Numeracy
Advancing Education in Quantitative Literacy

Volume 6 | Issue 1 Article 7

1-2-2013

The Problem of Too Many Statistical Tests: Subgroup Analyses in a Study Comparing the Effectiveness of Online and Live Lectures

David M. Lane
Rice University, lane@rice.edu

Recommended Citation
Lane, David M. (2013) "The Problem of Too Many Statistical Tests: Subgroup Analyses in a Study Comparing the Effectiveness of Online and Live Lectures," *Numeracy*: Vol. 6: Iss. 1, Article 7.
DOI: http://dx.doi.org/10.5038/1936-4660.6.1.7
Available at: http://scholarcommons.usf.edu/numeracy/vol6/iss1/art7

The Problem of Too Many Statistical Tests: Subgroup Analyses in a Study Comparing the Effectiveness of Online and Live Lectures

Abstract
The more statistical analyses performed in the analysis of research data, the more likely it is that one or more of the conclusions will be in error. Multiple statistical analyses can occur when the sample contains several subgroups and the researchers perform separate analyses for each subgroup. For example, separate analyses may be done for different ethnic groups, different levels of education, and/or for both genders. Media reports of research frequently omit information on the number of subgroup analyses performed thus leaving the reader with insufficient information to assess the validity of the conclusions. This article discusses the problems with a media report on research that was analyzed by conducting many subgroup analyses. The article concludes that the quantitatively literate reader should be skeptical of articles that report subgroup analyses without reporting the number of analyses that were done.

Keywords
multiple comparisons, quantitative literacy, online learning, subgroup analyses

Cover Page Footnote
David Lane is an associate professor of Psychology, Statistics, and Management at Rice University. His primary research interests are human-computer interaction and educational technology. His emphasis within educational technology is statistics education, and he is the main developer of the public domain projects: "Rice Virtual Lab in Statistics", http://onlinestatbook.com/rvls.html, and "Online Statistics Education: An Interactive Multimedia Course of Study," http://onlinestatbook.com/index.html.

Introduction

Performing multiple statistical comparisons greatly increases the probability of a false conclusion. Indeed, Shaffer (2010) argued that one of the major reasons that apparent scientific findings fail to replicate is that researchers fail to control adequately for the effects of performing multiple statistical tests.

The problem of multiple statistical tests is often encountered when a researcher performs several subgroup analyses. Suppose a researcher were interested in whether people have better memories on Mondays than on Tuesdays and this researcher gives half of the sample a memory test on Monday and the other half a test on Tuesday. The researcher decides more data is better than less data and therefore records each subject's age, gender, income, city of birth, and many other demographic variables. An extensive analysis is done in which the researcher compares Monday memory with Tuesday memory for each of the many subgroups (young versus old, male versus female, old male versus old female, etc.). Even if there were no real difference between Monday memory and Tuesday memory for any subgroup, it is not unlikely that one of the many tests of this difference would, by chance, falsely indicate a real difference. As an extreme example, if 100 tests were conducted using the 0.05 level of significance, it is almost certain that at least one of these tests would show a significant difference.

A striking real example of this problem is provided by Austin, Mamdani, Juurlinka, and Hux (2006) who classified patients according to astrological sign and looked for differences in the incidences of various diseases. They found Leos had a significantly higher probability of gastrointestinal hemorrhage ($p = 0.0041$) whereas Sagittarians had a significantly higher probability of humerus fracture ($p = 0.0458$). Austin et al. concluded that these analyses illustrate how the testing of multiple hypotheses increases the likelihood of detecting implausible associations.

A quantitatively literate person should be aware of the problem of performing multiple tests and be able to spot failures of researchers to take the number of tests conducted into account. Unfortunately, the media often does not present enough information about research findings to allow the reader to assess this problem.

A Widely Cited Study Using Subgroup Analyses

The present article focuses on a study (Figlio, Rush, and Lin, 2010) that highlights the problems of subgroup analyses. This study compared the effectiveness of online and live lectures and was reported widely in the media including the *New York Times* (Lohr , 2010). According to the *New York Times* report:

"Certain groups did notably worse online. Hispanic students online fell nearly a full grade lower than Hispanic students that took the course in class ..."

There are several problems with the way the *New York Times* presented the results. First, the *New York Times* failed to report that, overall, there was no credible evidence that live lectures are better than online lectures. Evidence for a difference between online and live lectures was found only after examining the difference separately for a variety of subgroups. An astute quantitatively literate reader could perhaps notice that the article did not mention the overall difference and that the discussion centered on the subgroups. However, at least some readers would likely assume that the study provided evidence that in-class performance is better than online performance when all students are considered.

Second, the article did not report the number of subgroup analyses that were or could have been performed. The authors gathered data on and could potentially have grouped students by University GPA, SAT, ACT, High School GPA, Gender, Ethnicity (Black, Asian, White, Hispanic), and/or Mother's Education (5 levels). Failure to present the reader with information about these potential statistical comparisons prevents the reader from being able to properly assess the difference between live and online lectures for the Hispanic subgroup. Figlio et al.'s procedure of conducting subgroup analyses without adjusting for having performed multiple statistical tests is not good statistical practice because, as discussed previously, it increases the chance of an incorrect conclusion. The problem is magnified when, as is the case here, there is not a significant overall effect of the independent variable (in-class versus online teaching) or evidence that the effect of teaching method differs significantly for different subgroups.

Finally, although the *New York Times* article reports that the study employed a respectably large total sample size of 312, it fails to report the sample sizes for the subgroup analyses. A reader of the *New York Times* article has no way of knowing that the major finding reported in the article concerning Hispanic students was based on a comparison of only eight Hispanic students who viewed online lectures with 25 students who attended live lectures. A proper interpretation of the key finding that the Hispanic students who viewed the lectures online achieved much lower grades than the Hispanic students who viewed the lectures in class would take into consideration the large margin of error that necessarily accompanies the very small sample size used.

Concluding Remarks

In general, the reporting of subgroup analyses without mentioning the overall effect should serve as a red flag that many statistical tests were performed. Another potential red flag is that the reported analysis is only one of many plausible analyses. For example, if a correlation is reported between a personality measure and the activation of a particular brain region, the reader should suspect that data on other personality measures and other brain regions were collected.

DOI: http://dx.doi.org/10.5038/1936-4660.6.1.7

Because the reader can often only guess about the number of statistical tests performed, the journalist writing the article bears most of the responsibility for reporting these details. It is typical for journalists to interview an author and the interview provides an excellent opportunity to explore the number of statistical tests conducted and what, if any, adjustments were made. This should be standard procedure for journalists, and statistical literacy courses for journalists should cover this in depth.

Acknowledgment

Partial support for this work was provided by the National Science Foundation's Division of Undergraduate Education through grant DUE-0919818. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Austin, P. C., M.M. Mamdani, D. N. Juurlinka, and J. E, Hux. 2006. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of Clinical Epidemiology*, 59 (9): 964-969. http://dx.doi.org/10.1016/j.jclinepi.2006.01.012

Figlio, D. N., M. Rush, M., and L. Yin. 2010. Is it live or is it Internet? Experimental estimates of the effects of online instruction on student learning. NBER Working Paper No. 16089, the National Bureau of Economic Research. http://www.nber.org/papers/w16089

Lohr, S. 2010. Second thoughts on online education. *New York Times*. Accessed from http://bits.blogs.nytimes.com/2010/09/08/second-thoughts-on-online-education/ on July 25, 2012.

Shaffer, J. P. 2010. Multiplicities and false-positive rates in science: Overview. Paper presented at the 176th annual meeting of the American Association for the Advancement of Science February, February, 2010, San Diego, CA.