Performing Inferential Statistics Prior to Data Collection

Educational and Psychological
Measurement

2017, Vol. 77(2) 204–219

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416659745

journals.sagepub.com/home/epm

David Trafimow¹ and Justin A. MacDonald¹

Abstract

Typically, in education and psychology research, the investigator collects data and subsequently performs descriptive and inferential statistics. For example, a researcher might compute group means and use the null hypothesis significance testing procedure to draw conclusions about the populations from which the groups were drawn. We propose an alternative inferential statistical procedure that is performed prior to data collection rather than afterwards. To use this procedure, the researcher specifies how *close* she or he desires the group means to be to their corresponding population means and how *confident* she or he wishes to be that this actually is so. We derive an equation that provides researchers with a way to determine the sample size needed to meet the specifications concerning closeness and confidence, regardless of the number of groups.

Keywords

a priori procedure, confidence, precision, complexity

In the context of structural equation modeling, Wolf, Harrington, Clark, and Miller (2013; also see Marcoulides & Saunders, 2006) have distinguished between inferential statistical approaches featuring operations performed prior to data collection or performed after data collection. Although the present focus is on sample means rather than on structural equation modeling, the distinction nevertheless is useful to keep in mind and it will feature strongly in the present article.

It is possible to identify different inferential statistics camps. Two of these camps fall within the “frequentist” way of thinking, where the basic assumption is that hypotheses are correct or incorrect but cannot take on probabilities other than 0 or 1.

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

Corresponding Author:

David Trafimow, New Mexico State University, PO Box 30001, Las Cruces, NM 88003-8001, USA.

Email: dtrafimo@nmsu.edu

One camp—comprising those who wish to use data to reject or fail to reject null hypotheses—favor the null hypothesis significance testing procedure. A second camp—comprising those who wish to use frequentist thinking for parameter estimation—favor confidence intervals. Yet a third camp—comprising those who believe that hypotheses can take on probabilities other than 0 or 1—favor the use of the famous theorem by Bayes.¹ It also is possible, perhaps, to identify a fourth camp that features equivalence testing, where the idea is to control the size of misspecification at a prespecified value so as to assess the goodness of the model at a desired level of confidence (Wellek, 2010).² Equivalence testing has been advocated, for example, when the goal is to show that a treatment is equivalent to some other treatment (Walker & Nowacki, 2011), though there are many other uses as well (see Yuan, Chan, Marcoulides, & Bentler, 2016, for a list). Arguably, especially in Yuan et al. (2016), the procedure used to determine the desired sample size for equivalence testing occurs prior to data collection, which distinguishes it from the other three camps.

There has been much disagreement across the camps, with equivalence testing perhaps being the exception (e.g., Bakan, 1966; Berkson, 1938; Cohen, 1994; Fidler & Loftus, 2009; Fisher, 1973; Gigerenzer, 1993; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Hogben, 1957; Loftus, 1996; Lykken, 1968; Mayo, 1996; Meehl, 1967, 1978; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Popper, 1983; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997; Suppes, 1994; Thompson, 1992; Trafimow, 2003, 2006; Trafimow & Rice, 2009). For present purposes, it is not necessary to discuss these controversies nor to advocate for any one camp or set of camps over any others.³ Instead, our goal is to follow-up on a particular type of inferential procedure recently suggested by Trafimow (2016), that is not equivalent to any of the foregoing camps, and address what we consider to be a serious limitation of that procedure. We emphasize that the procedure we propose can be used in conjunction with other procedures and so there is no necessary implication of a competition between procedures.

Based, in part, on a recent article by Trafimow (2016), our proposal differs from those of the other camps, with the possible exception of equivalence testing, in the following ways. First, our *a priori procedure* (APP) pertains to a different question, as we explain in the subsequent section.⁴ Second, as the label suggests, there is no need to collect any data whatsoever to answer that question. The APP is compatible with the other approaches but also works as a stand-alone procedure. Again, we emphasize that there is no necessity here for a competition between procedures.

Being Confident of Being Close

Suppose we pointed out to researchers that it would be much easier to obtain a single participant than to obtain a larger sample of participants. Based on this, we might ask, “Why collect a sample of participants?” After asking questions to get beyond issues such as the importance of getting statistical significance to publish, and

publishing to get tenure, and so on, we believe that eventually researchers would converge on the notion that a sample of participants aids researchers in feeling confident that the sample statistic is close to the population parameter, or at a minimum, the larger the sample, the more likely it is to resemble the population. Few researchers would be interested in sample statistics if they felt that these were completely unrepresentative of population parameters and because the most widely used sample statistic researchers use is the sample mean, as an estimate of the population mean, our focus will be on means.⁵

Well, then, if researchers wish to collect samples of participants so that they can be confident that their sample means are close to the corresponding population means, this implies two important issues: defining *close* and defining *confident*. In the case where only a single mean is at issue, there is a formula presented as Equation 1 that gives the number of participants needed (n) as a function of the fraction of a standard deviation defined as close (f) and the z -score (z_c) corresponding to the desired probability of obtaining a sample mean within the prescribed fraction of a standard deviation of the population mean.⁶

$$n = \left(\frac{z_c}{f} \right)^2 \quad (1)$$

To use the formula, the researcher decides, prior to data collection, how close he or she wishes the sample mean to be to the population mean and the confidence desired of actually being within that distance. For example, the researcher might desire to be within three tenths of a standard deviation of the mean and to have a probability of .95 of actually being within that distance. The z -score that corresponds to .95 is 1.96, and so the number of participants needed is $n = (1.96/0.3)^2 = 42.68$, or 43 rounded to the nearest whole number. We emphasize that the analysis is conducted *a priori*.

Equation 1 assumes a normally distributed population. To test the importance of adhering to this assumption, Trafimow (2016) performed computer simulations with decidedly nonnormal distributions such as a rectangular distribution, a right triangular distribution, and even an exponential distribution. He found that using distributions that differed substantially from the assumed normal distribution made very little difference in the results. We accept these simulations and the implication that the normality assumption is not an important problem for Equation 1. Nevertheless, we believe that there is an important limitation with respect to Equation 1, specifically, and the Trafimow (2016) article, generally. That is, Equation 1 is fine if there is only a single mean to be considered, but few researchers are concerned with only a single mean. Much more often, there is an experimental group and a control group, so that there are two means and neither Equation 1 nor anything in the Trafimow (2016) article accommodates two or more means. Fairly often, researchers use complex factorial designs with any number of means. Thus, our main goal is to expand the APP to include multiple means.

We also investigate closely related issues. For example, how does the APP differ from power analysis, which is sometimes conducted a priori? Finally, we address additional issues suggested by the foregoing ones, including the implications the APP has for complex experimental designs.

Multiple Means

It is possible to derive Equation 2, which gives the probability of obtaining one sample mean within a specified fraction of a standard deviation of the population mean or $p(1 \text{ Mean})$. In Equation 2, Φ is the cumulative distribution function (cdf) of the standard normal distribution, f is the precision, and n is the sample size.⁷ Equation 1 implies Equation 2, but we present an independent derivation in the appendix. To our knowledge, this is the first presentation of Equation 2.

$$p(1 \text{ Mean}) = 2 \cdot \Phi(f\sqrt{n}) - 1 \quad (2)$$

Suppose that we now imagine that there are two samples, with samples sizes n_1 and n_2 and that we also are concerned with the distance that each sample mean is from its corresponding population mean, that is, f_1 and f_2 . If we assume that the sampling errors $(M_1 - \mu_1)$ and $(M_2 - \mu_2)$ are independent, then the probability that both means are within their specified distances, denoted as $p(2 \text{ Means})$, is the product of the probabilities of each being within its specified distance, as Equation 3 shows. Note that in Equation 3, there are two conditions, and so there are two cell sample sizes (n_1 and n_2) and both sample means can be within specified distances of the corresponding population means (f_1 and f_2).

$$p(2 \text{ Means}) = (2 \cdot \Phi(f_1\sqrt{n_1}) - 1)(2 \cdot \Phi(f_2\sqrt{n_2}) - 1) \quad (3)$$

More generally, Equation 4 renders the probability of k means all being within their desired distances of the corresponding population means, where n_k denotes the sample size of the k th mean and f_k denotes the desired distance of the k th mean from its corresponding population mean.

$$p(k \text{ Means}) = \prod_1^k (2 \cdot \Phi(f_k\sqrt{n_k}) - 1) \quad (4)$$

In those cases where there is an equal sample size in each condition ($n_1 = n_2 = \dots = n_k = n$) and where there is an equal desired distance of each sample mean from its corresponding population mean ($f_1 = f_2 = \dots = f_k = f$), this can be simplified to the following.⁸

$$p(k \text{ Means}) = (2 \cdot \Phi(f\sqrt{n}) - 1)^k \quad (5)$$

Figure 1 explores the implications of Equation 5, where the probability that all of the means will be within f —that is, $p(k \text{ Means})$ —is given along the vertical axis as a function of the size of f (range is .01 to .5 along the horizontal axis) and the number

Figure 1. The probability that all of the sample means will be within f as a function of f , the number of means, and N .

of means (there are five curves representing 1, 2, 3, 4, or 5 means). Finally, there are four panels where the total sample size (N) = 100, 200, 300, or 400. Consider the top curve (1 mean) in the panel where $N = 100$. As one moves rightward from $f = .01$, the curve rises sharply before shifting toward asymptote. Thus, a small increase in imprecision along the horizontal axis implies a much larger probability that the mean is within the specified interval. In contrast, once asymptote is reached, even large increases in imprecision fail to substantially increase the probability that the mean is within the specified interval. As one peruses the other curves, where there are increasingly more means to be considered, asymptote is reached at increasingly larger levels of imprecision along the horizontal axis. As there are more means, the level of imprecision must increase to ensure that all of the means have a reasonable probability of being within the specified interval. Also, as there are more means, the total sample size is split among more groups, so that n decreases keeping N constant.

However, a look at the other three panels qualifies these conclusions. As N increases, asymptote is reached at lower levels of f along the horizontal axis. That is, increasing N allows the researcher to specify more stringent intervals in which there is a reasonable probability that all of the means are likely to be within them. More generally, the four panels of Figure 1 allow the researcher to see precisely how the desired precision, the number of means, and the total sample size interact to

influence the probability that all of the means are within the specified interval. However, a limitation of Figure 1 is that it is difficult to discern the number of participants that actually are needed for the researcher to be able to be confident that the sample means will be close to their corresponding population means.

Reversing Equation 5 and Its Implications

With the aid of Figure 1, we have seen how f , k , and number of participants interact to influence the probability that all of the sample means of interest will be within the desired distance of the corresponding population means. However, we have not yet explored how f , k , and the desired probability that the sample means are within the specified distance ($p(k \text{ Means})$) influence the number of participants needed per sample. To aid in this exploration, solving Equation 5 for n renders Equation 6.

$$n = \left(\frac{\Phi^{-1} \left(\frac{\sqrt[k]{p(k \text{ Means})} + 1}{2} \right)}{f} \right)^2 \quad (6)$$

Figure 2 provides a visual rendition of Equation 6, where n is expressed along the vertical axis as a function of f along the horizontal axis and the number of means (1, 2, 3, 4, or 5, as in Figure 1). In Figure 1, f ranged from .01 to .5, but using this range caused a problem with respect to Figure 2. To understand the problem, consider the case where one wishes to have a .95 probability that five sample means are within .01 standard deviations of their respective population means. The value of n , in this case, according to Equation 6, is 65,985. Not only is this practically unrealistic but it also stretches the vertical axis to the point of making it difficult to see the effects to be described subsequently. To avoid the problem, we used a range of .1 to .5 for f along the horizontal axis. Finally, as in Figure 1, there were four panels, where the probability that all means are within f was set at .65, .75, .85, and .95. Put another way, each panel specifies a different level of confidence that the researcher can have that all of the sample means will be within the specified distance of the population means.

Let us start with the first panel, where the confidence was set at .65, and with the lowest curve denoting the case where there is only a single mean to be considered. When f is at the minimum value of .1, thereby indicating impressive precision, the number of participants needed to have confidence at .65 is 87 (rounded to the nearest whole number). As one moves toward increasingly less precision, there is a steep drop in the number of participants needed at first but the drop becomes increasingly less steep as one becomes increasingly less precise. Perhaps another way to look at it is to go from right to left, where it can be seen that, at low levels of precision, substantial improvements in precision can be made at very little cost in n . In contrast, as one continues to move leftward along the horizontal axis, if one moves sufficiently far in that direction, even small improvements in precision have a substantial cost in

Figure 2. Figure displays n as a function of f , the number of means, and the desired confidence level.

n . This fact suggests that researchers might consider the concept of a “best buy,” that is, how much improvement in f is worth how much cost in n ?

Two other effects are worth mentioning. Most obviously, as there are more means to be considered, n must increase accordingly. But this effect of the number of means is qualified by f . Consider, for example, the difference between one mean and five means when f is .1. In this case, when there is one mean, the needed n is 87, but when there are five means, the needed n is 301, for a difference of 214. In contrast, as the level of imprecision increases, differences in the level of n needed to reach f attenuate dramatically. For example, when $f = .5$, the levels of n needed for one or five means are 3 and 12, respectively, for a difference of 9.

All of the foregoing effects for when confidence is set at .65 become increasingly more dramatic as the level of confidence increases to .75, .85, and .95. This is because, although the panels do not look very different from each other in terms of the shapes of the curves, the extent of the values along the vertical axis increases as the level of confidence increases. For example, we saw earlier that when there are five means and confidence is set at .65, $n = 301$. But when confidence is set at .95, this value becomes 660. More generally, as confidence increases, so does (a) the effect of the number of means, (b) the effect of precision, and (c) the interaction between the number of means and precision. Figure 3 provides a “blown-up”

Figure 3. Figure displays n as a function of f , the number of means, and the desired confidence level. However, n was capped at 100 along the vertical axis.

version of Figure 2 where the vertical axis is restricted to a maximum of 100 participants so as to allow the reader to better view the implications of Equation 6 at sample sizes typically used in research.

As dramatic as the foregoing effects might seem to be, they arguably provide underestimates because Figures 2 and 3 represent n along the vertical axis rather than N . To see why this might matter, consider that increasing the number of means necessitates that more participants are needed in two ways. First, as Figures 2 and 3 illustrate, increasing the number of means necessitates an increase in the number of participants in each condition. But, as Figures 2 and 3 fail to illustrate, increasing the number of means also indicates an increase in the number of conditions. Thus, although Figures 2 and 3 are fine for understanding how the number of conditions increases the number of participants needed in each of them, it is necessary to multiply n by the number of conditions (k) to understand the total effects. Figure 4 includes this change: N (rather than n) is represented along the vertical axis as a function of f , k , and the four levels of confidence used in Figure 2 (.65, .75, .85, and .95). As can be seen by attending to the actual values along the vertical axis, all of the effects described with respect to n are much more dramatic with respect to N .

Figure 4. Figure displays n as a function of f , the number of means, and the desired confidence level.

The APP and Power Analysis

The APP is similar to a priori power analysis in the sense that both are prior to data collection. However, there is a difference. The purpose of power analysis is to determine the N needed to have a reasonable chance of obtaining a statistically significant finding, given that there is an effect to be found. For those who favor confidence intervals over the null hypothesis significance testing procedure, the purpose of power analysis is to determine the N needed to obtain a confidence interval of a desired width. Either way, the purpose of the power analysis is to aid the researcher in what eventually will be an a posteriori analysis. In contrast, the APP can be used in isolation or in conjunction with a posteriori procedures. One starts by specifying how close one wishes the sample means to be to their corresponding population means, and the desired confidence that this actually will be so. From there, the APP provides the number of participants needed. Another way to think about the APP is that the closeness and confidence specifications indicate the conditions needed for the researcher to trust the data, so that as long as the researcher collects the required sample sizes to fulfill these conditions, the researcher can simply trust the resulting means without further inferential analyses, at least to the extent of the specified degree of confidence. The APP recognizes that the extent to which sample means

can be trusted as estimates of population means has nothing to do with what the findings actually are but rather depends solely on the size of the samples. We hasten to add, however, that because the APP pertains to drawing conclusions about samples rather than about populations, the researcher is not justified in concluding that the population mean is within the specified distance (f') of the obtained sample mean, with the specified degree of confidence.⁹

There is an additional difference that can be illustrated with an example. Suppose that Experimenter A and Experimenter B perform experiments with an experimental group and a control group. Based on previous research, Experimenter A can count on a very large effect size and Experimenter B anticipates a very small effect size. Because p is, to an important degree, influenced by the effect size, Experimenter A needs only a small sample size whereas Experimenter B needs a large sample size to have a reasonable chance of being able to reject the null hypothesis. In contrast, using the APP, the issue is not about obtaining a particular value for p but about having sample means that can be trusted to accurately estimate population means. From this point of view, it does not follow that Experimenter B needs to have a larger sample size than Experimenter A. From the point of view of trusting whether the sample means accurately estimate the population means, both researchers could use the same sample size and be able to trust their sample means equally.¹⁰

Complexity

Many researchers employ complex factorial designs. However, the APP suggests that there is an important problem with this approach that few researchers appreciate. Specifically, as more conditions are added, there are more means, and the probability that all of the means are close to the corresponding population means decreases. Put simply, as the design becomes increasingly complex, the researcher can place less trust in the cell means. But there is more than one reason for this.

Most obviously, keeping the overall sample size (N) constant, increasing the number of conditions necessitates that each cell mean will be based on fewer participants. In turn, fewer participants indicates that less trust can be placed in the resulting cell means. But beyond that, even if more participants are added, so that the sample size per condition remains the same, the foregoing equations imply that the probability that all of the sample means are within the specified distance of the corresponding population means decreases as the number of conditions increases. It is interesting to consider this mathematical fact in the context of the complex designs that are common in psychology. For example, psychologists often use $2 \times 2 \times 2$ designs, and even $2 \times 2 \times 2 \times 2$ designs are not particularly uncommon.

A few quick calculations are illuminating. When there is a single condition ($k = 1$) with 100 participants, there is a better than 95% chance that the sample mean will be within .2 standard deviations of the population mean. But suppose that the design is a $2 \times 2 \times 2$ design, so that there are eight condition ($k = 8$). Let us even make an extremely favorable stipulation that we use $N = 800$ participants so that the number

of participants in each condition remains constant at $n = 100$. Nevertheless, the probability that all of the means will be within .2 standard deviations of the corresponding population means is only .69. And matters become even worse if we consider a $2 \times 2 \times 2 \times 2$ design, so that $k = 16$. Even keeping n at 100 (so that $N = 1,600$) implies that the probability that all of the means will be within .2 standard deviations of the corresponding population means reduces to .47. Thus, the APP makes salient that there is an important problem with using complex designs, the nature of which otherwise would not be apparent. To reiterate, as the number of conditions increases, less trust can be placed in the obtained sample means.

Contrasting the Proposed Procedure With Other Procedures

Different procedures entail different issues. Our goal in the present section is to make clear how the APP asks a different question, and comes to a different sort of conclusion, than other procedures. Let us commence by considering the most common statistical procedure, namely, that concerned with testing null hypotheses. The idea of this procedure is to compute p and then reject or fail to reject the null hypothesis. As we pointed out earlier, many researchers have criticized this procedure, with much of the criticism based on the logical fact that one cannot validly draw conclusions about the probabilities of hypotheses given findings from the probabilities of findings given hypotheses. It would take us too far afield to discuss inverse inferences in detail and it is sufficient merely to note that there is an increasing trend for researchers to be concerned with this issue (see Trafimow, 2016, for such a discussion). Confidence intervals are closely related. In fact, it is possible to use confidence intervals as an alternative way to test null hypotheses. In addition, some researchers have suggested that confidence intervals can be used for parameter estimation. Critics of confidence intervals have argued that confidence intervals cannot validly be used for parameter estimation because there is no way to know the probability that the population mean (or difference between population means) is within the constructed interval. Both null hypothesis significance tests and confidence intervals have in common that the goal is to make inverse inferences about probabilities concerning hypotheses or population parameters given sample data. Critics have argued that this is not logically valid.

Bayesians are among the most vociferous claimants that null hypothesis significance tests and confidence intervals cannot validly be used to draw conclusions about populations from sample data. In contrast, the famous theorem by Bayes can be used to draw logically valid inferences about populations from sample data. However, frequentists are critical of Bayesian procedures on the grounds of unsound premises. They ask how one can know the prior probability of a hypothesis, what a prior distribution looks like, and how to handle the catch-all hypothesis of “not the null” (e.g., Mayo, 1996). Thus, critics of null hypothesis significance tests and confidence intervals tend to be dissatisfied based on the logical validity problems involved with inverse inferences, and critics of Bayesian methods tend to be dissatisfied based on what they consider to be questionable premises.

In contrast to null hypothesis significance tests, confidence intervals, and Bayesian procedures, which involve inverse inferences, the APP does not involve inverse inferences. Thus, the question of concern is not, “Can I reject the null hypothesis?” or “Can I conclude that the population mean is likely to be within a constructed interval?” Rather, the question is, “How can I be confident that the sample means are likely to be close to their corresponding population means?” As we suggested earlier, researchers can use the APP to address this question and still use one of the available a posteriori procedures to address questions about hypotheses or about population means. Alternatively, those researchers who believe that available a posteriori procedures are problematic can use the APP to answer the question about confidence and closeness and leave questions about hypotheses or population means unaddressed.

Conclusion

Although Trafimow (2016), with his emphasis on a priori inferential statistics, provided an advance, we noted an important limitation that his procedure only is applicable when there is a single mean at issue. Our expansion of the procedure to work with multiple means provides researchers with the opportunity to perform a priori inferential statistical analyses with a variety of designs, such as when there are experimental and control conditions, complex factorial designs, and so on.

In turn, however, our expansion suggests additional implications, such as the fact that complex research designs can be quite problematic from the point of view of precision. We also showed how the APP differs from other APPs, such as power analyses—the latter depends importantly on the effect size one expects to obtain whereas our APP is uninfluenced by the expected effect size. There also is a necessity to follow-up power analyses with traditional a posteriori analyses whereas this is not true of the APP. More generally, and in agreement with Wolf et al. (2013), we believe that researchers have much to gain by performing a priori inferential statistics, and we hope and expect that our expansion of the interaction of the concepts of closeness and confidence to apply to any number of conditions will aid in the future development of the areas of education and psychology.

Appendix

Assume:

$$M \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Then:

$$\begin{aligned}
 p(1\text{Mean}) &= (p(|M - \mu| < f\sigma)) \\
 &= p(-f\sigma < M - \mu < f\sigma) \\
 &= p\left(\frac{-f\sigma}{\sigma/\sqrt{n}} < \frac{M - \mu}{\sigma/\sqrt{n}} < \frac{f\sigma}{\sigma/\sqrt{n}}\right) \\
 &= p(-f\sqrt{n} < Z < f\sqrt{n})
 \end{aligned}$$

where

$$Z \sim N(0, 1)$$

Continuing,

$$\begin{aligned}
 &p(-f\sqrt{n} < Z < f\sqrt{n}) \\
 &= \Phi(f\sqrt{n}) - \Phi(-f\sqrt{n}) \\
 &= \Phi(f\sqrt{n}) - [1 - \Phi(f\sqrt{n})] \\
 &= 2 \cdot \Phi(f\sqrt{n}) - 1.
 \end{aligned}$$

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. One of the underlying disagreements between frequentists and Bayesians concerns the definition of probability. Frequentists often, but not always, define probability as a relative frequency, whereas Bayesians often, but not always, define it as a belief state.
2. We thank an anonymous reviewer for suggesting this.
3. This is not to say that we lack strong convictions about these issues only that there is no point in bogging down the present discussion with them.
4. Following Wolf et al. (2013), we could have used the term *proactive*. But as our procedure is based on Trafimow (2016), we decided to use the term *a priori*, as he did.
5. Some researchers have argued that perhaps the mean is too strongly emphasized, at least at particular times (e.g., Speelman & McGann, 2013). Nevertheless, whether justified or not, researchers favor means and so that is our focus here.
6. The formula is listed in many sources (e.g., Hays, 1994; also see Harris & Quade, 1992). Trafimow (2016) provided an accessible derivation.
7. Many statistics textbooks include cdf tables. In addition, there are many software- and Internet-based sources for this information. For example, Equation 2 can be implemented

in Microsoft Excel using the formula $=2*\text{NORM.S.DIST}(f*\text{SQRT}(N), \text{TRUE})-1$, where f and N are replaced with their desired numeric values.

8. Equations 3 to 5 assume independence. We thank an anonymous reviewer for suggesting that when there is dependence, there is a lower bound set by 1 minus the sum of the non-conditioned probabilities that each sample mean is not within the prescribed distance of its corresponding population mean (see Kutner, Nachtsheim, Neter, & Li, 2005, for a relevant discussion).
9. Confidence intervals do not justify this conclusion either. Although, for example, performing an experiment many times and constructing a 95% confidence interval each time will result in 95% of the constructed intervals enclosing the population mean, it is not the case that the population mean has a 95% probability of being enclosed by a particular obtained confidence interval. There is no way to know this probability.
10. Of course, if a researcher wishes to capitalize on a difference between means, the expected size of the difference likely will influence the imprecision the researcher is willing to tolerate. But this is a different issue than the issue of how much trust a researcher is willing to place in sample means as indications of corresponding population means.

References

- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437. 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. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- 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, 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.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157-1164. doi:10.3758/s13423-013-0572-3
- Hogben, L. (1957). *Statistical theory*. London, England: Allen & Unwin.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: WCB/McGraw-Hill.
- 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. doi:10.1111/1467-8721.ep11512376

- Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70(3, Pt 1), 151-159. <http://doi:10.1037/h0026141>
- Marcoulides, G. A., & Saunders, C. (2006). PLS: A silver bullet? *MIS Quarterly*, 30, iii-ix.
- 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(2), 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.
- 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. doi:10.3758/s13423-015-0947-8
- 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.
- 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.
- Speelman, C. P., & McGann, M. (2013). How mean is the mean? *Frontiers in Psychology*, 4, 451. doi:10.3389/fpsyg.2013.00451
- 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. (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. (2016). Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164416667977
- 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
- Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26, 192-196. doi:10.1007/s11606-010-1513-8
- Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety.

Educational and Psychological Measurement, 73, 913-934. doi:10.1177/0013164413495237

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, 23, 319-330. doi:10.1080/10705511.2015.1065414