable method is the Holm procedure (a.k.a. step-down Bonferroni procedure). To use this procedure you simply need to have the p-values for the individual tests and then apply a simple adjustment. PROC MULTEST in SAS will take these p-values and return an adjusted p-value that is compared to your nominal, say 0.05, α level. Statistics texts such as Neter, Kutner, Nachtsheim, and Wasserman's *Applied Linear Statistical Methods, 4th edition* (pp. 739-742) or Glantz and Slinker's *Applied Regression & Analysis of Variance, 2nd edition* (pp. 313-316) also discuss how to determine what α -levels you would compare each of your original p-values to if your statistical package will not perform the adjustment for you.

Failure to Report Effect Sizes and /or Confidence Intervals

The American Psychological Association's (APA) Task Force on Statistical Inference made several recommendations for statistical methodology in APA journals (Wilkinson & APA Task Force on Statistical Inference, 1999). Two notable recommendations are that effect sizes and confidence intervals should always be reported for primary outcomes. These estimates should be reported so that readers may judge the magnitude of an effect and evaluate the practical significance of the results. Cohen (1988), Cumming and Finch (2005), and Thompson (1996) are three useful references for computing and interpreting effect sizes and confidence intervals. Many other references may also be found in the literature.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.