METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

The Inaccuracy of Sample-Based Confidence Intervals to Estimate A Priori Ones

David Trafimow[a], Joshua Uhalt[a]

[a] *New Mexico State University, Las Cruces, NM, USA.*

Methodology, 2020, Vol. 16(2), 112–126, https://doi.org/10.5964/meth.2807

Received: 2018-07-22 • **Accepted:** 2019-09-14 • **Published (VoR):** 2020-06-18

Corresponding Author: David Trafimow, Department of Psychology, MSC 3452, New Mexico State University, P. O. Box 30001, Las Cruces, NM 88003-8001. E-mail: dtrafimo@nmsu.edu

Abstract

Confidence intervals (CIs) constitute the most popular alternative to widely criticized null hypothesis significance tests. CIs provide more information than significance tests and lend themselves well to visual displays. Although CIs are no better than significance tests when used solely as significance tests, researchers need not limit themselves to this use of CIs. Rather, CIs can be used to estimate the precision of the data, and it is the precision argument that may set CIs in a superior position to significance tests. We tested two versions of the precision argument by performing computer simulations to test how well sample-based CIs estimate a priori CIs. One version pertains to precision of width whereas the other version pertains to precision of location. Using both versions, sample-based CIs poorly estimate a priori CIs at typical sample sizes and perform better as sample sizes increase.

Keywords

a priori procedure, a priori confidence intervals, accuracy, width, location

The null hypothesis significance testing procedure is increasingly coming under attack (see Hubbard, 2016; Ziliak & McCloskey, 2016 for recent reviews) and it was widely criticized at the recent American Statistical Association Symposium on Statistical Inference (October 11-12, 2017) and in the recent (2019) special issue of *The American Statistician.* A popular alternative is for researchers to use confidence intervals (CIs) (see Cumming & Calin-Jageman, 2017 for a recent review). Aficionados of CIs point out that they contain more information than *p*-values, better lend themselves to visual displays than *p*-values and are less likely than *p*-values to be misused to draw unwarranted conclusions about hypotheses. To illustrate support for CIs, Harlow (1997), in an introductory chapter for the famous edited book, “What if there were no significance tests?” noted that although there was disagreement among the various chapter authors about the merits of null hypothesis significance testing, there was “unanimous support” for CIs (p. 5). Subsequent authorities have continued to support CIs (e.g., Cumming, 2014: Cumming & Finch, 2005; Fidler & Loftus, 2009; García-Pérez, 2005; Loftus, 1993, 1996; Ranstam, 2012; Young & Lewis, 1997). Meehl (1997) suggested a particularly interesting point of view with respect to CIs. On the one hand, consistent with the Harlow summary, Meehl favored CIs over significance tests. On the other hand, however, Meehl stated that the more important scientific problems had to do with epistemology, and the lack of researchers submitting their theories to risky predictions.

Although most statistically savvy researchers favor CIs over significance tests, CIs also can be criticized. The most popular use of CIs is as an alternative form of significance testing; if the critical value falls outside the CI, the finding is “significant.” When used in this way, CIs fail to improve on traditional significance tests. Alternatively, some have promoted CIs for parameter estimation, but this can be done in a naïve or sophisticated way. A naïve example would be when a researcher computes a sample mean, then constructs a 95% CI around the mean, and concludes that the population mean has a 95% chance of being within the constructed CI. The unfortunate fact is that there is no way to know this probability, and serious frequentists would argue that probabilities are irrelevant, as the parameter either is in the CI or is not. The researcher’s lack of knowledge about whether the parameter is in the interval fails to justify assigning a probability.

But if CIs should not be used as an alternative form of significance testing, nor to assign probabilities with respect to the placement of population parameters, what is the potential contribution? The usual answer given by CI sophisticates is that CIs provide researchers with information about the precision of the data (e.g., Cumming, 2014; Cumming & Finch, 2005; Fidler & Loftus, 2009; Loftus, 1993, 1996; Ranstam, 2012; Young & Lewis, 1997; but see Trafimow, 2018 for an exception). Wide CIs indicate less precision whereas narrow CIs indicate more precision.

There is empirical support for the precision argument. Cumming and Calin-Jageman (2017) described computer simulations keeping track of how often the 95% CI computed in one replication captured the mean in the following replication. They reported that 83% of simulated 95% CIs captured the mean in the following replication. This 83% figure contrasts with the 95% figure; the 95% figure refers to the percentage of 95% CIs that capture an unknown population mean whereas the 83% figure refers to the percentage of 95% CIs that capture a simulated sample mean in the following replication. Nevertheless, 83% might be considered pretty good, even if it is not as good as 95%.

But is the precision argument misdirected? Our intention is to argue that it is. Put briefly, we see the ability of sample-based CIs to capture sample means in following simulations as not very relevant because the estimation goal concerns population parameters and not sample statistics. The more relevant issue, as will become clear in the ensuing discussion, is whether sample-based CIs accurately estimate *a priori* CIs.

'A priori' Confidence Intervals

What are *a priori* CIs? The notion comes out of recent work by Trafimow and his colleagues (Trafimow, 2017; Trafimow & MacDonald, 2017; Trafimow, Wang, & Wang, 2019; Wang, Wang, Trafimow, & Myüz, 2019; see Trafimow, 2019 for a review). The idea is that a researcher can ask, prior to data collection, how close she wants her sample statistics to be to their corresponding population parameters, and what probability (confidence) she wants to have of being that close. For example, consider the simple case where there is one group, the descriptive statistic of interest is the sample mean, and the researcher is concerned about its probability of being close to the population mean. Assuming random sampling from a normal distribution, Trafimow (2017) provided an accessible derivation of Equation 1, where n is the sample size, f is the maximum distance of the sample mean from the population in standard deviation units, and where z_c is the z-score that corresponds to the confidence level one can have of obtaining a sample mean that is within f of the population mean.

$$n = \left(\frac{z_c}{f}\right)^2 \text{ or } f = \frac{z_c}{\sqrt{n}} \text{ or } z_c = f\sqrt{n} \qquad (1)$$

As a quick example, suppose $n = 100$, and we wish to be 95% confident of obtaining a sample mean within f of the population mean. What is f? The z-score that corresponds to 95% is 1.96; thus, $f = \frac{1.96}{\sqrt{100}} = .196$. In sum, when the sample size is 100, there is a 95% probability of obtaining a sample mean within .196 standard deviations of the population mean. Or, had we started with a desire to have a 95% probability of obtaining a sample mean within .196 standard deviations of the population mean, Equation 1 implies that we would need at least 100 participants. That it is possible to compute the sample size needed to achieve prior designations for confidence and precision caused Trafimow and MacDonald (2017) to term this the *a priori* procedure. Ideally, the researcher makes *a priori* specifications for closeness and confidence, uses an equation such as Equation 1 to compute the necessary sample size, and collects that sample size or a larger one. This procedure provides the capability of designing the experiment to have whatever is deemed a satisfactory degree of trust in the sample statistic to be obtained, such as the sample mean.

Although CIs provide the basis for the *a priori* procedure (see Appendices in Trafimow, 2017; Trafimow & MacDonald, 2017 for derivations of equations), these are not sample-based CIs.[1] Rather, the CIs do not depend on the data to be obtained, particularly the standard deviation to be obtained. This is because the standard deviation cancels out in the derivation of Equation 1. Consequently, when using the *a priori* procedure, there is no requirement for the researcher to have an intention to construct a

[1] More complex a priori equations have been developed since Trafimow (2017) and Trafimow and MacDonald (2017); such as by Trafimow, Wang, and Wang (2019), and Wang, Wang, Trafimow, and Myüz (2019).

sample-based CI. To reiterate, the relevant CI is used prior to data collection, to help the researcher design a study where she can trust that the sample statistic to be obtained is a good estimate of the population parameter of interest.

For a computer simulation employing the standard normal distribution (mean equals 0 and standard deviation equals 1), Equation 1 renders possible *a priori* CIs with which many sample-based CIs can be compared. How would one compute an *a priori* CI? As an example, we saw earlier that $n = 100$ implies $f = .196$. Thus, the 95% *a priori* CI, in which 95% of sample means will fall employing the standard normal distribution, when $n = 100$, is $0 \pm .196$.

How well will sample-based CIs approximate *a priori* CIs? If the approximation is good, this would provide a strong reason for researchers to favor sample-based CIs as an important inferential tool in estimation, consistent with the precision argument by CI sophisticates. In contrast, if the approximation is poor, an implication would be that sample-based CIs are not very relevant for researchers interested in estimating population parameters.

There are at least three ways in which sample-based CIs can do well or badly at approximating *a priori* CIs. First, the width of sample-based CIs might provide a good or poor approximation of the width of corresponding *a priori* CIs; this would directly address the precision issue with respect to *width*. Second, the lower limit of sample-based CIs might provide a good or poor approximation of the lower limit of population-based CIs. Third, the upper limit of sample-based CIs might provide a good or poor approximation of the upper limit of population-based CIs. Either of the latter two might be considered as addressing the precision issue with respect to *location*.

Method

The simulation was based on the manipulation of sample size. Sample sizes ranged from 10 to 1,000 increasing by 10 (i.e., 10, 20, ..., 1,000). For the simulation, pseudo-random data were obtained from the standard normal distribution with mean and variance equal to zero and one, respectively. A random seed was set to 12 to ensure the results could be perfectly replicated. The simulation ran 10,000 times for each sample size. Each sample was then subjected to a one-sample *t*-test analysis in MatLab R2015b, which provides sample-based confidence intervals as well as *t*-tests. Both the lower and upper limits of the sample-based confidence interval were recorded for each sample size, and a width was calculated from these two values (see Supplementary Materials).

Results

Before comparing sample-based CIs to *a priori* CIs, it was necessary to calculate *a priori* CIs for each sample size. This was accomplished using Equation 1, under the standard normal distribution, using 95% CIs throughout (so $z_c = 1.96$), as in the example. Subsequently, we arbitrarily selected conservative (2.5%), moderate (5.0%), and liberal (10.0%) criteria with respect to the percentage of sample-based CI widths and limits being reasonable approximations of corresponding *a priori* CI widths and limits. That is, for example, under the conservative criterion, if the width of a sample-based CI was less than 2.5% smaller, or less than 2.5% larger, than the width of the corresponding *a priori* CI, it was deemed “in range.” Figure 1 illustrates the percentages of sample-based confidence interval widths within 2.5%, 5%, or 10% of the corresponding *a priori* confidence interval widths.

Figure 1

The Percentages of Sample-Based Confidence Interval Widths Within 2.5%, 5%, or 10% of the Corresponding A Priori 95% Confidence Interval Widths are Expressed Along the Vertical Axis, Sample Sizes are Expressed Along the Horizontal Axis

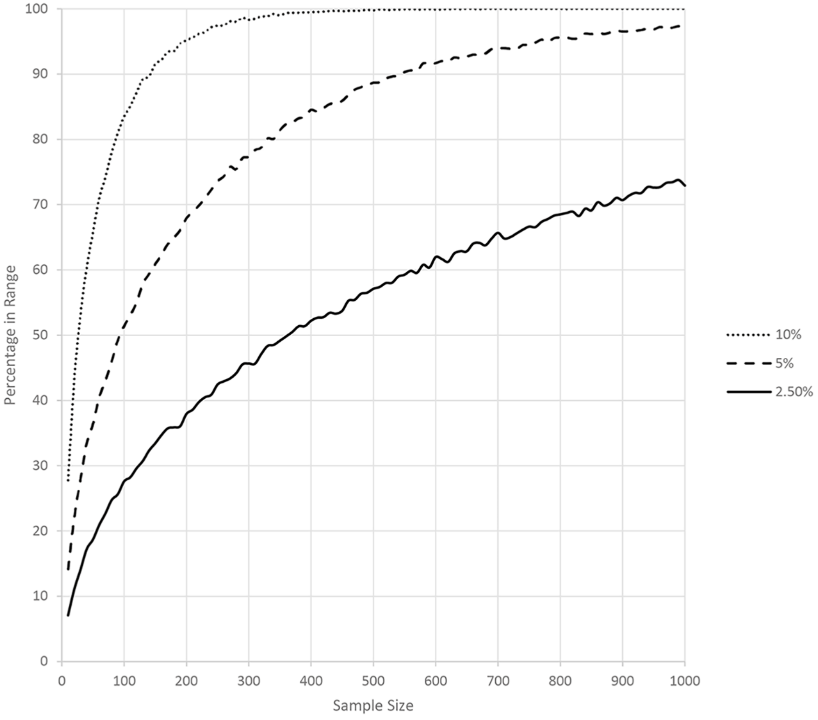
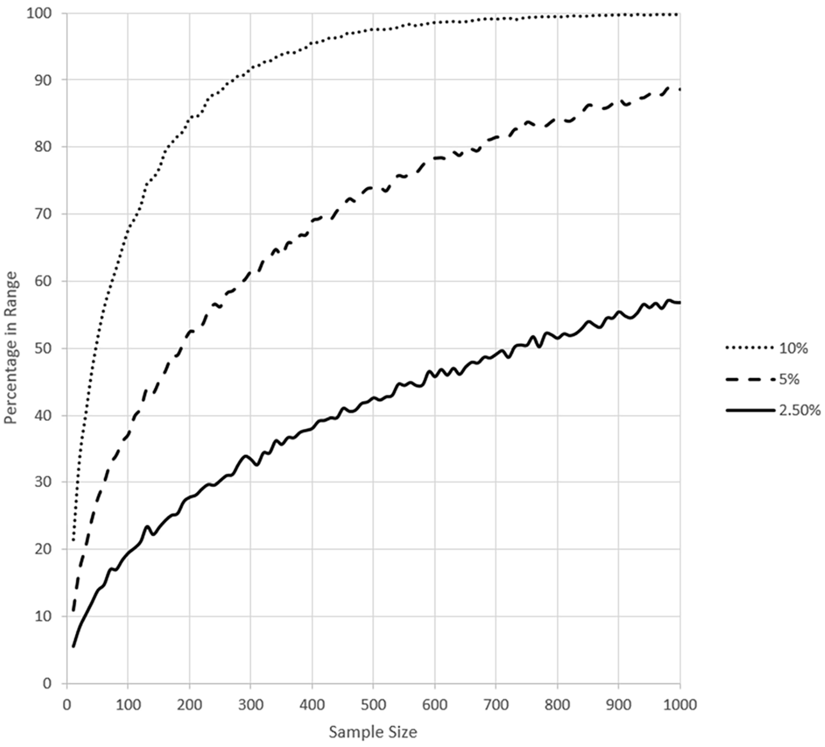


Figure 1 illustrates that under typical sample sizes, sample-based CI widths provide poor approximations of *a priori* CI widths. The estimations improve as sample sizes improve, though using a conservative criterion of 2.5%, even when the sample size reaches 1000, the percentage of sample-based CI widths within the criterion fails to exceed 73%. With respect to location, the data pertaining to lower limits and upper limits are very similar, so Figure 2 just illustrates lower limits.

Figure 2

The Percentages of Lower Limits of Sample-Based Confidence Intervals Within 2.5%, 5%, or 10% of the Lower Limits of Corresponding 95% A Priori Confidence Intervals are Expressed Along the Vertical Axis, Sample Sizes are Expressed Along the Horizontal Axis

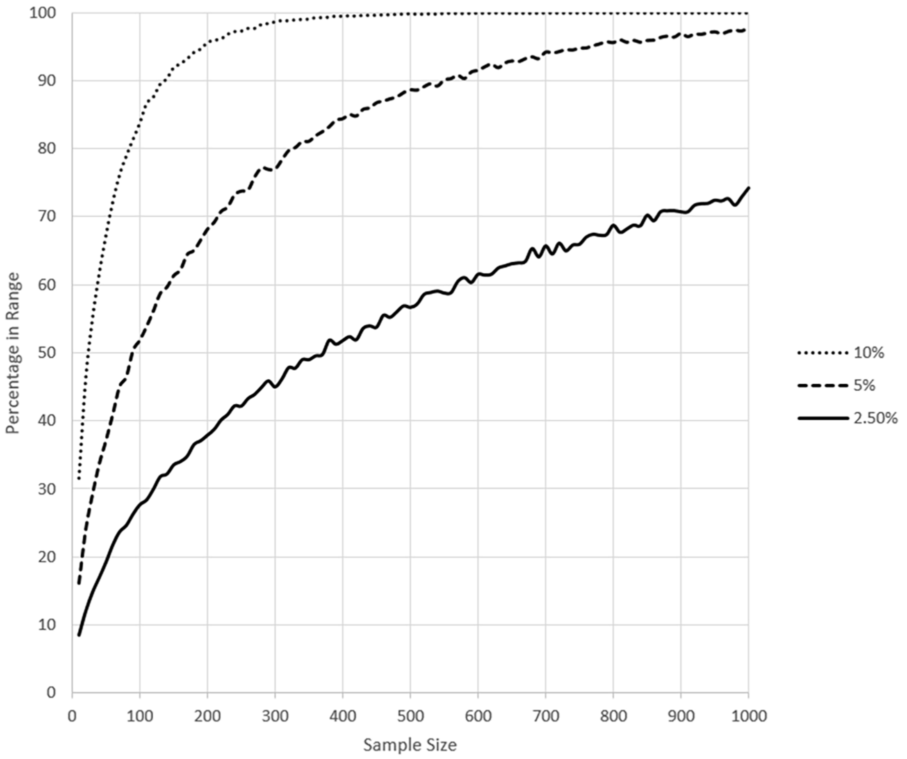


Also, because lower limits are single numbers, in contrast to widths being intervals, there was no way to calculate a percentage of a lower limit in the way that we did for widths, and we simply used absolute numbers to create ranges. For example, in the 2.5% case, we determined the percentage of sample-based lower limits between each *a priori* lower limit plus or minus .025. Figure 2 illustrates that, at typical sample sizes, sample-based locations poorly estimate *a priori* ones, with improvement as sample sizes increase. A cautionary note is that because the process of determining "in range" for widths and lower limits necessarily differed, Figure 2 should not be uncritically compared with Figure 1, though their main implications are similar.

A possible reason the findings were so unflattering to sample-based CIs is because we used 95% CIs that can be considered extreme.[2] To address this issue, we performed analyses resembling the foregoing; but using 50% CIs instead of 95% CIs. Figure 3 is analogous to Figure 1 and concerns CI widths; whereas Figure 4 is analogous to Figure 2 and concerns CI locations. The pessimistic implications of Figures 1 and 2 remain when considering Figures 3 and 4.

Figure 3

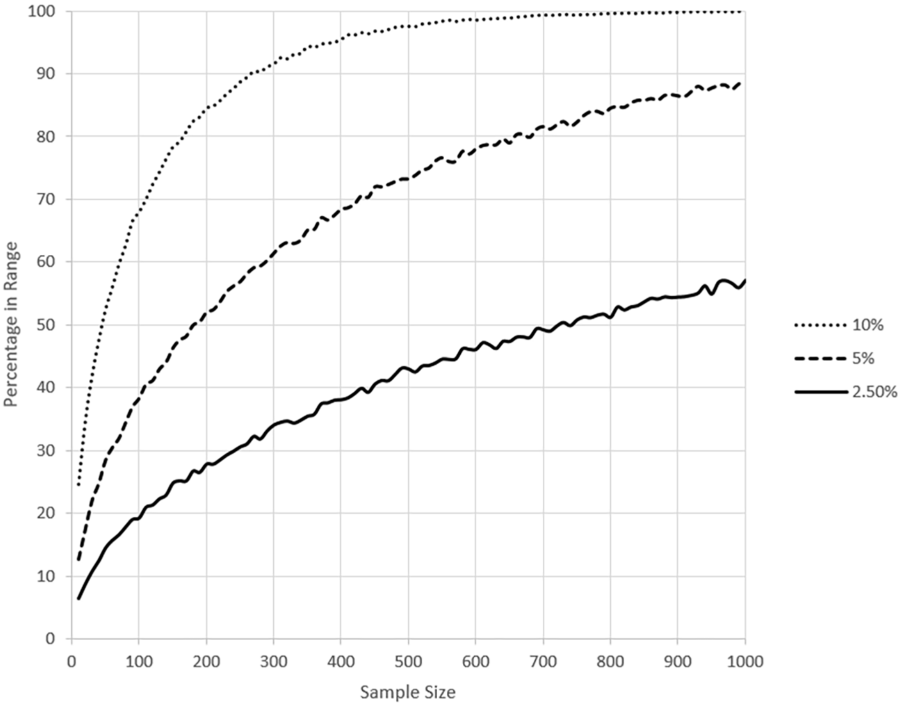
The Percentages of Sample-Based Confidence Interval Widths Within 2.5%, 5%, or 10% of the Corresponding A Priori 50% Confidence Interval Widths are Expressed Along the Vertical Axis, Sample Sizes are Expressed Along the Horizontal Axis



[2] We thank an anonymous reviewer for suggesting this possibility.

Figure 4

The Percentages of Lower Limits of Sample-Based Confidence Intervals Within 2.5%, 5%, or 10% of the Lower Limits of Corresponding 50% A Priori Confidence Intervals are Expressed Along the Vertical Axis, Sample Sizes are Expressed Along the Horizontal Axis



Although our main points have been made, there is one final matter. It might be useful to gain an idea of the effect of sample sizes on empirical distributions in a more general way than is conveyed by Figures 1, 2, 3, and 4.[3] One way to accomplish this is to consider the absolute value of the mean difference score, between each empirically generated range and the a priori range, within each sample size. The expectation is that mean difference scores should decrease as sample sizes increase. A second way is to consider the standard deviations of empirically generated ranges, within each sample size, which should decrease as sample sizes increase. Figure 5 illustrates how mean difference scores decrease as sample sizes increase whereas Figure 6 illustrates how standard deviations decrease as sample sizes increase.

[3] We thank an anonymous reviewer for suggesting this possibility.

Figure 5

Mean Difference Scores for Ranges are Expressed Along the Vertical Axis as a Function of Sample Sizes Along the Horizontal Axis

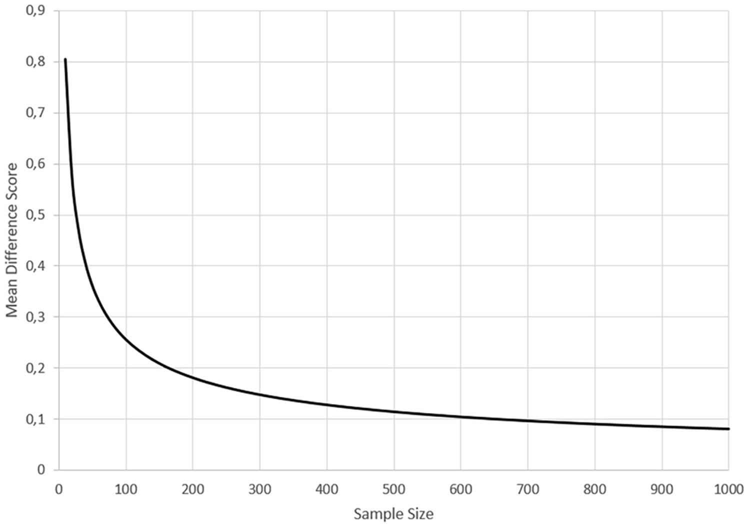
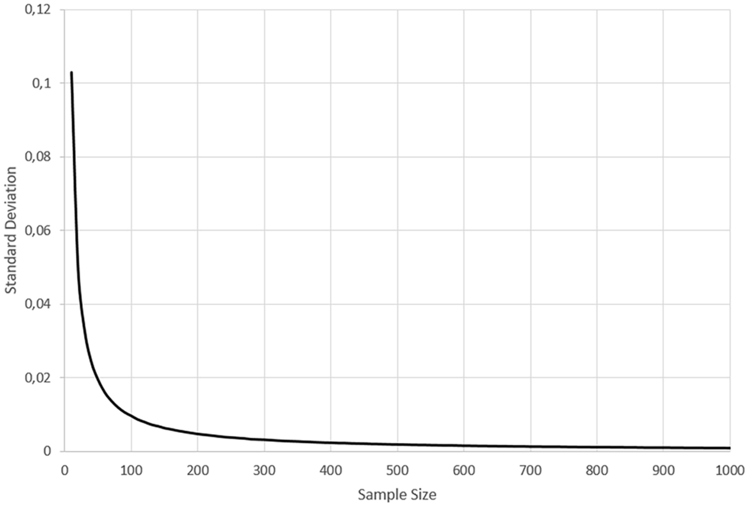


Figure 6

Standard Deviations for Empirical Ranges are Expressed Along the Vertical Axis as a Function of Sample Sizes Expressed Along the Horizontal Axis



Illustrations analogous to those in the foregoing paragraph can be applied to lower limits too.[4] We considered the difference between empirically generated lower limits and the lower limit of the a priori interval, at each sample size. Figure 7 illustrates how mean difference scores decrease as sample sizes increase.

Figure 7

Mean Difference Scores for Lower Limits are Expressed Along the Vertical Axis as a Function of Sample Sizes Along the Horizontal Axis

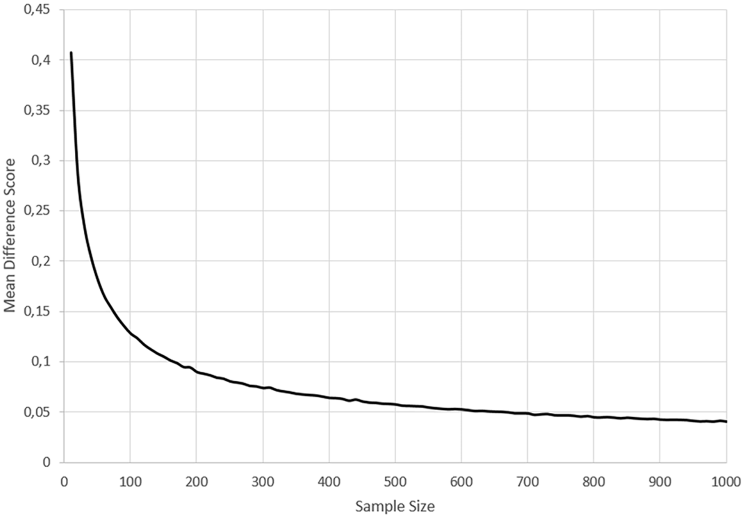
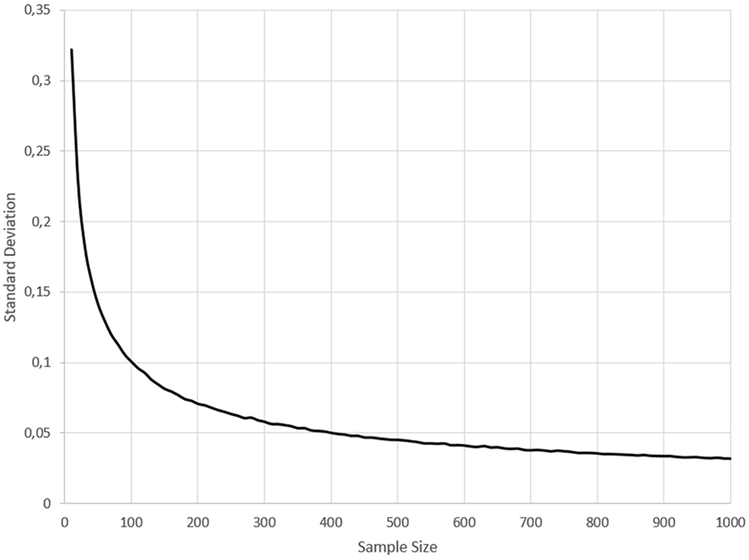


Figure 8 illustrates how standard deviations of empirically generated lower limits decrease as sample sizes increase. Together, Figures 5, 6, 7, and 8 provide a more general view of how increasing sample sizes benefits empirically generated distributions. This general picture is quite consistent with the implications of Figures 1, 2, 3, and 4.

[4] As before, lower limits and upper limits generate similar data, so we remained with lower limits.

Figure 8

Standard Deviations for Empirical Lower Limits are Expressed Along the Vertical Axis as a Function of Sample Sizes Expressed Along the Horizontal Axis



Finally, as a check on the programming, at all sample sizes, we calculated the percentages of CIs that enclosed the population mean using 95% sample-based CIs and 50% sample-based CIs. Supporting that the programming was valid, at all sample sizes, all percentages pertaining to 95% CIs and 50% CIs were very close to 95% and 50%, respectively.[5] However, we reiterate a point made earlier. Knowing the percentage of sample-based CIs that enclose the population mean does not justify drawing a conclusion about the probability of a population mean being within a single sample-based CI. There is no way to know this latter probability.

Discussion

CIs are not much of an improvement over significance tests if they are merely to be used as significance tests. Nor can CIs be used to estimate the probability that the population parameter of interest (e.g., the population mean) is within the constructed CI. Sophisticated users of CIs know these points and argue instead that CIs are useful for estimating the precision of the data. On the contrary, however, Figures 1 and 3 illustrate that,

[5] We also investigated medians but found nothing sufficiently interesting to be reported here.

under typical sample sizes, a distressingly small percentage of widths of sample-based CIs are within a range of 2.5% of widths of corresponding *a priori* CIs. And although we labeled 2.5% as "conservative," remember that the criterion refers to 2.5% as either an underestimate or an overestimate, for a spread of 5%. Thus, our "conservative" criterion could be argued to be quite traditional, and not particularly conservative. Figures 2 and 4 illustrate a similar implication, but with respect to locations rather than widths. More generally, Figures 1, 2, 3, and 4 cast serious doubt on the vaunted ability of sample-based CIs to provide good precision estimates in terms of widths or locations, at typical sample sizes. Furthermore, Figures 4, 5, 6, 7, and 8 show, more generally, how increasing sample sizes benefits empirically generated distributions. But Figures 4, 5, 6, 7, and 8 also imply, consistent with Figures 1, 2, 3, and 4, that sample-based confidence intervals are not very precise at typical sample sizes.

Well, then, if sample-based CIs do not work well for precision, how do they contribute to statistical inference? Our answer is that researchers should eschew sample-based CIs in favor of *a priori* thinking. That is, researchers should decide, before collecting data, how close they want their sample statistics to be to their corresponding population parameters; and what probability they wish to have of being that close. In the one-sample case, Equation 1 can be used to compute the necessary sample size, though more complex equations are needed for more complex designs or more complex comparisons (Trafimow & MacDonald, 2017; Trafimow, Wang, & Wang, 2019; Wang, Wang, Trafimow, & Myüz, 2019). Once the necessary minimum sample size is determined, researchers can obtain the sample, or a larger one, and then directly take the obtained sample statistics as being satisfactory estimates of the desired population parameters. There is no need to perform significance tests nor sample-based CIs. More generally, although we favor researchers thinking in terms of intervals, these should be *a priori*, not sample-based.

We end by admitting an important limitation of the *a priori* procedure. And that limitation is that there are commonly used analyses for which no *a priori* equations have yet been developed. For example, in medicine, researchers may be interested in cure rates under different conditions; but no *a priori* equations have yet been invented for proportions or for hazard ratios. In many business areas, such as management and economics, regression analyses, path analyses, or more complex types of causal modeling are common. But no *a priori* equations have been invented for these sorts of analyses. Finally, although Trafimow et al. (2019) and Wang et al. (2019) have invented *a priori* equations for skew-normal distributions, as opposed to settling for the smaller family of normal distributions, there are many other families of distributions for which no *a priori* equations have been invented. In cases such as these, where the *a priori* procedure is not yet an option; sample-based CIs may be the best frequentist inferential statistical option open, despite the present demonstrations of their inaccuracy. Alternatively, when there are no applicable *a priori* equations, researchers might decide not to perform inferential statistics. We make no attempt here to tell researchers what to do but merely stress that sample-based CIs are often inaccurate. Possibly, as more *a priori* equations are developed in the coming years, sample-based CIs will be gradually phased out.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: The authors have no support to report.

Supplementary Materials

MatLab syntax for the simulation is available via the PsychArchives repository (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Trafimow, D., & Uhalt, J. (2020). *Supplementary materials [code] to: The inaccuracy of sample-based confidence intervals to estimate a priori ones.* PsychOpen. https://doi.org/10.23668/psycharchives.3006

References

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open Science, and beyond.* New York, NY, USA: Taylor and Francis Group.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist, 60*(2), 170-180. https://doi.org/10.1037/0003-066X.60.2.170

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie. The Journal of Psychology, 217*(1), 27-37.

García-Pérez, M. A. (2005). On the confidence interval for the binomial parameter. *Quality & Quantity, 39*, 467-481. https://doi.org/10.1007/s11135-005-0233-3

Harlow, L. L. (1997). Significance testing introduction and overview. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1-17). Mahwah, NJ, USA: Erlbaum.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science.* Los Angeles, CA, USA: Sage Publications.

Loftus, G. R. (1993). A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instruments, & Computers, 25*(2), 250-256. https://doi.org/10.3758/BF03204506

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*(6), 161-171. https://doi.org/10.1111/1467-8721.ep11512376

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425) Mahwah, NJ, USA: Erlbaum.

Ranstam, J. (2012). Why the *p*-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage, 20*(8), 805-808. https://doi.org/10.1016/j.joca.2012.04.001

Trafimow, D. (2017). Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. *Educational and Psychological Measurement, 77*(5), 831-854. https://doi.org/10.1177/0013164416667977

Trafimow, D. (2018). Confidence intervals, precision and confounding. *New Ideas in Psychology, 50*, 48-53. https://doi.org/10.1016/j.newideapsych.2018.04.005

Trafimow, D. (2019). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics, 7*(2), Article 26. https://doi.org/10.3390/econometrics7020026

Trafimow, D., & MacDonald, J. A. (2017). Performing inferential statistics prior to data collection. *Educational and Psychological Measurement, 77*(2), 204-219. https://doi.org/10.1177/0013164416659745

Trafimow, D., Wang, T., & Wang, C. (2019). From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement, 79*(1), 129-150. https://doi.org/10.1177/0013164418764801

Young, K. D., & Lewis, R. J. (1997). What is confidence? Part 1: The use and interpretation of confidence intervals. *Annals of Emergency Medicine, 30*(3), 307-310. https://doi.org/10.1016/S0196-0644(97)70166-5

Wang, C., Wang, T., Trafimow, D., & Myüz, H. A. (2019). Desired sample size for estimating the skewness under skew normal settings. In V. Kreinovich & S. Sriboonchitta (Eds.), *Structural changes and their economic modeling* (pp. 152-162). Cham, Switzerland: Springer.

Ziliak, S. T., & McCloskey, D. N. (2016). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor, MI, USA: University of Michigan Press.

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology Information (ZPID), Germany.