What Is In a (Non-) Significant Finding?
Moving Beyond False Dichotomies

Marco R. Steenbergen

University of Zurich

Version February 8, 2019

Null hypothesis significance testing (NHST) is under attack and p-values, long seen as the quintessence of the scientific method, are receiving particularly bad press. The criticisms are not new (Berkson, 1942; Rozeboom, 1960) and they cover a variety of fields (Kline, 2013; Wasserstein and Lazar, 2016; Ziliak and McCloskey, 2008). They cover a variety of aspects, including misunderstandings of p-values (Goodman, 2008) and what one might call the .05 fetish (Benjamin et al., 2018; Yates, 1951). One other aspect of the criticisms is that NHST has contributed to black-and-white thinking about scientific findings. In this paper, I focus on this aspect and discuss some novel and not-so-novel approaches that permit for a more nuanced interpretation of p-values.

The p-Value Dichotomy

In scientific practice, p-values have come to be used as an arbiter of sorts—a statistical criterion resulting in a binary decision (for a rare exception, see Cox and Donnelly, 2011). In its mildest form, the resulting dichotomy distinguishes between more and less reliable effects. More often than not, however, p-value-based dichotomies go much further. Thus, p-values are used to decide between interesting and uninteresting findings, worthwhile and worthless experiments (in the broadest sense of that term), publishable and unpublishable findings, and, ultimately, truth and falsehood. (The latter represents a particularly egregious misunderstanding of p-values).

There are numerous problems with this practice. First, it is questionable whether p-values were ever intended to play this role. In itself, the p-value is a continuum that captures the consistency of empirical evidence with the null hypothesis. It became linked to a binary decision in the Neyman-Pearson framework on hypothesis testing, in particular (Neyman and Pearson, 1928*a*,*b*), although Fisher (1925) also contributed to it. It should not be forgotten, however, that Neyman and Pearson conceived their approach in terms of repeated experiments and not a single study, as is characteristic of much political science. In the latter context, the wisdom of reaching a dichotomous decision based on some significance level is questionable.

Second, the dichotomous use of p-values incentivizes scholars to engage in less than desirable scientific practices in order to reach significance. HARKing (hypothesizing after results are known) and p-hacking are textbook examples of practices that hamper reproducible science (Munafò et al., 2017). They do not contribute to a robust scientific canon and, indeed, can hurt in a world in which science deniers have gained a strong political voice.

Third, the practice results in the well-known publication bias (Munafò et al., 2017). Significant findings make it into the scientific discourse, whereas non-significant findings, no matter how valuable, land in filing drawers never to see the light of day. The baseline rate of failed experiments thus remains hidden, making it all the more difficult to assess the replicability of research.

I am certainly not the first one to lament the dichotomous use of p-values (see Goodman, 1999*a*; Stern, 2016). But the question is how to change scientific practice. One could abandon p-values altogether, as *Political Analysis* did for a short while and other journals still do. However, this may be throwing out the baby with the bathwater, at least if one agrees that the problem may not so much be p-values themselves as their use in scientific practice. One could also lower p-values but this might just shift the threshold at which dichotomous decisions are made (Benjamin et al., 2018).

My modest proposal is not to abandon p-values wholesale, which may anyway be difficult in light of the generations of scholars who have learnt (and often come to love) p-values. Rather, my plea is to move away from its dichotomous use. Fortunately, there exist both older and newer techniques that allow us to take a more nuanced view of p-values. Here, I focus on the Bayes factor (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995) and the analysis of credibility (Matthews, 2018), which both derive from Bayesian inference.

The Bayesian View of Inference

Shikano (this issue) provides an extensive discussion of the Bayesian perspective on hypothesis testing. Thus, it suffices to highlight a few aspects that we shall need to develop the Bayes factor and analysis of credibility. In general, the Bayesian approach can be summarized as follows (e.g., Jackman, 2011):

$$\text{Prior} + \text{Data} \quad \Rightarrow \quad \text{Posterior} \tag{1}$$

Work on the Bayes factor emphasizes the data aspect, whereas the analysis of credibility focuses on the prior.

In NHST à la Neyman-Pearson, we focus on a pair of hypotheses, H_0 and H_1. When testing means across treatment (index by 1) and control (indexed by 0) groups we may, for example, test $H_0 : \mu_1 = \mu_0$ against $\mu_1 \neq \mu_0$. We can now ask, a posteriori how much support do H_0 and H_1 receive. A standard result gives the posterior as (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995)

$$\frac{p(H_0|\text{Data})}{p(H_1|\text{Data})} = \text{BF}_{01}(\text{Data}) \cdot \frac{p(H_0)}{p(H_1)} \tag{2}$$

The left-hand side gives the relative beliefs in H_0 and H_1 after the experimental evidence has been collected. The second term on the right-hand side gives the beliefs in H_0 and H_1 prior to conducting the experiment. Finally, BF is the Bayes factor. This is the ratio of the marginal likelihoods for H_0 and H_1 and, as such, captures the empirical evidence vis-à-vis both hypotheses—what some have called the "weight of the evidence" (Good, 1950). It influences whether and how beliefs about the hypotheses should be adjusted.

Many test quantities asymptotically follow a normal distribution. This is true of means, differences in means, and log-odds ratios, for example. Imagine we have a normally distributed quantity Y with a known variance of σ^2. We are interested in the posterior distribution of the mean, μ, of Y. With a normal prior $\mu \sim \mathcal{N}(\mu_0, \phi_0)$, the posterior is given by

$$\begin{aligned} \mu|\text{Data} &\sim \mathcal{N}(\mu_P, \phi_P) \\ \mu_P &= \phi_P \left[\frac{\mu_0}{\phi_0} + \frac{\bar{y}}{\phi}\right] \\ \phi_P &= \left[\phi_0^{-1} + \phi^{-1}\right]^{-1} \end{aligned} \tag{3}$$

Here, $\phi = \sigma^2/n$ is the sampling variance of the sample mean, $\bar{y}$. We can simplify the posterior by dividing by ϕ_P so that $\mu_P/\phi_P = \mu_0/\phi_0 + \bar{y}/\phi$ and $1/\Phi_P = 1/\phi_0 + 1/\Phi$. These expressions play an important role in the analysis of credibility (Matthews, 2018).

The Bayes Factor

One of the major problems with NHST is that it cannot say anything about the relative merits of H_0 and H_1. After all, the p-value is computed under the assumption that H_0 is true (cf. Stern, 2016). The Bayes factor rectifies this problem. It is given by (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995)

$$\text{BF}_{01}(\text{Data}) = \frac{p(\text{Data}|H_0)}{p(\text{Data}|H_1)} = \frac{\int p(\text{Data}|\theta, H_0)p(\theta|H_0)d\theta}{\int p(\text{Data}|\theta, H_1)p(\theta|H_1)d\theta} \tag{4}$$

By focusing on the parameters θ, which drive the marginal likelihoods on the right-hand side of Equation 4, we shift our focus from hypothesis testing to estimation. Note that we can also state the Bayes factor as the weight of evidence in favor of H_1: $\text{BF}_{10}(\text{Data}) = 1/\text{BF}_{01}(\text{Data})$.

If our concern is the dichotomous interpretation of p-values, then the Bayes factor already offers a clear advantage because it typically is classified in a more nuanced manner. Table 1 shows four different classifications of the Bayes factor (Goodman, 1999*b*; Held and Ott, 2016; Jeffreys, 1998; Kass and Raftery, 1995). We see that there is more nuance in the manner in which evidence is interpreted. If dichotomous interpretations of the p-value are black and white, then the interpretation of the Bayes factor is in terms of shades of grey. We are forced to think about the empirical merits of the null and alternative hypotheses in a nuanced manner. Obviously, this is not the decisive selling point of the Bayes factor but it certainly helps to put empirical evidence against the null hypothesis into some perspective.

One could raise two objections at this point. First, there hardly seems to be any consensus on the classification of the Bayes factor. Sure, it may be nuanced but no one seems to agree on what those nuances are. I do not view this as a particularly troublesome argument. The different classifications vary on how much evidence they require against H_0 in order to upgrade the verbal designation of that evidence. An author or journal could settle on a particular scheme and take it from there. I will be using the scheme of Held and Ott (2016).

A second objection is that the Bayes factor is notoriously difficult to compute in most cases (see Kass, 1993). They also require choices about priors that may require expertise beyond the typical skill level among political scientists (again, see Kass, 1993). How can such a difficult technique ever be a practical replacement for the p-value?

In fact, it does not have to be. In the context of precise null hypotheses (e.g., $\mu_1 = \mu_0$), Berger and Delampady (1987) showed that it is possible to derive lower bounds on the Bayes factor from p-values. Vovk (1993), Sellke, Bayarri and Berger (2001),and Held and Ott (2018) formally derived the lower-bounds under different assumptions about the priors. These lower-bounds allow us to establish a link between p-values and the more nuanced interpretation that flows from Bayes factors.

The starting point for all of the approaches is to assume that the p-values a priori follow the uniform distribution $\mathcal{U}(0,1)$ under H_0. H_1 a priori favors low p-values, which can be captured by selecting an appropriate beta distribution. Vovk (1993) and Sellke, Bayarri and Berger (2001) assume $p \sim \mathcal{BE}(\alpha, 1)$, where $\alpha \leq 1$. The resulting priors are generally uninformative and produce

$$\min \text{BF}_{01}(p) = \begin{cases} -ep\ln p & \text{if } p < 1/e \\ 1 & \text{Otherwise} \end{cases} \tag{5}$$

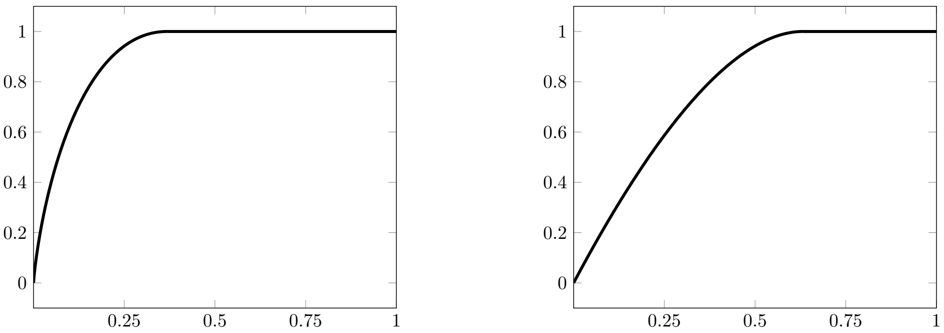
A graphic depiction is shown in the left panel of Figure 1. Arguing that the selected beta

Table 1: Classifying the Bayes Factor

Goodman		Held & Ott		Jeffreys		Kass & Raftery	
Range	Meaning	Range	Meaning	Range	Meaning	Range	Meaning
1 to 1/5	Weak	1 to 1/3	Weak	> 1	Negative	1 to 1/3	Not worth more than a bar mention
1/5 to 1/10	Moderate	1/3 to 1/10	Moderate	1 to $\approx$ 32/100	Barely worth mentioning	1/3 to 1/20	Positive
1/10 to 1/20	Moderate to strong	1/10 to 1/30	Substantial	$\approx$ 32/100 to 1/10	Substantial	1/20 to 1/150	Strong
1/20 to 1/100	Strong to very strong	1/30 to 1/100	Strong	1/10 to $\approx$ 3/100	Strong	$> 1/150$	Very strong
		1/100 to 1/300	Very strong	$\approx$ 3/100 to 1/100	Very strong		
		$< 1/300$	Decisive	$< 1/100$	Decisive		

Notes: Table shows the support for H_0 relative to H_1 (Goodman, 1999*b*; Held and Ott, 2016; Jeffreys, 1998; Kass and Raftery, 1995).

Figure 1: Minimum Bayes Factors from p-Values



(a) Vovk Formula

(b) Held & Ott Formula

prior is not always so uninformative, **?** favor $\mathcal{BE}(1, \beta)$ with $\beta \geq 1$. This results in

$$\min \mathrm{BF}_{01}(p) \quad = \quad \begin{cases} -e(1-p)\ln(1-p) & \text{if } p < 1 - 1/e \\ 1 & \text{Otherwise} \end{cases} \tag{6}$$

Since $\ln(1-p) \approx -p$, we may also write $\min \mathrm{BF}_{01}(p) = ep$. This is shown in the right panel of Figure 1. In the range between 0 and 0.5, the Held-Ott formula consistently produces more conservative estimates of the minimal Bayes factor than the Vovk formula. Indeed, to date **?** offer the most conservative minimal Bayes factor for p-values below 0.50.

Having defined the minimum Bayes factors in this manner, how could one offer a nuanced interpretation of a p-value? Table 2 classifies the Bayes factor according to **?**. It also shows the approximate ranges of p-values that fall into a particular class according to the Held-Ott formula. If we look at this table, we see that the conventional 0.05 significance level at best offers only *moderate* evidence against the null hypothesis. The minimum Bayes factor is 0.132, meaning that the evidence favors H_1 only by a factor of $1/0.132 \approx 7.5$. By contrast, the 0.005 significance level favored by Benjamin et al. (2018) yields strong support against H_0, with a minimum Bayes factor of $\min \mathrm{BF}_{01} = 0.014$ or $\min \mathrm{BF}_{10} = 73.761$.

Other cutoffs that we see in political science are 0.10 (sometimes indicated by a plus

Table 2: The Weight of Evidence Contained in Different p-Values

Designation	p-values
Decisive	0.000-0.001
Very Strong	0.001-0.004
Strong	0.004-0.012
Substantial	0.012-0.037
Moderate	0.037-0.132
Weak	0.132-1.000

Table 3: Alternative Presentation of Empirical Results

Parameter	Est	SE	p	Evidence Against Null	Maximum Weight Favoring Alternative
Left-Right	-0.629	0.066	0.000	Decisive	∞
Left-Right Squared	0.005	0.006	0.405	Weak	1.191
Economy	0.315	0.078	0.000	Decisive	6837.722
Protestant	0.265	0.216	0.220	Weak	1.899
Catholic	-0.073	0.214	0.733	Weak	1.000
Secular	0.181	0.226	0.423	Weak	1.159
Education	0.142	0.016	0.000	Decisive	∞
Income	0.008	0.020	0.689	Weak	1.000
Male	-0.786	0.100	0.000	Decisive	∞
Age	0.025	0.017	0.141	Weak	2.810
Age Squared	0.000	0.000	1.000	Weak	1.000
Urban	0.262	0.152	0.085	Moderate	4.538
German	0.377	0.325	0.246	Weak	1.728
French	0.701	0.344	0.042	Moderate	9.040
Constant	6.356	0.584	0.000	Decisive	∞

Notes: Based on Steenbergen (2010, p. 419, results for the Greens). Evidence Against Null shows the classification of p-values that follows from Held and Ott (2016) and shows, in qualitative terms, how much evidence there is against the null hypothesis of no effect. Maximum weight favoring the alternative gives $\max \mathrm{BF}_{10}(p)$. It shows the maximal factor by which prior beliefs about the hypotheses should be updated in favor of H_1. The symbol ∞ means that the Bayes factor approaches infinity.

symbol in tables) and 0.01 (sometimes indicated with two stars). Like the 0.05 threshold, 0.10 qualifies only as moderate support against H_0. The minimum Bayes factor is only 0.258, however, meaning that the evidence favors H_1 only by a factor of roughly 3.9, about half of what 0.05 can offer. A significance level of 0.01 again qualifies as strong evidence against H_0 with $\mathrm{BF}_{10} \approx 37$.

To show the reader how one might use the Bayes factor in a publication, I replicate Table 5 from Steenbergen (2010). This table lists results from a hierarchical linear model of vote propensities for left parties in Switzerland. For the sake of simplicity, I only report the fixed effects estimates and to safe space, I only report the results for one party: the Greens. Table 3 shows the replicated results with the kind of annotation one can derive from the p-based Bayes factor.

The original table contained five predictors for which $p < 0.05$: left-right self-placements, economic retrospections (economy), education, male, and domicile in the French-speaking part of Switzerland. As we can see in Table 3, however, the cluster of “significant” effects is quite heterogeneous. For four of the predictors, we have decisive evidence against the null hypothesis of a null effect. In one instance, the evidence against the null is less compelling, receiving the moniker of moderate. These nuances are also reflected in the Bayes factors, which range from about 9 (for the variable French) to values approaching infinity. Such differences matter. After conducting the research, I am much less convinced of the divide between the French-speaking part and the rest of Switzerland ($\max \mathrm{BF}_{10} \approx 18$) than a gender gap the divide between urban and rural areas, even when both divides are significant at the .05-level.

There is nuance, too, among the non-significant findings. Some can still be classified as moderate evidence against H_0, whereas in other cases the evidence is weak. For example, the urban-rural divide in Green party support fails to be statistically significant at the .05-level. Still, the Bayes factor here is clearly more suggestive of an effect than, for example, the divide between German-speaking and other cantons. It is useful to communicate such nuances because there is information even in non-significant findings.

The Analysis of Credibility

The idea that there is information in non-significant findings plays an important role in the analysis of credibility (AnCred; Matthews, 2018). AnCred effectively asks what kind of prior would be needed to challenge a significant finding a posteriori or to make a non-significant finding a posteriori significant. The answer flowing from such an exercise can then be evaluated in terms of its plausibility, for example, in the light of published results. The beauty of the approach is that it only requires knowledge of the confidence intervals, which many believe should be published anyway.

Consider again Equation (1). We create a posterior distribution such that the a posteriori $100 \cdot (1 - \alpha)$ credible interval barely includes a null effect. On the left-hand side of the equation, the information in the data are given by the confidence interval. Following Good (1950), we now work backwards to derive the prior. Specifically, we derive the prior credible interval (CPI) that would be needed to generate the posterior.

Obviously any significant finding can be rendered a posteriori void by setting a prior with (almost) all of its probability mass located at 0. Any non-significant finding can be rendered a posterior significant by putting the mass at a value different from 0. In both cases, however, we effectively rule out that the data could ever cause us to revise our beliefs concerning the null hypothesis. This would not constitute a fair-minded—one could also say, open-minded—challenge to the empirical evidence. As Matthews (2018) argues, the key is to launch fair-minded challenges and advocacies, i.e., ones that leave room for the data.

So what would a fair-minded challenge look like? The starting point is that we have obtained a statistically significant result. Specifically, given a significance level of α, we have obtained a $100 \cdot (1 - \alpha)$ percent confidence interval, (L, U) that excludes 0. Here, $L = \bar{y} - z\alpha/2\sqrt{\phi}$, $U = \bar{y} + z_{\alpha/2}\sqrt{\phi}$, and $z_{\alpha/2}$ is the critical value of the standard normal

Table 4: Applying Skepticism Limits to a Significant Finding

	95% Conf. Int.		95% CPI	
	L	U	-SL	SL
French vs. Italian Speakers	0.027	1.375	-2.369	2.369

distribution corresponding to α. We posit ourselves as skeptics, meaning that we do not believe in an effect a priori. Hence, under normality, our prior is $\mu \sim \mathcal{N}(0, \phi_0)$. The corresponding CPI is $(-\text{SL}, \text{SL})$, where $\text{SL} = z_{\alpha/2}\sqrt{\phi_0}$ is the *skepticism limit*.

The key is now to derive $\phi_0 > 0$. We define the posterior so that the $100 \cdot (1-\alpha)$ credible interval barely includes 0. This implies $\mu_P - z_{\alpha/2}\sqrt{\phi_P} \geq 0$ or $\mu_P \geq z_{\alpha/2}\sqrt{\phi_P}$. We can now use equation (3) to solve for ϕ_0. Substituting the implied values for μ_0 and μ_P, as well as the empirical estimates of $\bar{y}$ and ϕ, it is easily shown that

$$\phi_0 = \frac{(U-L)^4}{16 \cdot z^2_{\alpha/2} \cdot U \cdot L} \tag{7}$$

Consequently,

$$\text{SL} = \frac{(U-L)^2}{4 \cdot \sqrt{U \cdot L}} \tag{8}$$

Let us apply the idea of skepticism to one of the results in Table 3. Let us distinguish between two groups: French-speaking (F) versus Italian-speaking (I) voters. Our null hypothesis is that the two groups do not differ in their proclivity to vote for the Greens: $H_0 : \mu_F - \mu_I = 0$. The estimate of the mean difference is 0.701, has a standard error of 0.344, and is statistically significant at the .05-level. The 95 percent confidence interval runs from 0.027 to 1.375 (see Table 4). Applying Equation 8, we obtain SL = 2.369 and a CPI that runs from -2.369 to 2.369 (see Table 4).

How would we interpret this? If we want to turn the significant finding into an a posteriori insignificant one, we would need tp include 0 in the 95 percent posterior credible interval. In light of the evidence, we can accomplish this only by letting the a priori difference between French and Italian speakers to range between -2.369 and 2.369. This credible interval is rather wide, meaning that we have to allow all sorts of mean differences a priori. More specifically, a priori we would have to allow for the possibility that Italian speakers would give up to a 2.4 point higher vote propensity score to the Greens than their French speaking peers, holding all else constant. This does not seem very plausible because, for one, the Greens have never performed better in the Ticino than in the Romandie. Thus, we would conclude that it is difficult to challenge the statistically significant effect of the variable French in Table 3.

Let us now turn to the idea of advocating for a statistically non-significant finding.

Table 5: Applying Advocay Limits to a Non-Significant Finding

	95% Conf. Int.		95% CPI	
	L	U	0	AL
Urban versus Rural	-0.036	0.560	0.000	4.625

Imagine, we anticipate a positive effect. The *advocacy limit*, AL, is a value such that the CPI covers the open interval from 0 to AL. We can recover AL by again working backwards from the posterior. As with the skepticism limit, we have $\mu_P \geq z_{\alpha/2}\sqrt{\phi_P}$. We further know that $\mu_0 - z_{\alpha/2}\sqrt{\phi_0} = 0$ or $\mu_0 = z_{\alpha/2}\sqrt{\phi_0}$. Further, in mathematical terms AL $= \mu_0 + z_{\alpha/2}\sqrt{\phi_0}$. Substituting the result for μ_0, this means AL $= 2z_{\alpha/2}\sqrt{\phi_0}$.

The key is now once more to derive the expression for ϕ_0. Again using equation (3), it can be shown that, in the case of advocacy,

$$\phi_0 = \frac{(U+L)^2(U-L)^4}{16z^2_{\alpha/2}U^2L^2} \tag{9}$$

Consequently,

$$\text{AL} = -\frac{(U+L)(U-L)^2}{2LU} \tag{10}$$

(The negative multiplier is there because L and U are of opposing signs, rendering the denominator in equation (10) negative. To offset this, we retain only the negative root of ϕ_0, which ensures that AL > 0.) Once the advocacy limit is obtained, we once more assess its plausibility by considering relevant past evidence and theory.

As an illustration, consider the result for the predictor urban in Table 3. As Table 5 shows, the 95 percent confidence interval runs from -0.036 to 0.560. We expect that urban voters are more inclined to vote for the Greens than rural voters. Accordingly, our prior is that the urban effect is positive. But how positive should it be to turn over the lack of statistical significance? If we want to ensure that the 95 percent posterior credible interval excludes a null effect, then AL = 4.625. This means that the 95 percent CPI runs from 0 to 4.625. Any effect of urban in this prior region would suffice to turn over the non-significance a posteriori.

What does that mean? Those who believe in an urban-rural divide in Green party support can still credibly challenge the non-significance of the urban predictor by arguing that the anticipated effect is somewhere in the range between 0 and 4.6 points. Of course, they could also challenge the non-significance by claiming a much larger effect. However, such a claim would lack credibility since the CPI suggests that effects greater than 4.6 are a priori unlikely. In the present case, a credible challenge of the null effect seems possible. The value constellation of urban voters is likely to favor environmental causes and parties more than that of rural voters. At the same time, few would argue that the divide should amount to more than 5 points on a 0-10 scale such as the vote propensity scale.

The advocacy CPI also sheds light on the amount of information contained in an empirical finding. Imagine that $0 < \text{CPI} < \infty$. This would mean that nearly anything goes a priori against the lack of significance. In this case, we had better not put much stock in the failure to reject H_0. The tighter the CPI becomes, the smaller the range of possible values that would challenge non-significance and the more difficult it becomes to launch a credible claim at the behest of H_1.

Thus, we see that the analysis of credibility is another way of moving beyond the dichotomous interpretation of p-values. What looks to be statistically significant may actually be an easy target for a challenge. Conversely, what appears to be non-significant might plausibly be rescued, receiving the benefit of, for example, a second test with new data. The nice thing about analysis of credibility is that one can adjust the formulas easily to a variety of generalized linear models (Matthews, 2018). In this sense, it is as flexible as the use of Bayes factors.

Conclusions

Many reasons exist why one might challenge the current emphasis on p-values in scientific practice. Not least is the fact that such reliance has often been accompanied by the compulsion to divide research into worthwhile and worthless efforts, findings that should be sent out for review and those that should not, and papers that deserve to be published and those that do not. As a result, highly surprising and significant results receive a great deal of play, even when reproducibility frequently turns out to be problematic. By contrast, non-significant findings, even when they are extremely robust, may never see the light of day.

In this paper, I have shown that one can think about p-values in a much more nuanced way. So-called significant findings may contain less evidence against the null hypothesis than meets the eye, while non-significant findings might actually be more meaningful than one thinks. Bayes factors and analysis of credibility can be used to extract more information from p-values and to allow for a more nuanced view of scientific evidence. Authors should consider adding them to their tables and journal editors might want to ask for such information, if only to place research findings in a more realistic perspective and temper publication bias.

References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin

Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson. 2018. "Redefine statistical significance." *Nature Human Behaviour* 2(1):6–10.
URL: *http://www.nature.com/articles/s41562-017-0189-z*

Berger, James O. and Mohan Delampady. 1987. "Testing Precise Hypotheses." *Statistical Science* 2(3):317–335.
URL: *http://projecteuclid.org/euclid.ss/1177013238*

Berkson, Joseph. 1942. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37(219):325–335.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1942.10501760*

Cox, David R. and Christi A. Donnelly. 2011. *Principles of Applied Statistics.* Cambridge: Cambridge University Press.
URL: *Statistics*

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Good, Irving J. 1950. *Probability and the Weighting of Evidence.* London: Griffin.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology* 45(3):135–140.
URL: *https://www.sciencedirect.com/science/article/pii/S0037196308000620*

Goodman, Steven N. 1999*a*. "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy." *Annals of Internal Medicine* 130(12):995.
URL: *http://annals.org/article.aspx?doi=10.7326/0003-4819-130-12-199906150-00008*

Goodman, Steven N. 1999*b*. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." *Annals of Internal Medicine* 130(12):1005.
URL: *http://annals.org/article.aspx?doi=10.7326/0003-4819-130-12-199906150-00019*

Held, Leonhard and Manuela Ott. 2016. "How the Maximal Evidence of <i>P</i> -Values Against Point Null Hypotheses Depends on Sample Size." *The American Statistician* 70(4):335–341.
URL: *https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1209128*

Held, Leonhard and Manuela Ott. 2018. "On p-Values and Bayes Factors." *Annual Review of Statistics and Its Application* 5(1):393–419.
URL: *http://www.annualreviews.org/doi/10.1146/annurev-statistics-031017-100307*

Jackman, Simon. 2011. *Bayesian Analysis for the Social Sciences.* Chichester: Wiley.

Jeffreys, Harold. 1998. *Theory of Probability.* 3 ed. Oxford: Oxford University Press.

Kass, Robert E. 1993. "Bayes Factors in Practice." *Journal of the Royal Statistical Society, Series D (The Statistician)* 42(5):551–560.
URL: *https://www.jstor.org/stable/10.2307/2348679?origin=crossref*

Kass, Robert E. and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90(430):773–795.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572*

Kline, Rex B. 2013. *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences.* 2 ed. Washington, D.C.: American Psychological Association.

Matthews, Robert A. J. 2018. "Beyond 'significance': principles and practice of the Analysis of Credibility." *Royal Society Open Science* 5(1):171047.
URL: *http://rsos.royalsocietypublishing.org/lookup/doi/10.1098/rsos.171047*

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware and John P. A. Ioannidis. 2017. "A manifesto for reproducible science." *Nature Human Behaviour* 1(1):0021.
URL: *http://www.nature.com/articles/s41562-016-0021*

Neyman, J. and E. S. Pearson. 1928*a*. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I." *Biometrika* 20A(1/2):175–240.
URL: *https://www.jstor.org/stable/2331945?origin=crossref*

Neyman, J. and E. S. Pearson. 1928*b*. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II." *Biometrika* 20A(3/4):263–294.
URL: *https://www.jstor.org/stable/2332112?origin=crossref*

Rozeboom, William W. 1960. "The fallacy of the null-hypothesis significance test." *Psychological Bulletin* 57(5):416–428.
URL: *http://content.apa.org/journals/bul/57/5/416*

Sellke, Thomas, M. J Bayarri and James O Berger. 2001. "Calibration of <i>ρ</i> Values for Testing Precise Null Hypotheses." *The American Statistician* 55(1):62–71.
URL: *http://www.tandfonline.com/doi/abs/10.1198/000313001300339950*

Steenbergen, Marco R. 2010. "Decomposing the Vote: Individual, Communal, and Cantonal Sources of Voting Behavior in Switzerland." *Swiss Political Science Review* 16(3):403–424.
URL: *http://doi.wiley.com/10.1002/j.1662-6370.2010.tb00435.x*

Stern, Hal S. 2016. "A Test by Any Other Name: P Values, Bayes Factors, and Statistical Inference." *Multivariate Behavioral Research* 51(1):23–29.
URL: *https://www.tandfonline.com/doi/full/10.1080/00273171.2015.1099032*

Vovk, V G. 1993. "A Logic of Probability, with Application to the Foundations of Statistics." *Source: Journal of the Royal Statistical Society. Series B (Methodological)* 55(2):317–351.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2):129–133.
URL: *https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108*

Yates, Frank. 1951. "The Influence of <i>Statistical Methods for Research Workers</i> on the Development of the Science of Statistics." *Journal of the American Statistical Association* 46(253):19–34.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500764*

Ziliak, Stephen T. and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How Standard Errors Costs Us Jobs, Justice, and Lives.* Ann Arbor, MI: University of Michigan Press.