TEACHING NHST VS BAYESIAN INFERENCE IN POSTSECONDARY TECHNOLOGY PROGRAMS

John H. Mott[1] and Erin E. Bowen[2]
[1]Department of Aviation Technology, Purdue University, West Lafayette, IN, USA
[2]Department of Safety Science, Embry-Riddle Aeronautical University, Prescott, AZ, USA

While literature demonstrating the weaknesses inherent in null hypothesis significance testing (NHST) and Frequentist statistical analysis is extant, NHST is still the predominant statistical methodology employed for research in the social sciences. Although Bayesian inference as a means of statistical analysis has made inroads in the scholarly literature in some social science disciplines, the use of Bayesian data analysis in the area of technology is limited. This article examines the advantages and disadvantages of the introduction of Bayesian methods in postsecondary technology programs, and concludes that there are significant advantages to the teaching of such methods. The authors recommend a blended approach, whereby both techniques are taught and applied to practical problems.

INTRODUCTION

Although the Frequentist school of statistical thought led to the development of hypothesis testing in the early twentieth century by Fisher, Neyman, and Pearson and the accompanying methodology for statistical decision-making, the years that have passed since that development have given rise to objections to the foundation upon which the methodology is based (Levine, Weber, Hullett, Park, & Lindsey, 2008). Those objections, having become both more frequent and more strident, have led to the increasing popularity of Bayesian inference (Sohlberg & Andersson, 2005). While Bayesian methods have become somewhat more popular in the scientific literature in recent years, there is still significant resistance to their use as a means of making statistical decisions in the social sciences and, in particular, in the technology or applied engineering disciplines, where that use is infrequent. An anecdotal look at technology programs offered by higher education institutions indicates that courses where Bayesian inference is taught in either a theoretical or an applied context are limited in number. This causes one to pose the question as to why this is so. Could it be that Bayesian inference is not at least considered a valuable supplement to NHST? Or could there be other reasons? The current study is an effort to examine the weaknesses associated with null hypothesis testing, misinterpretations of NHST that are prevalent, and strengths that Bayesian inference can bring to statistical analysis. Furthermore, the study suggests a pedagogical method by which the two approaches can be combined in within the framework of postsecondary technology programs.

WEAKNESSES AND MISINTERPRETATIONS OF FREQUENTIST ANALYSES

The credibility of null hypothesis testing suffers from the problem that its results are dependent on the intentions of the researcher. Kruschke (2010b) provides an elegant example illustrating the problem. Suppose that a researcher desires to collect experimental data. Generally, researchers tend to begin the data collection process with an approximate idea of how many samples need to be collected, and there are, of course, statistical methods by which that approximation can be determined, given some knowledge of the nature of the population. Suppose, however, that the data collection process becomes governed, for whatever reason, by time constraints. In such cases, the collection of data will occur over a fixed period. It may be such that the number of samples collected in each case is precisely the same. The inherent problem, however, is that the subtle change to collection of data over a fixed period of time leads to a change in assumptions regarding the underlying sample distribution, since, for relatively small sample sizes, the data collected in a fixed period is governed by a Poisson distribution with a mean of λ, the rate parameter. Assuming a simple t-test for a two-level independent variable, it can be shown that, for a given sample size, the *t*-statistic differs between the fixed sample size case and the fixed time period case, even when the same number of samples is collected. The difference in the t statistic could easily lead to a difference in the decision to reject or not reject the null hypothesis as a result. If the person handling the data collection portion of the experiment is distinct from the primary

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

researcher and those two persons differ in their understanding of whether the data collection cutoff point was based on sample size or time (regardless that the sample size ends up being the same in both cases), one could easily conceive of a situation where, given the same data, one rejects the null hypothesis while the other fails to do so. Hence, it is clear that the intentions of the experimenter play a considerable role in the decision outcome for the data set.

Another problem with null hypothesis significance testing lies in the specification of the population mean. It would be attractive to be able to state that, given the sample data obtained in a particular experiment, the population mean lies, with a certainty of $(1 - \alpha)$, within a definite interval centered around the point estimate of that mean. Unfortunately, Frequentist approaches to interval estimation do not provide the researcher with that ability, and that fact is occasionally misunderstood by both students and instructors in the classroom. Rather, the confidence interval differs from sample to sample, and contains the parameter that is being estimated with probability $(1 - \alpha)$ only on an asymptotic basis, i.e., if the experiment is repeated *an infinite number of times*.

The Bayesian analogy to parameter estimation provides the researcher with a *credible interval*, the interval in which the parameter of interest lies with a specified certainty. Because Bayesian analysis allows the computation of a *posterior distribution*, the distribution of the parameter being estimated, calculating the credible interval is a relatively trivial task. If, for example, we are interested in a 95% credible interval, we simply select the interval on the posterior distribution for which 95% of the probability density of the distribution is exceeded, a selection that can be made easily using software. The fact that the posterior distribution is available to the Bayesian researcher implies that a wide and rich variety of post-hoc testing can be conducted. Relative frequency analyses, as a result of their limited availability of post-hoc tests, therefore suffer from derogation at the hand of Bayesian approaches.

A third difficulty with null hypothesis testing (Kruschke, 2010a), and it is, indeed, one that is problematic for Frequentist statisticians to resolve, is that the *p* value that is generally the result of the statistical test employed is sometimes misunderstood to be the probability that the null hypothesis is "true." In actuality, the *p* value is a conditional probability that is proportional to $P(D|H_0)$; i.e., the probability that the data occurs, given that the null hypothesis is true. It should be readily apparent that this is, in fact, not the quantity in which we are interested as researchers. Rather, the quantity of interest is $P(H_0|D)$, the probability that the null hypothesis is true, given that the data occurs. The lack of equivalence between $P(D|H_0)$ and $P(H_0|D)$ is a concept that is generally clear to students of basic probability theory, and is a direct result of, not-so-coincidentally, Bayes' Theorem. Therefore, the results from many Frequentist statistical tests are often not those which the researcher is actually seeking.

A rationalization that is commonly used in defense of the *p* value is that the two probabilities described above are close to equal, or are highly correlated. However, Monte Carlo simulations conducted by Trafimow and Rice (2009) indicate that the correlation between $P(D|H_0)$ and $P(H_0|D)$ is a relatively low .396. Even more disconcerting is the realization that $P(D|H_0)$ accounts only for less than 16% of the variance in $P(H_0|D)$, leaving more than 84% of the variance unaccounted for. It is apparent, then, that much of the presumed usefulness of the *p* value is, at best, misconstrued.

While students of probability theory may develop an understanding of the lack of equivalence between $P(D|H_0)$ and $P(H_0|D)$ in discussions of Bayes' Theorem and the underlying principles of probability, the subtlety of distinguishing between the two in the setting of post-secondary technology or applied engineering programs is regularly lost on them. One of the common features of statistical education in these settings is the focus on "real world" data analysis and applicability of interpretation. The focus on NHST as a 'tool' to be applied rather than a framework and methodology to be thoroughly understood in its varied implications often leads students using Frequentist analyses to make logical leaps unsupported by the limitations inherent in $P(D|H_0)$. The authors have noted that, even when Bayes' theorem is taught in the post-secondary technology classroom, it is limited in scope and isolated from the variety of Frequentist analytic methods presented later in the course. This separation creates a situation in which students master the basics of Bayes' Theorem but not the applicability of Bayesian methodology to settings in which they face themselves.

BAYESIAN INFERENTIAL APPROACHES TO ADDRESS NHST WEAKNESSES

It should be readily apparent from the preceding analysis that NHST has a number of inherent difficulties that are not present in Bayesian approaches. Bayesian methods, for example, do not depend on the intentions of the experimenter with regard to the conclusions drawn, nor are they subject to issues with misinterpretation and lack of utility of confidence intervals. Moreover, the posterior distribution available to the Bayesian analyst readily provides $P(H_0| D)$, the quantity in which the analyst is most interested, as opposed to the more ambiguous quantity $P(D|H_0)$.

Supporters of Frequentist methods may point out that Bayesian methods are not suitable for elementary courses and are themselves subjective in their reliance on prior distributions, but Berry (1997) counters the suitability argument in describing his implementation of those methods in an elementary statistics course. In addition, Fisher and Wolfe (2012) discuss a novel method of using spreadsheets to assist with students' comprehension of conditional probabilities. Berry also notes that science itself is subjective, and that the use of priors simply mirrors the effects of our own subjectivity in terms of the scientific experience we possess within our respective fields of endeavor.

Concerns over the use of appropriate prior distributions in the Bayesian equivalent of null hypothesis testing have been discounted by Lee and Wagenmakers (2005), who point out that the use of uninformative priors that are transformationally-invariant is a recent development in Bayesian data analysis that resolves such concerns.

BLENDED APPROACHES TO STATISTICAL ANALYSIS

Rodgers (2010) argues that a blended approach to the use of statistical analysis in social science and related fields (including technology programmatic education) has been quietly growing within certain circles of educators and practitioners amid a slow decline in the nearly exclusive emphasis on NHST in these arenas. Howard, Maxwell, and Fleming (2000) suggest that this over-reliance on NHST methods in the psychological disciplines has existed for a 60-year period, but that recent advances in analytical techniques have led to its decline.

Rodgers (2010) also suggests that statistical modeling techniques (structural equation modeling, multi-level modeling, etc.) provide a more comprehensive view of the nature of phenomena under investigation than is possible with Frequentist approaches; in addition, modeling approaches build upon the growing interest in technology and organizational theory regarding the nature of complex systems. A unique characteristic of a modeling-based rather than a Frequentist-based approach to post-secondary technology education programs is that the models have greater ability to incorporate multiple perspectives to statistical education. Subbian, Srinivasan, and Shanthi (2011) demonstrate how a Bayesian modeling approach to education data analysis can be successfully implemented in a higher education environment.

The primary author incorporated a blended approach in the teaching of a graduate-level course on Bayesian data analysis as applied in the aviation discipline in 2013. In a number of topical areas within that course, NHST methods were presented and were followed up with analogous Bayesian methods. For example, analysis of variance was explained and demonstrated using a simple example of Fisher's *F*-Test. Assumptions implicit in that test, i.e., homogeneity of variance and normality, were then discussed. The discussion was followed with a presentation of the Bayesian analog, in which it was demonstrated that the hierarchical model used could be modified to adapt the algorithm to take into account cases in which both non-normal and heteroscedastistic data were present. Students were then presented with homework designed to require them to apply the modifications to a set of aviation-related data. Anecdotal feedback indicated that this approach was successful in making the point that the Bayesian method provides both a fresh perspective of the problem and a set of tools that should prove useful for future application.

Ideally, statistics education in the post-secondary setting must provide future scholars and practitioners with a more comprehensive set of tools than is currently offered. Bayesian analysis should be integrated into the teaching of NHST rather than presented separately; and educators would be well served to teach NHST as only a basis for the development of statistical models. In particular, Albert (1995) suggests the teaching of subjective and conditional probability and then introducing the discrete Bayes approach to proportions to introduce basic concepts of inference.

A substantive weakness of much teaching in the post-secondary technology setting is the failure to teach students of the limits of NHST and to ensure their mastery of any but the most basic analytic techniques, including the limits of statistical power in Frequentist analyses (Ison, 2011). The presence and accessibility of sophisticated modeling and Bayesian analysis computer programs on most institutional campuses means that faculty should be able to devote more of their time to the nature of more complex analyses rather than a focus on the mechanics of hand-calculations. For example, Allenby and Rossi (2008) discuss how a Bayes package on an R platform was successfully utilized to solve nontrivial problems in a business statistics course; Lecoutre (2006) and Lecoutre, Lecoutre, and Grouin (2001) explain how analyses of variance can be taught in a Bayesian context using experimental data. However, the teaching of a blended method that synthesizes these varied approaches will only be as successful as the quality of the instructor – which means there must be significantly more devotion to the continuing education of statistics instructors in these programs.

REFERENCES

Albert, J. (1995). Teaching inference about proportions using Bayes and discrete models. *Journal of Statistics Education, 3*(3).

Allenby, G. M., & Rossi, P. E. (2008). Teaching Bayesian statistics to marketing and business students. *The American Statistician, 62*(3), 195-198. doi: 10.1198/000313008X330801

Berry, D. A. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician, 51*(3), 241-246. doi: 10.1080/00031305.1997.10473970

Fisher, C. R., & Wolfe, C. R. (2012). Teaching Bayesian parameter estimation, Bayesian model comparison and null hypothesis significance testing using spreadsheets. *Spreadsheets in Education, 5*(3). Retrieved from http://epublications.bond.edu.au/ejsie/vol5/iss3/3/

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *American Psychological Association, 5*(3), 315-332. doi: 10.1037/1082-989X.5.3.315

Ison, D. (2011). An analysis of statistical power in aviation research. *International Journal of Applied Aviation Studies, 11*(1), 67-84.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293-300.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(5), 658-676. doi: 10.1002/wcs.72

Lecoutre, B. (2006). Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science, 4*, 207-232.

Lecoutre, B., Lecoutre, M.-P., & Grouin, J.-M. (2001). A challenge for statistical instructors: Teaching Bayesian inference without discarding the "official" significance tests. *Bayesian methods with applications to science, policy and official statistics* (pp. 301-310). Luxembourg: Office for Official Publications of the European Communities.

Lee, M. D., & Wagenmakers, E. (2005). Bayesian statistical inference in psychology: Comment on Trafimow. *American Psychological Association, 112*(3), 662-668. doi: 10.1037/0033-295X.112.3.662

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*, 171–187.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *The American Psychologist, 65*(1), 1-12.

Sohlberg, S., & Andersson, G. (2005). Extracting a maximum of useful information from statistical research data. *Scandinavian Journal of Psychology, 46*(1), 69-77. doi: 10.1111/j.1467-9450.2005.00436.x

Subbiah, M., Srinivasan, M. R., & Shanthi, S. (2011). Revisiting higher education data analysis: A Bayesian perspective. *Int'l Journal of Science & Technology Education Research, 1*(2), 32-38.

Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *Journal of General Psychology, 136*, 261-269.