Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 48

11-1-2016

Reflections Concerning Recent Ban on NHST and Confidence Intervals

Grayson L. Baird
Lifespan Hospital System, grayson_baird@brown.edu

Sunny R. Duerr
SUNY New Paltz

Recommended Citation

Baird, Grayson L. and Duerr, Sunny R. (2016) "Reflections Concerning Recent Ban on NHST and Confidence Intervals," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 2 , Article 48.

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 821-824.
doi: 10.22237/jmasm/1478004360

ISSN 1538 − 9472

Reflections Concerning Recent Ban on NHST and Confidence Intervals

Grayson L. Baird
Lifespan Hospital System
Providence, RI

Sunny R. Duerr
SUNY New Paltz
New Paltz, NY

This letter addresses some of the immediate consequences of *Basic and Applied Social Psychology*'s (*BASP*) ban on null hypothesis significance testing (NHST) and confidence intervals. The letter concludes with three suggestions to improve research in general.

Keywords: NHST, NHSTP, BASP, basic and applied social psychology, ban on NHST, confidence intervals, null hypothesis significance testing

The editorial board of *Basic and Applied Social Psychology* (*BASP*) made a bold and unequivocal move by outright banning the use of Null Hypothesis Significance Testing (NHST) and confidence intervals, along with giving Bayesian methods at best conditional consideration (see Trafimow & Marks, 2015). *BASP*'s reasoning behind said ban is based on common concerns of Frequentist statistics in particular, though concerns of Bayesian statistics were also considered. The reasons for said ban are not of interest here, but rather *BASP*'s particular solution. They stated:

> …*BASP* will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem. (Trafimow & Marks, 2015, p. 1)

Although *BASP*'s intentions to improve the quality of research are commendable, what is not immediately evident is how the use of strong descriptive

Grayson L. Baird is a Research Statistician with the Lifespan Biostatistics Core. Email him at: grayson_baird@brown.edu.

statistics and larger sample sizes constitute a framework by which inference on a population may be made. By relying on descriptive statistics alone, *BASP* removed the notion of probability from their statistical methodology, save for the occasionally sanctioned Bayesian analysis. As a consequence, the scope of *BASP*'s scientific inquiry is therefore limited to the description of samples rather than inference to populations. The danger here is this limitation will not stop some readers from making inferences to populations, but will instead only remove the theoretical basis for doing so–thus blurring the distinction between interpretations of inferential and descriptive statistics.

What is especially curious about *BASP*'s aforementioned stance is their notion of sample size and its curative effects over the stability of descriptive statistics and the size of the sampling error (which are the foundational elements of the confidence interval, simply removing probability). Though descriptive statistics can become more stable and sampling error can decrease as sample size increases, this is only true in part.

The point can be illustrated by use of M&M's©. Assume there is a single 42oz "party-size bag" of M&M's and 20 bags of the regular 1.69oz store-size bags (totaling only 33.8oz), randomly sampled from different stores. Which would produce the better estimate of the color proportions from the factory machine settings: the proportions of the 42oz bag or the mean of the proportions of the 20 1.69oz bags? Granted, there is probably only a small difference between the two estimates, by design due to quality control. But if one machine goes out of control and fills a bag with too many green M&M's, the 42oz bag will be both large and largely biased, while the 1.69oz bag will be just one sample out of many.

The issue is clear: sample size alone cannot ensure better estimates of a population, the sampling methods by which a sample is procured are of the upmost importance. To quote former American Statistical Association president Peter Lachenbruch "A large *n* means nothing if the sampling is biased" (Cochran, 2015, p. 17). The nicety of the M&M's example is presumably the factory settings (population) are known and we could randomly sample the 1.69oz bags of M&M's, if we so desired. Unfortunately, this is difficult in the social and behavioral sciences, all the more reason why Morrison and Henkel (1969) asserted:

> …for statistical inference to be possible one must first specify the population and then probability sample from that population. The notions of sampling distribution and sampling error have no meaning in statistical inference apart from the assumption of randomness in the sample selection procedure–

> randomness being a central feature incorporated in all probability sampling designs. (p. 133)

Indeed, sampling error consists of two components: random and systematic. By increasing the sample size, only the random component of sampling error becomes less of a problem, while the systematic part remains unchanged. As for the stability of descriptive statistics, the law of large numbers ensures statistical consistency of estimates as sample size increases, but again, this property is predicated on the sampling being random; thus, biased sampling using large sample sizes can result in consistent and biased samples.

The point here is not to condemn research done without random sampling–random sampling is difficult if not prohibitive for many studies, more especially those in the behavioral, educational, and medical sciences; rather, it is to illustrate how *BASP*'s newly adopted methodology is not somehow more resilient to the aforementioned issues, relative to Frequentist and Bayesian methods. Ironically, by requiring authors to use large sample sizes and report descriptive statistics, *BASP*'s prescripts would only help these inferential frameworks, if they were allowed. Instead of removing inference from their methodology, *BASP* could improve the quality of research by requiring authors to do the three following things:

1. Clearly state the population of interest; not only does this help readers understand the scope of the research, it also provides useful information for conducting meta-analyses and replication studies.
2. Use random sampling methods, when possible. When random sampling is not possible, authors should be required to report what sampling methods were used, why random sampling could not be used, and likely sources of bias in the existing sample, including reporting detailed demographic statistics. This allows readers to evaluate the quality of the study sample, in reference to the population of interest.
3. Instead of relying on one all-or-nothing sample, authors should be required to collect multiple samples when possible (e.g., multiple schools, hospitals, etc.). In addition, *BASP* should promote publication of replication studies from existing research.

Implementation of these requirements could certainly be considered state of the art.

References

Cochran, J. (2015). [ASA Leaders Reminisce interview with Peter (Tony) Lachenbruch]. *Amstat News, 452*, 16-17. Retrieved from http://magazine.amstat.org/blog/2015/02/01/peterlachenbruch_feb2015/

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist*, *4*(2), 131-140. Available from http://www.jstor.org/stable/27701482

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. doi: 10.1080/01973533.2015.1012991