How and why alpha should depend on sample size: A Bayesian-frequentist compromise for significance testing

Jesper N. Wulff[*] Luke Taylor[*]

January 5, 2023

Abstract

The use of fixed alpha levels in statistical testing is prevalent in management research, but can lead to Lindley's paradox in highly powered studies. In this article, we propose a sample size-adjusted alpha level approach that combines the benefits of both frequentist and Bayesian statistics, enabling strict hypothesis testing with known error rates while also quantifying the evidence for a hypothesis. We present an R-package that can be used to set the sample size-adjusted alpha level for generalized linear models, including linear regression, logistic regression, and Poisson regression. This approach can help researchers stop relying on mindless defaults and avoid situations where they reject the null hypothesis when the evidence in the test actually favors the null hypothesis, improving the accuracy and robustness of statistical analysis in management research.

Keywords: Alpha level, significance testing, hypothesis testing, Bayesian-frequentist compromise

[*] Department of Economics and Business Economics, Aarhus University

"The real objection to P values is not that they usually are utter nonsense, but rather that they can be highly misleading, especially if the value of N is not also taken into account and is large." Good (1992, p. 600)

"The rule of thumb quite popular now, that is, setting the significance level arbitrarily to .05, is shown to be deficient in the sense that from every reasonable viewpoint the significance level should be a decreasing function of sample size." Leamer (1978, p. 92)

Introduction

Dichotomous questions play an important role in management research. Do acquisitions cause employees to leave? Do successions from a non-family CEO back to a family CEO reduce labor costs? Do university entrepreneurship programs promote entrepreneurship? To answer such questions empirically, researchers often rely on null hypothesis significance testing (NHST). NHST is a statistical procedure that can govern researchers' behavior towards whether a hypothesis is corroborated or not while ensuring that they are not wrong too often (Neyman et al., 1933). If researchers are interested in establishing claims about hypotheses, separating signal from noise and drawing valid conclusions on the basis of data while limiting long-run error rates, NHST is an important tool in a scientist's toolkit (Benjamini et al., 2021; Lakens, 2021).

An important - but often neglected - part of correctly applying and interpreting significance tests is to justify the significance threshold, i.e., alpha (Lakens et al., 2018). This choice should be justified before data collection and should not be based on the idea of *"one alpha to rule them all"* (Lakens et al., 2018, p. 169). Yet, justification for the alpha level is exceedingly rare in management research where researchers rely exclusively on arbitrary thresholds such as 0.10, 0.05, 0.01, and 0.001 (Aguines & Harden, 2009). These thresholds act as gatekeepers for whether a result is deemed valuable or not (Bettis et al., 2016). From 2002-2006, 99% of papers in top management journals relied on these conventional values for α, making management the business discipline that has most strongly embraced this tradition (Aguinis et al., 2010). It is fair to say that "*[p]articular p-values (*0.05, 0.01*, or* 0.001*) have been endowed with almost mythical properties*" (Bettis et al., 2016, p. 259).

Unfortunately, relying on a universal alpha level, such as 0.05, is problematic: in highly-powered studies, p-values lower than conventional alpha levels can be more likely when the null hypothesis is true than when the alternative is. This phenomenon - known as Lindley's paradox (Wagenmakers & Ly, 2021) - occurs because the distribution of p-values is a function of sample size (Cumming, 2008). Management researchers who subscribe to Bayesian statistical philosophy have already identified this as a major limitation of NHST: with a fixed alpha, "*[s]tatistical significance is an easy goal because any researcher can achieve it by adding more data and increasing the sample size*" (Starbuck, 2016, p. 61) or, put differently, "*[a] researcher who gathers a large enough sample can reject any point-null hypothesis*" (Schwab et al., 2011, p. 8).

In this paper, we propose a principled and practical way of lowering the alpha level as the sample size increases. This approach ensures the null is only ever rejected when it is less likely than the alternative. Researchers can thus enjoy the long-run Type I error rate guarantees of NHST while interpreting a significant test as evidence for the alternative hypothesis in a Bayesian fashion. Our solution to Lindley's Paradox can be seen as a frequentist/Bayesian compromise. Indeed, it brings Bayesian and frequentist statistics closer together by solving the large-n conflict that arises when a fixed alpha is used.

Our approach builds on the work of Maier and Lakens (2022), who recently proposed a Bayesian/frequentist compromise for justifying the alpha level in psychology research. We extend their method to linear and generalized linear regression models, which are far more prevalent in management research. Inspired by recent advances in the methodological development of Bayes factors (Mulder et al., 2021), our approach uses the approximate adjusted fractional Bayes factor (Gu et al., 2018). To allow immediate use, we develop an R-package that allows users to set alpha for all standard generalized linear models, including linear regression, logistic regression, and Poisson regression among others.

Every part of the research process should be clearly justified (Aguinis et al., 2021; Aguinis et al., 2018), yet discussions regarding the alpha level have been almost entirely absent from the management literature (for an exception, see Aguinis et al., 2010). With the era of big data creeping ever more into management research (Barnes et al., 2018; Wright, 2016), this omission becomes more serious; it is now standard to work with thousands, if not tens of thousands, of observations.[1] By providing researchers with a practical and principled way of reducing the alpha level as the sample size increases, we hope to ignite a paradigm shift in the management field where researchers justify their alpha level.

Significance testing and alpha levels

Null hypothesis significance tests are widely considered the dominant approach for statistical inference in quantitative management research (Lockett et al., 2014; van Witteloostuijn, 2020). In regression analysis, a typical null hypothesis is $\mathcal{H}_0 : \theta = 0$, where θ is a regression coefficient associated with some variable

[1] In management research, sample sizes have increased over time (Ketchen et al., 2008) while alpha has remained fixed, bringing the discipline to a situation where the average sample size has nearly 100% power with respect to small effects (Combs, 2010). More recent research confirms that large sample sizes in management research are common. For instance, Certo et al., 2016 report a median sample size of 500 in 64 *Strategic Management Journal* papers from 2005 to 2014, while Villadsen and Wulff, 2021a report a median sample size of 800 in 300 papers published in *Academy of Management Journal*, *Journal of International Business Studies*, *Journal of Management*, *Management Science*, *Organization Science*, *Research Policy*, and *Strategic Management Journal* from 2007 to 2016.

of interest. To test this hypothesis of no effect, the researcher computes the p-value, defined as the probability of observing a test statistic at least as extreme as the one observed in the sample if the null hypothesis were true (Greenland et al., 2016):

$$p = \Pr\left(|T| \geq |t|; \mathcal{H}_0 : \theta = 0\right), \tag{1}$$

where T is the test statistic that quantifies the incompatibility with $\mathcal{H}_0$ and t is the observed test statistic. If p is lower than the significance threshold, α, $\mathcal{H}_0$ is rejected.

This procedure, known as the Neyman-Pearson approach to statistical inference, allows researchers to make dichotomous claims while controlling the error rate (Lakens, 2021). If a researcher rejects $\mathcal{H}_0$ using a given α level, they claim an effect exists while acknowledging that they will be misled at most α% of the time on average. But what does it mean to be misled?

In any setting where incomplete information is used to make a claim, there is a chance of error. Here, a researcher wishes to make a claim about a population, yet they only have access to a sample, i.e., an incomplete picture of the population. In NHST, a researcher can be wrong in two ways: rejecting a true $\mathcal{H}_0$ (Type I error) and failing to reject a false $\mathcal{H}_0$ (Type II error). The alpha level is the Type I error rate. If a researcher sets α to 0.05, they are willing to accept a Type I error with 5% probability. β is the Type II error rate and is a function of the test statistic, the sample size, the true distribution of the test statistic (the null is false in this case, so it is different from the null distribution), and the α level. The Neyman-Pearson approach allows researchers to make claims while giving them full control of the Type I error rate; the Type II error rate is minimized subject to the Type I rate being fixed at α (Neyman et al., 1933).

The problem with universal thresholds

Conventional α levels can be traced back to Ronald A. Fisher, one of the fathers of frequentist hypothesis testing, who often used an α of 0.05 or 0.01. However, neither Fisher, Neyman, nor Pearson recommended a universal threshold (Maier & Lakens, 2022). For instance, Fisher, 1971 explains that *"[i]t is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result"*. Similarly, Neyman et al., 1933 made it clear that *"From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator"*. Either through lack of guidance or through believing it is better to be wrong with the crowd than to be right alone, researchers have ignored the recommendations of Fisher and Neyman, clinging to 0.05 like a ship to an anchor in rough waters.

Some have proposed to simply lower the conventional α level, for example, to 0.005, at least for new discoveries with low prior odds (Benjamin et al., 2018). Alas, this suggestion misses the root of the problem, which is not the size of alpha, but that it remains fixed across different sample sizes (Lakens et al., 2018). In this section, we explain the problems with a fixed alpha.

A Neyman-Pearson perspective

From a Neyman-Pearson perspective, it is logical that α should be a decreasing function of the sample size. As previously explained, Type I and Type II errors can occur when performing NHST. For a single study, the combined probability of a Type I or Type II error, ω, is the mean of α and β (Mudge et al., 2012).

The upper left of Figure 1 illustrates the relationship between α and the average error (ω) for a two-sample, two-sided t-test with a true effect size of 0.5 (equivalent to a regression on a single binary variable). In general, the average error rate falls as the α level decreases. However, below a certain point, the relationship reverses and the average error rate increases as α decreases. Furthermore, this change occurs at smaller α levels when the sample size is larger. Thus, for different sample sizes, we can identify the combination of α and β that minimizes the combined probability of a Type I and Type II error (Mudge et al., 2012). The upper right of Figure 1 shows how the α that minimizes the average error is a decreasing function of the sample size. For a sample size of 100, the optimal α is 0.0506, close to the conventional threshold of 0.05. At $n = 200$ the optimal α is 0.0085 and thus lower than conventional thresholds. Clearly, fixing α for different sample sizes is not optimal for overall error rates.

As an alternative to minimizing the average error, we can reach a similar conclusion by balancing Type I and Type II errors. The relationship between α and β implies that decreasing α decreases the power $(1-\beta)$ to detect deviations from the null (Mudge et al., 2012). Since a larger sample size means greater power, using a fixed α across sample sizes means that the Type I probability will often be orders of magnitude larger than the Type II probability. In the limit, the power for all consistent tests tends to 1 as $n \to \infty$; thus, the Type I probability becomes infinitely times larger than the Type II error, making the test severely biased towards Type I errors (Kim et al., 2018).

From Figure 2, we see if $n = 100$, the power is 94.4% resulting in a Type II error of 5.96% (100 – 94.4). So, by setting $\alpha = 0.05$, the two error types are relatively balanced. However, for $n = 150$, 200, and 250, the power is 99.08%, 99.88%, and 99.99%, respectively. So even for a modest sample size of 250, the Type II error rate is 500 times smaller than the Type I error rate of 5%. Unless a researcher has a compelling reason, it makes little sense to operate with this kind of error imbalance as a default. Instead, lowering α to 0.35% for $n = 250$ would still provide power of 99.59% (i.e. $\beta = 0.41\%$) while making the error rates almost balanced.[2]

[2] A similar example is provided by (Moshagen & Erdfelder, 2016), who suggests balancing

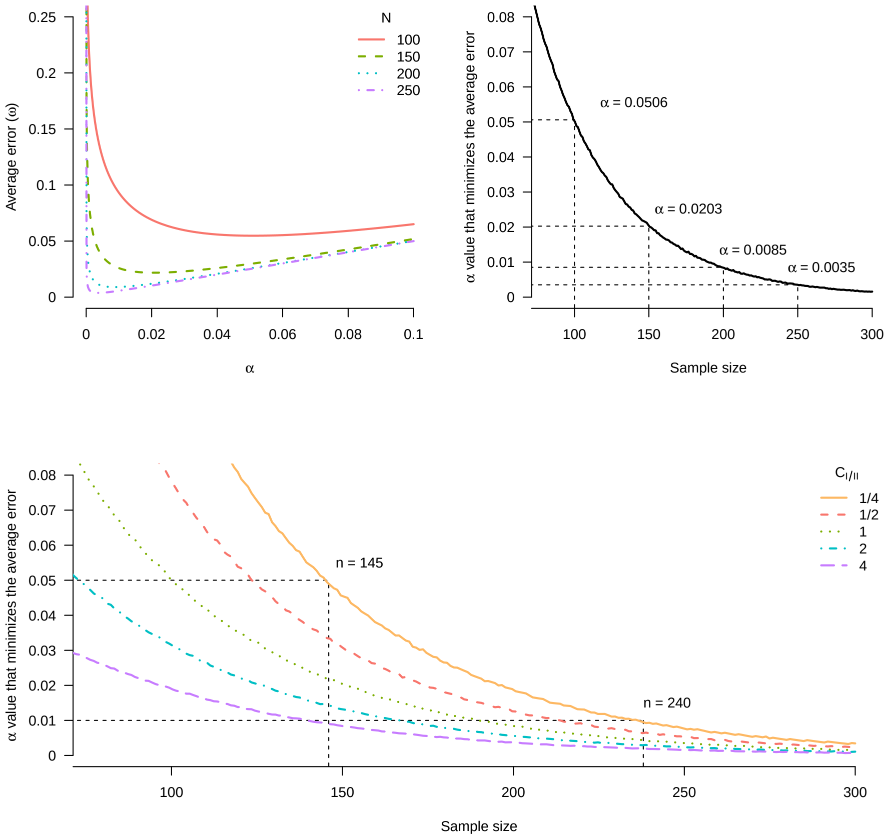

Figure 1: Minimizing the average error for a two-sample, two-sided independent t-test. Upper Left: Average error ω as a function of α for various sample sizes. Upper Right: Selecting the α that minimizes the average error as a function of sample size. Bottom: Selecting the α that minimizes the average error as a function of sample size for various relative costs of Type I and Type II errors $(C_{I/II})$

errors when testing model fit for structural equation models (SEM). If researchers estimate 1 billion models using a typical SEM study based on $n = 389$ (Jackson et al., 2009), 50 million correct models will be incorrectly rejected ($\alpha = 0.05$) if $\mathcal{H}_0$ is true, while only approximately two models will be wrongly accepted if $\mathcal{H}_0$ does not hold. This example illustrates how typically sized samples can result in vastly unbalanced error rates that are heavily biased toward Type I errors.

Unequal costs and priors

In some cases, it may be sensible to use unequal error rates depending on the relative costs of Type I and Type II errors and the base rate of true effects (J. Miller & Ulrich, 2019). For example, Aguinis et al. (2010) use the relationship between inter-divisional knowledge and invention impact (D. J. Miller et al., 2007) to illustrate how a Type I error can be more costly than a Type II error. Falsely concluding that such a relationship exists could lead firms to invest resources into knowledge transfer across divisions without any gains. On the other hand, a Type II error results in opportunity costs for firms by missing out on profitable investments. Aguinis et al. (2010) argue that in this case, a Type I error is more costly than a Type II.

Changing the relative error costs or the base rate of a true effect shifts the optimal α curve, but their effect on the optimal α diminishes as the sample size increases, as shown at the bottom of Figure 1. For example, the orange curve shows a scenario where the cost of making a Type II error is four times that of making a Type I error ($C_{I/II} = 1/4$). Beyond $n = 300$, the optimal α is almost entirely insensitive to the relative cost of a Type I error to a Type II. Thus, for surprisingly small sample sizes, unless researchers need to work with extremely unequal error costs and/or a very high probability that $\mathcal{H}_1$ is true, the relative costs and prior probabilities have almost no influence. This underscores that if we want to minimize or balance the weighted costs, the sample size is the key determining factor.

A Bayesian perspective

Reducing α as the sample size increases is also logical from a Bayesian perspective (Leamer, 1978). The p-value distribution is a function of statistical power (Cumming, 2008): higher power results in a more right-skewed distribution. Indeed, as statistical power increases, small p-values can be more likely when there is *no true effect* ($\mathcal{H}_0$) than when there *is an effect* ($\mathcal{H}_1$) (Maier & Lakens, 2022). Figure 1 displays this phenomenon, also known as Lindley's paradox (Wagenmakers & Ly, 2021). If there is no effect ($\mathcal{H}_0$ is true), p-values are distributed uniformly (solid line) irrespective of the sample size. When there is an effect (i.e., $\mathcal{H}_1$ is true), the p-value distribution becomes skewed, indicated by the dashed lines. Furthermore, a larger sample size produces a more right-skewed distribution since observing small p-values becomes even more likely.

When the solid line is above the dashed line, it is more likely to observe the p-value when there is no effect than when there is an effect. The point at which the lines cross (marked by circles) represents the point at which the null and alternative hypotheses are equally likely. As the sample size increases, the p-value at which the null and alternative hypotheses are equally likely decreases. If there are 200 observations in each group (these p-value are from two-sample t-tests), this point is 0.0106, which is well below the conventional 0.05. This means that if the observed p-value is between 0.0106 and 0.05, a researcher

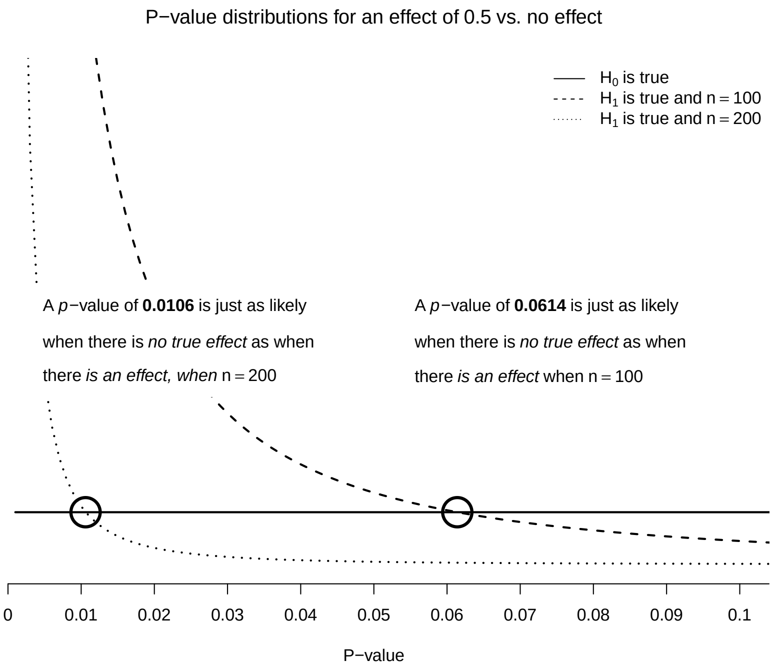

Figure 2: Illustration of Lindley's Paradox. *P*-value distributions for a two-sample, two-sided independent *t*-test with $n = 100$ and $n = 200$ in each group, respectively, shown for an effect size of 0.5 and of 0 (i.e. "no effect", solid line). The black circles mark which *p*-value is just as likely to be observed when there is *no true effect* as when there *is an effect*.

using $\alpha = 0.05$ will reject the $\mathcal{H}_0$ even though $\mathcal{H}_0$ is more likely than $\mathcal{H}_1$. This demonstrates how *p*-values of a given size do not indicate a fixed level of evidence for the alternative over the null (Royall, 1986).

Lowering alpha as a function of the sample size

We have seen that reducing α as a function of sample size is sensible. In large samples, a Neyman-Pearson perspective reveals we should trade off power for a lower probability of a Type I error (Wagenmakers & Ly, 2021), while a Bayesian perspective suggests we should lower α to avoid Lindley's Paradox. As eluded to previously, the two perspectives are closely related: minimizing the average error begins to align Bayesian and Neyman-Pearson testing procedures (Cornfield, 1966; Leamer, 1978). However, minimizing or balancing errors requires the researcher to estimate the statistical power, which includes knowing the effect size (Mudge et al., 2012), and to specify both the relative costs of Type I and

Type II errors and the relative probability of the null being true. While each of these parameters is challenging to determine, management researchers may find statistical power to be especially difficult. Indeed, power is rarely discussed in management research (Aguinis et al., 2009), perhaps due to the ubiquity of regression models, for which power is complex to estimate (Scherbaum & Ferreter, 2009).

To simplify the process of setting α, we propose a method that avoids power calculations and only requires researchers to specify the sample size. This approach avoids Lindley's Paradox by setting alpha such that a significant p-value only occurs when the alternative hypothesis is at least as likely as the null hypothesis (Maier & Lakens, 2022). In most cases, this will also lead to more balanced error rates than when using conventional α values. Indeed, our approach allows for the possibility to guarantee the Type-I and Type-II errors are equal.

Bayes factors

To set α to avoid Lindley's Paradox, we connect the p-value to an inference criterion that demands increasing evidence from the data as the sample size increases. One such criterion is the Bayes factor (BF) from Bayesian statistics (Kass & Raftery, 1995). The BF contrasts the probability of observing the data, y, under $\mathcal{H}_0$ to the probability of the data under $\mathcal{H}_1$.

$$\mathrm{BF}_{10} = \frac{\Pr(y|\mathcal{H}_1)}{\Pr(y|\mathcal{H}_0)}. \tag{2}$$

The BF expresses the evidence for $\mathcal{H}_1$ relative to $\mathcal{H}_0$ in the data, i.e., which of the two hypotheses is more likely to have generated the data. Unlike the p-value in 1, the BF takes into account both $\mathcal{H}_0$ and $\mathcal{H}_1$. By weighing the support for one model against the other, the BF quantifies the evidence for and against two competing statistical hypotheses (Andraszewicz et al., 2015).

A BF of 1 suggests equal evidence for $\mathcal{H}_0$ and $\mathcal{H}_1$, while a BF of 10 suggests the data are 10 times more likely under $\mathcal{H}_1$. BFs have a continuous scale, but Jeffreys (1939) suggested a series of discrete categories of evidential strength that can be useful to summarize the BF, a table can be found in Table 1.

Connecting Bayes factors to *p*-values

The Bayesian-Frequentist compromise we propose combines the evidential aspect of Bayesian statistics with the error control aspect of frequentist statistics. This compromise is achieved by connecting Bayes factors to p-values; transforming a p-value into a BF, ensuring α is set such that Lindley's paradox is avoided.

Several easy-to-calculate bounds for BFs exist in the literature; for example, $\mathrm{BF}_{10} < -1/e\,p\log p$ (Berger & Delampady, 1987; Sellke et al., 2001). Benjamin

Table 1: Common interpretation of Bayes factors from Lee and Wagenmakers (2013)

Value of BF_{10}	Evidence
>100	Extreme evidence for $\mathcal{H}_1$
30 – 100	Very strong evidence for $\mathcal{H}_1$
10 – 30	Strong evidence for $\mathcal{H}_1$
3 – 10	Moderate evidence for $\mathcal{H}_1$
1 – 3	Anecdotal evidence for $\mathcal{H}_1$
1	No evidence
1/3 – 1	Anecdotal evidence for $\mathcal{H}_0$
1/3 – 1/10	Moderate evidence for $\mathcal{H}_0$
1/10 – 1/30	Strong evidence for $\mathcal{H}_0$
1/30 – 1/100	Very strong evidence for $\mathcal{H}_0$
<1/100	Extreme evidence for $\mathcal{H}_0$

et al. (2018) used this bound to argue for setting the alpha level to 0.005 because a p-value of 0.005 implies a large-sample upper bound on the BF of 13.89. While this Volke-Sellke bound has been suggested as a valuable means for transforming p-values for coefficients in regression models to BFs (Harvey, 2017), the conversion does not take sample size into account. This is limiting because the evidence provided by a p-value against a point-null hypothesis depends on sample size (Held & Ott, 2016). Sample size dependent extensions exist (Held & Ott, 2016, 2018), but are valid only for a very narrow set of models. The same is true for the transformations developed by Faulkenberry (2019) and Rouder et al. (2009), which can only be used to calculate BFs for ANOVA and simple t-tests, respectively.

In this paper, we use the approximated adjusted fractional BF (AAFBF) of Gu et al. (2018). The AAFBF is sample size dependent and extends to testing hypotheses for generalized linear models such as linear, logistic, and Poisson regression models. The following section introduces the AAFBF - following closely the work of Gu et al. (2018) - before showing how it can be used to set α in a sensible manner.

Approximated adjusted fractional Bayes factors

As discussed, we concentrate on a single point-null hypothesis of significance, $\mathcal{H}_0 : \theta = 0$, against a two-sided alternative, $\mathcal{H}_1 : \theta \neq 0$, i.e. the negation of the null hypothesis. This form of alternative hypothesis is also known as an unconstrained alternative in the sense that it imposes the minimal possible constraint given that the null is false. When the alternative hypothesis is unconstrained, the encompassing prior method of Klugkist et al. (2005) can be used.

Specifying the prior is one of the biggest hurdles to overcome when conducting

Bayesian analysis, it captures information known to the researcher prior to data collection regarding the model parameters. For hypothesis testing, a prior is required for the parameters under both the null and alternative; however, the encompassing prior method sets the prior under the null as a function of the prior under the alternative, such that only one prior must be specified. In particular, following the notation of Gu et al. (2018), we write the prior distribution under the alternative as $\pi_1(\theta, \zeta)$, where ζ are the model parameters that do not appear in the null hypothesis, i.e. the regression coefficients whose significance are not being tested. The prior under the null, $\pi_0(\theta, \zeta)$, is then written as a constrained version of the prior under the alternative

$$\pi_0(\theta, \zeta) = \frac{\pi_1(\theta, \zeta)\mathcal{I}(\theta = 0)}{\int \pi_1(\theta = 0, \zeta) d\zeta}, \tag{3}$$

where $\mathcal{I}(\theta = 0)$ is an indicator function equal to 1 when $\theta = 0$ and 0 elsewhere. This form of $\pi_0(\theta, \zeta)$ can be viewed as placing mass at the point(s) where $\theta = 0$ and reweighting to ensure a properly defined distribution function.

Armed with these priors, the BF of Equation 2 can be written as the Savage-Dickey density ratio (Dickey, 1971)

$$\mathrm{BF}_{10} = \frac{\pi_1(\theta = 0)}{\pi_1(\theta = 0|y)}. \tag{4}$$

The denominator of this fraction is known as the posterior since it materializes only after the data has been observed; we will refer to the numerator as simply the prior. Each of these must be specified. If θ is estimated via maximum likelihood estimation (MLE) - as is the case for generalized linear models - standard asymptotic arguments show the posterior distribution is normal with mean $\hat{\theta}_{MLE}$ and variance $\hat{\sigma}^2$, where $\hat{\sigma}^2$ is an estimate of σ^2, the Cramer-Rao lower bound, i.e. the inverse of the Fisher information (see, for example, Gelman et al., 2013).

The prior distribution is specified as a fractional prior (O'Hagan, 1995). This constitutes updating an uninformative prior with a likelihood based on a fraction b_n of the data, where the subscript n is used to make explicit its dependence on the sample size; the choice of b_n will be discussed in detail in the following section. As well as avoiding an arbitrary choice of prior, BFs based on fractional priors enjoy several attractive properties, including consistency and invariance to linear transformations (O'Hagan, 1995). In our context, the fractional prior is normal with variance $\hat{\sigma}^2/b_n$, reflecting that the variance is inflated by a factor of $1/b_n$ relative to the posterior variance, which uses the full data. Ordinarily, the mean of the fractional prior would be equal to the mean of the posterior, $\hat{\theta}_{MLE}$, however, it is now common practice to adjust the mean to align with the null hypothesis (Mulder, 2014), i.e. for a test of significance, the mean is set to 0. This ensures that under the prior, values above the null hypothesis value are just as likely as values below. This adjustment, together with the large-sample approximation to normality, lead to this form of prior being known as an approximated adjusted fractional prior.

Plugging the posterior and the approximated adjusted fractional prior into the BF given in equation 4 gives the approximated adjusted fractional Bayes factor

$$\mathrm{AAFBF}_{10} = b_n^{1/2}\ \exp\left(\tfrac{1}{2}W\right), \tag{5}$$

where $W = (\hat{\theta}_{MLE}/\hat{\sigma})^2$, i.e. it is the square of the student t-statistic, also known as the Wald statistic. The Wald statistic has a chi-squared asymptotic distribution with one degree of freedom; thus, the p-value of a Wald statistic is given by $(1 - F_{\chi_1}(W))$, where F_{χ_1} is the cumulative distribution function of the Chi-squared distribution with one degree of freedom. This statistic can be used to test for significance in any generalized linear regression model, such as linear, logit, or Poison regression.[3] Using equation (5), we can finally connect the frequentist p-value to the Bayesian BF. A simple rearrangement of this equation yields

$$W = 2\mathrm{ln}\left(b_n^{-1/2}\mathrm{AAFBF}_{10}\right), \tag{6}$$

and, consequently, the p-value can be expressed as

$$p = 1 - F_{\chi_1}\left(2\mathrm{ln}\left(b_n^{-1/2}\mathrm{AAFBF}_{10}\right)\right). \tag{7}$$

Equation 7 shows that, for a given choice of b_n, there is a one-to-one mapping between the p-value and the AAFBF_{10}. Thus, instead of using an arbitrary 5% alpha level and potentially falling afoul of Lindley's paradox, set $\mathrm{AAFBF}_{10} = 1$ and use the corresponding p-value as the alpha level to ensure the null hypothesis is only ever rejected when the alternative is more likely.

Guidance for setting b_n

The sample size enters the p-value expression in equation 7 only through b_n, which itself fully determines the approximated adjusted fractional prior in our setting. To set the alpha level as a function of sample size, we need to choose b_n carefully. Bayes factors are highly sensitive to the choice of prior (Tendeiro & Kiers, 2019), so arbitrary choices of b_n should generally be avoided (Berger & Pericchi, 2001). If b_n is chosen too small, the prior variance will be large and the BF will be correspondingly small, leading to $\mathcal{H}_0$ being favored too frequently. Ideally, we should use a prior that has a sufficiently large variance to cover the parameter values thought to be plausible - but not excessively large such that

[3] A generalization of this test statistic can be used to test joint hypotheses, though we do not pursue this extension here.

$\mathcal{H}_0$ is always favored (Raftery, 1999). Because formulating informative priors that accurately reflect one's beliefs is challenging and time-consuming (Berger, 2006), we provide four sensible choices of b_n based on suggestions from the previous literature.

The priors we recommend are broadly applicable, computationally convenient, have desirable theoretical properties, and only require the researcher to specify the number of observations. Below, we provide guidance on how researchers may choose b_n depending on which properties they find most important. An overview is available in Table 2.

Table 2: Methods for setting b_n

Method	Characteristic	Recommendation	Expression (b)	Reference
b_{JAB}	Highly agnostic regarding prior information	Use if no strong preference	$1/n$	Wagenmakers (2022)
$b_{\min}$	Uses the minimum sample for a proper prior	Use if concerned about misspecifying the prior	$2/n$	Gu et al. (2018)
b_{robust}	Ensures $b_{\min}$ is robust to small samples	Use if your sample is small	$\max\{(2)/n, 1/\sqrt{n}\}$	O'Hagan (1995)
b_{balanced}	Balances Type I and II error rates	Use if Type-II errors are costly	$\int_0^1 \exp(-n\beta_e^2/4)\, d\beta_e$	Gu et al. (2016)

Default choices of b_n

A good baseline prior is the normal unit-information prior (Raftery, 1999). Informally, the unit-information prior[4] can be thought of as a prior distribution with the same amount of information as a single, typical observation. If $b_n = 1/n$, the AAFBF is equivalent to a default BF based on a unit-information prior (Kass & Raftery, 1995). In particular, it is equivalent to Jeffreys' approximate BF (JAB)[5]. For this reason, we refer to this option as b_{JAB}.

An alternative choice for b_n is to use the minimal training sample for the prior specification, leaving maximal information in the data for hypothesis testing (Berger & Pericchi, 1996; O'Hagan, 1995). In other words, we set b_n such that the minimum number of observations is used to specify the prior under a given null hypothesis. In the case of a single null hypothesis, we need two observations; for the AAFBF, this is achieved by setting $b_{\text{min}} = 2/n$ (Gu et al., 2018). This is a sensible choice of b_n when there are concerns regarding the specification of the prior.

Researchers could instead choose a version of b_{min} that is robust to small samples. When using the AAFBF, a normal approximation is used for the fractional prior. However, the AAFBF can be overly sensitive to the prior distribution when b_n is small; a small b_n implies that less information in the data is used for prior specification (Conigliani & O'Hagan, 2000). To protect against this, we can set $b_{\text{robust}} = \max\{(2)/n, 1/\sqrt{n}\}$ (O'Hagan, 1995). This prevents b_n from becoming "too small", improving robustness to small sample sizes.

The final choice for b_n ensures the Type-I and Type-II error rates are balanced. Type-I and Type-II errors can be quite different when using b_{JAB}, b_{min}, or b_{robust} (Gu et al., 2016). A frequentist, who would like to control error rates, would prefer to specify b_n such that the error probabilities are kept aligned. If we assume a range of realistic effect sizes $\beta_e \in [0, 1]$ and assume that every effect size is equally likely within this interval, the choice for b_n that equalizes error rates is given by $b_{\text{balanced}} = \int_0^1 \exp(-n\beta_e^2/4)\, d\beta_e$ (Gu et al., 2016). Unfortunately, it is not always possible for b_{balanced} to produce equal error rates[6], still, it is only in rare cases that the Type-I and Type-II errors are far apart (Gu et al., 2016).

[4] A unit-information prior scales the prior to match the sampling variability of a single observation (Cousins, 2017).

[5] Jeffreys' approximate BF is a simple expression that languished in relative obscurity for 85 years (Wagenmakers, 2022). It is asymptotically identical to the widely-used Bayesian information criterion (BIC; Schwarz, 1978); thus, one can think of using b_{JAB} as setting the alpha level based on BIC, at least when sample sizes are large.

[6] The reason b_{balanced} cannot always give equal error rates is that we constrain $b_{\text{balanced}} \in [b_{\text{min}}, 1/2]$. This is done to ensure the prior specification is a consistent BF (Morey et al., 2016). The lower bound ensures that we use at least the minimum number of observations to specify proper priors. The upper bound ensures that no more than half the likelihood is used for prior specification, which would be undesirable in Bayesian testing (Berger & Pericchi, 1996).

Comparing different choices for b_n

The smaller is b_n, the more pessimistic AAFBF is regarding $\mathcal{H}_1$; thus, b_{JAB} puts the least faith in $\mathcal{H}_1$ and gives a lower α than the other three methods, as seen in Figure 3. For instance, the top-left plot shows that, for a sample size of $1,000$ and a BF of 3, we must set $\alpha = 0.003$. This makes b_{JAB} a good choice for those who want to be conservative against small effects. Skepticism against small effects is often desirable when Type-I errors are costly or if factors such as measurement error or observational research designs make an exact point null unlikely (Orben & Lakens, 2020). Using b_{JAB} in large samples, means small effects will not be viewed as evidence against the null. Because b_{JAB} is the most conservative choice, we recommend it when researchers have no strong preferences over theoretical properties.[7]

For researchers concerned about the normality approximation of the prior, they may opt for either b_{min} or b_{robust}. Recall that the normality approximation is based on using MLE, which is asymptotically normal. However, in small samples, this approximation may be inaccurate, particularly if the likelihood is misspecified. For instance, in management research, bounded dependent variables such as fractions, counts, or non-count variables with lower bounds at zero are often estimated via linear regression (i.e. assuming normality) (Villadsen & Wulff, 2021a, 2021b). This quasi-MLE approach (using a normal likelihood despite the data potentially being non-normal) is still consistent and asymptotically normal, but the finite sample distribution may be further from normality than the finite sample distribution of MLE. In such cases, we should use b_{min} when the sample size is large, and b_{robust} when n is small. In the top right of Figure 3, we can see b_{min} and b_{robust} generally result in a higher α than b_{JAB} but smaller than b_{balanced}. For instance, for a BF of 3 with $1,000$ observations, we must set $\alpha = 0.004$ in the non-robust case and $\alpha = 0.017$ in the robust case.

Finally, if Type-II errors are relatively costly, researchers can use b_{balanced} to keep them in line with Type-I errors. b_{balanced} sets α higher relative to the other methods, losing the skepticism against small effects inherent in the other choices. As shown in Figure 3, b_{balanced} results in the highest α. For $1,000$ observations, we must set $\alpha = 0.024$ to achieve a BF of at least 3.

Examining α as a function of sample size

Figure 4 illustrates the consequences of setting α as a function of n for various BFs. A close examination reveals three key insights: 1) p-values of a given size do not indicate evidence of fixed strength, 2) $\alpha = 0.05$ is too high for the most common sample sizes in management, and 3) fixed thresholds are doomed to fail.

[7] b_{JAB} is the default option in the alpnaN R-package developed as a part of this paper. b_{JAB} is also implemented in the R-package BFpack for Bayesian hypothesis testing (Mulder et al., 2021).

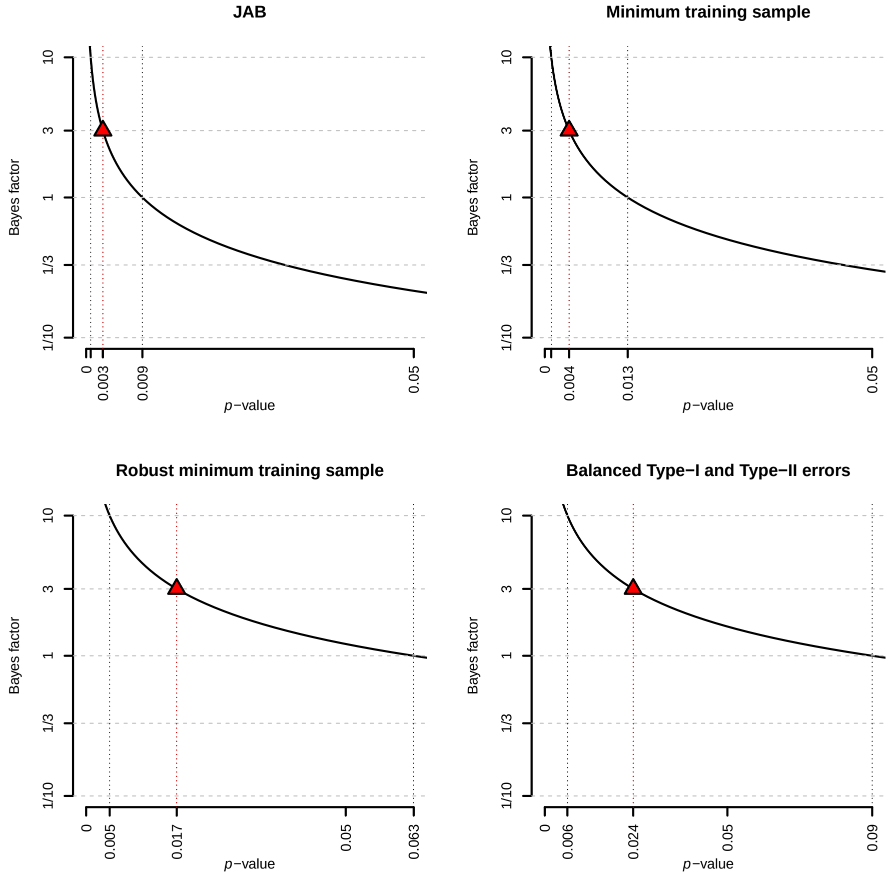

Figure 3: Setting α using the BF. Each plot illustrates a method for setting b_n: JAB (top left), minimum training sample (top right), robust minimum training sample (bottom left), and balanced Type-I and Type-II errors. The plots illustrate the relationship between the BF and the p-value for a sample size of $1,000$. The red triangle shows the p-value corresponding to a BF of 3.

p-values of a given size do not indicate evidence of fixed strength

Figure 4 illustrates the strong dependence of α on sample size. The larger the sample size, the smaller the p-value that corresponds to a given BF. Consequently, a p-value of a given size does not indicate evidence of fixed strength (Royall, 1986). In this light, it is problematic that management researchers rank their results according to which are “highly significant”, “significant”, or just “marginally significant” depending on the p-value (Aguinis et al., 2018) or use the p-value directly as a measure of the strength of a result (Bettis et al., 2016). Instead, the p-value can be viewed as an indirect measure of evidence

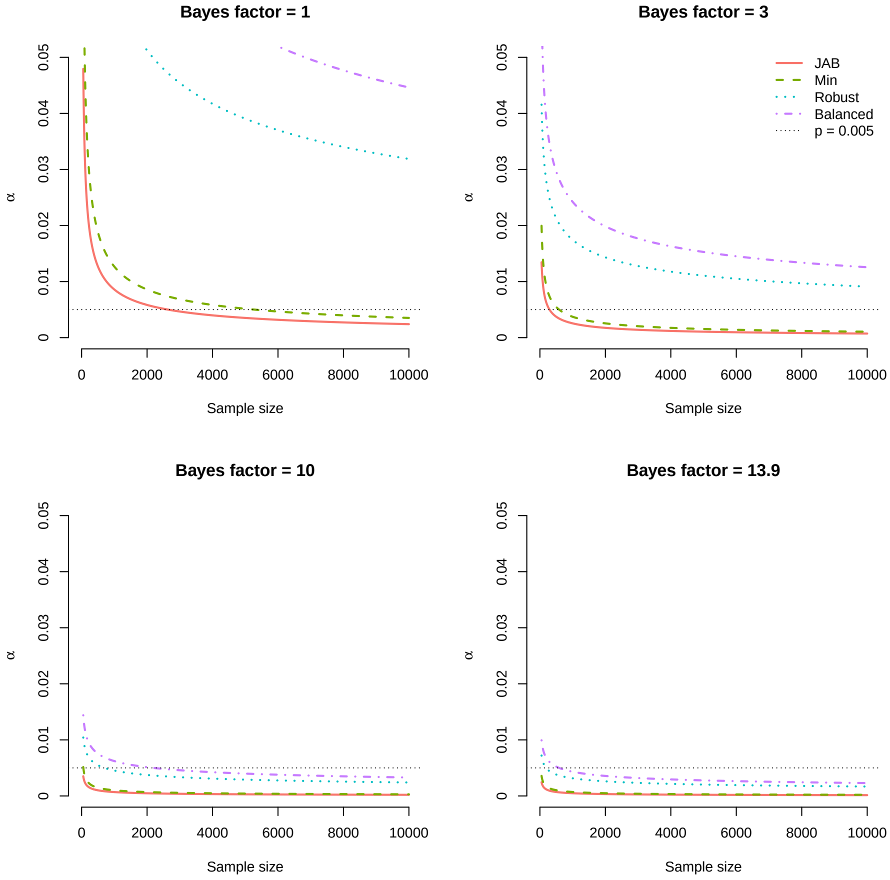

Figure 4: Examining α as a function of sample size. Each plot illustrates α as a function of n depending on the desired BF: $\mathcal{H}_0$ and $\mathcal{H}_1$ are equally likely (BF $= 1$, top left), moderate evidence (BF $= 3$, top right), strong evidence (BF $= 10$, bottom left), and the lower BF bound suggested by Benjamin et al. (2018) (BF $= 13.9$, bottom left).

whose magnitude must be judged in relation to the sample size used to compute it (Hartig & Barraquand, 2022; Lakens, 2022). For instance, a p-value of 0.002 maps to between moderate and strong evidence for $\mathcal{H}_1$ ($\text{BF}_{\text{JAB}} = 3.86, \text{BF}_{\text{balanced}} = 28.91$) for $n = 1,000$. However, if $n = 10$ million, the same p-value provides at best anecdotal evidence for $\mathcal{H}_1$ ($\text{BF}_{\text{balanced}} = 2.89$) and at worst strong evidence for $\mathcal{H}_0$ ($\text{BF}_{\text{JAB}} = 0.04$). This demonstrates how p-values are not a consistent measure of evidence (Hubbard & Lindsay, 2008).

The default $\alpha = 0.05$ is too high

Aguinis et al. (2010) surveyed papers published in *Administrative Science Quarterly (ASQ)*, *Academy of Management Journal (AMJ)*, and *Strategic Management Journal (SMJ)* from 2002 to 2006 and found 96% used $\alpha = 0.05$ to declare a result statistically significant. Assuming a sample size of 800, which was the median sample size reported in 300 management papers published between 2007 and 2016 (Villadsen & Wulff, 2021a), rejecting the null with $\alpha = 0.05$ would only ensure evidence of between 0.24 (JAB) and 1.76 (balanced); at best, only anecdotal evidence for $\mathcal{H}_1$. If researchers instead set α as a function of sample size, they could set $\alpha = 0.0029$ (for $n = 800$) to be conservative against small effects (JAB) or $\alpha = 0.0258$ when Type-II errors are costly. This demonstrates how researchers working with median-sized samples risk rejecting the null, even though - in the best case - there is only anecdotal evidence for $\mathcal{H}_1$.

Fixed thresholds are doomed to fail

Benjamin et al. (2018) suggested changing the default alpha level to 0.005 for claims of new discoveries. The argument for lowering α to 0.005 was based partly on a Bayesian argument, because, for a two-sided t-test, $p = 0.005$ implies a large-sample upper bound on the BF between 13.9 and 25.7 (Sellke et al., 2001). However, as argued earlier, the Volke-Selke bound is limited as it does not depend on the sample size.

In the bottom right of Figure 4, we set the desired level of evidence to 13.9, representing the lowest BF bound used by Benjamin et al. (2018). However, if we want to ensure BF ≥ 13.9, we require an α smaller than 0.005 whenever $n > 587$; moreover, this result is based on using $b_{balanced}$, i.e. the default prior resulting in the highest α. If we use JAB, to ensure BF ≥ 13.9, we must set α lower than 0.005 whenever $n > 14$. Thus, if researchers use $\alpha = 0.005$ for testing regression coefficients, they risk severely overestimating the evidence for a significant effect. The root of the problem is believing a fixed threshold gives consistent evidence. The reality, however, is that the level of evidence provided by an α level depends on sample size.

Emperical Demonstrations

Using our proposed method is easy: select a method for calculating b_n, choose a minimum desired level of evidence, then input the sample size. The resulting alpha is then preregistered before the analysis to ensure the interpretability of reported p-values (King et al., 2021; Wagenmakers et al., 2012). The researcher can then perform the desired regression and compare the p-value to the selected alpha. We can also quantify the evidence by calculating the BF from the computed p-value.

Here, we apply the proposed Bayesian-frequentist compromise to four published studies. Each study has been selected to demonstrate how researchers may decide on which method to use for choosing b_n. In all cases, we have chosen α to ensure a BF of at least 3, i.e. at least moderate evidence. The results are summarised in Table 3.

Example 1: Trivial effects and costly Type-I errors

In the first study, we use the JAB prior. This prior is relevant when factors such as measurement error or an observational research design make it highly unlikely that any effect is exactly zero. Yan et al. (2021) collect a large sample of observational data on firms' environmental performance and regress a subjective environmental score on various factors such as the proportion of green investing in the financial sector. In such a case, it is wise to be cautious of overstating the significance of trivial effects in large sample sizes. The JAB prior is an excellent default to safeguard against trivial findings that might be due to unintended factors, for example, measurement error in grading the environmental performance of firms. Finally, in this situation Type-I errors are costly. For instance, if there is no effect of green investing on environmental performance, a false-positive might lead to millions of dollars wasted on green investing. As discussed above, the JAB prior provides the lowest alpha of the four default priors, providing the most protection against Type-I errors.

In their original study, Yan et al. (2021) consider p-values lower than 0.1 to be significant. Given their sample size of $25,866$, we require $\alpha = 0.0006$ to achieve moderate evidence. The exact p-value of 1.5×10^{-7} related to green investing shows their result is still significant if we use the more stringent alpha level. In fact, the BF of 6029 suggests overwhelming evidence for the alternative hypothesis. The authors test a second hypothesis concerning environmental protection policy and report a p-value of 0.0143. This corresponds to a BF of 0.1254, suggesting the null is approximately eight times more likely than the alternative, despite the null being rejected in the original paper. In sum, the authors could have provided more impressive support for their hypothesis regarding green investing while avoiding claiming a significant effect for environmental protection policy in the face of substantial evidence to the contrary.

Example 2: Large-sample likelihood misspecification

For the second study, we use the minimum training sample prior. Recall that this prior is relevant when we suspect the MLE is misspecified. Kistruck et al. (2013) use ordinary least squares (OLS) to estimate the effect of diversification on efficiency in charitable organizations. Because efficiency is measured as a percentage, hence bounded between zero and 100, the normal likelihood cannot be correct (Villadsen & Wulff, 2021a). Although the likelihood is misspecified, OLS is still a quasi-MLE and therefore yields consistent estimates. In a finite sample, however, we might be concerned that the sample distribution is nonnormal (Wooldridge, 2010, CH 13). Because the authors use a large sample size of $17,860$, a minimum training sample prior should be sufficient to account for the misspecification of the prior.

Kistruck et al. (2013) hypothesize a u-shaped relationship between efficiency and two types of diversification. A sample size of $17,860$ requires $\alpha = 0.0008$ to achieve at least moderate evidence with a minimum training sample prior. Thus, authors relying on the most common fixed thresholds risk a significant result when in fact the null is more likely. Letting alpha depend on the sample size would have allowed Kistruck et al. (2013) to provide more impressive support for their hypothesis regarding geographic diversification where the alternative is at least three times more likely than the null. With respect to product diversification, our approach suggests that despite the significant result found in their paper ($p = 0.0057$), it is actually more than twice as likely that the null is true compared to the alternative (BF = 2.0597).

Example 3: Small-sample likelihood misspecification

In the third example, we use the robust minimum training sample prior as the MLE is misspecified and the sample is small. Jeong and Siegel (2018) hypothesize the threat of falling high status as a determinant of large-scale corporate bribery. The dependent variable is measured as the annual bribe amount in (KRW billion) - a non-count variable with a lower bound at zero for which a Poisson quasi-MLE regression model is appropriate (Villadsen & Wulff, 2021b). Because it is unlikely that the conditional distribution of the annual bribe amount follows a Poisson distribution, the likelihood is misspecified. The sample size is only 237, which raises concerns about the normality approximation; thus, we use the robust minimum training sample prior, which improves the robustness of the prior to small samples.

The paper's main result gives a 'significant' p-value of 0.0429 for the effect of the threat of falling high status. However, using the robust prior, we should set $\alpha = 0.0264$ to achieve at least moderate evidence. Thus, there is less than moderate evidence for the alternative (BF = 1.9791). This insignificant result does not give conclusive evidence of no effect (Lakens et al., 2020), it suggests that more data is needed to conclude whether the threat of falling high status is indeed a determinant of large-scale corporate bribery.

Table 3: Re-analyzing published p-values using a frequentist-Bayesian compromise

Prior	Claim	N	z-stat	p-values	BF	$\alpha_{\mathrm{BF}=3}$
JAB	Factors positively related to firm-level environmental performance (Yan et al., 2021)					
	Green investing in financial sector	25688	5.25	1.5×10^{-7}	6029	0.0006
	Country environmental protection policy	25688	2.45	0.0143	0.1254	0.0006
Min	U-shape between diversification and efficiency (Kistruck et al., 2013)					
	Geographic diversification (squared)	17860	3.4	0.0007	3.4261	0.0008
	Product diversification (squared)	17860	2.7662	0.0057	0.4855	0.0008
Robust	Threat of falling high status affects corporate bribery (Jeong & Siegel, 2018)					
	Threat of Falling High Status	237	2.03	0.0429	1.9791	0.0264
Balanced	CEOs with out-of-the-money options manipulate firm earnings (Zhang et al., 2008)					
	Main effect	1390	3.2346	0.0012	40.7818	0.0220
	Effect stronger for longer-tenured CEOs	1390	2.1667	0.0303	2.2799	0.0220

Notes: sample size (N), z-statistics and p-values collected from the listed papers. When z-statistics were not available, they were computed based on the p-values and vice-versa. The numbers in the z-statistics for two studies (Kistruck et al., 2013; Yan et al., 2021) are t-statistics. The calculations of BF and $\alpha_{\mathrm{BF}=3}$ are available in the appendix.

Example 4: Costly Type-II errors

Finally, Zhang et al. (2008) investigate whether CEOs with out-of-the-money options manipulate firm earnings. In this case, a prior that balances the error rates would be a wise choice since a Type-II error could be costly. Not detecting such behavior could cause serious problems: stock price decline, reputational damage, top management turnover, possible bankruptcies, and loss of investor confidence (Aguinis et al., 2010).

The study uses a sample size of $1,390$, so we set $\alpha = 0.0220$ to achieve at least moderate evidence. At this level, we still find a significant relationship ($p = 0.0012$) with strong evidence (BF = 40.7817). If the authors had set their alpha to achieve at least moderate evidence, they could have presented stronger support for their main claim without being too concerned about a Type-II error. To their surprise, Zhang et al. (2008) find that longer-tenured CEOs with greater value in out-of-the-money options are more likely to manipulate earnings ($p = 0.0303$). However, our re-analysis suggests this result is insignificant and only provides anecdotal evidence in favor of the alternative (BF = 2.2799). If the authors had set α to achieve at least moderate evidence, they could have presented stronger support for their main claim while avoiding interpreting a significant result as anything more than anecdotal evidence.

Discussion

There is significant debate and concern surrounding the use of significance testing and p-values, particularly in light of the replication crisis. Indeed, many scientists have argued that significance tests should be abandoned altogether (Anderson et al., 2000; Carver, 1993; Gill, 1999). This view has been echoed in management research, where some recommend we "let go of statistical significance once and for all" (van Witteloostuijn, 2020, p. 275), "escape the straightjacket of NHST" (Lockett et al., 2014, p. 870), "stop relying on NHSTs" (Schwab et al., 2011, p. 1106), or even that "[i]t would be better for journals to ban p-values as well" (Starbuck, 2016, p. 74). The misinterpretation of p-values has lead *SMJ* to no longer accept papers "that report or refer to cut-off levels of statistical significance" (Bettis et al., 2016, p. 261). This stance has led other journals, such as *Management and Organization Review*, to conclude that "the use of cut-off level of p-values to support or reject hypotheses is inappropriate" (Li et al., 2017, p. 440).

We agree that significance testing is often misused, but this does not warrant abandoning it; *abusus non tollit usum* - or, abuse does not cancel use. If all misused tools were abandoned, there would be few left to use; indeed, Bayesian methods are frequently misapplied due to incomplete prior reporting (van de Schoot et al., 2017) or BF misinterpretation (Tendeiro et al., 2022; Wong et al., 2022). Moreover, there are many potential costs if NHST were abandoned (Lakens, 2021): authors might overstate their conclusions more than with NHST, as seen after the 2016 ban of inferential statistics in *Basic and Applied Social Psychology*; without a threshold, there is no test of a claim (Mayo & Hand, 2022); error control is lost; and, NHST is one of the most studied and best-understood statistical procedures (Benjamini et al., 2021).

Instead of bans, researchers must improve the quality of their inference - both frequentist and Bayesian (Colling & Szűcs, 2021). We support the advice of the President's Task Force appointed by the board of the American Statistical Association (ASA): "P-values and significance testing, properly applied and interpreted, are important tools that should not be abandoned" (Benjamini et al., 2021, p 1084). This entails, among other things, justifying the alpha level. Our method provides a simple means to do this on the basis of sample size. Editors stress that researchers take n into account when evaluating statistical tests, but provide no guidance on how (Bettis et al., 2016; Combs, 2010; Hahn & Ang, 2017; Meyer et al., 2017). If researchers look towards standard textbooks in statistics, they will find little help there either: "Elementary statistics texts are not equipped to go into the matter; advanced texts are too preoccupied with the latest and fanciest statistical techniques to have space for anything so elementary. Thus the justifications for critical levels that are commonly offered are flimsy, superficial, and badly outdated" (Bross, 1971).

Our approach is a compromise between Bayesian and frequentist statistics; as with all compromises, we give up some benefits of each. These "pure" approaches might be attractive if researchers are comfortable providing more details e.g., by specifying relative costs of errors or informative priors. A pure frequentist approach is advisable when researchers (1) can specify the relative cost of Type-I and Type-II errors, (2) can justify the prior probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$, and (3) have enough information to perform a power analysis (Maier & Lakens, 2022). A frequentist approach is also recommended when researchers care only about controlling long-run error rates and not about an evidential interpretation of their test (Mudge et al., 2012). However, as demonstrated, a pure frequentist approach that balances error rates will also reduce alpha as n increases; so will often avoid Lindley's paradox too (Maier & Lakens, 2022).

A fully Bayesian approach is advisable when researchers (1) are comfortable specifying priors for their model parameters, (2) have well-specified alternatives, and (3) accept the computational burden from sampling from the posterior distribution (Harvey, 2017). While a fully Bayesian approach provides the probability of misleading evidence in the planning stage (Schönbrodt et al., 2017), it does not come with long-run error guarantees (Lakens, 2021).

There are strong norms to use fixed alpha levels in the management discipline. With this paper, we hope to persuade scholars to abandon the mindless use of fixed alphas and instead justify the alpha level as a function of sample size. Our explanations, demonstrations, empirical examples, and R-package hopefully make this adoption as straightforward as possible.

References

Aguines, H., & Harden, E. E. (2009). Sample size rules of thumb: Evaluating three common practices. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 269–288). Routledge.

Aguinis, H., Hill, N. S., & Bailey, J. R. (2021). Best Practices in Data Collection and Preparation: Recommendations for Reviewers, Editors, and Authors. *Organizational Research Methods*, *24*(4), 678–693. https://doi.org/10.1177/1094428119836485

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First Decade of Organizational Research Methods: Trends in Design, Measurement, and Data-Analysis Topics. *Organizational Research Methods*, *12*(1), 69–112. https://doi.org/10.1177/1094428108322641

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What You See Is What You Get? Enhancing Methodological Transparency in Management Research. *Academy of Management Annals*, *12*(1), 83–110. https://doi.org/10.5465/annals.2016.0011

Aguinis, H., Werner, S., Lanza Abbott, J., Angert, C., Joon Hyung Park, & Kohlhausen, D. (2010). Customer-Centric Science: Reporting Significant Research Results With Rigor, Relevance, and Practical Impact in Mind. *Organizational Research Methods*, *13*(3), 515–539. https://doi.org/10.1177/1094428109333339

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *The Journal of Wildlife Management*, *64*(4), 912–923. https://doi.org/10.2307/3803199

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*(2), 521–543.

Barnes, C. M., Dang, C. T., Leavitt, K., Guarana, C. L., & Uhlmann, E. L. (2018). Archival Data in Micro-Organizational Research: A Toolkit for Moving to a Broader Set of Topics. *Journal of Management*, *44*(4), 1453–1478. https://doi.org/10.1177/0149206315604188

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., Veaux, R. D. D., Efron, B., Evans, S., Glickman, M., Graubard, B. I., He, X., Meng, X.-L., Reid, N., Stigler, S. M., Vardeman, S. B., Wikle, C. K., Wright, T., Young, L. J., & Kafadar, K. (2021). The ASA president's task force statement on statistical significance and replicability. *The Annals of Applied Statistics*, *15*(3), 1084–1085. https://doi.org/10.1214/21-AOAS1501

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317–352.

Berger, J. O., & Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, *91*(433), 109–122. https://doi.org/10.2307/2291387

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. In P. Lahiri (Ed.), *Model selection* (pp. 135–208). Institute of Mathematical Statistics. Retrieved December 1, 2022, from https://projecteuclid.org/ebooks/institute-of-mathematical-statistics-lecture-notes-monograph-series/Model-selection/chapter/Objective-Bayesian-Methods-for-Model-Selection-Introduction-and-Comparison/10.1214/lnms/1215540968

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*, *37*(2), 257–261. https://doi.org/10.1002/smj.2477

Bross, I. D. (1971). Critical levels, statistical language and scientific inference. In G. Sprott (Ed.), *Foundations of statistical inference* (pp. 500–513). Holt, Rinehart; Winston.

Carver, R. P. (1993). The Case against Statistical Significance Testing, Revisited. *The Journal of Experimental Education*, *61*(4), 287–292. Retrieved December 15, 2022, from https://www.jstor.org/stable/20152382

Certo, S. T., Busenbark, J. R., Woo, H.-s., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639–2657. https://doi.org/10.1002/smj.2475

Colling, L. J., & Szűcs, D. (2021). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*, *12*(1), 121–147. https://doi.org/10.1007/s13164-018-0421-4

Combs, J. G. (2010). Big Samples and Small Effects: Let's Not Trade Relevance and Rigor for Power. *Academy of Management Journal*, *53*(1), 9–13. https://doi.org/10.5465/amj.2010.48036305

Conigliani, C., & O'Hagan, A. (2000). Sensitivity of the Fractional Bayes Factor to Prior Distributions. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *28*(2), 343–352. https://doi.org/10.2307/3315983

Cornfield, J. (1966). Sequential Trials, Sequential Analysis and the Likelihood Principle. *The American Statistician*, *20*(2), 18–23. https://doi.org/10.2307/2682711

Cousins, R. D. (2017). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, *194*(2), 395–432. https://doi.org/10.1007/s11229-014-0525-z

Cumming, G. (2008). Replication and p Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Dickey, J. M. (1971). The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. https://doi.org/10.1214/aoms/1177693507

Faulkenberry, T. J. (2019). Estimating Evidential Value From Analysis of Variance Summaries: A Comment on Ly et al. (2018). *Advances in Methods and Practices in Psychological Science*, *2*(4), 406–409. https://doi.org/10.1177/2515245919872960

Fisher, R. A. (1971). *The Design of Experiments* (9th ed.). Macmillan Pub Co.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall.

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly*, *52*(3), 647. https://doi.org/10.2307/449153

Good, I. J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, *87*(419), 597–606. https://doi.org/10.1080/01621459.1992.10475256

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *The British journal of mathematical and statistical psychology*, *71*(2). https://doi.org/10.1111/bmsp.12110

Hahn, E. D., & Ang, S. H. (2017). From the editors: New directions in the reporting of statistical results in the Journal of World Business. *Journal of World Business*, *52*(2), 125–126. https://doi.org/10.1016/j.jwb.2016.12.003

Hartig, F., & Barraquand, F. (2022). The evidence contained in the P-value is context dependent. *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2022.02.011

Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance*, *72*(4), 1399–1440. https://doi.org/10.1111/jofi.12530

Held, L., & Ott, M. (2016). How the Maximal Evidence of P-Values Against Point Null Hypotheses Depends on Sample Size. *The American Statistician*, *70*(4), 335–341. https://doi.org/10.1080/00031305.2016.1209128

Held, L., & Ott, M. (2018). On p-Values and Bayes Factors. *Annual Review of Statistics and Its Application*, *5*(1), 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

Hubbard, R., & Lindsay, R. M. (2008). Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology*, *18*(1), 69–88. https://doi.org/10.1177/0959354307086923

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

Jeffreys, H. (1939). *Theory of probability* (3rd ed.). Oxford University Press.

Jeong, Y., & Siegel, J. I. (2018). Threat of falling high status and corporate bribery: Evidence from the revealed accounting records of two South Korean presidents. *Strategic Management Journal*, *39*(4), 1083–1111. https://doi.org/10.1002/smj.2747

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, *90*(430), 773–795.

Ketchen, D. J., Boyd, B. K., & Bergh, D. D. (2008). Research methodology in strategic management: Past accomplishments and future challenges. *Organizational Research Methods*, *11*(4), 643–658. https://doi.org/10.1177/1094428108319843

Kim, J. H., Ahmed, K., & Ji, P. I. (2018). Significance Testing in Accounting Research: A Critical Evaluation Based on Evidence. *Abacus*, *54*(4), 524–546. https://doi.org/10.1111/abac.12141

King, A., Goldfarb, B., & Simcoe, T. (2021). Learning from Testimony on Quantitative Research in Management. *Academy of Management Review*, *46*(3), 465–488. https://doi.org/10.5465/amr.2018.0421

Kistruck, G. M., Qureshi, I., & Beamish, P. W. (2013). Geographic and Product Diversification in Charitable Organizations. *Journal of Management*, *39*(2), 496–530. https://doi.org/10.1177/0149206311398135

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Lakens, D. (2021). The practical alternative to the p-value is the correctly used p-value. *Perspectives on Psychological Science*, in press.

Lakens, D. (2022). Why P values are not measures of evidence. *Trends in Ecology & Evolution*, *37*(4), 289–290. https://doi.org/10.1016/j.tree.2021.12.006

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S. C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., . . . Zwaan, R. A. (2018). Justify your alpha. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests (D. Isaacowitz, Ed.). *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1st ed.). Wiley.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Li, M., Sharp, B. M., & Bergh, D. D. (2017). Assessing Statistical Results in MOR Articles: An Essay on Verifiability and Ways to Enhance It. *Management and Organization Review*, *13*(2), 431–441. https://doi.org/10.1017/mor.2017.18

Lockett, A., McWilliams, A., & Van Fleet, D. D. (2014). Reordering Our Priorities by Putting Phenomena before Design: Escaping the Straitjacket of Null Hypothesis Significance Testing. *British Journal of Management*, *25*(4), 863–873. https://doi.org/10.1111/1467-8551.12063

Maier, M., & Lakens, D. (2022). Justify Your Alpha: A Primer on Two Practical Approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221080396. https://doi.org/10.1177/25152459221080396

Mayo, D. G., & Hand, D. (2022). Statistical significance and its critics: Practicing damaging science, or damaging scientific practice? *Synthese*, *200*(3), 220. https://doi.org/10.1007/s11229-022-03692-0

Meyer, K. E., van Witteloostuijn, A., & Beugelsdijk, S. (2017). What's in a p? Reassessing best practices for conducting and reporting hypothesis-testing research. *Journal of International Business Studies*, *48*(5), 535–551. https://doi.org/10.1057/s41267-017-0078-8

Miller, D. J., Fern, M. J., & Cardinal, L. B. (2007). The Use of Knowledge for Technological Innovation within Diversified Firms. *The Academy of Management Journal*, *50*(2), 308–326. https://doi.org/10.2307/20159856

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*(1), e0208631. https://doi.org/10.1371/journal.pone.0208631

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. https://doi.org/10.1080/00273171.2015.1052710

Moshagen, M., & Erdfelder, E. (2016). A New Strategy for Testing Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink-Bosman, M., Menke, J., Aert, R. v., Fox, J.-P., Hoijtink, H., Rosseel, Y., Wagenmakers, E.-J., & Lissa, C. v. (2021). BFpack: Flexible Bayes Factor Testing of Scientific Theories in R. *Journal of Statistical Software*, *100*, 1–63. https://doi.org/10.18637/jss.v100.i18

Neyman, J., Pearson, E. S., & Pearson, K. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 99–138. Retrieved November 28, 2022, from https://www.jstor.org/stable/2346088

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Raftery, A. E. (1999). Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods & Research*, *27*(3), 411–427. https://doi.org/10.1177/0049124199027003005

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. https://doi.org/10.3758/PBR.16.2.225

Royall, R. M. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician*, *40*(4), 313–315. https://doi.org/10.2307/2684616

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating Statistical Power and Required Sample Sizes for Organizational Research Using Multilevel Modeling. *Organizational Research Methods*, *12*(2), 347–367. https://doi.org/10.1177/1094428107308906

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. https://doi.org/10.1037/met0000061

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests. *Organization Science*, *22*(4), 1105–1120.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71. Retrieved April 5, 2022, from https://www.jstor.org/stable/2685531

Starbuck, W. H. (2016). 60th Anniversary Essay. *Administrative Science Quarterly*, *61*(2), 165–183. https://doi.org/10.1177/0001839216629644

Tendeiro, J., Kiers, H., Hoekstra, R., Wong, T. K., & Morey, R. D. (2022). Diagnosing the Use of the Bayes factor in Applied Research. https://doi.org/10.31234/osf.io/du3fc

Tendeiro, J., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

van Witteloostuijn, A. (2020). New-day statistical thinking: A bold proposal for a radical change in practices. *Journal of International Business Studies*, *51*(2), 274–278. https://doi.org/10.1057/s41267-019-00288-8

Villadsen, A. R., & Wulff, J. N. (2021a). Are you 110% sure? Modeling of fractions and proportions in strategy and management research. *Strategic Organization*, *19*(2), 312–337. https://doi.org/10.1177/1476127019854966

Villadsen, A. R., & Wulff, J. N. (2021b). Statistical Myths About Log-Transformed Dependent Variables and How to Better Estimate Exponential Models. *British Journal of Management*, *32*(3), 779–796. https://doi.org/10.1111/1467-8551.12431

Wagenmakers, E.-J. (2022). Approximate Objective Bayes Factors From P-Values and Sample Size: The 3pn Rule. *psyarxiv*. https://doi.org/10.31234/osf.io/egydq

Wagenmakers, E.-J., & Ly, A. (2021). History and Nature of the Jeffreys-Lindley Paradox. *arXiv:2111.10191 [math, stat]*. Retrieved February 17, 2022, from http://arxiv.org/abs/2111.10191

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. *Collabra: Psychology*, *8*(1), 36357. https://doi.org/10.1525/collabra.36357

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (Vol. 2nd). The MIT Press.

Wright, P. M. (2016). Ensuring Research Integrity: An Editor's Perspective. *Journal of Management*, *42*(5), 1037–1043. https://doi.org/10.1177/0149206316643931

Yan, S., Almandoz, J. (, & Ferraro, F. (2021). The Impact of Logic (In)Compatibility: Green Investing, State Policy, and Corporate Environmental Performance. *Administrative Science Quarterly*, *66*(4), 903–944. https://doi.org/10.1177/00018392211005756

Zhang, X., Bartol, K. M., Smith, K. G., Pfarrer, M. D., & Khanin, D. M. (2008). Ceos On the Edge: Earnings Manipulation and Stock-Based Incentive Misalignment. *Academy of Management Journal*, *51*(2), 241–258. https://doi.org/10.5465/amj.2008.31767230

Appendix: Code for empirical demonstrations

Figures 3 and 4 and the results contained in Table 3 were all computed using the `alphaN` package that was developed for this paper. Information about the `alphaN` package including its code and vignettes demonstrating its use are available at https://github.com/jespernwulff/alphaN

Code for the empirical demonstrations

You can install the development version of alphaN from GitHub with:

```
# install.packages("devtools")
devtools::install_github("jespernwulff/alphaN")
```

Load the `alphaN` package

```
library(alphaN)
```

1. Environmental performance

```
# Set alpha
alphaN(25688, BF = 3)
```

```
## [1] 0.0004407494
```

```
# Compute BF for Green investing
JABt(25688,5.25)
```

```
## [1] 6029.144
```

```
# Compute BF for Protection policy
JABt(25688,2.45)
```

```
## [1] 0.1254761
```

2. Diversification in Charitable Organizations

```
# Set alpha
alphaN(17860, BF = 3, method = "min")
```

```
## [1] 0.0007774138
```

```r
# Compute BF for Geographic diversification
JABt(17860, 0.017/0.005, method="min")
```

```
## [1] 3.426071
```

```r
# Compute BF for Product diversification
JABt(17860, 0.639/0.231, method="min")
```

```
## [1] 0.4855072
```

```r
# Compute BF for Product diversification in favor of H0
1/JABt(17860, 0.639/0.231, method="min")
```

```
## [1] 2.059702
```

3. Threat of falling high status and corporate bribery

```r
# Set alpha
alphaN(237, BF = 3, method = "robust")
```

```
## [1] 0.02637516
```

```r
# Compute BF for Threat of FHS Definition 2
JABt(237, 0.164/0.081, method="robust")
```

```
## [1] 1.97916
```

4. CEO Earnings manipulation

```r
# Set alpha
alphaN(1390, BF =3, method = "balanced")
```

```
## [1] 0.02203032
```

```r
# Compute BF for main effect
JABt(1390, 0.33/1.32, method="balanced")
```

```
## [1] 0.2249601
```

```r
# Compute BF for tenure interaction effect
JABt(1390, 1.04/0.48, method="balanced")
```

```
## [1] 2.279919
```