CREMA

Center for Research in Economics, Management and the Arts

A Non-Bayesian Approach to Scientific Inference on Treatment-Effects

Working Paper No. 2020-14

CREMA Südstrasse 11 CH - 8008 Zürich www.crema-research.ch

A Non-Bayesian Approach to Scientific Inference on Treatment-Effects [*]

Subrato Banerjee[†] Benno Torgler[‡]

Abstract

Because the use of *p*-values in statistical inference often involves the rejection of a hypothesis on the basis of a number that itself assumes the hypothesis to be true, many in the scientific community argue that inference should instead be based on the hypothesis' actual probability conditional on supporting data. In this study, therefore, we propose a non-Bayesian approach to achieving statistical inference independent of any prior beliefs about hypothesis probability, which are frequently subject to human bias. In doing so, we offer an important statistical tool to biology, medicine, and any other academic field that employs experimental methodology.

Keywords: Statistical inference, experimental science, hypothesis testing, conditional probability

[*] This paper has received valuable feedback from Zhiyong Johnny Zhang, KR Parthasarathy, Kyle M Lang, Domenico Marinucci, Karen Kafadar, Tony Beatton, JV Meenakshi, and all the active members (notably Nathaniel Stevens, David Bristol, Chauncey Dayton, Richard Potthoff, JS (Steve) Marron, Andrew Ekstrom, Eugene Komaroff, and Robert O'Brien) in the public discussion forums of the American Statistical Association (ASA). I extend my thanks to Jason Connor and Douglas Landsittel (in the *ASA discussion forum*) for their interest and enthusiasm.

[†] Corresponding author. Australia India Institute, University of Melbourne; Centre for Behavioural Economics, Society and Technology (BEST), Queensland University of Technology; and Visiting faculty, Manipal University, Jaipur. Email: subrato.banerjee@unimelb.edu.au. ORCID: https://orcid.org/0000-0003-1218-8533.

[‡] School of Economics and Finance; Centre for Behavioural Economics, Society and Technology (BEST), Queensland University of Technology; CREMA—Center for Research in Economics, Management, and the Arts, Switzerland. Email: benno.torgler@qut.edu.au. ORCID: https://orcid.org/0000-0002-9809-963X.

1. Introduction

Scientific inquiry has traditionally relied heavily on statistical inference methodology, with biological, medical, and ecological sciences all overemphasizing the role of significance testing (Meehl 1967, Yoccoz 1991). Not only was such reliance criticized as "overused, misused, and often inappropriate" by a Biometrics Working Group at the Wildlife Society's 1998 annual conference (Johnson 1999, p. 763), but many academic journals, particularly in psychology (e.g. Trafimow and Marks 2015), have recently banned the use of *p*-values. Many scholars, however, view these editorial policies as illegitimate censorship (Meehl 1997), and *p*-values remain the most widely reported measure in statistical inference across diverse fields of experimental sciences, including clinical trials. The central objection to their use is that, whereas science demands the *probably that a hypothesis is true* conditional on support from data, the *p*-value provides none of the probability information necessary to strongly reject the null hypothesis, conveying only the probability that the finding is valid if the null hypothesis holds.

The resulting search for alternative methods of statistical inference is particularly important for the biological and medical sciences because of their core mission (among others) of measuring and comparing organism phenotypes (Nakagawa and Schielzeth 2010) relies heavily on scientific experimentation. Clinical experiments, for example, directly determine whether a new drug should be introduced to the market. Yet when biological applications need to combine results from independent tests whose raw data cannot be pooled (Rice 1990), a *p*-value cannot be a final verdict on whether or not to reject the null hypothesis. If, for example, a study achieves a *p*-value of 0.01, the probability that an *exact* replication will also produce a *p*-value of 0.01 is only 50% and not 99% as many assume (Nuzzo 2014).[1] Obviously, scientists should not

[1]It should be noted that even Bonferroni *p*-value corrections do not fully address this problem.

rely on a process shown to be so blatantly invalid and incapable of accurately reflecting the quality of evidence, making it also prone to false positives.

Nor is such criticism new: scholars have frequently expressed concerns about significance tests since at least the 1950s (e.g., Hogben 1957, Bakan 1966, Morrison and Henkel 1970). Yet not until 2019 did the *American Statistician* dedicate an entire issue to encouraging the search for alternate measures beyond *p*-values (Wasserstein et al. 2019) to facilitate our investigation into realities that can often only be inferred. In this paper, we offer just such a powerful alternative by demonstrating that given any sample-size *n* and observed (treatment) effect-size *t* (i.e., Cohen's *d*), the probability of the truthfulness of *exactly* one of the two competing (mutually exclusive and exhaustive) hypotheses has a lower bound given by

$$\theta_{LB} = \frac{t^2 n - 4}{t^2 n}$$

In other words, one of the two competing (null and alternate) hypotheses will have a probability of being true that always exceeds the above expression.

2. Beyond a Bayesian Approach

The battle between Bayesians and frequentists is an old and recurring one,[2] with Fisher (1950) convinced that "the theory of inverse probability," being based on an error, "must be wholly rejected" (p. 9) and Gigerenzer (1993) calling the inverse probability maneuver "Bayesian Id's wishful thinking" (p. 330). Neyman (1957) also criticized

> "the dogmatism … occasionally apparent in the application of Bayes' formula when the probabilities a priori are not implied by the problem treated and an author attempts to

[2]For an excellent overview of the strongest non-Bayesian arguments, see Gelman (2008).

> impose on the consumer of statistical methods the particular a priori probabilities invented by himself for this particular purpose." (p. 19).

Bayesians, however, stress the Bayesian test's power to allow quantification of the evidence in favor of the null hypothesis (Wagenmakers et al. 2011, p. 429), believing Bayesian estimation to be richer, more informative, and meaningful than null hypothesis significance testing (Kruschke 2013). Frequentist subjectivity, in contrast, is "hidden from view," being "carefully locked up" in the minds of those who compiled the data set (Wagenmakers et al. 2008, p. 1999). Yet, according to Gigerenzer (1993), "Fisher both rejected the Bayesian cake and wanted to eat it, too … [seeing]…the level of significance as a measure of the degree of confidence in a hypothesis", a measure that some academics have since well-nigh transformed into a type of "Bayesian posterior probability" (p. 330). In fact, because of its simplicity, the Bayes factor may end up being used in a similarly dogmatic way as *p*-values.

Another criticism of the Bayesian approach is its heavy reliance on prior beliefs about hypothesis probabilities (Berger 1985, Ghosh et al. 2006), making it highly susceptible to human bias, although the biases and fallacies documented by scholars like Kahneman and Tversky frequently disappear when subjects are asked for frequency judgements rather than single-event probabilities (Gigerenzer 1991). Bayesians, on the other hand, argue that Bayes factors are not oversensitive to reasonable variation in priors (Rouder et al. 2009), so if an appropriate distribution over priors is used, there is relatively weak impact on Bayesian parameter estimation (Schönbrodt et al. 2017). Nevertheless, because the Bayesian approach is still dependent on judgment, increasing researcher degrees of freedom can produce additional ambiguity (Simmons et al. 2011) or subjectivity (Halsey 2019) even when transparent rather than hidden (Rouder et al.

2009, Wagenmakers et al. 2008). The existence of two different Bayesian schools of thought[3] around the selection of priors underscores these latter as a substantial Achilles heel of Bayesian statistics.

Hence, in proposing a suitable alternative to the *p*-value for use in statistical inference, we take a completely different approach by obtaining the probabilities of the underlying hypotheses (conditional on supporting data) *without any use of prior beliefs.* In doing so, we address several contemporary objections to the *p*-value problem that were actually identified back in the 1920s. Our proposed method thus has useful implications for experiments in various fields ranging from clinical trials to tests aimed at shaping public policy. Our method does not resort to beliefs but rather provides a statistical rule derived from problem conditions and observed data. Such general features, as Neyman (1957) points out, are at the core of Gauss' brilliant foundation for the least squares method.

3. Preliminaries

As a first step, we define the null hypothesis H_0, and the alternate hypothesis H_1 as follows:

H_0: *A*

H_1: *B*

In the interest of generality, *A* and *B* could be any two competing *positive* (i.e., non-normative) statements (see Example 1). Then, for any statistical test that uses a predetermined decision rule to differentiate the hypotheses based on data, we define two competing events D_A (data supports *Statement A*) and D_B (data supports *Statement B*) for some specified decision rule. Thus, event D_B leads to the rejection of H_0 while event D_A leads to the rejection of H_1. In the context of

[3] Whereas the subjective Bayes school posits that priors should reflect the analyst's a priori beliefs, the objective school argues that they should reflect as few assumptions as possible (Rouder et al. 2009, p. 229).

clinical trials, for instance, statement A could simply mean "ineffective" and statement B, "effective".

Example 1: First, assuming a *Statement A* that "the success probability of a binomial trial equals 0.70" and a negating *Statement B* that "the success probability of a binomial trial does not equal 0.70," we define a random variable X that represents the number of successes in 10 trials. If we then apply the decision rule "do not reject H_0 if $6 \leq X \leq 8$, and reject otherwise," we count exactly 7 successes as a realization of event D_A, and (say) 2 successes as a realization of event D_B. Clearly, this decision rule implies that $D_A = \{x: 6 \leq x \leq 8\}$, and $D_B = \{x: not(D_A)\}$.

In the resulting setup, the conditional probability $P(D_B|A)$ is α, the probability of a Type I error, while $P(D_A|B)$ is β, the probability of a Type II error. At the same time, because the data can support exactly one hypothesis at a time (i.e. either A or B), $P(D_A) + P(D_B) = 1$. Similarly, because we assume that statements A and B are two mutually exclusive and exhaustive possibilities, $P(A) + P(B) = 1$. Based on these preliminaries, the conditional probability $P(D_A|A)$ equals $1 - \alpha$, and the conditional probability $P(D_B|B)$ equals $1 - \beta$. Thus, using our notation,

$$\frac{P(D_B|A)}{P(D_A|A)} \cdot \frac{P(D_A|B)}{P(D_B|B)} \equiv \left(\frac{\alpha}{1-\alpha}\right)\left(\frac{\beta}{1-\beta}\right) \equiv \frac{\alpha\beta}{(1-\alpha)(1-\beta)} = \mu \qquad (1)$$

where μ merely designates the right-hand-side of (1). We can then derive our first proposition in relation to conditional probabilities:

Lemma 1. *The conditional probability reversal rule (CPRR):*

$$\frac{P(D_B|A)}{P(D_A|A)} \cdot \frac{P(D_A|B)}{P(D_B|B)} = \frac{P(A|D_B)}{P(A|D_A)} \cdot \frac{P(B|D_A)}{P(B|D_B)} \qquad (2)$$

Proof: The above statement is true because both terms represent (and are equal to) the same expression below:

$$\frac{P(D_B \cap A)}{P(D_A \cap A)} \cdot \frac{P(D_A \cap B)}{P(D_B \cap B)}$$

This completes the proof. ■

It is important to note that equation (2) is an identity, whose usefulness lies in the fact that each of the conditional probabilities on the left-hand side is reversed on the right-hand side. In short, one side involves probabilities conditional on the hypotheses, and the other, probabilities conditional on the supporting data. Understanding the latter involves Bayesian methods whose potential bias (from heavy reliance on prior beliefs about each hypothesis' probability) has critical implications for fields like clinical research and disaster-related policy-making. Our method, in contrast offers a more bias-free tool for sensitive analyses.[4]

3. Solution to the *p*-value problem

3.1. The constant probability approach (CPA)

If we accept that the conditional probability of a hypothesis given support from data is a constant *regardless of the (null or alternate) hypothesis,* then the solution to the *p*-value problem is trivial. This *constant* probability, denoted by θ, equals $1/(1 + \sqrt{\mu})$, where μ is as defined in (1):

Lemma 2. If the conditional probability of a hypothesis, given that it is supported by observed data, is a constant, then it is uniquely equal to $\theta = 1/(1 + \sqrt{\mu})$.

Proof: Combining equalities (1) and (2) yields

$$\frac{P(A|D_B)}{P(A|D_A)} \cdot \frac{P(B|D_A)}{P(B|D_B)} \equiv \frac{\alpha\beta}{(1-\alpha)(1-\beta)} = \mu$$

Rearranging the left-hand terms with $P(B|D_A) = 1 - P(A|D_A)$, and $P(A|D_B) = 1 - P(B|D_B)$, gives

[4]Fisher (1973) provides concrete examples of when such prior assumptions on the probabilities of the hypotheses may be justified. However, as many researchers agree, the existence of such a rationale is the exception rather than the rule.

$$\left[\frac{1-P(A|D_A)}{P(A|D_A)}\right]\cdot\left[\frac{1-P(B|D_B)}{P(B|D_B)}\right]\equiv\left[\frac{1}{P(A|D_A)}-1\right]\cdot\left[\frac{1}{P(B|D_B)}-1\right]\equiv\mu \qquad (3)$$

Since the conditional probability of a hypothesis (given that the observed data supports it), is (assumed to be) a constant, $\theta = P(A|D_A) = P(B|D_B)$, the above equation reduces to $\left[\frac{1}{\theta}-1\right]^2 \equiv \mu$, and solving for θ completes the proof. ■

It should additionally be noted that although the above assumption of constant conditional probability is arguably very strong, even without it, the value of θ is useful as the most precise measure of test strength:

Definition 1. *Test strength* is denoted by $\theta = 1/(1 + \sqrt{\mu})$.

In this case, because $\theta = 1$ only when $\alpha = 0$ or when $\beta = 0$, the closer θ comes to unity, the greater the strength of the underlying test. At the same time, because θ is obtained from the intersection of the (45°) line $P(A|D_A) = P(B|D_B)$ and the inverse relation (3), neither $P(A|D_A)$, nor $P(B|D_B)$ can be simultaneously less than θ. The use of θ as a measure of inference strength is thus not blatantly invalid (as often argued of *p*-values) given its direct relation to the probability of an underlying hypothesis (absent any prior judgment) conditional on the existence of supportive data. As a result, θ is of more immediate scientific interest than a *p*-value.

In addition, even when $P(A|D_A) \neq P(B|D_B)$, the information on θ is useful for a crucial understanding of the scientific principle of *falsifiability* (Popper 1934). In our setting, a null hypothesis is falsifiable if the data support the alternative and vice versa. To illustrate using Example 1, *Statement A* is *falsifiable* since $P(D_B|A) > 0$, and *Statement B* is *falsifiable* since $P(D_A|B) > 0$. Because $P(D_B|A) = \alpha$, and $P(D_A|B) = \beta$, requiring that each of the two competing

hypotheses be falsifiable translates to $\alpha\beta > 0$, and thus $\mu > 0$, which in turn implies that $\theta < 1$. This latter condition is the basis for our next definition:

Definition 2. *Falsifiability* is summarized by the condition that $\theta < 1$ (with a weaker notion requiring that at least one of the two competing hypotheses be falsifiable, so that $\theta \leq 1$).

Example 2: For a binomial trial, one can decide that the probability of success is 0.90 (H_0) and not 0.60 (H_1) if X, the observed number of successes, in $n = 20$ trials, is greater than 14 (thus D_A: $X > 14$, and D_B: $X \leq 14$). For this experiment, $P(D_B|A) = \alpha = 0.0114$, and $P(D_A|B) = \beta = 0.1255$ (our Type I and Type II errors respectively), giving us $\mu = 0.0017$, and the strength of this test $\theta = 96.09\%$. Thus, if $P(A|D_A) < 0.9609$, then $P(B|D_B) > 0.9609$, (and conversely). This means that the conditional probabilities of both the hypotheses (given support from the data) cannot be simultaneously less than 96.09%. This test clearly has a strong design.

For decision rules like the above, it is clear that α and β are inversely related, meaning that μ, and thus θ (our measure of inference strength), is relatively *invariant* to changes in Type I and Type II errors, making it *more stable* than either. This feature of θ has great scientific value given that *p*-values are seldom replicable (Nuzzo 2014). If, for instance, we alter the decision rule of Example 2 with one that rejects the null hypothesis when $X \leq 15$, then $P(D_B|A) = \alpha$ almost quadruples to 0.0433, while $P(D_A|B) = \beta$ falls to 0.0509 (less than half the initial value), yet the inference strength only marginally lowers to $\theta = 95.30\%$. In Example 2, therefore, the rejection region could be chosen so that α and β maximize θ instead of minimizing one error after fixing the other. It should further be noted that based on our decision rules, both the above hypotheses are *falsifiable*.

Example 3: To demonstrate that Type II errors are inconsequential when *p*-values are too strong, if we assume a 1% power with $\beta = 0.99$, a value of $\alpha = 0.0001$ yields $\theta = 90.95\%$, which increases to $\theta = 99.01\%$ when $\beta = 0.50$.

3.2. Large sample properties of inference strength θ

To determine the large sample properties of inference strength, we first introduce two monotonic transformations, $a = 1/\sqrt{\alpha}$ and $b = 1/\sqrt{\beta}$ so that

$$\mu = \frac{1}{\left(\frac{1}{\alpha} - 1\right)\left(\frac{1}{\beta} - 1\right)} = \frac{1}{(a^2 - 1)(b^2 - 1)} \tag{4}$$

Because θ is bounded by 1, we focus on the minimum possible inference strength θ attainable by a given sample size *n*. In the interest of generality, we allow for sampling procedures that may or may not satisfy the conditions for the validity of the central limit theorem (CLT). We thus consider the sample mean $\bar{X}$ of a random sample of size *n* from one of two possible populations: $\mu = \mu_0$ (under H_0) and $\mu = \mu_1$ (under H_1). Assuming $\mu_1 > \mu_0$ (without loss of generality), our decision rule for realizations $\bar{x}$ of the sample mean and a critical value *c* (> *0*) is

$$d(\overline{x}) = \begin{cases} \mu_0 \; for \; \overline{x} \le \mu_0 + c \\ \mu_1 \; for \; \overline{x} > \mu_0 + c \end{cases}$$

We then define a general expression for *n* that simultaneously contains both the Type I and Type II errors within specified upper limits α and β without relying on the CLT (*Lemmas 3* and *4*).

Lemma 3. The probability of a Type I error does not exceed α when we fix α equal to $\sigma_X^2/(nc^2)$.

Proof: Recognizing that $\underbrace{\mathrm{P}(\overline{X} \le \mu_0 + c) \ge \mathrm{P}(\mu_0 - c \le \overline{X} \le \mu_0 + c)}_{LHS\ spans\ more\ values} \ge \mathrm{P}\left(\mu_0 - c < \overline{X} < \mu_0 + c\right) =$

$\underbrace{\mathrm{P}\left(\left|\overline{X} - \mu_0\right| < c\right) \ge 1 - \frac{\sigma_X^2}{nc^2}}_{Chebyshev's\ inequality}$, we combine these two inequalities, yielding$\mathrm{P}(\overline{X} \le \mu_0 + c) \ge 1 - \frac{\sigma_X^2}{nc^2}$,

and then subtract each side from 1 to give $P(\overline{X} - \mu_0 > c|\mu = \mu_0) \leq \frac{\sigma_X^2}{nc^2}$. Because the LHS here is the probability of a Type I error, fixing α equal to $\sigma_X^2/(nc^2)$ completes the proof. ■

Lemma 4. The Type II error does not exceed β when we fix β equal to $\sigma_X^2/n(\mu_1 - \mu_0 - c)^2$.

Proof: The same steps as for *Lemma 3* above. ■

It is important to note that because the above lemmas make no assumptions about the functional forms of the underlying population distributions, the errors obtained from *any specified pair of distributions* are necessarily contained within these bounds, an observation we address below

Theorem 1. If sample size is determined according to the following rule

$$n = \frac{\sigma_X^2}{\left(\mu_1 - \mu_0\right)^2}\left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right)^2 \tag{5}$$

then the statements *P(Type I error)* $\leq \alpha$ and *P(Type II error)* $\leq \beta$ are simultaneously true regardless of the functional forms of the underlying densities assumed under the hypotheses.

Proof: Solving for *c* in *Lemma 3* yields $c = \sigma_X/\sqrt{\alpha n}$, which when placed into the expression for β in *Lemma 4* gives $\beta = \frac{\sigma_X^2}{n\left(\mu_1 - \mu_0 - \frac{\sigma_X}{\sqrt{\alpha n}}\right)^2}$, after which solving for *n* completes the proof. ■

To assess the minimum *strength* θ guaranteed by the above sample-size, we first define $t = (\mu_1 - \mu_0)/\sigma_X$ as our *treatment effect*[5] and then, using our transformations $a = 1/\sqrt{\alpha}$ and $b = 1/\sqrt{\beta}$, write (5) as $a + b = t\sqrt{n}$. Because test strength θ is inversely related to μ (*Definition* 1), minimizing θ is the same as maximizing μ, which in turn is the same as minimizing $1/\mu$ (subject

[5]This term is referred to as Cohen's *d* (effect size) in the literature (Cohen 1977). For a detailed discussion of *Lemmas* 3 and 4, see Banerjee (2015).

to sample size). We therefore use equation (4) to formulate our minimization problem using the following Lagrangian

$$\mathcal{L} = (a^2 - 1)(b^2 - 1) - \lambda[\mathrm{a} + \mathrm{b} - t\sqrt{n}]$$

where *a* and *b* are the minimizing variables and λ is the Lagrangian multiplier.

Theorem 2. The *strength* attainable by sample-size *n* for a given treatment effect *t* always exceeds

$$\theta_{LB} = \frac{t^2 n - 4}{t^2 n} \tag{6}$$

Proof: Solving the first order conditions ($\partial\mathcal{L}/\partial a = 0$; $\partial\mathcal{L}/\partial b = 0$; and $\partial\mathcal{L}/\partial\lambda = 0$) of our minimization problem for *a* and *b* gives $a = b = t\sqrt{n}/2$, which translates to $\alpha = \beta = 4/(t^2 n)$, yielding $\mu = 16/(t^2 n - 4)^2$ and thus the desired θ_{LB}. ■

Example 4: For *any* values of α and β, the test strength will be at least θ_{LB}. For cases in which CLT conditions are satisfied, a sample size expression such as $n = \frac{\sigma_X^2}{(\mu_1 - \mu_0)^2}\left(z_\alpha + z_\beta\right)^2$, where z_α and z_β are the critical values (of the standard normal variate) associated with the *specified* error sizes, will replace the constraint in (5). Because the Type I and Type II errors are lower when derived from a specified distribution, this replacement will result in a θ_{LB} greater than that of (6). The fact that this outcome is a consequence of using the *upper bounds* on errors rather than the actual errors in *Lemmas 3* and *4* means that *Theorem 3* in fact represents the weakest scenario.

Corollary 1. *Test strength* approaches *certainty* as the sample size increases indefinitely.

Proof: Given that the test strength for any *n* is clearly bounded in the interval $\theta_{LB} \leq \theta \leq 1$, recognizing from (6) that $\lim_{n\to\infty} \theta_{LB} = 1$ completes the proof. ■

4. Inferential Remarks

Somehow, there is always a *concluding note* to the end of research papers that rely heavily on statistics. We intend to change that to subtly emphasize (by example) that statistics is always about inference and hardly ever about conclusions. Given science's historical reliance on statistical inference and its wide application in myriad disciplines, understanding the techniques associated with inferential analysis is critical. Hence, our proposal for an alternate measure of statistical inference (strength) assuming a constant hypothesis probability conditional on supporting data makes a valuable contribution to all disciplines that employ scientific experimentation to enhance human understanding of our realities. We are then able to determine the large sample properties of our measure of *inference strength* θ, which makes it a valuable addition to the inferential statistics that facilitate good science. In particular, although θ provides no specific threshold (to allow for the varying standards in different disciplines), it does address the *p*-value related problems of probability data insufficiency and subjectively biased a priori beliefs identified almost a century ago (Kennedy-Shaffer 2019) but equally relevant today.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423-437.

Banerjee S. (2015). Power analysis and sample sizes: A binding frontier approach. Indian Statistical Institute, Planning Unit, New Delhi Discussion Papers 15-04, Indian Statistical Institute, New Delhi, India.

Berger, J. O. (1985). Statistical decision theory and Bayesian analysis. Springer-Verlag.

Cohen J. (1977). Statistical power analysis for the behavioral sciences. Academic Press.

Everett III, H. (1957). "Relative state" formulation of quantum mechanics. *Reviews of Modern Physics*, *29*(3), 454-462.

Fisher, R. A. (1922). Statistical methods for research workers. Oliver and Boyd.

Fisher R. A. (1973). Statistical methods and scientific inference (3rd Edition). Collin Macmillan.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*(3), 445-449.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83-115.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In: G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*, 311-339.

Ghosh J.K., Delampady M., & Samanta T. 2006. An introduction to Bayesian analysis. Springer.

Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum?. *Biology Letters*, *15*(5), 20190174.

Hogben, L. (1957). Statistical theory. London: Allen & Unwin.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103-115.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In: L. L. Harlow, S. A.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: Areader*. New Brunswick.

Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*. Psychological Press.

Nakagawa, S., &Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: Apractical guide for biologists. *Biological Reviews*, *85*(4), 935-956.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique*, *25*(1/3), 7-22.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*,*506*(7487), 150.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, *63*(3), 763-772.

Kennedy-Shaffer, L. (2019). Before $p< 0.05$ to beyond $p< 0.05$: Using history to contextualize p-values and significance testing. *American Statistician*, *73*(sup1), 82-90.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Popper K. (1934). *The logic of scientific discovery*. Routledge.

Rice, W. R. (1990). A consensus combined P-value test and the family-wide significance of component tests. *Biometrics*, 303-308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237.

Trafimow, D., & Marks, M. (2015). Banning null hypothesis significance testing procedures (editorial). *Basic and Applied Social Psychology*, *37*(1), 1-2.

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., &Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322-339.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181-207). Springer, New York, NY.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011).*Journal of Personality and Social Psychology*, *100*(3), 426-432.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05" (editorial).*American Statistician, 73*(51), 1- 19.

Yoccoz, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, *72*(2), 106-111.