econometrics

MDPI

Article

A Frequentist Alternative to Significance Testing, *p*-Values, and Confidence Intervals

David Trafimow

Department of Psychology, MSC 3452, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003-8001, USA; dtrafimo@nmsu.edu

Received: 27 March 2019; Accepted: 24 May 2019; Published: 4 June 2019

Abstract: There has been much debate about null hypothesis significance testing, *p*-values without null hypothesis significance testing, and confidence intervals. The first major section of the present article addresses some of the main reasons these procedures are problematic. The conclusion is that none of them are satisfactory. However, there is a new procedure, termed the a priori procedure (APP), that validly aids researchers in obtaining sample statistics that have acceptable probabilities of being close to their corresponding population parameters. The second major section provides a description and review of APP advances. Not only does the APP avoid the problems that plague other inferential statistical procedures, but it is easy to perform too. Although the APP can be performed in conjunction with other procedures, the present recommendation is that it be used alone.

Keywords: a priori procedure; null hypothesis significance testing; confidence intervals; *p*-values; estimation; hypothesis testing

1. A Frequentist Alternative to Significance Testing, *p*-Values, and Confidence Intervals

Consistent with the purposes of the *Econometrics* special issue, my goal is to explain some of the problems with significance testing, point out that these problems are not solved satisfactorily using *p*-values without significance testing, and show that confidence intervals are problematic too. The second major section presents a frequentist alternative. The alternative can be used on its own or in conjunction with significance testing, *p*-values, or confidence intervals. However, my preference is for the alternative to be used on its own.

2. Discontent with Significance Testing, *p*-Values, and Confidence Intervals

2.1. Significance Testing

Researchers use widely the null hypothesis significance testing (NHST) procedure, whereby the researcher computes a *p*-value, and if that value is under a threshold (usually 0.05), the result is declared statistically significant. Once the declaration has been made, the typical response is to conclude that the null hypothesis is unlikely to be true, reject the null hypothesis based on that conclusion, and accept the alternative hypothesis instead (Nickerson 2000). It is well-known that this sort of reasoning invokes a logical fallacy. That is, one cannot validly make an inverse inference from the probability of the obtained effect size or a more extreme one, given the null hypothesis; to the probability of the null hypothesis, given the obtained effect size (e.g., Cohen 1994; Fisher 1973; Nickerson 2000; Trafimow 2003).[1] The error is so common that it has a name: the modus tollens fallacy.

[1] This is an oversimplification. In fact, the *p*-value is computed from a whole model, which includes the null hypothesis as well as countless inferential assumptions. That the whole model is involved in computing a *p*-value will be addressed carefully later. For now, we need not consider the whole model to bring out the logical issue at play.

To see the reason for the name, consider that if the probability of the obtained effect size or a more extreme one given the null hypothesis were zero; then obtaining the effect size would guarantee that the null hypothesis is not true, by the logic of modus tollens (also termed denying the consequent). However, modus tollens does not work with probabilities other than zero and the unpleasant fact of the matter is that there is no frequentist way to compute the probability of the null hypothesis conditional upon the data. It is possible to use the famous theorem by Bayes; but most frequentists are unwilling to do that. Trafimow (2003, 2005) performed the Bayesian calculations and, not surprisingly, obtained very different findings from those obtained via the modus tollens error.[2] Thus, there is not only a logical invalidity; but a numerical one too.

An additional difficulty with making the modus tollens error is that to a frequentist, hypotheses are either true or false, and do not have probabilities between zero and unity. From this perspective, the researcher is simply wrong to assume that $p < 0.05$ implies that the probability of the null hypothesis given the data also is less than 0.05 (Nickerson 2000).

To be sure, researchers need not commit the modus tollens error. They can simply define a threshold level, such as 0.05, often termed an alpha level; and reject the null hypothesis whenever the obtained p-value is below that level and fail to reject the null hypothesis whenever the obtained p-value is not below that level. There is no necessity to assume anything about the probability of the null hypothesis.

But there are problems with threshold levels (see (Trafimow and Earp 2017) for a review). An important problem is that, under the null hypothesis, p-values have a uniform distribution between zero and one (0, 1). Therefore, whether the researcher obtains a p-value under the alpha level is partly a matter of luck. Just by getting lucky, the researcher may obtain a large sample effect size and a small p-value, thereby enabling publication. But in that event, the finding may be unlikely to replicate. Although this point is obvious from a statistical regression standpoint, the Open Science Collaboration effort (2015) showed empirically that the average effect size in the replication cohort of studies was less than half that in the original cohort of studies (from 0.403 to 0.197). Locascio (2017a) has argued that the most important disadvantage of NHST is that it results in scientific literatures replete with inflated effect sizes.[3]

Some have favored reducing the alpha level to lower values, such as 0.01 or 0.005 (e.g., Melton 1962; Benjamin et al. 2018); but these suggestions are problematic too. The most obvious problem is that having a more stringent alpha level merely increases statistical regression effects so that published sample effect sizes become even more inflated (Trafimow et al. 2018). Furthermore, more stringent thresholds need not increase replication probabilities. As Trafimow et al. (2018) pointed out, applying a more stringent alpha level to both the original and replication studies would render replication even more difficult. To increase the replication probability, the researcher would need to apply the more stringent alpha level to the original study and a less stringent alpha level to the replication study. And then there is the issue of what justifies the application of different alpha levels to the two categories of studies.

There are many problems with NHST, of which the present is only a small sampling. But this small sampling should be enough to render the reader highly suspicious. Trafimow and Earp (2017) presented a much larger list of problems and the subsequent section includes yet additional problems. Critique of the use of Fisherian p-values and Neyman-Pearson hypothesis testing for model selection has even made its way into (graduate) statistics textbooks; see, e.g., Paolella (2018, sct. 2.8).

---

[2] Also see Kim and Ji (2015) for Bayesian calculations pertaining to significance tests in empirical finance.

[3] The interested reader should consult the larger discussion of the issue in the pages of *Basic and Applied Social Psychology* (Grice 2017; Hyman 2017; Kline 2017; Locascio 2017a, 2017b; Marks 2017).

2.2. *p*-Values without Significance Testing

The American Statistical Association (Wasserstein and Lazar 2016) admits that *p*-values provide a poor justification for drawing conclusions about hypotheses. But might *p*-values be useful for something else? An often-touted possibility is to use *p*-values to obtain evidence against the statistical model of which the null hypothesis—or test hypothesis more generally—is a subset. It is a widely known fact—though not widely attended—that *p*-values depend on the whole model and not just on the test hypothesis. There are many assumptions that go into the full inferential statistical model, in addition to the test hypothesis, such as that the researcher sampled randomly and independently from a defined population (see (Trafimow), for a taxonomy of model assumptions). As Berk and Freedman (2003) pointed out, this assumption is practically never true in the soft sciences. Worse yet, there are many additional assumptions too numerous to list here (Armhein et al. 2019; Berk and Freedman 2003; Trafimow). As Armhein et al. (2019) concluded (p. 263): "Thus, statistical models imply countless assumptions about the underlying reality." It cannot be overemphasized that *p*-values pertain to models—and their countless assumptions—as opposed to just hypotheses.

Let us pretend, for the moment, that it is a valuable exercise for the researcher to obtain evidence against the model.[4] How well do *p*-values fulfill that objective? One problem with *p*-values is that they are well known to be unreliable (e.g., Halsey et al. 2015), and thus cannot provide strong evidence with respect to models or to the hypotheses that are subsets of models.[5] In addition to conceptual demonstrations of this point (Halsey et al. 2015), an empirical demonstration also is possible, thanks to the data file the Open Science Collaboration (2015) displayed online (https://osf.io/fgjvw/). The data file contains a wealth of data pertaining to the original cohort of published studies and the replication cohort. After downloading the data file, I obtained exact *p*-values for each study in the cohort of original studies; and for each study in the cohort of replication studies. After correlating the two columns of *p*-values, I obtained a correlation coefficient of 0.0035.[6] This empirical demonstration of the lack of reliability of *p*-values buttresses previous demonstrations involving mathematical or computer simulations.

And yet, there is a potential way out. Greenland (2019) suggested performing a logarithmic transformation of *p*-values: $(-1)\cdot\log_2(p)$. The logarithmic transformation causes *p*-values to be expressed as the number of bits of information against the model. For example, suppose that $p = 0.05$. Applying the logarithmic transformation implies that there are approximately four (the exact number is 4.32) bits of information against the model. An advantage of the transformation, as opposed to the untransformed *p*-value, is that the untransformed *p*-value has endpoints at 0 and 1, with the restriction of range potentially reducing the correlation that can be obtained between original and cohort sets of *p*-values. As an empirical demonstration, when I used the logarithmic transformation on the two columns of *p*-values obtained from the Open Science Collaboration data file, the original *p*-value correlation of 0.0035 jumped to 0.62! Thus, not only do transformed *p*-values have a more straightforward interpretation than do untransformed *p*-values; but they replicate better too. Clearly, if one insists on using a *p*-value to index evidence against the model, it would be better to use a transformed one than an untransformed one. However, there nonetheless remains the issue of whether it is worthwhile to gather evidence against the model in the first place.

Let us return to the issue that a *p*-value—even a transformed *p*-value—is conditioned not only on the test hypothesis; but rather on the whole model. For basic research, where one is testing a

---

[4] As will become clear later, this is not a valuable exercise; but the pretense is nevertheless useful to make an important point about logarithmic transformations of *p*-values.

[5] Sometimes *p*-value apologists admit *p*-value unreliability but point out that such unreliability has been known from the start. Although this contention is correct, it fails to justify *p*-values. That a procedure has been known from the start to be unreliable does not justify its use!

[6] This correlation, rounded to 0.004, was mentioned by Trafimow and de Boer (2018); but these researchers did not assess transformed *p*-values.

theory, there are assumptions in the theory (theoretical assumptions) as well as auxiliary assumptions used to connect non-observational terms in the theory to observational terms in empirical hypotheses (Trafimow). If the researcher is to employ descriptive statistics, such as means, standard deviations, and so on; there are statistical assumptions pertaining to issues such as whether means should be used or whether some other location statistic should be used, whether standard deviations should be used or whether some other dispersion statistic should be used, and so on (Trafimow 2019a). Finally, there are inferential assumptions such as those pointed out by Berk and Freedman (2003), especially concerning random and independent sampling. As suggested earlier, it is a practical certainty that not all the assumptions are precisely true, which means that the model is wrong (Armhein et al. 2019; Berk and Freedman 2003; Trafimow 2019b, forthcoming). The question then arises: What is the point in gathering evidence against a model that is already known to be wrong? The lack of a good answer to this question is a strong point against using statistics of any sort, even transformed *p*-values, to gather evidence against the model.

A counter can be generated out of the cliché that although all models are wrong, they might be close enough to correct to be useful (e.g., Box and Draper 1987). As an example of the cliché, suppose that based on the researcher's statistical model, she considers sample means as appropriate to estimate population means. In addition, suppose that the sample means really are close to their corresponding population means, though not precisely correct. It would be reasonable to argue that the sample means are useful, despite not being precisely correct, because they give the researcher a good approximation of the population means. Can this argument be extended to *p*-values or transformed *p*-values?

The answer is in the negative. To see why, consider that in the case of a *p*-value or a transformed *p*-value, there is no corresponding population parameter; the researcher is not estimating anything! And if the researcher is not estimating anything, closeness is irrelevant. Being close counts for much if the goal concerns estimation; but being close is irrelevant in the context of making accept/not accept decisions about hypotheses. In the case of such binary decisions, one cannot be close; one can only be correct or incorrect. It is worthwhile to reiterate. The Box and Draper (1987) quotation makes sense in estimation contexts; but it fails to save *p*-values or transformed *p*-values because nothing is being estimated. Obtaining evidence against a known wrong model fails to provide useful information.

2.3. Confidence Intervals

Do the foregoing criticisms of *p*-values, either with or without NHST, provide a strong case for using confidence intervals (CIs) instead? Not necessarily.

To commence, most researchers use CIs like they use *p*-values. That is, if the critical value falls outside the CI, it is statistically significant; and if the critical value does not fall outside the CI, it is not statistically significant. Used in this way, CIs are plagued with all the problems that plague *p*-values when used for NHST.

Alternatively, researchers may use CIs for parameter estimation. For example, many researchers believe that constructing a 95% CI around the sample mean indicates that the population mean has a 95% chance of being within the constructed interval. However, this is simply false. There is no way to know the probability that the population mean is within the constructed interval. To understand what a 95% CI really entails, it is necessary to imagine the experiment performed an indefinite number of times, with the researcher constructing a 95% CI each time. In this hypothetical scenario, 95% of the constructed 95% CIs would enclose the population mean but there is no way to know the probability that any single constructed interval contains the population mean. Interpreting a CI as giving the probability that the population parameter is within the constructed interval constitutes another way of making an inverse inference error, not dissimilar to that discussed earlier with respect to *p*-values. Furthermore, if one is a frequentist, it does not even make sense to talk about such a probability as the population parameter is either in the interval or not, and lack of knowledge on the part of the researcher does not justify the assignment of a probability.

Sophisticated CI aficionados understand the foregoing CI misinterpretations and argue instead that the proper use of CIs is to provide information about the precision of the data. A narrow CI implies high precision and a wide CI implies low precision. But there is an important problem with this argument. Trafimow (2018b) showed that there are three types of precision that influence the size of CIs. There is sampling precision: larger sample sizes imply greater sampling precision under the usual assumptions. There is measurement precision: the more the random measurement error, the lower the measurement precision. Finally, there is precision of homogeneity: the more similar the people in the sample are to each other, the easier it is to discern the effect of the manipulation on the dependent variable. CIs confound the three types of precision.[7]

The obvious counter is that even a confounded precision index might be better than no precision index whatsoever. But Trafimow (2018b, 2019a) showed that provided the researcher has assessed the reliability of the dependent variable, it is possible to estimate the three kinds of precision separately, which provides superior precision information than a triply confounded CI. The fact that the three types of precision can be estimated separately places the CI aficionado in a dilemma. On the one hand, if the aficionado is honestly interested in precision, she should take the trouble to assess the reliability of the dependent variable and estimate the three types of precision separately. On the other hand, if the aficionado is not interested in precision, there is no reason for her to compute a CI anyhow. Thus, either way, there is no reason to compute a CI.

A further problem with CIs, as is well-known, is that they fluctuate from sample to sample. Put another way, CIs are unreliable just as *p*-values are unreliable. Cumming and Calin-Jageman (2017) have attempted to justify that this is okay because most CIs overlap with each other. But unfortunately, the extent to which CIs overlap with each other is not the issue. Rather, the issue—in addition to the foregoing precision issue—is whether sample CIs are good estimates of the CI that applies to the population. Of course, in normal research, the CI that applies to the population is unknown. But it is possible to perform computer simulations on user-defined population values. Trafimow and Uhalt (under submission) performed this operation on CI widths, as well as upper and lower CI limits. The good news is that as sample sizes increase, sample CI accuracy also increases. The bad news is that unless the sample size is much greater than those typically used, the accuracy of CI ranges and limits is very poor.

In summary, CIs are triply confounded precision indexes, they tend not to be accurate, and they are not useful for estimating population values. To what valid use CIs can be put is far from clear.

2.4. Bayesian Thinking

There are many ways to "go Bayesian." In fact, Good (1983) suggested that there are at least 46,656 ways![8] Consequently, there is no feasible way to do justice to Bayesian statistical philosophy in a short paragraph. The interested reader can consult Gillies (2000), who examined some objectivist and subjectivist Bayesian views, and described important dilemmas associated with each of them. There is no attempt here to provide a critical review, except to say that not only do Bayesians disagree with frequentists; but Bayesians often disagree with each other too. For researchers who are not Bayesians, it would be useful to have a frequentist alternative that is not susceptible of the problems discussed

---

[7] To understand why, consider that CIs are based largely on the standard error. In turn, the standard error is based on the standard deviation and the sample size. Finally, the standard deviation is influenced by random measurement error but also by systematic differences between people. Thus, the standard deviation in the numerator of the standard error calculation is influenced by both measurement precision and precision of homogeneity; and the denominator of the standard error includes the sample size, thereby implicating the importance of sampling precision. Thus, all three types of precision influence the standard error. This triple confound is problematic for interpreting CIs.

[8] Good (1983) stated that Bayesians can make a variety of choices with respect to a variety of facets. In calculating that the number of Bayesian categories equals 46,656, Good also pointed out that this is larger than the number of professional statisticians so there are empty categories (p. 21).

with respect to NHST, *p*-values, and CIs. Even Bayesians might value such an alternative. We go there next.

3. The A Priori Procedure (APP)

The APP takes seriously that the researcher wishes to estimate population parameters based on sample statistics. To see quickly the importance of having sample statistics that are at least somewhat indicative of corresponding population parameters, imagine Laplace's Demon, who knows everything, and who warns researchers that their sample means have absolutely nothing to do with corresponding population means. Panic would ensue in science because there would be no point in obtaining sample means.

In contrast to the Demon's disastrous pronouncement, let us imagine a different pronouncement. Suppose the Demon offered researchers the opportunity to specify how close they wish their sample statistics to be to corresponding population parameters; and the probability of being that close. And the Demon would provide the necessary sample sizes to achieve specifications. This would be of obvious use, though not definitive. That is, researchers could tell the Demon their specifications, the Demon could answer with necessary sample sizes, and researchers could then go out and obtain those sample sizes. Upon obtaining the samples, researchers could compute their descriptive statistics of interest under the comforting assurance that they have acceptable probabilities of being within acceptable distances of the population values. After all, it is the researcher who told the Demon the specifications for closeness and probability. Because the researcher is assured that the sample statistics have acceptable probabilities of being within acceptable distances of corresponding population parameters, the need for NHST, *p*-values, or CIs is obviated.

Of course, there is no Demon; but the APP can take the Demon's place. As a demonstration, consider the simplest possible case where the researcher is interested in a single sample with a single mean, and where participants are randomly and independently sampled from a normally distributed population. That these assumptions are rarely true will be addressed later. For now, though, let us continue with this ideal and simple case to illustrate how APP thinking works.

Trafimow (2017) provided an accessible proof of Equation (1) below:

$$n = \left(\frac{Z_C}{f}\right)^2, \tag{1}$$

where

- f is the fraction of a standard deviation the researcher defines as sufficiently "close,"
- Z_C is the z-score that corresponds to the desired probability of being close, and
- n is the minimum sample size necessary to meet specifications for closeness and confidence.

For example, suppose the researcher wishes to have a 95% probability of obtaining a sample mean that is within a quarter of a standard deviation of the population mean. Because the z-score that corresponds to 95% confidence is approximately 1.96, Equation (1) can be solved as follows: $n = \left(\frac{1.96}{0.25}\right)^2 = 61.47 \approx 62$. In other words, the researcher will need to recruit 62 participants to meet specifications for closeness and confidence.[9]

The researcher now can see the reason for the name, a priori procedure. All the inferential work is performed ahead of data collection and no knowledge of sample statistics is necessary. But the fact that the APP is an a priori procedure does not preclude its use in a posteriori fashion. To see that this is indeed possible, suppose that a researcher had already performed the study and a second researcher wishes to estimate the closeness of the reported sample mean to the population mean, under

[9] Because participants do not come in fractions, it is customary to round upwards to the nearest whole number.

the typical value of 95% for confidence. Equation (2) provides an algebraic rearrangement of Equation (1), to obtain a value for f given that n has already been published:

$$f = \frac{Z_C}{\sqrt{n}}. \tag{2}$$

For example, suppose that the researcher had 200 participants. What is the closeness? Using Equation (2) implies the following: $f = \frac{1.96}{\sqrt{200}} = 0.14$. In words, the researcher's sample mean has a 95% probability of being within 0.14 standard deviations of the population mean.

3.1. *j* Groups

An obvious complaint to be made about Equations (1) and (2) is that they only work for a single mean. But it is possible to extend to as many groups as the researcher wishes. Trafimow and MacDonald (2017) derived Equation (3), that allows researchers to calculate the sample size per condition needed to ensure that all (j) sample means are within specifications for closeness and probability:

$$n = \left(\frac{\Phi^{-1}\left(\frac{\sqrt[j]{p(j\ means)}+1}{2} \right)}{f} \right)^2 \tag{3}$$

where

- j is the number of groups,
- $p(j\ means)$ is the probability of meeting the closeness specification with respect to the j groups,
- and Φ^{-1} is the inverse of the cumulative distribution function of the normal distribution.

Algebraic rearrangement of Equation (3) renders Equation (4) that can be used to estimate the precision of previously published research:

$$f = \left(\frac{\Phi^{-1}\left(\frac{\sqrt[j]{p(j\ means)}+1}{2} \right)}{\sqrt{n}} \right) \tag{4}$$

(Trafimow and Myüz (forthcoming) used Equation (4) to analyze a large sample of published papers in lower-tier and upper-tier journals in five areas of psychology. They found that although precision was unimpressive in all five areas of psychology, it was worst in cognitive psychology and least bad in developmental and social psychology.

3.2. Differences in Means

In some research contexts, researchers may be more interested in having the difference in two sample means be close to the population difference, than in having each individual mean be close to its corresponding population mean. Researchers who are interested in differences in means may use matched or independent samples. If the samples are matched, Trafimow et al. (forthcoming) derived the requisite equation:

$$t_{\frac{\alpha}{2},\, n-1} \leq \sqrt{n} f, \tag{5}$$

where $t_{\frac{\alpha}{2},\, n-1}$ is the critical t-score, analogous to the use of the z-score from Equation (1). Unfortunately, Equation (5) cannot be used in the simple manner that Equations (1)–(4) can be used. For instance, suppose a researcher wishes to specify $f = 0.2$ at 95% confidence for matched samples. The researcher might try $n = 99$, so the right side of Equation (2) is 1.99, which satisfies Equation (5). That is, $t_{\frac{\alpha}{2},\, n-1} = 1.81 \leq \sqrt{99} \cdot 0.2 = 1.99$. Alternatively, the researcher might try $n = 98$, which does not satisfy

Equation (5): $t_{\frac{\alpha}{2},\, n-1} = 1.9847 \not\leq \sqrt{99}\cdot 0.2 = 1.9799$. Because $n = 99$ satisfies Equation (5) whereas $n = 98$ does not, the minimum sample size necessary to meet specifications for precision and confidence is $n = 99$. Equation (5) is best handled with a computer that is programmed to try different values until convergence on the smallest sample size that fulfils requirements.

Equation (5) can be algebraically rearranged to give closeness, as Equation (6) shows:

$$f \geq \frac{t_{\frac{\alpha}{2},\, n-1}}{\sqrt{n}} \tag{6}$$

If the researcher uses independent samples, as opposed to matched samples, Equation (5) will not work and it is necessary instead to use Equation (7). When there are independent samples, there is no guarantee that the sample sizes will be equal, and it is convenient to designate that there are n participants in the smaller group and m participants in the larger group, where $k = \frac{n}{m}$. Using k, Trafimow et al. (forthcoming) derived Equation (7):

$$t_{\frac{\alpha}{2},\, q} \leq \sqrt{\frac{n}{k+1}} f, \tag{7}$$

where $t_{\frac{\alpha}{2},q}$ is the critical t-score that corresponds to the level of confidence level $1 - \alpha$ and degrees of freedom $q = n + \left[\frac{n}{k}\right] - 2$ in which $\left[\frac{n}{k}\right]$ is rounded to the nearest upper integer.

If the researcher has equal sample sizes, Equation (3) reduces to Equation (4):

$$t_{\frac{\alpha}{2},\, 2(n-1)} \leq \sqrt{\frac{n}{2}} f. \tag{8}$$

Like Equation (5), Equation (7) or Equation (8) is best handled using a computer to try out different sample sizes. Again, the lowest sample sizes for which the equations remain true are those required to meet specifications.

Alternatively, if the researcher is interested in the closeness of already published data, Equation (7) can be algebraically rearranged to render Equation (9); and Equation (8) can be algebraically rearranged to render Equation (10).

$$f \geq \frac{t_{\frac{\alpha}{2},\, q}}{\sqrt{\frac{n}{k+1}}} \tag{9}$$

$$f \geq \frac{t_{\frac{\alpha}{2},\, 2(n-1)}}{\sqrt{\frac{n}{2}}} \tag{10}$$

3.3. Skew-Normal Distributions

Most researchers assume normality and consequently would use Equations (1)–(10). But many distributions are skewed (Blanca et al. 2013; Ho and Yu 2015; Micceri 1989), and so Equations (1)–(10) may overestimate the sample sizes needed to meet specifications for closeness and confidence. The family of skew-normal distributions is more generally applicable than the family of normal distributions. This is because the family of skew-normal distributions employs three, rather than two, parameters. Let us first consider the two parameters of normal distributions: mean μ and standard deviation σ. For skew-normal distributions, these are replaced by the location parameter ξ and scale parameter ω, respectively. Finally, skew-normal distributions also include a shape parameter λ. When $\lambda = 0$, the distribution is normal, and $\xi = \mu$ and $\omega = \sigma$. But when $\lambda \neq 0$, then the distribution is skew-normal, and $\xi \neq \mu$ and $\omega \neq \sigma$. Although the mathematics are too complex to render here, skew-normal equations have been derived analogous to Equations (1)–(10) (e.g., Trafimow et al. 2019). In addition, Wang et al. (2019a) have shown how to find the number of participants necessary to meet specifications for closeness and confidence with respect to estimating the shape parameter. Finally, Wang et al. (2019b)

have shown how to find the number of participants necessary to meet specifications for closeness and confidence with respect to estimating the scale parameter (or standard deviation if normality is assumed).

3.4. Limitations

Although much progress has been made in the approximately two years since the APP was invented, there nevertheless remain limitations. The most important limitations are conceptual. Unlike many other inferential statistical procedures, the APP does not dictate what hypotheses to accept or reject. For those researchers who believe that other inferential statistical procedures, such as NHST, really do validly dictate what hypotheses to accept or reject, this is an important limitation. Hopefully, the remarks in the first major section of the present article have disabused the reader that any procedure is valid for making decisions about hypotheses. If the reader is convinced, then although the limitation remains serious in the absolute sense that it would be nice to have an inferential procedure that validly dictates what hypotheses to accept or reject; the limitation is not serious in the relative sense that other inferential procedures that make the promise fail to deliver, so nothing is lost by using the APP.[10]

A second conceptual limitation is suggested by one of the arguments against *p*-values. That is, the model is known wrong and so there is no point using *p*-values to gather evidence against it. It is tempting to apply model wrongness to the APP, where the models again will not be precisely correct. However, as we pointed out earlier, closeness counts heavily with respect to estimation; but does not count at all for binary decisions, that are either correct or incorrect. Because the APP is for estimation, if the model is reasonably close to being correct, though it cannot be precisely correct, the estimate should be reasonably good. And this point can be explained in more specific terms. Suppose that the model is close but not perfect, thereby resulting in a sample size that is slightly larger or slightly smaller than the precisely correct sample size necessary to meet specifications. In the case of the larger sample size, the result will be that the researcher will have slightly better closeness, and little harm is done, except that the researcher will have put greater than optimal effort into data collection. In the case of the smaller sample size, the researcher's sample statistics of interest will not be quite as close to their corresponding population parameters as desired; but may nevertheless be close enough to be useful. Therefore, although model wrongness always is important to consider, it need not be fatal for the APP whereas it generally is fatal for *p*-values.

There are also practical limitations. Although work is in progress concerning complex contrasts involving multiple means (assuming a normal distribution) or multiple locations (assuming a skew-normal distribution), the requisite APP equations do not yet exist. Similarly, for researchers interested in correlations, regression weights, and so on; although work is in progress; the requisite APP equations do not yet exist. Another practical limitation is the lack of an APP computer program that will allow researchers to perform the calculations without having to do their own programming. And yet, work proceeds and we hope to be able to address these practical limitations in the very near future.

3.5. APP versus Power Analysis

To some, the APP may seem like merely an advanced way to perform power analysis. However, this is not so as can be shown in both a general sense and in two specific senses. Speaking generally, the APP and power analysis have very different goals. The goal of power analysis is to find the number of participants needed to have a good chance of obtaining a *p*-value that comes in under threshold

---

[10] I thank an anonymous reviewer for pointing out that, due to the lack of cutoff points, this limitation can be considered a strength. According to the reviewer, "There is no cutoff point, so potentially all estimates could be viable."

(e.g., $p < 0.05$) when the null hypothesis is meaningfully violated.[11] In contrast, the APP goal is to find the number of participants necessary to reach specifications for closeness and confidence.

This general difference results in specific mathematical differences too.[12] First, power analysis depends importantly on the anticipated effect size. If the anticipated effect size is small, a power analysis will indicate that many participants are necessary for adequate power; but if the anticipated effect size is large, a power analysis will indicate that only a small sample size is necessary for adequate power. In contrast, the anticipated effect size plays no part whatsoever in APP calculations. For example, suppose that anticipated effect size for a single sample experiment is 0.80. A power analysis would show that only 13 participants are needed for power = 0.80; but an APP calculation would nevertheless demonstrate that the closeness value is a woeful 0.54.[13]

A second specific difference is that APP calculations are influenced, importantly, by the desired level of closeness. In contrast, power calculations are completely uninfluenced by the desired level of closeness. Moreover, absent APP thinking, few researchers would even consider the issue of the desired level of closeness. In summary, the APP is very different from power analysis, both with respect to general goals and with respect to specific factors that influence how the calculations are performed.

3.6. The Relationship between the APP and Idealized Replication

Much recent attention concerns replication probabilities across sciences. For example, the Open Science Foundation (2015) publication indicates that most published papers in top psychology journals failed to replicate. One of the many disadvantages of both *p*-values and CIs is that they fail to say much about the extent to which experiments would be likely or unlikely to replicate. In contrast, as Trafimow (2018a) explained in detail, the results of the APP strongly relate to reproducibility.

To understand the relationship, it is necessary to make two preliminary points. The first point is philosophical and concerns what we would expect a successful replication to entail. Because of scientists' addiction to NHST, most consider a successful replication to entail statistically significant findings in the same direction, in both the original and replication studies. However, once NHST is admitted as problematic, defining a successful replication in terms of NHST is similarly problematic. But the present argument extends beyond NHST to effect sizes more generally.

Consider the famous Michelson and Morley (1887) experiment that disconfirmed the presence of the luminiferous ether that researchers had supposed to permeate the universe.[14] The surprise was that the effect size was near zero, thereby suggesting that there is no luminiferous ether, after all.[15] Suppose a researcher today wished to replicate. Because larger effect sizes correspond with lower *p*-values, it should be clear that going by replication conceptions involving *p*-values, it is much more difficult to replicate smaller effect sizes than larger ones.[16] Thus, according to traditional NHST thinking, it should be extremely difficult to replicate Michelson and Morley (1887), though physicists do not find it so. This is one reason it is a mistake to let effect sizes dictate replication probabilities. In contrast, using APP thinking, a straightforward conceptualization of a successful replication is if the descriptive statistics of concern are close to their corresponding population parameters in both the original and replication studies.[17] An advantage of this conceptualization is that it treats large and

---

[11] For those who prefer CIs, an alternative goal would be to find the number of participants required to obtain sample CIs of desired widths.

[12] For elaborated mathematical discussions of the differences, see Trafimow and Myüz (Trafimow and Myüz) and Trafimow (2019b).

[13] See (Trafimow and Myüz (forthcoming) for details.

[14] Michelson received his Nobel Prize in 1907.

[15] It is interesting that Carver (1993) reanalyzed the data using NHST and obtained a statistically significant effect due to the large number of data points. As Carver pointed out, had Michelson and Morley used NHST, the existence of the luminiferous ether would have been supported, with incalculable consequences for physics (also see Trafimow and Rice 2009).

[16] A counter might be to use equivalence testing; but this is extremely problematic because it involves the computation of at least two *p*-values, whereas we already have seen that even one *p*-value is problematic.

[17] If specifications are not met in one of the two studies, that constitutes a failure to replicate.

small effect sizes equally. What matters is not the size of the effect; but rather how close the sample effect is to the population effect, in both the original and replication studies.

The second point is to imagine an idealized universe, where all systematic factors are the same in both the original and replication study. Thus, the only differences between the original and replication study are due to randomness.

Remembering that our new conceptualization of a successful replication pertains to the sample statistics of interest being close to their corresponding population parameters, in both the original and replication studies, invoking an idealized universe suggests a simple way to calculate the probability of replication. Specifically, the probability of replication in the idealized universe is simply the square of the probability of being close in a single experiment (Trafimow 2018a). We have already seen that all APP equations can be algebraically rearranged to yield closeness given the sample size used in the original study. Well, then, it is equally possible to fix closeness at some level and algebraically rearrange APP equations to yield the probability of meeting the closeness specification given the sample size used. Once this has been accomplished, the researcher merely squares that probability to obtain the probability of replication in the idealized universe. Trafimow (2018a) described the mathematics in detail and showed how specifications for closeness and sample size influence the probability of replication.

A way to attack the usefulness of the APP conceptualization of a successful replication is to focus on the necessity to invoke an idealized universe to carry through the calculations. But the attack can be countered in both general and specific ways. The general counter is that scientists often have suggested idealized universes, such as the ideal gas law in chemistry, Newton's idealized universe devoid of friction, and so on. The history of science shows that idealized conceptions often have been useful, though not strictly correct (Cartwright 1983). More specifically, however, consider that the difference between the APP idealized universe and real universe is that only random factors can hinder replication in the idealized universe whereas both random and systematic factors can hinder replication in the real universe. Because there is more that can go wrong in the real universe than in the idealized universe, it should be clear that the probability of replication in the idealized universe sets an upper limit on the probability of replication in the real universe. Because Trafimow (2018a) showed that most research has a low probability of replication even in the idealized universe, the probability of replication in the real universe must be even lower. In summary, whereas *p*-values and CIs have little to say about the probability of replication, the APP has much to say about it.

3.7. Criteria[18]

An unaddressed issue concerns the setting of criteria with respect to closeness and the probability of being close. The temptation is to set cutoffs; for example, that closeness must be at the 0.02 level or better, at 95% probability or better, to justify publication. However, this temptation must be resisted, lest the APP degenerate into dichotomous thinking that is currently problematic in the sciences. Instead of cutoffs, it would be better to have graduated verbal descriptions for what constitutes different levels of closeness and probabilities of being close. However, even graduated verbal descriptions may be problematic because researchers in different fields, or areas within fields, might justifiably differ with respect to what constitutes suitable verbal descriptions. For example, closeness at the 0.40 level might be "poor" in some fields or areas, and "acceptable" in others. It would be a mistake to impose such criteria from outside.[19]

One way to address the issue would be to have conferences, workshops, or symposia where people in similar fields and areas meet; become familiar with the APP, with a solid understanding about closeness and the probability of being close; and engage in serious discussions about criteria for

---

[18] I thank an anonymous reviewer for suggesting this issue.

[19] Trafimow (2018a) used some graduated verbal descriptions but also emphasized that these should not be taken very seriously.

graduated verbal descriptions. An alternative possibility would be to use the social media. Yet another alternative would be for editors of substantive journals to promulgate special issues devoted to setting criteria for graduated verbal descriptions such that substantive experts can provide perspectives.

Given academia's publish or perish culture, it cannot be overemphasized what a mistake it would be to turn APP thinking dichotomous, with publication thresholds for closeness and probability of being close. The ability of researchers to meet criteria for various verbal descriptions should only be one consideration for publication. Many factors should influence publication decisions, including the worth of the theory, the execution of the study, the feasibility of obtaining participants, the writing clarity, and others. Hopefully, having graduated verbal descriptions, instead of dichotomous cutoffs, that differ across fields and areas; will facilitate journal editors and reviewers to engage in more nuanced thinking that weighs many relevant factors.

4. Conclusions

The first major section considered NHST, p-values without NHST, and CIs. All were found wanting. Consequently, the second major section focused on a new way to think: the APP. The APP differs from the others because all the inferential work is performed before data collection. This is not to say that the others involve no work, whatsoever, before data collection. The setting of threshold levels, for instance, is work that is done before data collection when one uses NHST. But there nevertheless is a strong difference. With traditional procedures, once the data have been collected, it is still necessary to calculate p-values or CIs. In contrast, using the APP, the only inferential work that needs to be done after data collection is to acknowledge the results of the descriptive work. The researcher can be assured that the descriptive statistics have acceptable probabilities of being acceptably close to corresponding population parameters. After all, it is the researcher who decided what constitutes an acceptable probability and an acceptable degree of closeness and collected the requisite sample size to meet specifications. And if reviewers or editors believe that the investigator was too liberal in setting specifications, they have the option to reject the manuscript or insist that the researcher augment the sample. For example, if the researcher uses $f = 0.4$ and the editor favors $f = 0.1$, it is transparent how to calculate how much the researcher needs to augment the sample to reach the editor's specification.

A reasonable person might agree that the APP is a good thing; but also argue that NHST, p-values without NHST, or CIs are good too. As was stated earlier, there is nothing about APP calculations performed before data collection that renders impossible the calculation of p-values or CIs after data collection. Thus, it is possible to use all the procedures for the same study. This possibility need not be inconvenient for setting the APP apart from other procedures. If researchers were to routinely use the APP, they also would become accustomed to APP thinking. In turn, this would result in their eventually perceiving just how barren p-values and CIs are if one wishes to advance science. This is not to say that scientists should not test hypotheses. They should. But they should not depend on automatized decision-makers such as p-values and CIs to do it. Instead, researchers should perform much of their thinking up front and make a priori specifications for closeness and confidence. Then they should take their descriptive results seriously; with such seriousness being warranted by a priori specifications of acceptable closeness at acceptable probabilities. Of course, there are other factors that also influence the trust researchers place in descriptive statistics, such as the worth of the theory, the validity of the auxiliary assumptions, and so on. Whether or how much to believe substantive hypotheses, or the larger theories from which they are derived, is a process that cannot be automated. There will always remain an important role for expert judgment. The APP recognizes this explicitly.

Funding: This research received no external funding.

Conflicts of Interest: The author declare no conflict of interest.

References

Armhein, Valentin, David Trafimow, and S. Sander Greenland. 2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician* 73: 262–70.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, and Colin Camerer. 2018. Redefine statistical significance. *Nature Human Behavior* 33: 6–10. [CrossRef]

Berk, Richard. A., and David A. Freedman. 2003. Statistical assumptions as empirical commitments. In *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, 2nd ed. Edited by Thomas G. Blomberg and Sheldon Cohen. New York: Aldine de Gruyter, pp. 235–54.

Blanca, Maria, Jaume J. Arnau, Dolores López-Montiel, Roser Bono, and Rebecca Bendayan. 2013. Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 9: 78–84. [CrossRef]

Box, George E. P., and Norman R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. New York: John Wiley and Sons.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Carver, Ronald P. 1993. The case against statistical significance testing, revisited. *Journal of Experimental Education* 61: 287–92. [CrossRef]

Cohen, Jacob. 1994. The earth is round ($p < 0.05$). *American Psychologist* 49: 997–1003.

Cumming, Geoff, and Robert Calin-Jageman. 2017. *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. New York: Taylor and Francis Group.

Fisher, Ronald A. 1973. *Statistical Methods and Scientific Inference*, 3rd ed. London: Collier Macmillan.

Gillies, Donald. 2000. *Philosophical Theories of Probability*. London: Taylor and Francis.

Good, Irving J. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Greenland, Sander. 2019. The Unconditional Information in *p*-values, and Its Refutational Interpretation via S-values. *The American Statistician* 73: 106–14. [CrossRef]

Grice, James W. 2017. Comment on Locascio's results blind manuscript evaluation proposal. *Basic and Applied Social Psychology* 39: 254–55. [CrossRef]

Halsey, Lewis G., Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. 2015. The fickle *P* value generates irreproducible results. *Nature Methods* 12: 179–85. [CrossRef]

Ho, Andrew D., and Carol C. Yu. 2015. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement* 75: 365–88. [CrossRef]

Hyman, Michael. 2017. Can 'results blind manuscript evaluation' assuage 'publication bias'? *Basic and Applied Social Psychology* 39: 247–51. [CrossRef]

Kim, Jae H., and Philip I. Ji. 2015. Significance testing in empirical finance: A critical review and empirical assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Kline, Rex. 2017. Comment on Locascio, results blind science publishing. *Basic and Applied Social Psychology* 39: 256–57. [CrossRef]

Locascio, Joseph. 2017a. Results blind publishing. *Basic and Applied Social Psychology* 39: 239–46. [CrossRef] [PubMed]

Locascio, Joseph. 2017b. Rejoinder to responses to "results blind publishing". *Basic and Applied Social Psychology* 39: 258–61. [CrossRef] [PubMed]

Marks, Michael J. 2017. Commentary on Locascio. *Basic and Applied Social Psychology* 39: 252–53. [CrossRef]

Melton, Arthur. 1962. Editorial. *Journal of Experimental Psychology* 64: 553–57. [CrossRef]

Micceri, Theodore. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105: 156–66. [CrossRef]

Michelson, Albert A., and Edward W. Morley. 1887. On the relative motion of earth and Luminiferous ether. *American Journal of Science, Third Series* 34: 333–45. [CrossRef]

Nickerson, Raymond S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5: 241–301. [CrossRef]

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716. [CrossRef]

Paolella, Marc S. 2018. *Fundamental Statistical Inference: A Computational Approach*. Chichester: John Wiley and Sons.
Trafimow, David. 2003. Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review* 110: 526–35. [CrossRef] [PubMed]
Trafimow, David. 2005. The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review* 112: 669–74. [CrossRef]
Trafimow, David. 2017. Using the coefficient of confidence to make the philosophical switch from *a posteriori* to *a priori* inferential statistics. *Educational and Psychological Measurement* 77: 831–54. [CrossRef] [PubMed]
Trafimow, David. 2018a. An *a priori* solution to the replication crisis. *Philosophical Psychology* 31: 1188–214. [CrossRef]
Trafimow, David. 2018b. Confidence intervals, precision and confounding. *New Ideas in Psychology* 50: 48–53. [CrossRef]
Trafimow, David. 2019a. My ban on null hypothesis significance testing and confidence intervals. In *Structural Changes and Their Economic Modeling*. Edited by Vladik Kreinovich and Songsak Sriboonchitta. Cham: Springer, pp. 35–48.
Trafimow, David. 2019b. What to do instead of null hypothesis significance testing or confidence intervals. In *Beyond Traditional Probabilistic Methods in Econometrics*. Edited by Vladik Kreinovich, Nguyen Ngoc Thack, Nguyen Duc Trung and Dang Van Thanh. Cham: Springer, pp. 113–28.
Trafimow, David, and Michiel de Boer. 2018. Measuring the strength of the evidence. *Biomedical Journal of Scientific and Technical Research* 6: 1–7. [CrossRef]
Trafimow, David, and Brian D. Earp. 2017. Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology* 45: 19–27. [CrossRef]
Trafimow, David, and Justin A. MacDonald. 2017. Performing inferential statistics prior to data collection. *Educational and Psychological Measurement* 77: 204–19. [CrossRef]
Trafimow, David, and Hunter A. Myüz. Forthcoming. The sampling precision of research in five major areas of psychology. *Behavior Research Methods*.
Trafimow, David, and Stephen Rice. 2009. What if social scientists had reviewed great scientific works of the past? *Perspectives in Psychological Science* 4: 65–78. [CrossRef]
Trafimow, David, Valentin Amrhein, Corson N. Areshenkoff, Carlos J. Barrera-Causil, Eric J. Beh, Yusef K. Bilgic, Roser Bono, Michael T. Bradley, William M. Briggs, Héctor A Cepeda-Freyre, and et al. 2018. Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology* 9: 699. [CrossRef] [PubMed]
Trafimow, David, Tonhui Wang, and Cong Wang. 2019. From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement* 79: 129–50. [CrossRef] [PubMed]
Trafimow, David, Cong Wang, and Tonghui Wang. Forthcoming. Making the a priori procedure (APP) work for differences between means. *Educational and Psychological Measurement*.
Trafimow, David. Forthcoming. A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*. [CrossRef]
Wang, Cong, Tonghui Wang, David Trafimow, and Hunter A. Myüz. 2019a. Desired sample size for estimating the skewness under skew normal settings. In *Structural Changes and Their Economic Modeling*. Edited by Vladik Kreinovich and Songsak Sriboonchitta. Cham: Springer, pp. 152–62.
Wang, Cong, Tonghui Wang, David Trafimow, and Xiaoting Zhang. 2019b. Necessary sample size for estimating the scale parameter with specified closeness and confidence. *International Journal of Intelligent Technologies and Applied Statistics* 12: 17–29.
Wasserstein, Ronald. L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70: 129–33. [CrossRef]