To cite published article: van Zyl, C. J. J. (2018). Frequentist and Bayesian inference: A conceptual primer. *New Ideas in Psychology*, 51, 44-49.

Frequentist and Bayesian inference: A conceptual primer

Casper J J van Zyl

University of Johannesburg

Abstract

In recent years, there has been a crisis of confidence in many empirical fields including psychology, regarding the reproducibility of scientific findings. Among several causes thought to have contributed to this situation, the inferential basis of traditional, or so-called frequentist statistics, is arguably chief among them. Of particular concern is null hypothesis significance testing (NHST), which inadvertently became the de facto basis of scientific inference in the frequentist paradigm. The objective of this paper is to describe some of the most prominent issues plaguing frequentist inference, including NHST. In addition, some Bayesian benefits are introduced to show that it offers solutions to several problems inherent in frequentist statistics. The overall aim is to provide a non-threatening, conceptual overview of these concerns. The hope is that this will facilitate greater awareness and understanding of the need to address these matters in empirical psychology.

Keywords: Frequentist, Bayes, statistics, Null Hypothesis Significance Testing, reproducibility crisis, inference

1. Introduction

In recent years, a crisis of confidence has emerged in psychology (Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012; Pashler & Wagenmakers, 2012; Simmons, Nelson, & Simonsohn, 2011). However, psychology is not unique in this sense, given that the veracity of scientific findings have been questioned in several fields of study (Begley & Ellis, 2012; Button et al., 2013; Osherovich, 2011).

The problem is multifaceted with a number of factors contributing to reproducibility issues in empirical science. For instance, one issue is publication bottlenecks that result from too much research competing for space in too few outlets. Another is bias toward aesthetically pleasing findings and presentation over genuine concern for truth (Giner-Sorolla, 2012). HARKing, or "hypothesizing after the results are known" (Kerr, 1998), is one example where aesthetic concern over 'truth' has creeped into our science. This happens when findings are presented with a perfect fit between one's hypothesis and results, however, the hypothesis was in fact, formulated or amended after seeing the data. One reason for this is the notion that being wrong has somehow become a weakness in scientific papers, which only obfuscates the work, and confuses readers (Giner-Sorolla, 2012). Indeed, a very small proportion of psychology papers report findings that disconfirm their initial hypotheses (Bones, 2012).

In addition, there is the well-known file-drawer problem arising from journal-based bias toward novel findings, along with an unwillingness to publish null results (Ferguson & Heene, 2012). Even high quality research yielding null results have tacitly been relegated to outlets created for this purpose, for example, PLOS Missing Pieces and the Journal of Articles in Support of the Null Hypothesis. While it is no doubt positive that such journals do exist, it underscores the fact that null results are not typically deemed suitable for publication in mainstream journals.

However, one of the most problematic practices that have likely contributed to the confidence crisis in psychology is the overreliance and misuse of null hypothesis significance testing (Szucs & Ioannides, 2017). This, despite substantial and prolonged criticism and calls for reform (Cohen, 1994; Halsey, Curran-Everett, Vowler, & Drummond, 2015; Johnson, 2013; Nuzzo, 2014). A cursory inspection of empirical psychology journals reveal that classical (or so-called frequentist) statistics, and the practice of null hypothesis significance testing in particular, is still the default and dominant paradigm used for empirical research, and subsequent basis of scientific inference.

Unfortunately, frequentist inference suffer from a number of weaknesses that are the cause of considerable issues in psychological science. The nature of these problems constitute strong reason to consider seriously these shortcomings and the potential of Bayesian inference to advance the way we do empirical research in psychology. The aim of this paper is not to 'convert' frequentist researchers to Bayesians, but to discuss some of the most salient criticisms of the frequentist paradigm and to briefly point to some advantages of Bayesian inference (see also Dienes, 2008; 2011; Lambert, 2018; Wagenmakers, Lee, Lodewyckx, & Iverson, 2009; Wagenmakers et al., 2017).

The goal is to facilitate conceptual understanding of the inherent problems of frequentist statistics, so that researchers get a better 'feel' for the reason to change our practice, rather than simply being told that they should. This paper in no way represents a comprehensive discussion of a large and at times obscure issue, but seeks to prime a recognition of the need to change our practice and to provide some guideposts in that direction.

2. Null hypothesis significance testing

Broadly speaking, in classical (or frequentist) based research, scientific inference proceeds by designing studies in which data are collected and corresponding probabilities (*p*-values) generated. These *p*-values are the conditional probabilities of obtaining the observed data or

more extreme data given the null hypothesis (H_0).The *p*-values are compared to a set criterion (level of significance), which is used as a decision mechanism regarding hypotheses of interest. Typically, this entails rejecting the H_0 when $p < .05$. Failure to reject the H_0 occurs when the *p*-value is larger than the set threshold ($p > .05$). This *p*-value driven practice is known as null hypothesis significance testing (NHST). Although the use of *p*-values is essentially routine practice in psychological research, it is commonly misunderstood by students and researchers alike (Haller & Krauss, 2002; Oakes, 1986). The inferential basis of conclusions drawn from data following this routine procedure in the frequentist paradigm has been on the receiving end of substantial criticism for decades (Grigerenzer, 1998; 2004).

According to Cohen (1994), the first book length treatment of this issue appeared as far back as 1957. By 1970 an edited book titled 'The Significance Test Controversy' was published in which NHST was again criticized across the board (Morrison & Henkel, 1970). In it, one of the authors, Paul Meehl, portrayed NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (p. 265). These legitimate criticisms have not subsided over the years and continues to this day. This is because the practice of NHST has continued largely unabated since then. The recent replication crisis (Pashler & Wagenmakers, 2012) in psychology again brought the issue to the fore.

The subsequent 2015 Science article by the Open Science Collaboration, which reported results of a large-scale replication effort in psychology, provided empirical support of these long-standing criticisms. For the first time the magnitude of the problem was investigated empirically and could it now be quantified to some degree. In fairness, NHST is only one of several issues contributing to the reproducibility crisis. So, what then is the problem with NHST? This question is considered next, followed by a discussion of some additional frequentist problems. The paper concludes by presenting selected benefits of Bayesian inference as a viable alternative to frequentist inference.

2.1. Problems with NHST

As already mentioned, NHST functions as a decision mechanism in frequentist statistics (Dienes, 2008) where we compare an obtained *p*-value to a set criterion (commonly .05 or .01). Conventionally, if the observed $p > .05$, we fail to reject a null hypothesis, and if $p < .05$ we reject the null hypothesis. This is the way researchers typically report results in journal articles and what we teach our students. However, failing to reject the null hypothesis does not constitute evidence for it, and neither does rejecting the null hypothesis (H_0) necessarily mean we found evidence to support the alternative hypothesis (H_1). The inferential decision procedure based on this routine and 'mindless' use of *p*-values lies at the heart of the issue (Dienes, 2008; Grigerenzer, 2004).

2.2. *p*-values

The problem with *p*-values is that they are highly susceptible to misinterpretation. Common misconceptions include the notion that a *p*-value reflects the probability that an observed result is due to sampling error or a chance effect; the probability that the null hypothesis is true based on the data; or the probability that the alternative hypothesis is true given the data (Kline, 2004). In addition, *p*-values are believed to reflect the magnitude of an effect, in the sense that small *p*-values are thought to reflect large effects and vice versa; rejection of the H_0 is interpreted as confirmation of the H_1; and failure to reject the H_0 is considered evidence in support of it (Kline, 2004). If none of the above represents a correct interpretation, what then is a *p*-value?

A *p*-value is the conditional probability of encountering a test statistic as extreme, or more extreme than the one observed, assuming the null hypothesis is true. It takes the form $p(D \mid H_0)$ which is the probability (*p*) of the observed data (D) given (|) the null hypothesis (H_0). Important to note is that a *p*-value is a conditional probability that takes the truth of the H_0 as given. As such, no hypothesis is being tested. Neither the probability of the null $p(H_0 \mid D)$ or the probability of the alternative hypothesis $p(H_1 \mid D)$ is being evaluated. The null is

considered true a priori, and what is expected under the alternative hypothesis is just not actually considered (Wagenmakers et al., 2017).

At this point, a frequentist reader may object by arguing that when we design our studies, we are really trying to show evidence of some effect, which would typically be our alternative hypothesis. If our data suggest that we should reject the null ($p < .05$), surely we have then obtained evidence for the alternative hypothesis? Unfortunately, it does not follow logically that if the null hypothesis is extremely unlikely, that the alternative hypothesis must therefore be true. Consider the following syllogisms that Wagenmakers et al. (2017) use to demonstrate the error contained this view (see also Pollard & Richardson, 1987 and Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016).

Example A:

(Premise) If Tracy is an American then it is very unlikely that she is a US congresswoman;

(Premise) Tracy is a US congresswoman;

(Conclusion) It is very likely that Tracy is not an American.

or

Example B:

(Premise) If an individual is a man, he is unlikely to be the Pope;

(Premise) Francis is the Pope;

(Conclusion) Francis is probably not a man.

In both examples, it should be clear that despite the probability of the hypothesis being very unlikely, it clearly would make no sense to infer the opposite when the unlikely event is realized. To the contrary, it is impossible to become a congresswoman if one is not an American, or to become Pope if one is female (Pollard & Richardson, 1987).

These syllogisms would however, be valid forms of propositional logic, specifically *modus tollens* arguments, if the terms related to likelihood (i.e., unlikely, very likely, probably) were removed. *Modus tollens* is a well-known valid form of deductive reasoning. The problem is that the conclusion in a deductive argument is only valid when the premises are completely true (Manktelow, 2012). Unfortunately, neither human reasoning nor scientific inquiry proceeds on the basis of deductive reasoning. Rather, it is probabilistic and inductive (Chater & Oaksford, 2008; Manktelow, 2012). The unfortunate irony is that this erroneous form of reasoning became conventional practice and the basis of scientific inference in frequentist statistics. NHST-based inference using *p*-values takes exactly the same form as these syllogistic arguments (Pollard & Richardson, 1987; Rouder et al., 2016). Expressed as a syllogism, NHST takes this form:

(Premise) If the H_0 is true, then is it very unlikely that I will observe result X

(Premise) I observe result X

(Conclusion) Therefore, it is very likely that the H_0 is not true

While this may not immediately strike one as problematic, examples A and B above show the error contained in this form of reasoning. Importantly, these syllogisms show that it is possible for the observed data to be extreme under both the null and the alternative hypotheses at the same time. In fact, the data can simultaneously be extreme under the null hypothesis, and impossible under the alternative hypothesis (Szucs & Ioannidis, 2017).

Another inferential error of NHST occurs when we infer one conditional probability from knowing its inverse, because, we believe it to be the same thing. For instance, in NHST we compute the conditional probability of the observed data under the null hypothesis. When this probability is sufficiently small (i.e., $p < .05$), we reject the null hypothesis. Thus we are interpreting this result as its inverse, as the probability of the null hypothesis given the observed data $p(H_0 \mid D)$. However, the null hypothesis cannot be rejected, because it is assumed to be

true. At best, we can conclude that the data is very unlikely under the null hypothesis. The probability of the data given the null hypothesis is not the same as the probability of the null hypothesis given the data $p(D \mid H_0) \neq p(H_0 \mid D)$. Dienes (2008) clearly illustrates the problem that arises when one infers a conditional probability from its inverse as follows:

> The probability of being dead given that a shark has bitten one's head clean off, *p(*dead | head bitten clean off by shark), is 1. But the probability that a shark has bitten one's head clean off given that one is dead, *p(*head bitten clean off by shark | dead), is very close to zero. Most people die of other causes. (p. 276)

This example shows how problematic it is to equate the probability of the observed data given that the null hypothesis is true $p(D \mid H_0)$, with the probability that the null hypothesis is true given the observed data $p(H_0) \mid D)$. The probability of the latter $p(H_0 \mid D)$ is never considered in the frequentist paradigm. Accordingly, one cannot make pronouncements regarding the likelihood of the null hypothesis in the frequentist approach. Such an inference, can however, be made in the Bayesian framework (see section on Bayesian inference below).

Indeed, the logic of *p*-values is convoluted, in the sense that we really want to show the existence of some effect (H_1), so we postulate that there is no effect ($H_0 = 0$), hoping that the data will show that this cannot be the case (H_0 = rejected because $p < .05$), so that we can then infer that the effect we are interested in must therefore exist (H_1 = True). Not only is the inferential process tortured, it is not logically coherent, as we saw above.

A further irony is that despite the ubiquity of NHST in many fields, including psychology, it fails to provide the information that researchers presumably seek. Wagenmakers (2007) argues that researchers are not actually interested in knowing the probability of encountering a statistic as extreme or more than the one observed, given that the null hypothesis is true (i.e., the information obtained from NHST). What they really want to know is how much evidence the observed data provides for one hypothesis relative to another hypothesis.

So how did NHST using *p*-values as a decision procedure come about? Over time the evidential notion of *p*-values as conceptualized by Sir Ronald Fisher, along with Jersey Neyman and Egon Pearson's decision procedure for managing Type I and II error rates (α, β; also introducing the concepts of power and H_1), somehow became conflated (Szucs & Ioannides, 2017). However, these ideas are to a large degree, incompatible (see Hubbard & Bayarri, 2003 for more detail on this issue). The inadvertent merger of these two different procedures eventually became a *fait accompli*. This outcome has been lamented by many commentators (Hubbard & Bayarri, 2003; Christensen, 2005; Gigerenzer, 1993, 1998, 2004; Grigerenzer Krauss, & Vitouch, 2004), arguing that the existing approach has essentially become meaningless in applied research (Wagenmakers, Lee et al., 2009).

3. Further issues with frequentist inference

In frequentist statistics, evidence is not quantified. According to Wagenmakers, Lee et al. (2009), Fisher himself was of the opinion that *p*-values constitute evidence against the null hypothesis. However, this would require that *p*-values should at minimum conform to the requirement of consistency when used to evaluate evidence. This would necessitate a *p*-value of say, .05 to reflect the same amount of evidence in a sample with 12 observations and in a sample with 1200 observations. Unfortunately, it does not. Despite considerable debates regarding the evidential load of *p*-values in different sample sizes it turns out that a p-value of .05 actually reflect more evidence for some effect in a small sample than in a large sample (Wagenmakers, Lee et al., 2009). However, this does not mean that a statistically significant effect observed in a small sample will be robust. This is another common misinterpretation of NHST. In fact, it is much more likely that an effect will be overestimated in small samples compared to large samples due to increased measurement error. This is an error which Loken and Gelman (2017) refers to as the "what does not kill statistical significance makes it stronger" fallacy (p. 584).

A further underappreciated, but critical issue, is the fact that frequentist inference depends on the subjective intentions of the researcher. This refers to the fact that the sampling intentions of the researcher is a critical determinant of the final *p*-value upon which conclusions will be drawn. For example, different stopping intentions in the data collection process can yield different *p*-values for exactly the same data. Berger and Wolpert's classic story involving a frequentist statistician and a naïve scientist (1988, p. 30-33) makes the bizarre consequences of this practice evident. The story is repeated here in full to retain its impact:

> The naïve scientist has obtained 100 independent observations that are assumed to originate from a normal distribution with mean θ and standard deviation 1. In order to test the null hypothesis that $\theta = 0$, the scientist consults a frequentist statistician. The mean of the observations is 0.2, and hence the p-value is a little smaller than .05, which leads to a rejection of the null hypothesis. However, the statistician decides to probe deeper into the problem and asks the scientist what he would have done in the fictional case that the experiment had not yielded a significant result after 100 observations. The scientist replies that he would have collected another 100 observations. Thus, it may be hypothesized that the implicit sampling plan was not to collect 100 observation and stop; instead, the implicit sampling plan was to first take 100 observations and check whether $p < .05$. When the check is successful, the experiment stops, but when the check fails, another 100 observations are collected and added to the first 100, after which the experiment stops. The statistician then succeeds in convincing the scientist that use of the implicit sampling plan requires a correction in order to keep the Type I error rate at $p = .05$. Unfortunately, this correction for planning multiple tests now leads to a p-value that is no longer significant. Therefore, the puzzled scientist is forced to continue the experiment and collect an additional 100 observations. Note that the interpretation of the data (i.e., significant or not significant), depends on what the scientist was planning

> to do in a situation that did not actually occur. If the very same data had been collected by a scientist who had answered the statistician's question by saying, whether truthfully or not, "I would not have collected any more observations", then the data would have been judged to be significant. Same data, different inference. But the story becomes even more peculiar. Assume that the scientist collects the next 100 observations, and sets up another meeting with the statistician. The data are now significant. The statistician, however, persists and asks what the scientist would have done in case the experiment had not yielded a significant result after 200 observations. Suppose that the scientist now answers: This would have depended on the status of my grant renewal; If my grant is renewed, I would have had enough funds to test another 100 observations. If my grant is not renewed, I would have had to stop the experiment. Not that this matters, of course, because the data were significant anyway". The frequentist statistician then explains that the inference depends on the grant renewal; if the grant is not renewed, the sampling plan stands and no correction is necessary. But if the grant is renewed, the scientist could have collected more data, in the fictional case that the data would not have been significant after 200 observations. This calls for a correction for planning multiple tests, similar to the first one. The story concludes with the scientist resolving to never again share with the statistician the options he considers under different conditions.

In addition to exact researcher intentions, this story also shows how *p*-values are dependent on unobserved data and decisions that were never made. This refers to the fact that *p*-values are affected by data that were never observed (i.e., the hypothetical sampling distribution). This however, is argued to be a violation of the conditionality principle that statistical conclusions should only be based on actual observed data (see Wagenmakers 2007 and Berger & Wolpert, 1988 for detailed discussions of the issue).

Arguably, the most troubling aspect of NHST, is the fact that a statistically significant result can *always* be obtained (whether .05 or .01). This can be achieved by continually calculating *p*-values as the data comes in and stopping as soon as it drops below the set significance level. In frequentist statistics, this is guaranteed to happen eventually, even if the null hypothesis is known to be true (Armitage, McPherson & Rowe, 1969; Dienes, 2011; 2016; Meehl, 1990; Wagenmakers 2007). Not only will 5% of findings be statistically significant in the long run when the H_0 is true (Szucs & Ioannides, 2017), the probability of false positives is further compounded by so-called researcher degrees of freedom.

This refers to a researcher's decision flexibility during a study. Researchers have substantial discretion regarding the hypotheses to be tested, the design of the study, the analyses conducted and the reporting of results, which can each have a substantial and untoward influence on the final *p*-value obtained (Simmons et al., 2011). Although researcher degrees of freedom is something that has to be dealt with in any statistical paradigm, in NHST the problem has culminated in so-called *p*-hacking (Ioannides, 2005). This refers to the opportunistic use of researcher degrees of freedom in an effort to obtain statistically significant results, since non-significant results are unlikely to get published in peer-reviewed journals (Simmons et al., 2011). Common examples include the decision to run some additional participants when faced with a non-significant result, or making use of multiple comparisons that were never part of the initial analysis plan. There are a myriad of such seemingly benign decisions which greatly inflate the chance of finding false positive results (Wicherts et al. 2016). While pre-registration have in recent years gained much traction as a mechanism to reduce the adverse effect of researcher degrees of freedom, by making a clear distinction between exploratory and confirmatory research (Wagenmakers, Wetsels, Borsboom, Van der Maas & Kievit, 2012), it remains a widespread problem. Fortunately, many journals have now implemented compulsory pre-registration practices for confirmatory research and the list is growing (Wicherts et al.

2016). In the next section we consider Bayesian inference, an alternative paradigm that provides solutions to several of the issues plaguing frequentist statistics.

4. Bayesian inference

Bayesian inference is described only insofar as it enables discussion of selected Bayesian benefits over some frequentist problems highlighted above. The objective here is not to provide a comprehensive introduction to Bayesian statistics, or to fully explicate it (for more comprehensive treatments of Bayesian inference see e.g., Bernardo & Smith, 1994; Jaynes, 2003; Jeffreys, 1961; Lambert, 2018). Rather, the goal is to draw attention to some of the solutions that Bayesian statistics provide to several of the issues in NHST that impedes proper scientific inference.

First, it is important to note that there is an important philosophical distinction between Bayesian and frequentist statistics. According to Dienes (2008) Bayesian statistics uses probability to quantify uncertainty, or degree of belief. Thus, in the Bayesian paradigm, probability distributions are used to represent states of belief. This requires the use of priors. Priors are probability distributions used to represent what we believe or know about some state of the world before we observe the data. They are explicitly modeled in Bayesian statistics. Although the subjective nature of priors are criticized by frequentists, Bayesians point out there is subjectivity present in all analyses. In fact, it is considered a strength that subjectivity is made explicit in Bayesian analysis (Lambert, 2018), and that existing knowledge is formally incorporated into new conclusions (Nuzzo, 2014).

These prior beliefs are then updated via the likelihood to a posterior set of beliefs. Simply stated, it is the degree to which we should change our prior beliefs in the face of the present data (likelihood), to a new and updated set of beliefs (posterior distribution). Bayes' rule provides an optimal way with which to update prior beliefs in the face of evidence or

observed data. Accordingly, “Bayes' rule states that the posterior distribution is proportional to the product of the prior and the likelihood” (Wagenmakers, Lee et al., 2009, p. 10).

This stands in contrast to the frequentist assumption of probability as long-run frequency. In this paradigm, probability is used to inform us about the relative frequency of an event over the long run, that is, how often something is expected to happen in an infinite series of exact replications. When, for example, we want to examine the relationship between two variables, X and Y, by computing a correlation coefficient in frequentist statistics, we are in effect asking “assuming the null hypothesis that r = 0, what is the probability of obtaining our observed effect, or one that is even more extreme?

For frequentists, uncertainty resides in the hypothetical sampling distribution with the population parameter(s) being fixed (Zepher & Oswald, 2015). In this case, that the correlation is zero. In contrast, in the Bayesian paradigm probability is directed to the parameter(s) of interest, where uncertainty about the parameter is quantified by a range of probabilistic values based on actual observed data (Zepher & Oswald, 2015). Thus, in frequentist statistics, the parameters are fixed (null hypothesis assumed true) and the data considered variable (one sample from a hypothetically infinite sample possibilities), whereas, in Bayesian statistics, the parameters are variable and the data is considered constant or fixed (Zepher & Oswald, 2015).

5. Advantages of Bayesian inference

Arguably, one of the most useful features of Bayesian statistics is the ability to quantify evidence. For instance, the Bayesian answer to the fact that evidence is not quantified in frequentist statistics is the Bayes factor, “the primary tool used in Bayesian inference for hypothesis testing and model selection” (Berger, 2006, p. 378). Bayes factors quantify evidence from observed data on a continuous scale for and against the null and alternative hypothesis (Wagenmakers et al. 2017). Bayes factors indicate the degree to which the data supports the

null hypothesis, the alternative hypothesis, or neither hypothesis (Berger, 2006; see also Wagenmakers, Love et al. 2018 for suggested Bayes factor interpretation guidelines).

In contrast to NHST where no hypothesis is directly being tested because the null is assumed to be true, Bayes factors provide direct support for each hypothesis under consideration. Importantly, this includes the null hypothesis. For example, the Bayes factor could show that the H_0 is 5 times more likely than the H_1 under the present data, or that the data is 14 times more likely under the H_1 than under the H_0, or that the data is insensitive and does not provide clear evidence in support of either hypothesis (Dienes, 2016). Notice, the probability of both the H_0 and the H_1 is explicitly being evaluated. Thus, we are not asking about the likelihood of getting the data we did while accepting the truth of the null $p(\text{data} \mid H_0)$, we are asking how much evidence the present data provides for one hypothesis relative to another hypothesis or $p(H_0 \mid \text{data})$ vs $p(H_1 \mid \text{data})$. We might also want to quantify the confidence in parameters. For instance, we could determine how much more likely one parameter estimate is over another, say .25 over .35, using the Savage Dickey density ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2009). We could also determine the posterior probability of a certain parameter value, or determine which parameter values the data and priors make 95% probable, in a similar sense as the frequentist confidence interval.

Although Bayesian credible intervals are somewhat similar to well-known confidence intervals used in frequentist statistics, it is important to note that there is a critical conceptual difference between them. As a reminder, in frequentist statistics, a 95% confidence interval for an estimated parameter refers to an interval that in the case of repeated sampling will have a 95% probability of containing the true value of the population parameter (for a detailed discussion see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). As such, it is not possible to know if a true value falls within the 95% confidence interval of any given sample in the frequentist paradigm. In contrast, one can directly determine the 95% probability of a

true value falling in a Bayesian credible interval (Morey et al., 2016). Thus, Bayesian inference using credible intervals tells us what we want to know, but cannot know when using frequentist confidence intervals.

Another important benefit of Bayesian statistics is that evidence can be continually computed and updated as the data comes in. This is possible because all inferences in Bayesian statistics are based on actual observed data (Wagenmakers et al., 2017) This is a major advantage of Bayesian over frequentist methods because inference is not dependent on data that was never observed (e.g., hypothetical sampling distribution), or the exact intentions with which data was collected (Dienes, 2008).

Lastly, Wagenmakers et al. (2017) note that in contrast to frequentist statistics, Bayesian inference is logically coherent and internally consistent. Thus, none of the consistency and coherence issues plaguing frequentist statistics are present in the Bayesian paradigm. This is guaranteed because Bayesian inference conforms to the axioms of probability theory, which is argued to be the cornerstone of Bayesian statistics (Lindley, 1985, 2000, 2006).

6. Conclusion

Some of the most problematic and pervasive problems afflicting scientific inference using frequentist statistics were described in this paper. It should, hopefully, be evident how the inferential shortcomings of frequentist statistics could contribute to reproducibility challenges in psychology. In addition, it becomes apparent that the choices we make as methodologists and substantive researchers is a critical determinant of this process. For researchers wishing to do meaningful and robust work, these issues are unavoidable. Moreover, if we want our scientific efforts to be, and remain credible, we will also have to address them explicitly in our teaching. The selection of frequentist problems described in this paper was by no means exhaustive and only selected Bayesian benefits were described. Readers are encouraged to

peruse the references for work that offers comprehensive and technically detailed treatments of the points raised in this paper.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A,* 132, 235-244.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483,* 531-533.

Berger, J. O. (2006). "Bayes Factors." In Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., and Johnson, N. L. (eds.), *Encyclopedia of Statistical Sciences, vol. 1* (2nd ed.), 378–386. Hoboken, NJ: Wiley.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward (CA): Institute of Mathematical Statistics.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory.* Chichester, New York: John Wiley & Sons.

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition—A satire in one part. *Perspectives on Psychological Science*, 7,307–309

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1-12.

Chater, N., & Oaksford, M. R. (eds.) (2008). *The probabilistic mind. Prospects for a Bayesian cognitive science.* Oxford: Oxford University Press

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59,* 121-126.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* London: Palgrave MacMillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6(3), 274-290.*

Dienes, Z. (2016). How Bayes factors change our scientific practice. *Journal of Mathematical Psychology, 72,* 78-89.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspective in Psychological Science, 7*(6), 555-561. doi: 10.1177|1745691612459059.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In: Keren, G., Lewis, C. (eds) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Erlbaum, NY: Hillsdale

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21,* 199-200.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33,* 587-606.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7,* 562–571.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1-20.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12,* 179-185.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing(with comments). *The American Statistician*, 57, 171–82.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696-701.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23,* 524-532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 19313-19317.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lambert, B. (2018). *A student's guide to Bayesian statistics.* London: Sage.

Lindley, D. V. (1985). *Making decisions* (2nd ed.). London: Wiley.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician, 49,* 293-337.

Lindley, D. V. (2006). *Understanding uncertainty.* Hoboken: Wiley.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585.

Manktelow, K. (2012). *Thinking and reasoning: An introduction to the psychology of reason, judgment, and decision making.* New York: Psychology Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244.

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.

Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The Significance Test Controversy - A Reader*. London: Butterworths.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217-243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Nuzzo, R. (2014). Statistical errors. *Nature, 506,* 150-152.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Hoboken, NJ: Wiley.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

Osherovich, L. (2011). Hedging against academic risk. *Science-Business eXchange, 4*(15).

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7,* 528-530.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin, 102,* 159–163.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M. Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8,* 520-547.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366.

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience, 11*:390. doi: 10.3389|fnhum.2017.00390

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M. & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Frontiers of Psychology. 7:1832. doi: 10.3389/fpsyg.2016.01832

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804.

Wagenmakers, E. J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2009). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, and P. A. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses*, pp. 181-207. Springer: New York.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2009). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60,* 158-189.

Wagenmakers, E. J., Love, J., & Marsman, M. et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review,* 25, 58-76. doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E. J., Marsman, M., Jamil, T. et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review,* 25, 35-57. doi.org|10.3758|s13423-017-1343-3

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke1, D., Steingroever, H., Jeff N., Rouder, J. N., & Morey, R. (2017). The Need for Bayesian Hypothesis Testing in Psychological Science. In S.O. Lillienfeld and I.D. Waldman (Eds.), Psychological science under scrutiny. (pp. 123-138). Chichester, UK: John Wiley& Sons Inc.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science,* 7, 632-638.

Zepher, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41*(2), 390-420.