Using the Coefficient of Confidence to Make the Philosophical Switch From A Posteriori to A Priori Inferential Statistics

Educational and Psychological
Measurement

2017, Vol. 77(5) 831–854

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416667977

journals.sagepub.com/home/epm

David Trafimow¹

Abstract

There has been much controversy over the null hypothesis significance testing procedure, with much of the criticism centered on the problem of inverse inference. Specifically, p gives the probability of the finding (or one more extreme) given the null hypothesis, whereas the null hypothesis significance testing procedure involves drawing a conclusion about the null hypothesis given the finding. Many critics have called for null hypothesis significance tests to be replaced with confidence intervals. However, confidence intervals also suffer from a version of the inverse inference problem. The only known solution to the inverse inference problem is to use the famous theorem by Bayes, but this involves commitments that many researchers are not willing to make. However, it is possible to ask a useful question for which inverse inference is not a problem and that leads to the computation of the coefficient of confidence. In turn, and much more important, using the coefficient of confidence implies the desirability of switching from the current emphasis on a posteriori inferential statistics to an emphasis on a priori inferential statistics.

Keywords

coefficient of confidence, squared coefficient of confidence, confidence interval, standard error of the mean, a posteriori statistics, a priori statistics

¹New Mexico State University, Las Cruces, NM, USA

Corresponding Author:

David Trafimow, Department of Psychology, New Mexico State University, MSC 3452, PO Box 30001, Las Cruces, NM 88003-8001, USA.

Email: dtrafimo@nmsu.edu

Many researchers have objected to the null hypothesis significance testing procedure (NHSTP) on the grounds that it involves an unjustified inverse inference. Researchers compute p , which is the probability of the finding (or one more extreme) given the null hypothesis but make a rejection decision about the null hypothesis given p . Because the probability of A given B is not necessarily the same as the probability of B given A , inverse inferences are generally logically invalid. In the case of “statistically significant” findings, obtaining a low value for p does not justify an inverse inference that the null hypothesis is unlikely to be true. Often, alongside objections to the NHSTP, there are admonitions to place greater emphasis on confidence intervals. However, confidence intervals also are problematic because of a similar inverse inference problem. I propose an alternative way to think about intervals that is similar, but not exactly the same, as current practice. I hope to demonstrate that a subtle difference in orientation, from a posteriori to a priori inferential statistical thinking, can have dramatic consequences for how researchers use statistical procedures. In general, frequentist statistical theory involves reasoning from assumptions about populations to inferences about repeated samples. Therefore, whenever frequentist statistical theory is used in the inverse direction—from known facts about a sample to inferences about a population—one necessarily commits an inverse inference fallacy. The NHSTP and confidence intervals exemplify the wrong-headed use of frequentist statistical theory to reason from samples to populations, and force the researcher into an inverse inference fallacy.

The Null Hypothesis Significance Testing Procedure

Many others have inveighed against the NHSTP and so I will present an abbreviated discussion here (e.g., Bakan, 1966; Berkson, 1938; Cohen, 1990, 1994; Fidler & Loftus, 2009; Gigerenzer, 1993; Hogben, 1957; Loftus, 1996; Lykken, 1968; Meehl, 1967, 1978; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997; Thompson, 1992; Trafimow, 2003, 2005, 2006; Trafimow & Rice, 2009). Possibly, the most often made objection is that the NHSTP is logically invalid. This argument takes at least two forms. In what might be termed the *modus tollens* fallacy, the NHSTP takes a logical form that rests on deductive reasoning when only probabilistic reasoning is possible. To see how this works, imagine the following syllogism:

1. If the null hypothesis were true, it would be impossible to obtain the finding.
2. The finding is obtained.
- C. Therefore, the null hypothesis *cannot* be true.

In this case, obtaining the finding does in fact validly disconfirm the null hypothesis by the logic of *modus tollens*. But now consider the following syllogism:

1. If the null hypothesis were true, it would be *unlikely* to obtain the finding.
2. The finding is obtained.

C. Therefore, the null hypothesis is *unlikely* to be true.

Because *modus tollens* does not apply to probabilities, this latter syllogism is logically invalid. To illustrate the invalidity in a dramatic way, consider the following example.

1. If a person is an American, that person is *unlikely* to be president
2. Obama is president
3. Therefore, it is *unlikely* that Obama is an American

A second way to describe the logical invalidity of the NHSTP is via the inverse inference fallacy. According to this fallacy, although a low p value renders the finding (or one more extreme) unlikely given that the null hypothesis is true, it is an unwarranted inference to conclude that the null hypothesis is unlikely to be true because one has obtained a finding that is associated with a low p value. Stated more generally, obtaining the probability of A given B does not logically permit a conclusion to be drawn about the probability of B given A . Whether described via the *modus tollens* fallacy or the inverse inference fallacy, the point is that the information that p values carry with them is insufficient to permit a logically valid inference about the probability of the null hypothesis being true or false. And if there is no way to make a valid inference about the probability of the null hypothesis being true or false, valid acceptance or rejection of the null hypothesis is thereby rendered problematic.

A possible way out of this trap might be to argue that hypotheses are either true or false but do not have probabilities (other than 0 or 1). The main problem with this counterargument is that although it provides a nice starting point to criticize approaches that depend on hypotheses having probabilities (e.g., Bayesian approaches), it fails to provide a valid reason for accepting or rejecting hypotheses. If hypotheses do not have probabilities, how does one traverse the logical gap between the probability of the finding given the null hypothesis, and rejecting or not rejecting the null hypothesis? There is an additional problem. If one listens for a sufficiently long time, some researchers who argue that hypotheses do not have probabilities will eventually make a statement such as the following: "As the p value decreases, the null hypothesis is less likely to be true" (citation omitted to protect the guilty). But this is tantamount to associating probabilities with hypotheses after asserting that this is improper.

In addition to the basic problem of logical invalidity, the NHSTP has been criticized on auxiliary grounds. One problem that has received much attention is the use of so called "point" null hypotheses (e.g., Loftus, 1996; Meehl, 1978). It is common practice for researchers to propose a null hypothesis of no effect or an effect of a specific value (i.e., a "point") against an alternative hypothesis that specifies a range of values. Because there is only one way for the null hypothesis to be correct and many ways for the alternative hypothesis to be correct, performing the NHSTP in this way stacks the deck in favor of the researcher's alternative hypothesis (Meehl, 1978).

Arguably, this problem is not inherent in the NHSTP itself but rather in how it is used. In principle, it would be possible for researchers to use a null hypothesis that specifies a range rather than a point, and in this manner provide a fair contest between the null and alternative hypothesis. But point null hypotheses remain common practice.

Even if researchers were not to stack the deck (or where there would be other reasons why the null hypothesis actually does have a reasonable chance of being true), there nevertheless remains the logical issue explained earlier, whether expressed as the inverse inference fallacy or the *modus tollens* fallacy. In addition, using Bayes' theorem, it is easy to show that even when the p value is an arbitrarily small number, the probability of the null hypothesis can be any value between 0 and 1, depending on the prior probability of the null hypothesis and the probability of the finding given that the null hypothesis is not true (see especially Trafimow, 2003, Figure 1). So in addition to the inverse inference or *modus tollens* fallacy, there also is quantitative invalidity.

Worse yet, because a small prior probability of the null hypothesis is necessary for a small posterior probability of the null hypothesis, using the NHSTP puts the researcher in a dilemma. Consider that information gain can be defined as the difference between the prior and posterior probability of the null hypothesis. Well, then, if the prior probability already is small, there can be little change in the downward direction. Consequently, validity of inference and information gain pull in opposite directions (see Trafimow, 2003, Figure 2). That is, a small prior probability is necessary for validly rejecting the null hypothesis but then information gain is minimized. Or, if the prior probability of the null hypothesis is a large number, information gain increases but at the cost of being unable to validly reject the null hypothesis.

Finally, Krueger (2001) has argued that despite the invalid logic of the NHSTP, p values are at least correlated with probabilities of null hypotheses. Although this is true, Trafimow and Rice (2009) showed that this correlation is small and becomes smaller still when dichotomous accept-reject decisions are made. In summary, the NHSTP is difficult to defend.

Confidence Intervals

Are confidence intervals a good alternative? As mentioned in almost all introductory statistics textbooks, confidence intervals provide another way to perform the NHSTP. If one computes a confidence interval, and the value specified by the null hypothesis lies outside the computed interval, this is equivalent to a test of the null hypothesis. Used in this way, confidence intervals suffer from problems similar to those engendered by the NHSTP, especially the inverse inference fallacy (Fidler & Loftus, 2009).

But confidence intervals do not have to be used in this way. They can be used for parameter estimation, though this can mean different things to different people. For example, Fidler and Loftus (2009) noted that the usual interpretation of "parameter estimation" is that obtaining a 95% confidence interval implies that the population

parameter (usually the population mean) has a 95% chance of being in the obtained interval. However, Fidler and Loftus also indicated that this interpretation is unsound. In the case of repeated random samples obtained from a population, confidence intervals jump around whereas the population parameters remain constant. From the point of view that the population parameter either is or is not in the interval obtained from a random sample, such that talk of probabilities is inappropriate, it is obvious that no conclusion about the probability that the parameter is within an interval can be drawn. For those with less strict conceptualizations of probability, the probability that the parameter is within an interval is meaningful. However, the determination of that probability requires a specifiable prior probability distribution of population parameters. So, again, the 95% confidence interval fails to validly permit a conclusion about the probability of the population parameter being within the obtained interval. Thus, using confidence intervals to draw conclusions about the probability that the parameter of interest is within the stated interval constitutes yet another instance of the inverse inference fallacy.

An alternative is to argue that confidence intervals provide the researcher with an idea about the precision with which the sample statistic estimates the population parameter. As the vast majority of researchers focus on means, a specific version of this argument would be that a confidence interval around the sample mean provides the researcher with an idea about the precision with which the population mean is estimated. A small confidence interval suggests that the sample mean is a more precise estimate of the population mean than a large confidence interval would suggest.

All else being equal, as the sample size increases, the standard error of the mean decreases, and so does the spread of confidence intervals. Doubtless, having a larger sample size and smaller standard error of the mean is better than having a smaller sample size and larger standard error of the mean, all else being equal. But making the leap of faith to the conclusion that confidence intervals constitute a fully valid measure of precision of an estimate is not justified.

Morey et al. (2016) provided a nice description of why confidence intervals are invalid as a measure of estimate precision and I quote this below at length:

There is no necessary connection between the precision of an estimate and the size of a confidence interval. One way to see this is to imagine two researchers—a senior researcher and a PhD student—are analyzing data of 50 participants from an experiment. As an exercise for the PhD student's benefit, the senior researcher decides to randomly divide the participants into two sets of 25 so that they can separately analyze half the data set. In a subsequent meeting, the two share with one another their Student's *t* confidence intervals for the mean. The PhD student's 95% CI is 52 ± 2 , and the senior researcher's 95% CI is 53 ± 4 . The senior researcher notes that their results are broadly consistent, and that they could use the equally weighted mean of their two respective point estimates, 52.5, as an overall estimate of the true mean. The PhD student, however, argues that their two means should not be evenly weighted: she notes that her CI is half as wide and argues that her estimate is more precise and should thus be weighted more heavily. Her advisor notes that this cannot be correct, because the estimate from unevenly weighting the two means would be different

from the estimate from analyzing the complete data set, which must be 52.5. The PhD student's mistake is assuming that CIs directly indicate post-data precision. Later, we will provide several examples where the width of a CI and the uncertainty with which a parameter is estimated are in one case inversely related, and in another not related at all. (p. 105)

In summary, many of the researchers who have objected to the NHSTP have supported the use of confidence intervals based on the following:

- They can be used as a substitute for the NHSTP
- They provide an interval where the population mean is likely to be with a known probability
- They provide a valid way to index the precision to which one has estimated a population parameter

However, we have seen that none of these touted advantages actually is so.

Bayesian Procedures

There is only one way known to solve the inverse inference problem and that is to use the famous theorem by Bayes (e.g., Edwards, Lindman, & Savage, 1963). If one uses this theorem to find the posterior probability of a hypothesis (e.g., the null hypothesis), one has to specify a prior probability of that hypothesis to make the Bayesian machinery run. Or to use the theorem to find an interval that contains the population parameter with a known probability, one has to specify a prior probability distribution. Either way, unless the requisite prior data are available (e.g., by virtue of a relevant epidemiological study), the prior probability specification necessarily will be either subjective (in the sense that it is based on a person's opinion) or artificial (in the sense that it is based on a procedure for generating numbers rather than on actual knowledge of relative frequencies). For researchers who are not troubled by subjectivity or artificiality, a Bayesian procedure could be ideal because it solves the inverse inference problem. However, researchers who consider Bayesian prior probability specifications to be problematic are unlikely to be impressed with Bayesian solutions. As the cliché has it, one can have good premises and bad logic associated with the NHSTP or one can have bad premises and good logic associated with Bayesian procedures.

Of course, this is an oversimplification. There are different ways of being subjective or artificial, and it is possible to argue that some of these ways are better than others. A complex discussion of the different ways of making prior probability specifications, and the advantages and disadvantages of each of them, is beyond the present scope. At present, without making a commitment to the desirability of different proposals for making Bayesian prior probability specifications, it is sufficient merely to point out that many researchers do not accept any of them (e.g., Fisher, 1973; Mayo, 1996; Popper, 1983; Suppes, 1994), and so it would be useful to have a statistical procedure that is not Bayesian.

Finally, there are Bayesian procedures that do not depend on prior probability specifications. For example, there are Bayes' factors that provide the researcher with an idea of the strength of the evidence in favor of one hypothesis versus another hypothesis. An advantage of Bayes' factors is that they do not depend on prior probability specifications. A disadvantage is that they do not address the inverse inference problem, which, arguably, removes the big gains that many consider to be the promise of Bayesian reasoning. Another potential disadvantage is that Bayes' factors can be influenced importantly by how the hypotheses are stated (Mayo, 1996; Trafimow, 2006). As with the prior probability specification issue, I will remain noncommittal with respect to Bayes' factors and be content merely to indicate that they are controversial.

The Coefficient of Confidence

We have seen that the NHSTP and confidence intervals are plagued by the inverse inference problem. Bayesian procedures solve the inverse inference problem but at a price in questionable prior probability specifications that many researchers are not willing to pay. Is there a way out? In one sense there is not; there is no solution to the inverse inference problem without using a Bayesian procedure. In another sense there is a way out if we could find a different but useful question to ask. And there is a different and useful question. To move in the direction of finding this question, we might commence by asking why we collect a reasonably sized sample of data in the first place. An answer might be that we hope that the sample statistic will be close to the population parameter. Typically, the population parameter of interest is the mean and so the hope is that the sample mean is close to the population mean. To see that this matters, suppose that Laplace's demon (who knows everything) appeared and pronounced that our sample statistics are, and would be into the indefinite future, unrelated to the population parameters they are supposed to estimate so that sample means have nothing to do with population means. It would be no exaggeration to say that the demon's pronouncement would constitute the most important crisis in the history of social science.

Fortunately, Laplace's demon has not appeared nor made a pronouncement. But considering the possibility of the demon's appearance and pronouncement dramatizes the importance that researchers place in being able to be confident that sample means are close to population means. So we are left with a question: "How can researchers be *confident* that the sample mean is *close* to the population mean?" But this justification can be parsed into two questions. "How confident is 'confident'?" "How close is 'close'?"

To see how these questions suggest a shift in how we think about confidence intervals, let us commence with the question, "How close is 'close'?" Likely, this will depend on the substantive research area but I emphasize that the researcher ought to define this *a priori* (before collecting the data). In the scheme I am proposing, the researcher specifies closeness in terms of a fraction of the population standard

deviation within which the researcher wishes the sample mean to be close to the population mean. For example, the researcher might consider it acceptable for the sample mean to be within three tenths of a standard deviation of the population mean.

Let us now consider the question, “How confident is ‘confident’?” Again, in the scheme I am proposing, the onus is on the researcher to specify a desired level of confidence that the sample mean is within the specified distance of the population mean. Thus, before collecting any data, the researcher specifies the fraction of a standard deviation that he or she will consider to be “close” and also the level of confidence desired that the distance between the sample mean and the population mean will be within the specified fraction of a standard deviation.

Given that one has specified the desired fraction of a standard deviation and the desired confidence that the distance between the sample mean and population mean is within that value, it is easy to progress further using trivially easy mathematics based on the usual assumption of a normal distribution. This seems like a good place to state that the coefficient of confidence to be derived (see especially Equations 7 and 8) is not new but has been in textbooks for many years (e.g., Hays, 1994; also see Harris & Quade, 1992). The contribution I wish to make is not the introduction of the coefficient of confidence *per se* but rather the philosophical shift based on it that, until now, statisticians and philosophers have failed to see. The necessity of this philosophical shift cannot be understood without understanding the coefficient of confidence in great detail and so the philosophical presentation is delayed until the Discussion section.

Derivation of the Coefficient of Confidence

Based on random sampling from a normally distributed population, suppose that we wish to set an interval around $(M - \mu)$ with boundaries based on fractions (f) of a standard deviation (σ) designated as c in Equation 1:

$$p(-f\sigma < M - \mu < f\sigma) = c \quad (1)$$

That $M - \mu$ is normally distributed implies that the mean of $M - \mu$ is 0 and the standard deviation is $\frac{\sigma}{\sqrt{N}}$ where N is the sample size (e.g., Hays, 1994). Letting $z_c = \Phi^{-1}\left(\frac{1+c}{2}\right)$, Equation 2 follows:

$$p\left(-z_c \frac{\sigma}{\sqrt{n}} < M - \mu < z_c \frac{\sigma}{\sqrt{n}}\right) = c \quad (2)$$

In words, z_c is the z-score corresponding to the upper bound of the $(100*c)\%$ confidence interval around $(M - \mu)$. Because the left half of Equations 1 and 2 both equal c , they can be set equal to each other to arrive at Equation 3:

$$p(-f\sigma < M - \mu < f\sigma) = p\left(-z_c \frac{\sigma}{\sqrt{n}} < M - \mu < z_c \frac{\sigma}{\sqrt{n}}\right) \quad (3)$$

Equations 4 and 5 follow:

$$f\sigma = z_c \frac{\sigma}{\sqrt{n}} \quad (4)$$

$$-f\sigma = -z_c \frac{\sigma}{\sqrt{n}} \quad (5)$$

By cancelling σ , we arrive at Equation 6:

$$f = \frac{z_c}{\sqrt{N}} \quad (6)$$

Algebraic rearrangement gives Equation 7:

$$\sqrt{N} = \frac{z_c}{f} \quad (7)$$

And squaring both sides of Equation 7 renders Equation 8:

$$N = \left(\frac{z_c}{f} \right)^2 \quad (8)$$

It will be convenient to be able to refer to the quantities $\frac{z_c}{f}$ and $\left(\frac{z_c}{f} \right)^2$ in words, so these will be termed the *coefficient of confidence* (COC) and the *squared coefficient of confidence* (SCOC), respectively. Using these terms, Equation 8 states that the SCOC gives the necessary value of N needed to have a specified probability that the sample mean is within the interval decided upon by the researcher.

Consider an example. Suppose that the researcher wishes to know the sample size needed to have a 97% probability that the sample mean will be within half a standard deviation of the population mean. The z value that corresponds to a confidence interval of 97% is 2.17009 and $f = .5$. The SCOC based on Equation 6 is $\left(\frac{2.17009}{.5} \right)^2 = 19$ (to the nearest whole number) and so that is the sample size that is necessary. For another example, I could have used $f = .1$ and a 99% probability level, in which case the z -value is 2.57583 and the SCOC is 663.

Illustrations

It can be useful to graph equations. Unfortunately, because standard deviation units are scaled very differently from sample sizes, such figures are necessarily misleading. To mitigate this problem somewhat, I provided three figures to illustrate Equation 8. Figure 1 presents a graph with confidence ranging between 50% and 90%, with curves that represent 10% differences (50%, 60%, 70%, 80%, and 90%). It is interesting to read from right to left (as in many Semitic languages) starting at $f = .5$ and moving to increasingly smaller values for f along the horizontal axis. Consider moving from $f = .5$ to $f = .4$ for the curve representing 90% confidence (top curve). It is

Figure 1. Sample size (N) expressed as a function of specified fraction of a standard deviation (f) ranging from .01 to 0.5, and the desired confidence interval. There are different curves, from bottom to top, representing the 50%, 60%, 70%, 80%, and 90% confidence intervals, respectively.

possible to obtain a gain of .1 in precision with very little cost in terms of N (an increase from 11 to 17 participants, for a cost of 6 participants rounded to the nearest whole number). As we move to smaller values for f , the cost in terms of N continues to increase for similar increments in f . For example, when moving from $f = .2$ to $f = .1$, the cost is $272 - 68 = 204$ assuming we remain with the 90% confidence level. Figure 1 also shows that as the confidence level decreases, this effect becomes less pronounced, and we will observe this effect again in Figures 2 and 3.

Figure 2 assumes, consistent with current trends, that researchers are primarily interested in having confidence at a very high level, from 95% to 99% (with 95% being the most popular), with curves that represent 1% differences (95%, 96%, 97%, 98%, and 99%). Note that, relative to Figure 1, the vertical axis in Figure 2 maximizes at a much larger value than in Figure 1. The reason for this is that, at any particular level of precision (f), as the confidence level increases, Equation 8 implies that N also must increase. Note that the effect described in the foregoing paragraph is augmented in Figure 2. As an example, consider the 99% confidence interval in Figure 2 (top curve). When improving precision from .5 to .4, the cost in terms of N is $42 - 27 = 15$. But when improving precision from .2 to .1, the cost in terms of N is $666 - 167 = 499$.

Figure 3 provides what might be considered to be an old-fashioned range of lower levels of confidence, from 50% to 60%, with curves that represent 2% differences (50%, 52%, 54%, 56%, 58%, and 60%). Relative to Figures 1 and 2, the vertical axis maximizes at a much smaller value. In Figure 3, there remains an “elbow” just as in Figures 1 and 2. For example, considering the smallest confidence level illustrated

Figure 2. Sample size (N) expressed as a function of specified fraction of a standard deviation (f) ranging from .01 to 0.5, and the desired confidence interval. There are different curves, from bottom to top, representing the 95%, 96%, 97%, 98%, and 99% confidence intervals, respectively.

Figure 3. Sample size (N) expressed as a function of specified fraction of a standard deviation (f) ranging from .01 to 0.5, and the desired confidence interval. There are different curves, from bottom to top, representing the 50%, 52%, 54%, 56%, 58%, and 60% confidence intervals, respectively.

(50%), the cost in terms of N when improving precision from .5 to .4 is $3 - 2 = 1$. This cost in terms of N when improving precision from .2 to .1 is $45 - 11 = 34$. Thus, the effect we observed in Figures 1 and 2 occurs in Figure 3 as well. However, the effect is greatly attenuated compared to the other figures. In general, as is illustrated

both within and across Figures 1 to 3, the lower the confidence level, the more attenuated the effect and the higher the confidence level, the more augmented the effect.

Arbitrary Cutoffs

The fact that all curves in Figures 1 to 3 have an elbow suggests that it might be worthwhile to consider the notion of a “best buy.” How much is each increment in precision worth considering the increasing price that needs to be paid in terms of N ? One way to consider the issue is in terms of first and second derivatives. The first derivative gives the slope of the curve at any level of precision and confidence one specifies whereas the second derivative gives the change in slope. As one traverses the horizontal axis from lesser precision to greater precision (right to left), the slopes continue to become increasingly negative and the slope change (acceleration) continues to increase. Table 1 provides first and second derivatives for the 95% confidence level as a function of the level of precision. The implications for the required N also are presented. Unfortunately, it is not necessarily clear what constitutes a “best buy” or even a “sufficient buy.” Much might depend on the type of research under consideration. For example, it is easy to obtain many participants in internet surveys, the difficulty increases in laboratory experiments, and the difficulty may increase further yet for research pertaining to rare populations. Arguably, then, the social sciences should be willing to tolerate lesser levels of confidence or precision for some types of research than for other types of research. As an example, using the 95% confidence interval, we might insist that the second derivative exceed 10,000 for internet surveys, 5,000 for laboratory experiments, and 1,000 for experiments involving special populations. The implied levels of precision are .21, .26, and .39, respectively; and the implied levels of N are 87, 57, and 25, respectively. These values are by way of example and should not be taken too seriously. The setting of arbitrary cutoff points, if they turn out to be needed, is something that, perhaps, should be agreed upon by researchers in different social science areas. Finally, in research where differences between means or increasingly complex interactions are of interest, it might be reasonable to demand increasingly better precision (smaller fractions of a standard deviation), which implies increased sample sizes if the level of confidence remains constant.

What If the Population Is Not Normally Distributed?

Because Equations 7 and 8 provide the basis for the argument I eventually will make for a priori statistical procedures, that constitute a major shift in statistical thinking, it is important first to address the obvious objection that these equations do not apply to nonnormal populations. As is well-known, the central limit theorem is a powerful theorem that indicates that even if this assumption is wrong, the sampling distribution of means becomes increasingly normal as the sample size increases. The proposed COC procedure exploits the power of the central limit theorem to the fullest. What sample size is necessary to take advantage of the central limit theorem in the context of the COC procedure? I investigated this issue by performing computer simulations using a

Table 1. First derivatives [$f'(x)$] and second derivatives [$f''(x)$] are presented as a function of the precision when the confidence level is set at 95%.

Precision	$f'(x)$	$f''(x)$	N
0.1	-7,683	230,496	384
0.11	-5,773	157,432	317
0.12	-4,446	111,157	267
0.13	-3,497	80,703	227
0.14	-2,800	60,000	196
0.15	-2,276	45,530	171
0.16	-1,876	35,170	150
0.17	-1,564	27,597	133
0.18	-1,317	21,957	119
0.19	-1,120	17,687	106
0.2	-960	14,406	96
0.21	-830	11,852	87
0.22	-722	9,839	79
0.23	-631	8,236	73
0.24	-556	6,947	67
0.25	-492	5,901	61
0.26	-437	5,044	57
0.27	-390	4,337	53
0.28	-350	3,750	49
0.29	-315	3,259	46
0.3	-285	2,846	43
0.31	-258	2,496	40
0.32	-234	2,198	38
0.33	-214	1,944	35
0.34	-195	1,725	33
0.35	-179	1,536	31
0.36	-165	1,372	30
0.37	-152	1,230	28
0.38	-140	1,105	27
0.39	-130	996	25
0.4	-120	900	24

Note. The last column is the implied number of participants (N). All values are given to the nearest whole number except for the precision.

uniform (rectangular), skewed triangular, and exponential distribution because they are quite different from the normal distribution, as Figure 4 illustrates. I also used $z_c = 1.96$ because it corresponds to the traditional 95% confidence interval that is popular in the social sciences. The idea was to determine the N needed so that approximately .95 of the distances between the sample means and the population mean would be within the specified level of precision (f), despite using distributions that differ substantially from the assumed normal distribution.

In the computer simulation process, it was convenient to use Equation 6, repeated for convenience:

Figure 4. Illustration of the standard normal distribution, the uniform distribution, the right triangular distribution, and the exponential distribution used in a set of computer simulations.

$$f = \frac{z_c}{\sqrt{N}} \tag{6}$$

Because the z-score that corresponds to a confidence interval of .95 is 1.96, Equation 6 implies that the precision (f) is .8765, .6198, .5061, .4383, .392, and .3578 when $N = 5, 10, 15, 20, 25$, or 30 , respectively. The simulations to be reported provided the percentages of sample means that were within these distances of the population mean.

Uniform (Rectangular) Distribution. Uniform distributions have two defining parameters. These are α and β , which are the two endpoints of the distribution. I assigned a value of 0 for α and $\sqrt{12}$ for β . The reason for these user-defined values was to enjoy the convenience of having the standard deviation and variance equal unity. Equations 9, 10, and 11 provide the mean (μ), variance (σ^2), and standard deviation (σ) of a uniform distribution, respectively, including the instantiation of the user-defined values for α and β .

$$\mu = \frac{(\beta - \alpha)}{2} = \frac{\sqrt{12} - 0}{2} = 1.7321 \tag{9}$$

$$\sigma^2 = \frac{(\beta - \alpha)^2}{12} = \frac{(\sqrt{12} - 0)^2}{12} = 1 \tag{10}$$

$$\sigma = \frac{(\beta - \alpha)}{\sqrt{12}} = \frac{\sqrt{12} - 0}{\sqrt{12}} = 1 \quad (11)$$

Based on 10,000 samples for each case where $N = 5, 10, 15, 20, 25$, and 30 , I expected approximately .95 of these sample means to be within the specified fraction of a standard deviation of the population mean (1.7321). In fact, the proportions were as follows for $N = 5, 10, 15, 20, 25$, and 30 , respectively: .9524, .9532, .9517, .9534, .9495, and .9505. As all of these are close to the desired proportion of .95, the uniform distribution simulations support that the proposed procedure works well even when the population is uniform rather than normal.

Right Triangular Distribution. One potential criticism of the uniform distribution simulations is that these simulations did not involve any skewness. In contrast to uniform distributions, it is possible to have a skewed triangular distribution. Triangular distributions have three parameters. These are the minimum (a), maximum (b), and peak or modal (c) values. To maximize skewness, it is possible to set the peak value as being equal to the minimum or maximum value so that the triangular distribution has the shape of a right triangle (one angle is at 90 degrees), as Figure 4 illustrates. For the present simulations, the user defined values were as follows: $a = 0, b = \sqrt{18}, c = 0$. The reason for setting the maximum at $\sqrt{18}$ was so that the standard deviation would equal unity. The reason for setting the peak value at the same level as the minimum value was to obtain a maximally positive skew. The equations that determine the mean, standard deviation, and skewness (γ), as well as instantiated values, are given below:

$$\mu = \frac{a+b+c}{3} = \frac{0+\sqrt{18}+0}{3} = \frac{\sqrt{18}}{3} = 1.4142 \quad (12)$$

$$\sigma^2 = \frac{a^2+b^2+c^2-ab-bc-ca}{18} = \frac{0^2+\sqrt{18}^2+0^2-0-0-0}{18} = 1 \quad (13)$$

$$\sigma = \sqrt{\frac{a^2+b^2+c^2-ab-bc-ca}{18}} = \sqrt{\frac{0^2+\sqrt{18}^2+0^2-0-0-0}{18}} = 1 \quad (14)$$

$$\begin{aligned} \gamma &= \frac{\sqrt{2}(a+b-2c)(2a-b-c)(a-2b+c)}{5(a^2+b^2+c^2-ab-bc-ca)^{\frac{3}{2}}} \\ &= \frac{\sqrt{2}(0+\sqrt{18}-2(0))(2(0)-\sqrt{18}-0)(0-2\sqrt{18}+0)}{5(0^2+\sqrt{18}^2+0^2-0\sqrt{18}-\sqrt{18}(0)-0(0))^{\frac{3}{2}}} = .5657 \end{aligned} \quad (15)$$

Based on Equation 6, we have seen that the precision is .8765, .6198, .5061, .4383, .392, and .3578 when $N = 5, 10, 15, 20, 25$, or 30 , respectively. The

simulations to be reported in the subsequent paragraph provided the percentages of sample means that were within these distances of the population mean.

As in the previous simulation, based on 10,000 samples per simulation, where $N = 5, 10, 15, 20, 25$, and 30 , I expected approximately .95 of these sample means to be within the specified fraction of a standard deviation of the population mean (1.4142). In fact, the proportions were as follows for $N = 5, 10, 15, 20, 25$, and 30 , respectively: .9510, .9539, .9511, .9478, .9476, and .9485. As all of these are close to the desired proportion of .95, the triangular distribution simulations support that the COC procedure works well even when the population is substantially nonnormal and skewed.

Exponential Distribution. I performed a final set of simulations using an exponential distribution as a more extreme departure from normality, and one where skewness increases (see Figure 4). Exponential distributions only have one parameter, λ , and the user-defined value was unity so as to keep the standard deviation and variance at unity, as in the foregoing simulations. The equations that render the mean, variance, standard deviation, skewness, and the instantiated values, are given as Equations 16 to 19:

$$\mu = \frac{1}{\lambda} = \frac{1}{1} = 1 \quad (16)$$

$$\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{1^2} = 1 \quad (17)$$

$$\sigma = \frac{1}{\lambda} = \frac{1}{1} = 1 \quad (18)$$

$$\gamma = 2 \quad (19)$$

As usual, the precision is .8765, .6198, .5061, .4383, .392, and .3578 when $N = 5, 10, 15, 20, 25$, or 30 , respectively. The exponential distribution simulations provided the percentages of sample means that were within these distances of the population mean.

As in the previous simulations, based on 10,000 samples per simulation, where $N = 5, 10, 15, 20, 25$, and 30 , I expected approximately .95 of these sample means to be within the specified fraction of a standard deviation of the population mean. In fact, the proportions were as follows for $N = 5, 10, 15, 20, 25$, and 30 , respectively: .9565, .9548, .9554, .9519, .9519, and .9529. As all of these are close to the desired proportion of .95, the exponential distribution simulations support that the COC procedure works well even when the population is exponential.

Discussion

The proposed COC procedure implies a very different statistical philosophy than that which is implied by alternative procedures. When researchers use traditional

procedures, they collect the data and then perform the inferential statistical analyses to come to a conclusion about a hypothesis or the likely placement of the population mean. These are a posteriori procedures in the sense that they are put into practice after the data have been collected. However, we already have seen that these a posteriori procedures are associated with the inverse inference fallacy.

In contrast, the proposed COC procedure addresses a different question and one that obviates the need for any sort of a posteriori procedures. Hence, the researcher avoids the inverse inference fallacy. There is an a priori decision about how confident the researcher wishes to be that the sample mean will be within the specified distance of the population mean, Equation 8 gives the required N , and that is it! Provided that the researcher actually follows through by obtaining the computed N , she then trusts the data (provisionally of course based on later empirical work), computes her descriptive statistics, and makes her case. Contrasting the proposed a priori thinking versus the usual a posteriori procedures suggests important advantages for the former over the latter.

In the first place, note that there is no standard deviation in Equation 8 although the standard deviation was present in earlier equations. The reason for the absence in Equation 8 is that the standard deviation cancelled out in proceeding from Equations 4 and 5 to Equation 6. A consequence of this cancellation is that there is no need to estimate the standard deviation. And if there is no need to estimate the standard deviation, there likewise is no need to use a t distribution, and so the more convenient normal distribution can be used. A second and related advantage is that without the necessity to estimate the standard deviation, a source of error is removed.

A third advantage, and possibly the most important advantage, is the elimination of the inverse inference fallacy. Instead of attempting to reject the null hypothesis or estimate the probability that the population mean is within an interval—both of which encourage the inverse inference fallacy—the proposed COC procedure answers the question put to it. That is, given the desired level of closeness at the desired level of confidence, the COC specifies the number of participants that meet the two criteria. There is no pressure to commit the inverse inference fallacy.

A fourth advantage is ease of interpretation. As we have seen, in the case of traditional confidence intervals, there is no way to know, given a particular confidence interval that was obtained, the probability that the population mean is within that interval. In fact, it is difficult to describe, in words, precisely what a traditional confidence interval provides. This difficulty is doubtless an important reason why researchers so often misinterpret the meaning of traditional confidence intervals. In contrast, it is easy to describe, in words, what the COC procedure provides. It provides the number of participants needed to obtain a sample mean that is within a specified distance of the population mean under a specified probability. And that's it!

A fifth but related advantage is that the COC procedure simplifies the task of the researcher who is interested in substantive issues. Provided that the researcher has obtained the required N , the descriptive statistics need not be augmented by a posteriori inferential statistics. Because the criteria for closeness and confidence to justify

trusting the data are decided a priori, the researcher collects and trusts the data with no need of further inferential analyses. Those who read articles also benefit by not having to plow through myriad t tests, F tests, and p values.

A sixth advantage pertains not only to the present crisis in confidence in psychology but also in other sciences (Ioannidis, 2005; Loscalzo, 2012; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). There is suspicion that researchers engage in questionable practices such as including a few more participants until the magic level of $p < .05$ is reached, eliminating outliers that impede obtaining $p < .05$ without specifying ahead of time what counts as an outlier, and so on (e.g., Ioannidis, 2012; Simmons, Nelson, & Simonsohn, 2011).¹ The COC procedure renders these techniques irrelevant because all of the work done by Equation 8 is a priori. Once the levels of precision and confidence have been decided, and Equation 8 used to compute N , there is no need to use questionable practices for inferential statistical purposes because the inferential work already has been accomplished. To be sure, the COC procedure is not proof against descriptive statistical cheating, such as fabricating means but this kind of cheating requires extra sinfulness that most researchers hopefully do not have (Trafimow, 2013).

Finally, the a priori thinking I advocate is more consistent with frequentist philosophy than is the NHSTP or computation of confidence intervals. To see that this is so, consider that frequentist theory is, in the words of Morey et al. (2016), “pre-data.” Frequentist theory gives long-run probabilities based on repeated sampling (Jaynes, 2003; Mayo, 1996). Put another way, frequentist theory gives the probability of obtaining sample characteristics given assumed population characteristics. But frequentist theory does not—and I cannot emphasize this too strongly—give inverse probabilities of population characteristics given sample characteristics. This is the inverse inference fallacy. Well then, if frequentist theory gives probabilities of sample characteristics given assumed population characteristics whereas frequentist theory does not give probabilities of population characteristics given sample characteristics, the a priori thinking implied by the SCOC is much more sensible and consistent with frequentist theory than is the invalid a posteriori thinking employed by the NHSTP and confidence intervals.²

Remaining Issues

We have seen that the COC procedure has several important advantages over other statistical procedures. We also have seen that the COC procedure is robust to reasonable violations of normality, even when the sample size is as low as $N = 5$. Nevertheless, there are unaddressed issues and I mention some of these in the following subsections.

The Issue of Units. Although the COC procedure gives the probability of obtaining a sample mean within a specified distance of the population mean, the specified distance is given in terms of a fraction of a standard deviation, and the researcher does not know what this distance is in actual units. If the researcher needs to have an

interval in practically relevant units, this will necessitate using the sample standard deviation to estimate the population standard deviation.

On the other hand, in much, perhaps even most, research in psychology, there are no practically relevant units anyway. For example, in social psychology, what is an attitude unit, a prejudice unit, or a self-affirmation unit? In clinical psychology, what is a depression, anxiety, or psychopathy unit? In education, what is a knowledge unit? Similar questions can be asked with respect to measurement in many areas. And if there are no practically relevant units, precision in standard deviation units might be the best that is obtainable. In the event that practically relevant units exist, standard deviation units and practically relevant units might be useful for different purposes.

The Order Issue. An important philosophical difference between the present proposal and the traditional procedure for constructing confidence intervals is that researchers who use the COC procedure specify the desired distance and confidence a priori whereas traditional procedures invokes a posteriori reasoning (i.e., using the data to infer a p value or the spread of an interval). The COC procedure, with its a priori emphasis, provides important advantages, as I have stated earlier. But does the a priori emphasis mean that the COC procedure cannot be used in an a posteriori way?

Au contraire! The COC procedure can be used in an a posteriori way too, although, in general, a priori is better than a posteriori. Imagine that a researcher collects a random sample but does not determine N based on specified levels of distance and confidence. It is possible for the researcher to use the COC procedure anyhow to gain an idea about how much confidence one should have that the sample mean is within a specified distance of the population mean but there is an ambiguity. To see the ambiguity, consider that the fraction of a standard deviation can be expressed as a function of the z -score corresponding to the desired confidence level and N or that the z score can be expressed as a function of a specified fraction of a standard deviation and N . Equation 20 illustrates the ambiguity.

$$f = \frac{z_c}{\sqrt{N}} \quad \text{or} \quad z_c = f\sqrt{N} \quad (20)$$

Consider an example where a researcher has obtained 25 participants. Suppose we use a 95% confidence interval so that $z_c = 1.96$. In that case, $f = \frac{1.96}{\sqrt{25}} = .392$. On the other hand, suppose that we specify that the value we want for f is .20. In that case, the z score is $z_c = .20\sqrt{25} = 1.0$. The confidence level that corresponds to this z score is 63%. Therefore, we could state either of the following. First, we could state that there is a 95% probability that the sample mean is within .392 of a standard deviation of the population mean. But we also could state that there is a 63% probability that the sample mean is within .20 of a standard deviation of the population mean. Nor are these the only possibilities. A disadvantage of failing to specify z_c and f a priori is that there are many ways to interpret the data. An interesting philosophical discussion could ensue over the fact that even with a priori specification, someone else later can

assert a different combination of confidence and distance that is equally consistent with the obtained N .

A potential way to address the ambiguity might be appeal to conventionality. On the one hand, it could be argued that 95% intervals are conventional in the social sciences and therefore the z score should be set at 1.96. This appeal to conventionality favors an interpretation that the precision is .392 of a standard deviation. On the other hand, an appeal to conventionality might not be considered to be intellectually satisfying.

Conventionality and the 95% Number. Social scientists are used to conventions. There are conventions about the acceptable p value, power, designations for small, medium, and large effect sizes, reliability, and so on. An advantage of conventions is that they provide interpretive consensus. A disadvantage is that researchers become slaves to the conventions and there may be cases where the conventions are ill suited for particular purposes.

An example might be the convention of using 95% confidence intervals. Historically, confidence intervals have been used as another way to perform the NHSTP. And the NHSTP convention of using .05 for the cutoff value for significance goes very naturally with a convention of using confidence intervals of 95%. But given that the inverse inference fallacy renders the NHSTP invalid, it is far from clear why the convention of using 95% confidence intervals should perpetuate into the indefinite future. For example, Schmidt and Hunter (1997) have suggested that prior to the NHSTP, researchers often used the “‘probable error’—the 50% confidence interval” (p. 51). I have no wish to attempt to impose an arbitrary confidence level but merely point out that there is no compelling reason to remain with a convention that is a vestige of the discredited NHSTP.

Also, regarding conventions, I pointed out earlier that it is not clear what levels of confidence and precision the researcher ought to specify. I suggested using either the first or second derivative as one way to help decide on such conventions but I do not insist on this. In fact, I would rather have researchers devote effort to a consideration of the relevant factors at play in their respective areas, and use their collective wisdom for this purpose. Alternatively, authors of manuscripts could justify to their audience (and reviewers and editors) why they chose particular levels of confidence and precision. It would then be up to journal editors to make the ultimate decision about the merits of the justifications. As I suggested earlier, justifications could include considerations of the difficulty of obtaining data, the theoretical or practical importance of the research, the methodology employed, the complexity of the desired conclusion, and so on.

New Directions. Most researchers are interested in means and so this was the present focus. But researchers potentially could be interested in correlations, proportions, standard deviations, and others and generalizing the COC procedure to these could be a useful direction for future research.

Another new direction might be to test alternative distributions that are even more different from the normal distribution than the rectangular, right triangular, or exponential distributions I tested. There might be distributions that differ so extremely from a normal distribution that the proposed COC procedure does not work well. Or there might be distributions that require enhanced sample sizes to enable the proposed COC procedure to work well. The development of new equations to handle such cases might be a useful direction for future research.

A third possible direction for future research concerns increasingly complex comparisons. In the present work, I was concerned with a single mean. But researchers might be interested in differences between means, or differences between differences between means (e.g., 2×2 interaction), or comparisons of yet higher orders. Should researchers insist on increasingly more precise values for f as the comparisons of interest are of increasingly higher orders? If so, how much extra precision is warranted for each increase in complexity? A potential argument in the opposite direction is that more complex designs tend to require more participants anyhow and so the additional participants needed to meet an added requirement of increased precision might constitute an undesirable barrier for researchers to have to pass to use complex designs.

Conclusion

The NHSTP has been widely criticized. Many critics prefer traditional confidence intervals but these also fail to provide much useful information, whether used as an alternative way to conduct significance tests, for parameter estimation, or for precision estimation. In contrast to the NHSTP and confidence intervals, Bayesian methods permit logically valid inverse inferences about probabilities of hypotheses or probabilities of population means being in stated intervals but there is an issue about the permissibility of subjective or artificial prior probability specifications. Therefore, there are two logically valid inferential statistical options and each comes with a price tag. First, one can go the Bayesian route and enjoy the advantages that come with logically valid inverse inferences but at the price of subjective or artificial prior probability specifications. Second, one can use a priori inferential statistics advocated here that avoid the inverse inference fallacy and do not depend on subjective or artificial prior probability specifications but at the price of not being able to make inverse inferences. Possibly, future researchers will find ways to combine these approaches but at present, these are the only two possibilities that do not result in logical invalidity.

Acknowledgments

I thank Justin A. MacDonald, Uli Widmaier, and two anonymous reviewers for their helpful comments on an earlier version of this article. I really appreciate the help!

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This is not to say that outliers do not matter with the COC procedure. Rather, it is an argument that the COC procedure, in contrast to the NHSTP, does not push the researcher to make questionable decisions about outliers to obtain a desired value for p .
2. Harris and Quade (1992) used similar mathematics to that employed here to suggest a type of power analysis. Their approach differs philosophically from the COC procedure because their goal is to find the number of participants needed to obtain particular values for p . Another difference is that Harris and Quade were concerned with detecting a minimally important difference between conditions whereas the present focus is on being confident that the obtained mean is close to the population mean.

References

- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 1-29. doi:10.1037/h0020412
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-542. doi: 10.2307/2279690
- Cohen, J. (1990). Things I have learned (so far). *The American Psychologist*, 45, 1304-1312. doi:10.1037/0003-066X.45.12.1304
- Cohen, J. (1994). The earth is round ($p < .05$). *The American Psychologist*, 49, 997-1003. Retrieved from <http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&uid=1995-12080-001>
- Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242. doi.org/10.1037/h0044139
- Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology*, 217(1), 27-37. doi:10.1027/0044-3409.217.1.27
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). London, England: Collier Macmillan.
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.
- Harris, R. J., & Quade, D. (1992). The minimally important difference significant criterion for sample size. *Journal of Educational Statistics*, 17(1), 27-49. doi:10.3102/10769986017001027
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace College.

- Hogben, L. (1957). *Statistical theory*. London, England: Allen & Unwin.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. doi:10.1371/journal.pmed.0020124
- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7, 645-654. doi:10.1177/1745691612464056
- Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.
- Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *The American Psychologist*, 56(1), 16-26. doi:10.1037//0003-066X.56.1.16
- 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, 161-171. doi: 10.1111/1467-8721.ep11512376
- Loscalzo, J. (2012). Irreproducible experimental results: Causes, (mis)interpretations, and consequences. *Circulation*, 125, 1211-1214. doi:10.1161/CIRCULATIONAHA.112.098244
- Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159. doi:10.1037/h0026141
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115. Retrieved from <http://www.jstor.org/stable/186099>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834. Retrieved from <http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1979-25042-001>
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23, 103-123.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). doi:10.1126/science.aac4716
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528-530. doi:10.1177/1745691612465253
- Popper, K. R. (1983). *Realism and the aim of science*. London, England: Routledge.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428. doi:10.1037/h0042040
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1, 115-129. Retrieved from <http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1996-04469-001>
- Schmidt, F. L., & Hunter, J. E. (1997). Eight objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366. doi:10.1177/0956797611417632
- Suppes, P. (1994). Qualitative theory of subjective probability. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 17-38). Chichester, England: Wiley.

- Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, 70, 434-438. doi:10.1002/j.1556-6676.1992.tb01631.x
- Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, 110, 526-535. doi: 10.1037/0033-295X.110.3.526
- Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review*, 112, 669-674. doi:10.1037/0033-295X.112.3.669
- Trafimow, D. (2006). Using epistemic ratios to evaluate hypotheses: An imprecision penalty for imprecise hypotheses. *Genetic, Social, and General Psychology Monographs*, 132, 431-462. doi:10.3200/MONO.132.4.431-462
- Trafimow, D. (2013). Descriptive versus inferential cheating. *Frontiers in Theoretical and Philosophical Psychology*, 4, 627. doi:10.3389/fpsyg.2013.00627
- Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *Journal of General Psychology*, 136, 261-269. doi:10.3200/GENP.136.3.261-270