MIS Quarterly

RESEARCH NOTE

STATISTICAL INFERENCE WITH PLSc USING BOOTSTRAP CONFIDENCE INTERVALS

Miguel I. Aguirre-Urreta
Florida International University, RB 258A, 11200 SW 8th Street,
Miami, FL 33199 U.S.A. {miguel.aguirreurreta@fiu.edu}

Mikko Rönkkö
Department of Computer Science and Information Systems, University of Jyväskylä, P.O. Box 35,
FI-40014 Jyväskylä FINLAND and
Department of Industrial Engineering and Management, Aalto University School of Science,
P.O. Box 11000, FI-0076 Aalto FINLAND {mikko.ronkko@jyu.fi}

Appendix A

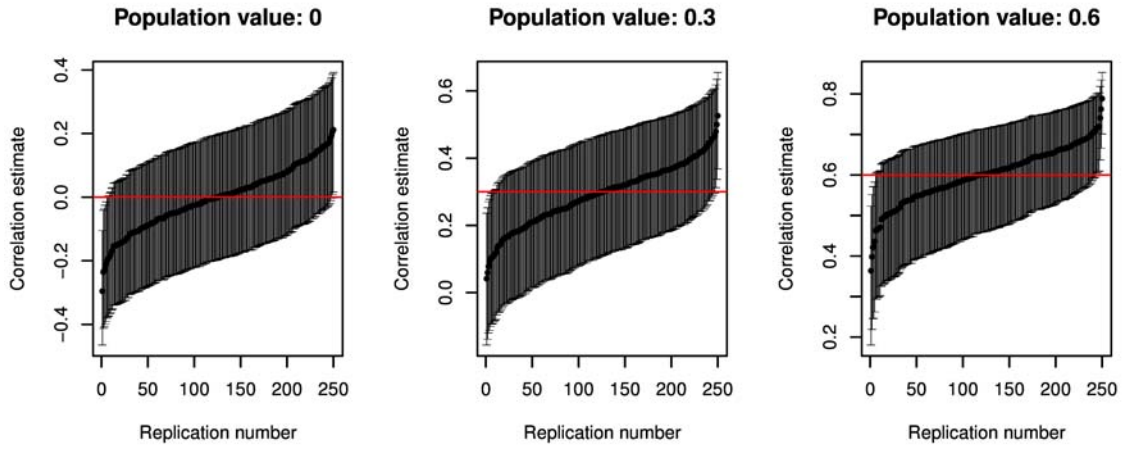
Bootstrapped Confidence Intervals

Confidence intervals (CIs) are an alternative to null hypothesis significance testing (NHST) (Nickerson 2000; Wood 2005); both techniques essentially convey information about how certain we are of an estimate. A CI consists of two values, upper and lower confidence limits, associated with a confidence level, typically 95%. Valid CIs should satisfy two important conditions (DiCiccio and Efron 1996). First, a CI should contain the population value of the parameter under estimation with the stated degree of confidence over a large number of repeated samples. For example, a 95% CI should contain the population value of a parameter 95% of the time, with the true value of the parameter falling outside of the interval only in 5% of the cases. Second, in those cases where the true value of the parameter falls beyond the boundaries of the interval, it should do so in a balanced way. Using again the 95% CI as an example, the population value should be higher than the upper boundary in 2.5% of the samples and lower than the lower boundary of the interval 2.5% of the time.

We illustrate these two properties of CIs in Figure A1. The figure shows 250 correlation estimates drawn from three different populations (where the true value of the correlation is 0, 0.3, or 0.6, respectively), ordered from smallest to largest, and their 95% CIs. In this scenario, the population value falls outside the CI only about 5% of the time and does so in a balanced way, such that the population value lies above the CI 2.5% of the time and below the CI 2.5% of the time. The figure also shows that both the variance of the estimates and the width of the CIs depend on the population value of the correlation; when the population correlation is zero, the difference between the largest and smallest estimate is close to 0.5, but when the population value is 0.6 this difference decreases to about 0.35. Similarly, the CIs are narrower for larger estimates. This is an important feature of CIs that unfortunately complicates their calculation, as we discuss later.

The CI of a correlation has a valid closed form solution, but estimating CIs for more complex scenarios is a non-trivial problem. The most straightforward way to estimate CIs is to use a known theoretical distribution. We refer to these as *parametric* approaches. When the distribution of the estimates is not known, as is the case with those obtained from PLSc, CIs based on bootstrapping provide an attractive alternative (Wood 2005). In these approaches, which we refer to as *empirical*, the endpoints of the CIs are not taken from a known statistical distribution, but rather the values are obtained from the empirically approximated bootstrap distribution.

Bootstrapping means that we draw a large number of samples from our original data and calculate the statistic for each sample. The samples are drawn with replacement, which means that each observation in the original sample can be included in each bootstrap sample multiple times. While bootstrapping can be useful when working with statistics whose sampling distributions are unknown, it is not a silver bullet. A key



Note: N = 100

Figure A1. Confidence Intervals of 250 Correlation Estimates

assumption behind bootstrapping is that the bootstrap estimates follow the same distribution as the original statistic, but this is not always the case, and can lead to incorrect inference unless corrections are applied. Generally, many properties of bootstrap estimators have been proven only for the asymptotic case (Davison and Hinkley 1997, pp. 37-41) and may therefore work well only with large samples.

Empirical CIs can be calculated from bootstrapped estimates in many different ways. Three such approaches are the *percentile*, *bias-corrected*, and *bias-corrected and accelerated* CIs, which we review next. More technical discussions can be found in DiCiccio and Efron (1996), Efron and Tibshirani (1993), or Davidson and Hinkley (1997). All three types of CIs are calculated from the same bootstrap replication data. In the *percentile* approach, all bootstrap estimates of a parameter of interest are ordered from smallest to largest. Then, two particular estimates are selected to represent the boundaries of the CI with a desired coverage (e.g., the 250th and 9750th replicates for a 95% interval based on 10,000 bootstrap replications). This is the simplest way of creating an empirical CI from the bootstrap replications, and works best when the distribution is symmetrical and centered on the original estimate. While simple, this approach is often suboptimal because of two problems.

First, bootstrap estimates are generally biased in small samples. The bootstrap bias is illustrated in Figure A2, which shows the sampling distribution of the mean from a normally distributed population and three bootstrapped distributions for four different sample sizes. The bootstrapped distributions are roughly the right shape and width, but are off sideways.

Second, the sampling variance of a statistic can depend on the population value, as briefly mentioned in reference to Figure A1 earlier. In bootstrapping, the original sample is essentially treated as a population from which samples are drawn, and this makes the variance of the bootstrap replications dependent on the value of the original estimate, particularly if the original statistic is itself skewed. The problem of uneven variance is illustrated in Figure A3, which shows PLS estimates from a simple example with two exogenous latent variables, which are uncorrelated, and one endogenous latent variable. All latent variables are measured with three indicators with loadings of 0.7 and errors of 0.51. The paths relating the latent variables are 0.4 and 0, thus explaining 16% of the variance in the endogenous latent variable. The sample size for this example is 100 and we replicated the analysis 1,000 times. For this example, we are focusing on the nonzero path coefficient (population value of 0.4) for the case of normal data.

Using these data in PLS estimation resulted in a substantially left-skewed sampling distribution of the estimates. To demonstrate the effects that a skewed distribution has on the bootstrap, we chose six replications producing a wide range of estimates. These six selected samples were each bootstrapped with 1,000 replications and we estimated the distribution of the PLS estimates with these six sets of resamples. We then centered the six resulting bootstrap distributions to the mean of the original distribution, thus completely eliminating bootstrap bias, and plotted the unbiased distributions over the original distribution in Figure A3. The figure shows a clear skew in the original estimates and consequently the variance of the bootstrap estimates is uneven depending on the value of the original estimate. As a result, the 95% intervals of the distributions calculated from the bootstrap replications do not always match those calculated from the original distribution, and any CIs based on these distributions would be problematic.

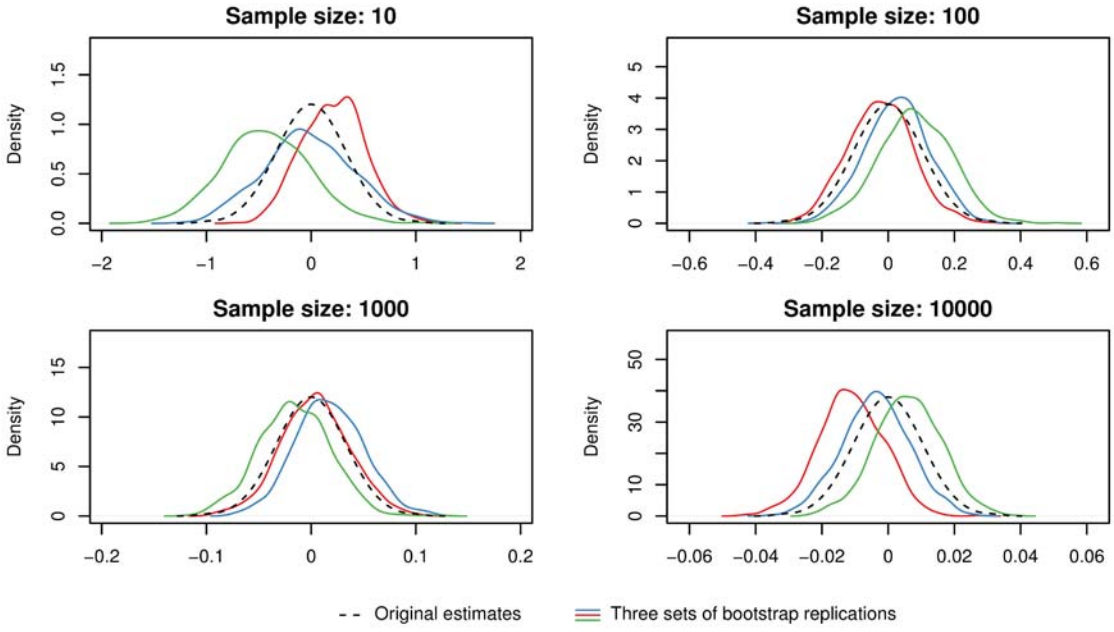


Figure A2. Sampling Distribution of the Mean of a Normally Distributed Population and Three Bootstrap Approximations for Four Sample Sizes

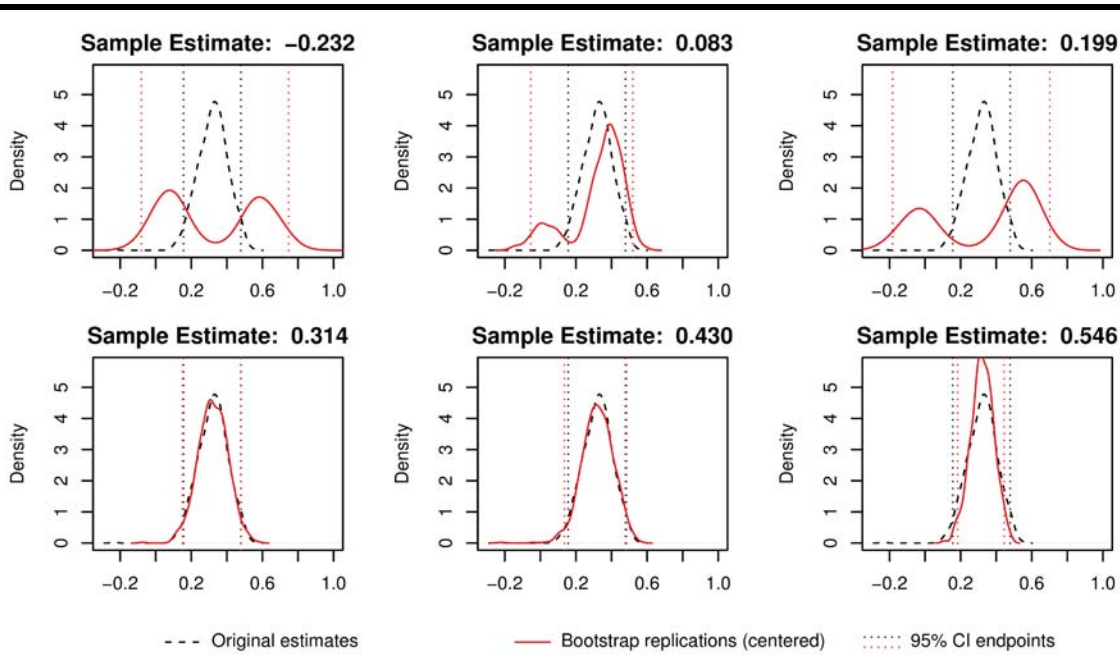


Figure A3. Bias and Uneven Variance of Bootstrap Replicates

Bias-corrected and accelerated (BCa) CIs were developed to simultaneously address both of these problems, and have been shown to significantly improve interval coverage and balance (DiCiccio and Efron 1996; Efron and Tibshirani 1993). Like percentile intervals, these are obtained by choosing replicates from the bootstrap distribution as the boundaries of the CIs, but instead of using fixed replicates the choice is based on two statistics that correct for the median bias and uneven variance of the bootstrap distribution.

We will now explain these corrections in more detail. The *bias-corrected* (BC) CI is best understood as a special case of the more general bias-corrected and accelerated interval. Therefore, we explain the more complex BCa correction first. Let θ be the original estimate, $\theta^{*}(b)$ the estimate for the same parameter obtained from the b^{th} bootstrap sample, and B the number of bootstrap samples. The bias correction adjustment is calculated as (DiCiccio and Efron 1996, eq. 2.8, p. 193):

$$z = \phi^{-1}\left(\frac{\#\left\{\theta^{*}(b) < \theta\right\}}{B}\right)$$

Where *#{condition}* is the number of samples that meet the condition in parentheses, in this case being lower than the original estimate, and ϕ^{-1} is the inverse function of the cumulative distribution function of the standard normal distribution.

To address the issue of uneven variance, an acceleration coefficient is calculated to measure how quickly the standard error of the estimate is changing on a normalized scale. Efron and Tibshirani (1993) suggested that this *acceleration coefficient* be calculated using the jackknife ("keep one out") estimates of the parameter of interest. Let $\theta_{(i)}$ be the estimate obtained on the sample with the i^{th} case removed, n the number of jackknife samples, and $\theta_{(*)}$ be the mean of the estimates from the jackknife samples. The acceleration coefficient is calculated as (DiCiccio and Efron 1996, eq. 6.6):[1]

$$a = \frac{\sum_{i=i}^{n}\left(\theta_{(*)} - \theta_{(i)}\right)^{3}}{6\left\{\sum_{i=1}^{n}\left(\theta_{(*)} - \theta_{(i)}\right)^{2}\right\}^{3/2}}$$

The bias-correction and acceleration constants are then used to select which bootstrap replicates will be used as the boundaries of the CI. Let z be the bias-correction coefficient and a the acceleration coefficient (as defined above), B the number of bootstrap replications, ϕ the standard normal cumulative distribution, ϕ^{-1} the inverse function of the cumulative distribution function of the standard normal distribution, and α the chosen significance level for the CI (5% throughout this research). We can then calculate coefficients α_1 and α_2 as (Davison and Hinkley 1997, eq. 5.21, p. 204):

$$\alpha_1 = \phi\left(z + \frac{z + \phi^{-1}(\alpha/2)}{1 - a\left(z + \phi^{-1}(\alpha/2)\right)}\right)$$

$$\alpha_2 = \phi\left(z + \frac{z + \phi^{-1}(1-\alpha/2)}{1 - a\left(z + \phi^{-1}(1-\alpha/2)\right)}\right)$$

These two coefficients are then multiplied by the number of bootstrap replications to identify the two replicates to be used as the lower and upper boundaries for the CI (LB = Lower Boundary, UB = Upper Boundary) (Davison and Hinkley 1997):

[1]Alternatively, empirical influence values can be obtained with regression techniques using the bootstrap replications without the use of jackknife (Davison and Hinkley 1997, sec. 2.7.4).

$$LB = B * \alpha_1$$
$$UB = B * \alpha_2$$

When the acceleration statistic is zero, this simplifies to the *bias-corrected* interval. When both the acceleration and bias-correction statistics are zero, the chosen replications will be the same as in the simpler case of the *percentile* interval discussed before.

Appendix B

Null Hypothesis Significance Testing

Criticisms of Null Hypothesis Significance Testing (NHST)

The practice of NHST, even though commonplace in the social and behavioral sciences, has been the subject of much criticism for more than 40 years (Balluerka et al. 2005); for comprehensive treatments of these discussions, see Oakes (1986), Kline (2004), or Ziliak and McCloskey (2008); for a recent commentary, see Antonakis (2017). In order to ground our discussion, the NHST as performed by applied researchers can be summarized in the following sequence of activities (Balluerka et al. 2005; Gigerenzer 2004; Levine et al. 2008). First, a null hypothesis is proposed regarding the value of a parameter in the population. Although not strictly necessary, this hypothesis is most commonly that of no effect, also labeled the null-nil hypothesis. Second, a test statistic—typically the ratio of an estimate and its standard error—is chosen. Third, a random sample of data is obtained and the statistic is calculated in this sample. Fourth, the probability of obtaining a test statistic, which is as large as or larger than the value actually obtained, given the sample size and assuming that the null hypothesis is true in the population, known as the *p* value, is calculated. Finally, the obtained probability is compared against a predefined criterion (most commonly $\alpha = 0.05$). If the obtained probability is less than the criterion, the null hypothesis is rejected, and this is taken to provide support for the alternative hypothesis (though this is potentially problematic, as discussed later). There are two sets of criticisms commonly levelled against NHST, which we cover in more detail in the next two sections.

Misinterpretations of NHST or Misconceptions about the Process

The first set of issues has to do with the presence of major misconceptions about how NHST works and how its results should be interpreted. Our discussion is based on Nickerson (2000), who covers the issues in great detail. For additional discussion on these see Hubbard and Lindsay (2008), Gliner et al. (2002), Gigerenzer (2004), Gigerenzer et al. (2004), or Greenwald et al. (1996). These problems include (page references are to the work of Nickerson (2000), which discusses them in detail) the belief that *p* is the probability that the null is true and that, by extension, $1 - p$ is the probability that the alternative hypothesis is true (p. 246), the belief that rejection of the null hypothesis establishes the truth of a theory that predicts it to be false in the first place (p. 254), the belief that a small *p* value is evidence that results are replicable (p. 256), the belief that a small *p* value indicates an effect size of large magnitude (p. 257), the belief that statistical significance implies theoretical or practical significance (or relevance) (p. 257), the belief that α is the probability of committing a type I error if the null hypothesis has been rejected (p. 258), the belief that the value at which α is set for a given experiment or a large set of experiments is the probability of type I error for those (p. 259), the belief that failing to reject the null hypothesis is tantamount to demonstrating it to be true (p. 260), or the belief that failing to reject the null hypothesis is evidence of poorly conducted research (p. 261).

All these represent misunderstandings or misconceptions about NHST and therefore do not, as such, present any fundamental challenges to practice (unlike those reviewed next). However, these are still worrisome, in that they reflect long-standing misunderstandings of a very widespread approach in the social and behavioral sciences, and should be cause of concern for the different disciplines that make extensive use of NHST.

Fundamental Challenges to NHST

There are also a number of issues that present fundamental challenges to the practice of NHST and which highlight the need for alternative approaches to the assessment of statistical estimates, with the use of CIs being one of the central recommendations arising from this literature. These challenges are reviewed here in more detail below; for an accessible introduction see Nuzzo (2014).

First, several researchers have challenged the practice of NHST—particularly when coupled with null-nil hypotheses of no effect—on the grounds that the null hypothesis is never strictly true, at least in the social and behavioral sciences (Balluerka et al. 2005, p. 58; Levine et al. 2008, p. 176; Nickerson 2000, p. 263; Schmidt and Hunter 2002). From this position, rejecting something that is not true in the first place is neither informative nor useful (Abelson 1997; Cohen 1994). A direct corollary is that all that is required to obtain significance is a large enough sample size, as there is always some underlying effect to detect (the issue of whether these effects are of any relevant magnitude is a different problem). In most cases, researchers are not strictly interested in whether an effect is precisely zero, but rather in whether it is close enough to zero to be of no practical interest.

Second, NHST has been heavily criticized for its sensitivity to sample size (Levine et al. 2008, p. 176; Nickerson 2000, p. 265). When the sample size is small, even strong—and arguably practically relevant—effects will fail to reach significance. Alternatively, large samples can make even smaller— and inconsequential—effects be significant. As a result, the *p* values obtained from NHST reflect not only the magnitude of the effect but also the investment in resources expended by researchers in collecting data. This can lead to ignoring important effects arising from small samples and embracing minor ones solely because a large enough sample was collected.[2] This is not a desirable property for the scientific enterprise (Levine et al. 2008).

Third, NHST has been criticized on the grounds that it is illogical (Balluerka et al. 2005, p. 57; Nickerson 2000, p. 267). This refers to a rule of logic called *modus tollens*, which is valid when statements are categorical, but not so when those are probabilistic, as is the case when dealing with empirical tests of a theory (Cohen 1994; Falk and Greenbaum 1995; Macdonald 1997). The general form of the statement "*If P then Q; not Q, therefore not P*" is not necessarily valid when the premises are probabilistic, that is *"If P then probably Q; not Q, therefore probably not P"*; or, in NHST terms, *"If H_0 is true then probably $p > .05$; $p < .05$, therefore probably H_0 is false"* (Nickerson 2000). Carver (1978) labeled this the "odds-against-chance fantasy" (p. 382).

Finally, other challenges to NHST focus on its value, or usefulness, for the practice of theory testing and, more generally, for the advancement of scientific knowledge. Among these, the following are likely those of most importance. First, NSHT tests provide relatively little information about the outcomes of a research study (Balluerka et al. 2005, p. 58; Nickerson 2000, p. 268). One such position is that the aims of NHST and that of statistical inference—the latter more in line with the aims of researchers—are substantially different. That is, researchers are interested in the probability that the null hypothesis is true based on the observed results [more formally, $p(H_0|D)$], whereas NHST only informs the probability of obtaining observed results if the case of the null hypothesis being true [that is, $p(D_0|H_0)$], which is not the same thing. Second, the "all or nothing" nature of NHST means that a *p* value that is only slightly greater than the α level is lumped together with those where they value is much greater. Many researchers find this practice problematic, which is generally expressed in statements about "marginally significant," "borderline significant," or "significance at the more liberal level of (for example) $p < .10$." Moreover, nonsignificant results are largely ignored which, coupled with a strong bias toward the publication of significant results, inhibits the accumulation of knowledge. Closely related to this is the reification of an arbitrary decision criterion as the arbiter of significance (e.g., $\alpha = .05$) (Nickerson 2000, p. 269). Finally, but of central importance, is the argument that NHST does not actually enable the testing of our theories (Balluerka et al. 2005, p. 58). In current practice, a researcher develops a set of arguments for why a relationship would be expected to be observed between certain variables (or, in experimental designs, why a difference between groups would be expected). Null and alternative hypotheses are presented as competing and mutually exclusive, and rejection of the null hypothesis is seen as providing support for the alternative one, which in turn provides evidence in support of the theory. There is, however, an important disconnect between these last two steps. That is, the rejection of the null-nil hypothesis of no effect does not automatically imply that the observed effects are due to the reasons postulated by researchers in their theoretical developments. Rather, this inference from observed results to theory can only be made when care has been taken to design and execute a research design that excludes other competing theoretical reasons that could explain the observed results, as well as any methodological confounds that could be in operation. None of this, however, can be ascertained solely by considering the significance (or not) of the results.

Moving Forward

Given these extensive criticisms of past research practice with regard to NHST, recent years have witnessed the emergence of a number of published recommendations for the reporting and interpretation of research findings that seek to ameliorate many of these issues; the spirit of much of this renewed emphasis on interpretation of research findings in a more extensive light has been captured in the recommendations provided by Cumming (2014, see Table 1). To be fair, some of these criticisms of NHST have also been challenged in the literature. It could be argued that we need some criteria on what can be published and $p < .05$ presents a fairly low bar that is necessary for ruling out chance as an explanation, but should not be interpreted as being sufficient evidence of a meaningful finding (Cortina and Landis 2011). However, most

[2]There is some recognition of this scenario in the IS literature proper, in that the ability to collect large datasets can lead researchers to support results of no practical significance (Lin et al. 2013).

importantly, there is no reason to believe that a researcher who misused or misinterpreted the results of NHST would not misuse or misinterpret CIs as well (Cortina and Landis 2011) and CIs and NHST are both based on the same statistical theory, so therefore CIs cannot solve all issues with NHST (Trafimow and Marks 2015). Thompson (2001, pp. 82-83), in an often cited quote, summarizes the discussion bluntly "if people interpret effect sizes with the same rigidity with which $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric." Nevertheless, we argue that, for example, when a statistic that is normally distributed in the sample has a point estimate of 0.2, stating that the p value is less than 0.05 is more likely to lead to false conclusions about the accuracy of the estimate than stating that the 95% CI for the estimate is 0.004–0.396. If one wants to make a dichotomous conclusion about a theoretical hypothesis based on a statistical analysis, the choice between NHST and normal approximation CIs is in many cases largely a personal preference, as both lead to the same conclusion. However, the same does not apply to the empirical CIs that we discuss in this research.

To address the issues presented above, a number of different approaches have been taken in the literature. First, some journals, such as *Epidemiology* or *Basic and Applied Social Psychology*, have taken an extreme view on the issue and instituted editorial guidelines outright banning the use of NHST in their published research. For the latter case, for example, the editors have noted that, while submissions that include mention of p values and the like will not be subject to automatic desk rejection, all such indications of the use of NHST (such as p values, t values, F values, statement about "significant" results, etc.) will need to be removed prior to publication. The ban is also extended to the use of CIs as well (Trafimow and Marks 2015).

Second, while not going to the same lengths, there have been renewed calls to move away from the binary decisions (significant or not) underlying the practice of NHST and to focus more on providing estimates of the magnitude of an effect that consider both the magnitude of the estimate as well as its precision. The use of effect sizes, together with CIs, has been strongly advocated for that purpose. CIs are seen as accommodating a more varied gamut of interpretations of the results, can potentially provide more insight into the findings as they provide a much more clear perspective on the range of possible outcomes that are implied by the collected data and theoretical model tested, and more heavily encourage replication in pursuit of subsequent narrowing of that range, thus leading to potentially greater insight into the research question (Balluerka et al. 2005; Kline 2004). At the very least, and taking into consideration the caveats discussed above, CIs provide no less information than what is made available under current reporting practices (whether an estimate is different from zero) but, by making clear the range of possible outcomes that go together with that binary assessment, give a more complete picture of the results.

Third, researchers have also started to advocate that, together with the reporting of CIs, the estimates obtained from the analysis of the data and research model are expressed in the form of effect sizes as well as in the original metric of the study (Thompson 1999). Likely even more important than publishing CIs for the original results, however, is also the inclusion of CIs for the effect sizes as well (Thompson 2007). This practice has the distinct advantage of moving researchers away from dichotomous assessments and into discussions of the magnitude of reported effects (though it could be argued most research practice in this area is rudimentary, still relying on rules of thumb as to what constitutes a large, medium, or small effect). This, in turn, leads naturally into considerations of the practical relevance of research findings, as opposed to its statistical significance. With the increasing availability of very large datasets, that distinction between results that are statistically significant but practically negligible, and those that are both statistically significant and practically relevant is more likely to take on a central role in our discussion of the implications of our research (Lin et al. 2013).

Fourth, together with a renewed emphasis on reporting more as well as different information about the research findings (e.g., effect sizes and CIs), new recommendations have arisen for how to best describe and present data (in both tabular and graphical form) in order to enhance our understanding of the findings (Valentine et al. 2015). These include the usage of a number of alternative ways in which results can be presented graphically, in terms of the types of results and variables being depicted, as well as the embodiment of basic principles such as "never a center without a spread," which captures the notion that any information presented about a distribution should be accompanied by information about the variability of such distribution (Fidler and Loftus 2009; Valentine et al. 2015) (which is notably absent from the plots and figures currently employed in research in the IS discipline).

Finally, and perhaps of greatest importance for the conceptual and theoretical development of a discipline, it can be argued that the research practice of assessing the outcome of a study by means of binary significant/nonsignificant decisions has been developed in parallel with the equally problematic practice of stating research hypotheses also in binary terms, for example, the presence or absence of an effect (the direction of causality between both practices, in historical terms, may be debatable; that the practice of NHST lead to hypotheses that reflect their testing may seem somewhat more likely). This, however, need not be the case. Edwards and Berry (2010, see Table 1 for a summary) recently sought to call attention to a number of practices designed to improve theoretical precision in social science research. Their main contention is that increased methodological rigor leads to hypotheses—in their simple, directional form—that are not likely to be seriously challenged, therefore putting theories at a lower risk of falsification, which is an undesirable side-effect. This argument also has implications for the practice of NHST (indeed, the authors note NHST as a barrier to increased precision in theoretical developments). In practice this means postulating more carefully crafted hypotheses, which not only take into account the presence or absence of an effect or relationship, or its simple directionality, but also the specific functional form relating the variables in question or the expected upper and lower boundaries of the effect (to name two

examples). While this kind of refined hypotheses are also arguably more interesting from a theoretical standpoint, it also shifts the focus to a wider range of statistical tests and models in order to ascertain the support for their theory and away from the simple binary decisions that currently characterize the practice of NHST. Doing so would lead not only to more precise theories, the main point of Edwards and Berry, but arguably also to more interesting ones.

Appendix C

Detailed Discussion of Results: Simple Model

There is an important anomaly in the results of the first variant (e.g., uncorrelated LVs) of the simple model, which influences all results plots for the model as all figures are calculated using all data and differ only in terms of how this data were divided between the panels. Understanding the causes of this anomaly helps understanding the differences among the techniques and presents a boundary conditions for applicability of the results. This anomaly occurs in the case of the weak path configuration, as shown in the third plot in Figure 3, and manifests as a marked downward angle in the lines corresponding to the upper boundaries of the BC and BCa CIs. The results show that in these conditions the proportion of CIs with upper boundaries that are smaller than the population values is much larger than expected. This effect is the result of a combination of a less well-known feature of PLS and the bias correction implemented in the BC and BCa approaches to CI construction. First, PLS weights have been shown to be susceptible to capitalization on chance, such that the regression paths are larger than would otherwise be the case (Goodhue et al. 2015; Rönkkö 2014; Rönkkö et al. 2015) and this effect carries over to PLSc as well (Rönkkö et al. 2016). While this effect is not yet fully understood, it seems to be related to sample size (negatively), number of indicators (positively), strength of factor loadings (negatively), and the strength of the paths between the latent variables in the population (negatively). When sample size is small and the effects between two latent variables are weak (as is the case here), the PLS estimates are clearly bimodal such that small negative (e.g., -0.1) or small positive (e.g. 0.1) values are much more likely than zero. This is consistent with the idea that PLS weights attempt to maximize the inner model R^2 (Hair et al. 2014), because paths with zero coefficients do not increase R^2 at all. In the weak path values scenario, the three paths that had an effect on the dependent latent variable were just 0.114, representing only about 1% of shared variance between the endogenous and each exogenous latent variable. Consequently, the estimates were strongly bimodal in small samples. This effect decreases with increasing sample size and strength of the effect between the latent variables so that the secondary mode on the negative side of the y axis first decreases up to a point where it appears just a long negative tail right before all estimates become positive (see Henseler et al. 2014, Figure 5).

When sample size increases, the accuracy of the bootstrap distribution as an estimate of the true distribution improves accordingly. In these scenarios, a small number of replications contained the original value in the negative tail and in this case the majority of the bootstrap replications were larger than the original replication. Because the bias correction factor is calculated based on the count of bootstrap replications that exceed the original replication, this results in very large negative bias correction values. The large bias correction values then result in rather extreme intervals. For example if the original estimate is in the negative tail (which is the more common case because the tail is heavier) it is possible to observe that 99% of bootstrap replications (i.e., 9,900 out of 10,000) are larger than the original estimate. In this scenario the bias correction coefficient receives the value of -2.33, leading to choosing the smallest and 35th smallest replication (out of 10,000) as the confidence limits of the resulting CI. In this extreme scenario even the original estimate, let alone the population value of the parameter, fall outside the limits of the CI, which creates a bias in the upper limit. We examined a small subset of conditions using sample sizes larger than 1,000 and concluded that the effect was present when the sampling distribution of the original estimates crossed zero, but disappeared when sample size grew large enough so that all estimates were always positive. This represents an important boundary condition for the application of CIs to results obtained from PLSc. Fortunately, identifying these occurrences is straightforward by inspecting the indices used for selecting the replications to be used as confidence limits. While these indices are not directly reported in any of the commercial PLS software,[3] they can be calculated by comparing the reported values of bootstrap replications against the confidence limits counting how many replications fall below the upper and lower limits.

[3]The indices are reported by the boot R package that implement empirical CIs in matrixpls.

References

Abelson, R. P. 1997. "A Retrospective on the Significance Test Ban of 1999 (If There Were No Significance Tests, They Would Be Invented)," in *What If There Were No Significance Tests?*, L. L. Harlow, S. A. Mulaik, and J. H. Steiger (eds.), Mahwah, NJ: Lawrence Erlbaum Associates, pp. 117-141.

Antonakis, J. 2017. "On Doing Better Science: From Thrill of Discovery to Policy Implications," *Leadership Quarterly* (28:1), pp. 5-21.

Balluerka, N., Gómez, J., and Hidalgo, D. 2005. "The Controversy over Null Hypothesis Significance Testing Revisited," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, (1:2), pp. 55-70 (https://doi.org/http://dx.doi.org/10.1027/1614-1881.1.2.55).

Carver, R. 1978. "The Case Against Statistical Significance Testing," *Harvard Educational Review* (48:3), pp. 378-399.

Cohen, J. 1994. "The Earth Is Round ($p < .05$)," *American Psychologist* (49:12), pp. 997-1003 (https://doi.org/10.1037/0003-066X.49.12.997).

Cortina, J. M., and Landis, R. S. 2011. "The Earth Is Not Round ($p = .00$)," *Organizational Research Methods* (14:2), pp. 332-349 (https://doi.org/10.1177/1094428110391542).

Cumming, G. 2014. "The New Statistics: Why and How," *Psychological Science* (25:1), pp. 7-29 (https://doi.org/10.1177/0956797613504966).

Davison, A. C., and Hinkley, D. V. 1997. *Bootstrap Methods and Their Application*, Cambridge, UK: Cambridge University Press.

DiCiccio, T. J., and Efron, B. 1996. "Bootstrap Confidence Intervals," *Statistical Science* (11:3), pp. 189-212 (https://doi.org/10.2307/2246110).

Edwards, J. R., and Berry, J. W. 2010. "The Presence of Something or the Absence of Nothing: Increasing Theoretical Precision in Management Research," *Organizational Research Methods* (13:4), pp. 668-689 (https://doi.org/10.1177/1094428110380467).

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*, New York: Chapman and Hall/CRC.

Falk, R., and Greenbaum, C. W. 1995. "Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception," *Theory and Psychology* (5:1), pp. 75-98.

Fidler, F., and Loftus, G. R. 2009. "Why Figures with Error Bars Should Replace *p* Values," *Zeitschrift Für Psychologie (Journal of Psychology)* (217:1), pp. 27-37 (https://doi.org/10.1027/0044-3409.217.1.27).

Gigerenzer, G. 2004. "Mindless Statistics," *The Journal of Socio-Economics* (33:5), pp. 587-606 (https://doi.org/10.1016/j.socec.2004.09.033).

Gigerenzer, G., Krauss, S., Vitouch, O., and Kaplan, D. 2004. "The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask," in *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan (ed.), Thousand Oaks, CA: Sage Publishing, pp. 391-408.

Gliner, J. A., Leech, N. L., and Morgan, G. A. 2002. "Problems with Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say?," *The Journal of Experimental Education* (71:1), pp. 83-92.

Goodhue, D. L., Lewis, W., and Thompson, R. 2015. "PLS Pluses and Minuses in Path Estimation Accuracy," in *Proceedings of the 21st Americas Conference on Information Systems*, San Juan, Puerto Rico (http://aisel.aisnet.org/amcis2015/ISPhil/ GeneralPresentations/3).

Greenwald, A., Gonzalez, R., Harris, R., and Guthrie, D. 1996. "Effect Sizes and *p* Values: What Should Be Reported and What Should Be Replicated?," *Psychophysiology* (33:2), pp. 175-183.

Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. 2014. *A Primer on Partial Least Squares Structural Equations Modeling (PLS-SEM)*, Thousand Oaks, CA: Sage Publishing.

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T., and Calantone, R. J. 2014. "Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013)," *Organizational Research Methods* (17:2), pp. 182-209 (https://doi.org/10.1177/1094428114526928).

Hubbard, R., and Lindsay, R. M. 2008. "Why *p* Values Are Not a Useful Measure of Evidence in Statistical Significance Testing," *Theory and Psychology* (18:1), pp. 69-88.

Kline, R. B. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*, Washington, DC: American Psychological Association.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., and Lindsey, L. L. M. 2008. "A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research," *Human Communication Research* (34:2), pp. 171-187.

Lin, M., Lucas, H. C., and Shmueli, G. 2013. "Research Commentary—Too Big to Fail: Large Samples and the *p*-Value Problem," *Information Systems Research* (24:4), pp. 906-917 (https://doi.org/10.1287/isre.2013.0480).

Macdonald, R. R. 1997. "On Statistical Testing in Psychology," *British Journal of Psychology* (88:2), pp. 333-347 (https://doi.org/10.1111/j.2044-8295.1997.tb02638.x).

Nickerson, R. S. 2000. "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods* (5:2), pp. 241-301 (https://doi.org/10.1037//1082-989X.5.2.241).

Nuzzo, R. 2014. "Scientific Method: Statistical Errors," *Nature* (506:7487), pp. 150-152 (https://doi.org/10.1038/506150a).

Oakes, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*, New York: Wiley.

Rönkkö, M. 2014. "The Effects of Chance Correlations on Partial Least Squares Path Modeling," *Organizational Research Methods* (17:2), pp. 164-181 (https://doi.org/10.1177/1094428114525667).

Rönkkö, M., McIntosh, C. N., and Aguirre-Urreta, M. I. 2016. "Improvements to PLSc: Remaining Problems and Simple Solutions," unpublished working paper, Aalto University (retrieved from http://urn.fi/URN:NBN:fi:aalto-201603051463).

Rönkkö, M., McIntosh, C. N., and Antonakis, J. 2015. "On the Adoption of Partial Least Squares in Psychological Research: Caveat Emptor," *Personality and Individual Differences* (87), pp. 76-84 (https://doi.org/10.1016/j.paid.2015.07.019).

Schmidt, F. L., and Hunter, J. 2002. "Are There Benefits from NHST?," *American Psychologist* (57:1), pp. 65-66 (https://doi.org/10.1037/0003-066X.57.1.65).

Thompson, B. 1999. "Improving Research Clarity and Usefulness with Effect Size Indices as Supplements to Statistical Significance Tests," *Exceptional Children* (65:3), pp. 329-337.

Thompson, B. 2001. "Significance, Effect Sizes, Stepwise Methods, and Other Issues: Strong Arguments Move the Field," *Journal of Experimental Education* (70:1), pp. 80-93.

Thompson, B. 2007. "Effect Sizes, Confidence Intervals, and Confidence Intervals for Effect Sizes," *Psychology in the Schools* (44:5), pp. 423-432 (https://doi.org/10.1002/pits.20234).

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

Valentine, J. C., Aloe, A. M., and Lau, T. S. 2015. "Life after NHST: How to Describe Your Data Without 'p-ing' Everywhere," *Basic and Applied Social Psychology* (37:5), pp. 260-273.

Wood, M. 2005. "Bootstrapped Confidence Intervals as an Approach to Statistical Inference," *Organizational Research Methods* (8:4), pp. 454-470 (https://doi.org/10.1177/1094428105280059).

Ziliak, S. T., and McCloskey, D. N. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press.