Published in final edited form as:
J Am Stat Assoc. 2017 ; 112(519): 902–904. doi:10.1080/01621459.2017.1311265.

Statistical Significance and the Dichotomization of Evidence: The Relevance of the *ASA Statement on Statistical Significance and p-values* for Statisticians

Eric B. Laber[*] and
Department of Statistics, NC State University

Kerby Shedden
Department of Statistics, University of Michigan

1 Introduction

Empirical efforts to document the practices and thought processes of data analysis are a promising way to understanding the real-world impact of statistical methodology. The empirical findings of McShane and Gal provide new insights into long-standing debates about dichotomization and NHST. We appreciate having the opportunity to comment on this important work. Our discussion is divided into two sections. First, we comment narrowly on the new empirical findings. Second, we discuss in broader terms the interpretations drawn from these studies, and present some of our own points of view about *p*-values, dichotomization, and NHST in applied research.

2 The McShane and Gal empirical studies

The principle argument of studies 1 and 2 is that statistical researchers endorse incorrect statistical statements about an applied study's findings, and more critically, that these incorrect assertions disproportionately arise when the evidence level crosses the $p = 0.05$ threshold. We especially appreciate the uniqueness and novelty of the latter finding. We do however have some reservations about the context in which these findings were elicited, and would be interested to see if they persist elsewhere.

A notable aspect of study 1 is that the correct line of reasoning in all cases is to simply state the numerical characteristics of a sample, avoiding any consideration of whether the results generalize to a population. While a literal reading of the study questions supports this approach, commenting only on the sample runs against our usual practice, especially when assessing a randomized trial. This presents each participant with a dilemma – can the question truly mean what it seems to say, given that there is almost no conceivable setting in which the stated fact (that the numerical means of the survival times differ between the groups) would provide a justified basis for action?

[*] The authors gratefully acknowledge funding from the National Science Foundation (DMS-1555141, DMS-1557733, DMS-1513579, DMS-1317631) and the National Institutes of Health (P01 CA142538).

Challenges in communication are particularly prominent when we aim to maintain a sharp distinction between the sample and the population. Many applied researchers choose to speak primarily about their data, but usually recognize that there is a difference between their data and “the truth.” Statisticians have extensive experience with the rhetorical means of maintaining a distinction between the sample and population. Effective collaboration may require us to take a somewhat figurative view of the language employed by applied researchers. Even expressions such as “speaking only of the subjects in this study,” which reads as an unmistakable cue once we are aware of the study’s intentions, can have a different impact if read slightly less literally. Acknowledging that McShane and Gal asked the participants for their own interpretation of the hypothetical research findings, the reality is that we are constantly engaged in interactions with applied researchers, and these interactions impact our communication. Thinking of our many discussions with applied researchers, “speaking only of the subjects in this study” could be taken to indicate that the *inferences* are to be made based only on the data observed in this study, not, for example, utilizing other studies of the same treatment.

Study 2 raises different, more subtle issues. As written, question 1 appears to ask about the values of unknown population parameters, whereas the intended focus may have been the direction of evidence in the observed data. Under the former interpretation, the correct answer is arguably ’d’ (cannot be determined). The authors reason that because respondents were more likely to select option ’d’ if the p-value was above 0.05, respondents were interpreting the question as intended, and their answers are evidence of dichotomous thinking. An alternative explanation is that respondents first noted that the question, as stated, could not be answered using the observed data and therefore opted to ‘read between the lines’ and answer the question they felt the investigators meant to ask. Statisticians often have the role of properly qualifying study findings that are imprecisely reported, e.g., by deeming a result statistically significant or not. It seems conceivable that some respondents may have opted to answer a question about the statistical significance of the findings rather than the direction of evidence.

While we appreciate the importance of using terminology appropriately, and share to some extent the authors’ dismay at the number of research statisticians who “failed” these tests, it is not clear to us that statistics as a discipline will have its greatest possible positive impact if we default to taking the most literal interpretation of statements made by researchers about their data, especially when such statements are at odds with the best practices of our field. Fortunately, in the real world we are rarely placed in a position where our response is limited to a list of pre-defined choices. By engaging researchers in a discussion about their data and the scientific context of their research, misstatements relating to uncertainty and evidence can usually be avoided.

3 Varying points of view about NHST and dichotomization of evidence

3.1 Holistic interpretation of evidence

A consistent theme in McShane and Gal’s article is that when interpreting statistical evidence, one should take “a more holistic and integrative view,” suggesting that additional factors to consider should include “prior and related evidence,” the “type of problem being evaluated," the "quality of the data," the model specification, the effect size, and real world costs and benefits. However even when taking an integrated view there will be pressure to make a binary decision to accept or reject a study's claims, and inevitably there will be near hits and near misses. Thus, while holistic interpretation is a laudable practice, we view this as a largely distinct issue from the problem of dichotomization.

Practices vary among scientific disciplines, here we speak of our own experiences. In years of working with life science researchers, especially around investigations of the molecular mechanisms of human diseases, we have noted that most scientists are intensely concerned about many of the evidentiary factors cited above. The relationship between novel and prior work, formulation of the hypotheses, and especially issues of data quality are all of great concern to many scientists. While statisticians are usually the first voice in the room when discussing statistical uncertainty, unfortunately we are too often excluded (or self-exclude) from discussions of other important aspects of data analysis, perhaps because these discussions tend to involve more specialized aspects of the subject area.

We have the following concern: if we raise doubts about the value of our narrow contributions relating to formalized statistical inference, and at the same time fail to engage seriously in other aspects of research, statisticians will lose a great deal of hard-won standing. It is appropriate to flag misuse of NHST, and sometimes to counsel against inappropriate dichotomous thinking. At the same time, we need to intensively seek out other contributions we can make to the practice of data-driven research, and to train the next generation of statistical researchers to think beyond stylized hypothesis testing.

3.2 Continuous evidence and actionability

There has been a great deal of discussion over the years about deficiencies of various statistical frameworks, but we do no believe any existing framework performs so flawlessly that it automates the process of reasoning with uncertainty. The specific emphasis of McShane's and Gal's argument is the notion of "dichotomous thinking," specifically using p-values and arbitrary thresholds to make binary decisions. McShane and Gal encourage us to think continuously about evidence. We are strongly in agreement to the extent that evidence does, nearly always, arrive to us in a continuous form. Many features of data or of a model are not of decisive importance, and can be presented simply as an estimate with standard error or other uncertainty measure. Nevertheless, decisions do arise that cannot be made continuously. When making a binary decision, there will inevitably be near-hits and near misses. This arbitrariness is inevitable in any setting where discrete actions must be taken.

One often-proposed way to resolve this difficulty is to work from a cost-based perspective, in which the decision, while (often) binary, is based on both the weight of evidence, and the costs of the two types of errors that can be committed. Ultimately this is still a binary decision, albeit using a threshold that is adapted to the context of the problem. Again there will be near hits and near misses, even when costs are taken into account. Furthermore, the "crisis of reproducibility" in science has most often been discussed in the context of basic research. In that setting, what is the cost of presenting a result that is not true, or that is only true in limited and difficult-to-replicate settings?

3.3 Adaptability of p-values and NHST

It is notable how novel and sophisticated developments and extensions of the NHST framework continue to regularly arise. For example, the rapidly growing toolbox of false discovery rate (FDR) methods has in our view been very successful at addressing concerns about multiple hypothesis testing and inference for exploratory analysis, in spite of being built on the binary notion of discoveries being either “false” or “true.” Along these lines, the work of Efron [2] and others on empirical null distributions, and the recent “knockoff” approach of Barber and Candès [3] are elegant and powerful approaches to inference that rest on the NHST framework.

In our view, most failures of statistical inference result from poor understanding of the sources of variation in the systems being studied, not from generic failures of inferential tools. Insights from domain-specific research including “cryptic relatedness” [8] and other forms of population structure in genetics, subtle placebo effects in clinical research [6], batch effects in genomic research [7], and false positives deriving from complex spatial noise in brain imaging [4] have provided us with a mechanistic basis for understanding previous inference failures. These insights do not mainly posit a failure of methods or of practitioners, but rather advance novel and fundamental mechanisms of variation that clarify the basis for past failures of NHST. Arguably, each of these mechanistic factors would affect “dichotomous” or “continuous” reasoning in statistical inference to similar degrees.

The most salient critique of NHST, in our view, is that rejecting a “straw man” null hypothesis resting on a simplistic model does not provide much evidence in favor of any particular alternative. But when the default “null model” is a rich and complex model, fit using efficient methods to large and carefully-modeled data, a NHST targeted to the effects of interest can become quite compelling. As a case in point, in linguistics, there has been much discussion lately about specification of mixed effects models, with one community suggesting to take the “maximal random effects structure” [5], meaning that every plausible random effect should be included. This nearly saturates the correlation model, with the view being that any parameter contrasts that appear strong against this correlational backdrop stand a good chance of being real.

There is good reason to be optimistic about the future of statistical inference as a relevant tool for discovery in science, including frequentist and NHST-based inference. As has been intensively discussed elsewhere, we are likely to be increasingly working with extensive volumes of fine-scale data on the systems we study. It has also been noted that “big data needs big models” [1]. These big models, including models derived from machine learning methods, as well as flexible procedures deriving from classical statistics such as semiparametric, empirical likelihood, dimension reduction, and localized methods, can be powerful tools for improving the properties of NHST. Recent work on high dimensional inference is providing new tools to build such models while not saturating the models to the point where parameter estimates become meaningless. However most applied researchers and many statisticians are not using these new tools to their full potential. The findings of McShane and Gal make clear that in terms of communication, training, and methods development, there is still a lot of room to grow.

References

1. Gelman, A. Big data needs big model. 2014. http://andrewgelman.com/2014/05/22/big-data-needs-big-model
2. Efron B. Large scale simultaneous hypothesis testing: the choice of a null hypothesis. JASA. 2004; 99(465)
3. Candès E, Barber R. Controlling the false discovery rate via knockoffs. Ann Statist. 2015; 43(5): 2055–2085.
4. Knutsson H, Eklund A, Nichols TE. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. PNAS. 2016; 113(28)
5. Barr DJ, et al. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language. 2013
6. Howick J, et al. Are treatments more effective than placebos? a systematic review and meta-analysis. PLoS One. 2013; 11(1)
7. Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010; 11:733–739.
8. Pritchard JK, Voight BF. Confounding from cryptic relatedness in case-control association studies. PLoS Genetics. 2005; 1