Common Statistical Analyses: A Primer

Joshua Schutts, The University of West Florida

"Stats" – the very word may cause anxiety and trepidation in the minds of many in the fraternity and sorority profession. To meet our responsibilities, it is important that we become research-literate and critique findings presented to us. In many "results" sections of manuscripts, significance testing is presented. While a myriad of techniques exist, several will be presented and discussed as a framework for increasing statistical self-efficacy within our discipline.

Testing for significance is important because the rigor involved in making those judgments provides us greater confidence that whatever results/difference/effect were not observed by chance. For us to make valid claims about group differences, program effects, and interventions, a framework must exist whereby we can draw a sample from the much larger population, test it, and then draw conclusions. Significance testing allows for just that. We must be diligent not to accept differences in scores (e.g. mean differences between fraternity/sorority students and unaffiliated students) at face value. Significance testing must be performed in order to make more general inferences about the entire population of students (i.e. all fraternity and sorority members and non-members across the country). It bears noting, however, that if the entire population were sampled, then a comparison of means directly would demonstrate difference without the need for significance testing. An example of this might be if you surveyed every fraternity/sorority member on a campus, all of them completed the questionnaire, and you only sought to make inferences about your campus, then significance testing would not be necessary.

Foundational to significance testing is the "*p*-value." This value is the probability of having found a value that size if there were no real difference or effect. In statistics, we assume that groups are not different until information arises that indicates the contrary. Small *p*-values correspond to small probabilities that the difference, effect, or relationship observed in the sample can be reasonably supported for the larger population. You might have seen something like this: $t(23) = 3.24$, $p = .01$. This is the correct way to report a *t*-test, which will be covered more in the subsequent paragraphs. The *p*-value in this case can be interpreted that there is a 1% chance the difference observed was by chance alone. Traditionally, we observe an *alpha* (α) level of 0.05 as the criteria for determining statistical significance. This is the threshold with which we are traditionally comfortable making conclusions about chance. At an *alpha* of .05, we accept a 5% chance that the difference observed was <u>not</u> real in the population. As a profession, we must also be conscious not to be entirely rigid with our *alpha*-level criteria. After all, we are not talking about amputations of limbs to save lives (in which case, we may want a more strict alpha level!). More commonly in educational research, *p*-values between .05 and .10 are traditionally considered "practically significant" and should always be reported with an effect size.

Before any inferences are drawn, a wise critic would begin with an exploration of the research methodology. The "gold standard" for research is experimental design, which involves random assignment to conditions and random sampling of students. Most often, we perform purposive or stratified sampling. In those cases, care should be given when drawing inferences because it might be safe to consider that certain 'types' of students participate in research and those types, by their very nature, are different than those students who would not. As a best practice, all sampling that is non-experimental should be discussed within the limitations section.

There are two types of analysis that directly involve testing for mean differences: a *t*-test and an analysis of variance (ANOVA). *T*-tests and ANOVAs are appropriate tests when the dependent variable of measure is captured with at minimum an interval scale of measurement. Interval data has roughly equivalent spacing between levels, and contains at least five possible values. Generally, five-point or higher Likert-type items (Strongly Disagree to Strongly Agree) can be treated as interval data. In the case of the *t*-test, the independent variable of study must be dichotomous, meaning it has only two levels. Examples of these could be fraternity/sorority, sex, "affiliated/unaffiliated," "treatment/non-treatment."

Similarly, ANOVA extends the analysis for usage with independent variables that are captured with more than two levels. Examples of these could be demographics such as classification, ethnicity, or religion. Both analyses permit for the study of multiple independent variables. In a program like SPSS, this is accomplished by dragging more variables into the appropriate box. A word of caution: *alpha* levels, as previously mentioned, are related to the "error" of measurement and considers that we will be accepting some level of potential error in our judgment (traditionally, $\alpha = .05$). When we perform multiple tests (one test per hypothesis), we are multi-dipping into the error pool. In these cases, a more conservative evaluation criteria (*alpha*) should be employed to determine if a difference is statistically significant. Bonferroni offers a correction to *alpha* that controls for what is known as "family wise error." This correction is calculated by dividing *alpha* by the number of hypotheses tested.

Multiple hypotheses could arise from several separate tests of dependent variables (i.e. the researcher wants to know if differences exist within ACT score, alcohol consumption, and term GPA based on classification), or from several independent variables tested against one dependent variable (i.e. the researcher wants to know if differences exist on ACT score given a student's socio-economic status, gender, and fraternity/sorority affiliation). In both cases, there are three hypotheses proposed, so the resulting Bonferroni correction to *alpha* would be .05/3, or .02. Therefore, determinations of statistical significance based on an *alpha* of .02 should be made.

In closing, this review can help a fraternity/sorority professional interpret common analytical techniques with greater confidence in our work. It is the intent of the article that readers will consider these analyses in their assessment and evaluation practices and implement them as they are able.