

Justify Your Alpha: A Practical Guide

Maximilian Maier¹ & Daniel Lakens²

¹ University of Amsterdam, The Netherlands

² Eindhoven University of Technology, The Netherlands

This unpublished manuscript is submitted for peer review

The default use of an alpha level of 0.05 is sub-optimal for two reasons. First, decisions based on data can be made more efficiently by choosing an alpha level that minimizes the combined Type 1 and Type 2 error rate. Second, in studies with very high statistical power p -values lower than the alpha level can actually be support for the null hypothesis, instead of for the alternative hypothesis. Because it is difficult to abandon a bad practice without providing an alternative, this manuscript explains two approaches that can be used to justify your alpha. The first approach is based on the idea to either minimize or balance Type 1 and Type 2 error rates. The second approach lowers the alpha level as a function of the sample size. Software is provided to perform the required calculations. Both approaches have their limitations (e.g., the challenge of specifying relative costs and priors, or the slightly arbitrary nature of how the alpha level should decrease as the sample size increases) but they nevertheless provide a clear improvement compared to current practices. The use of alpha levels that have a better justification should improve statistical inferences and increase the efficiency of scientific research.

Keywords: hypothesis testing, Type 1 error, Type 2 error, statistical power

Word count: 4890 words

Researchers often rely on data to decide how to act. In a Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1933) studies are designed such that erroneous decisions that determine how we act are controlled in the long run at some desired maximum level. If resources were infinite we could collect enough data to make the chance of a wrong decision incredibly small. But resources are limited, which means that researchers need to decide how to choose the rate at which they are willing to make errors. After data is collected researchers can incorrectly act as if there is an effect when there is no true effect (a Type 1 error) or incorrectly act as if there is no effect when there is a true effect (a Type 2 error). Given the same amount of data, a reduction in the Type 1 error rate will increase the Type 2 error rate (and vice versa) .

The question how error rates should be set in any study requires careful consideration of the relative costs of a Type 1 error or a Type 2 error. Regrettably researchers rarely provide such a justification, and predominantly

use a Type 1 error rate of 5%. In the past the strong convention to use a 5% alpha level might have functioned as a de facto prespecification of the alpha level, which needs to be before the data is analyzed (Tunç, Tunç, & Lakens, 2021). Nowadays researchers can transparently preregister a statistical analysis plan in an online repository, which makes it possible to specify more appropriate but less conventional alpha levels. Even though it is possible to preregister non-conventional alpha levels, there is relatively little guidance on how to choose an alpha level for a study. This article explains why error rates need to be justified, and provides two practical approaches that can be used to justify the alpha level. In the first approach, the Type I and Type II error rates are balanced or minimized, and in the second approach the alpha level is lowered as a function of the sample size.

Why do we use a 5% alpha level and 80% power?

We might naively assume that when all researchers do something, there must be a good reason for such an established practice. An important step towards maturity as a scholar is the realization that this is not the case. Neither Fisher nor Neyman, two statistical giants largely responsible for the widespread reliance on hypothesis tests in the social sciences, recommended the universal use of any specific threshold. Fisher (1935)

Correspondence concerning this article should be addressed to Daniel Lakens, ATLAS 9.402, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

writes: “It 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 and Pearson (1933) write: “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.”

Even though in *theory* alpha levels should be justified, in *practice* researchers tend to imitate others. Fisher writes in 1926: “Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.” This sentence is preceded by the statement “If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point).” Indeed, in his examples Fisher often uses an alpha of 0.01. Nevertheless, researchers seem to have copied the value Fisher preferred, instead of his more important take-home message that the significance level should be set by the experimenter. The default use of an alpha level of 0.05 seems to originate from the early work of Gosset on the *t*-distribution (Cowles & Davis, 1982; Kennedy-Shaffer, 2019), who believed that a difference of two standard deviations (a *z*-score of 2) was sufficiently rare.

The default use of 80% power (or a 20% Type 2, or beta (b) error) is similarly based on personal preferences by Cohen (1988), who writes: “It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that beta is set at .20. This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for alpha of .05. The beta of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc.”

We see that conventions are built on conventions: the norm to aim for 80% power is built on the norm to set the alpha level at 5%. However, the real lesson Cohen was teaching us is to determine the relative seriousness of Type 1 and Type 2 errors, and to balance both types of errors when a study is designed. If a Type 1 error is considered to be four times as serious as a Type 2 error, the

weighted error rates in the study are balanced. Instead of imitating the values chosen by Cohen, researchers should aim to imitate the approach he used to justify error rates.

Justifying the alpha level

In 1957 Neyman wrote: “it appears desirable to determine the level of significance in accordance with quite a few circumstances that vary from one particular problem to the next.” The mindless application of null hypothesis significance tests, including setting the alpha level at 5% for all tests, has been criticized for more than half a century (Bakan, 1966; Gigerenzer, 2018). But it is difficult to abandon a mediocre research practice without an alternative.

There are two main reasons to abandon the universal use of a 5% alpha level. The first reason to carefully choose an alpha level is that decision making becomes more efficient. If researchers use hypothesis tests to make dichotomous decisions from a methodological falsificationist approach to statistical inferences (Tunç, Tunç, & Lakens, 2021), it is typically possible to make decisions more efficiently, where the combined Type 1 and Type 2 error rate is minimized, by setting the alpha level at a different value than 0.05. If we aim to either minimize or balance Type 1 and Type 2 error rates for a given sample size and effect size, the alpha level should be set not based on convention, but by weighting the relative cost of both types of errors.

The second reason is that as the statistical power increases, some *p*-values below 0.05 (e.g., $p = 0.04$) can be more likely when there is *no* effect than when there is an effect. This is known as Lindley’s paradox (Cousins, 2017; Lindley, 1957). The distribution of *p*-values is a function of the statistical power (Cumming, 2008), and the higher the power, the more right-skewed the distribution becomes (i.e. the more low *p*-values are observed). When there is no true effect *p*-values are uniformly distributed, and 1% of observed *p*-values fall between 0.04 and 0.05. When the statistical power is extremely high, not only will most *p*-values fall below 0.05, most will also fall below 0.01. In Figure 1 we see that with high power very low *p*-values are more likely to be observed when there *is* an effect than when there is *no* effect (e.g., the black curve representing *p*-values when the alternative is true falls above the dashed horizontal line for a *p*-value of 0.01). But observing a *p*-value of 0.04 is more likely when the null hypothesis (*H*₀) is true than when the alternative hypothesis (*H*₁) is true and we have very high power (the horizontal dashed line falls above the black curve for *p*-values larger than ~0.025).

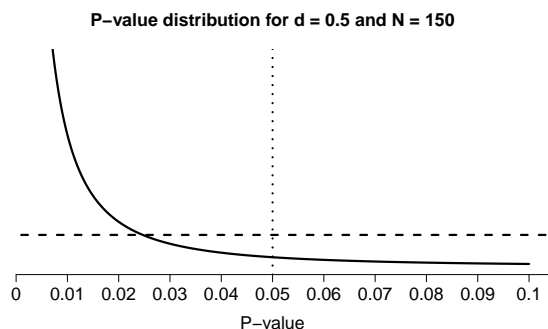


Figure 1. *P*-value distributions for a two-sided independent *t*-test with $N = 150$ and $d = 0.5$ (black curve) or $d = 0$ (horizontal dashed line) which illustrates how *p*-values just below 0.05 can be more likely when there is no effect than when there is an effect.

Researchers often want to interpret a significant test result as ‘support’ for the alternative hypothesis. In other words, in addition to controlling the *error rate*, researchers might be interested in interpreting the *relative evidence* in the data for null hypothesis over the alternative hypothesis. If so, it makes sense to choose the alpha level such that when a significant *p*-value is observed, the *p*-value is actually more likely when the alternative hypothesis is true than when the null hypothesis is true. This means that when statistical power is very high (e.g., the sample size is very large), the alpha level should be reduced. For example, if the alpha level in Figure 1 is lowered to 0.02 then the alternative hypothesis is more likely than the null hypothesis for all significant *p*-values that would be observed. This approach to justifying the alpha level can be seen as a frequentist/Bayesian compromise (Good, 1992). The error rate is controlled, but the alpha level is also set at a value that guarantees that whenever we reject the null hypothesis, the data is more likely under the alternative hypothesis, than under the null.

Minimizing or Balancing Type 1 and Type 2 Error Rates

If both Type 1 as Type 2 errors are costly, then it makes sense to optimally reduce both errors as you design studies. This leads to studies where you make decisions most efficiently. Researchers can choose to design a study with a statistical power and alpha level that minimizes the *combined error rate*. For example, assuming H_0 and H_1 are a-priori equally probable (the prior probability for both is 0.5). With a Type 1 error rate of 0.05 and statistical power of 0.80 the combined error rate is $(0.5 \times 0.05 + 0.5 \times 0.20) = 0.125$. We could have chosen to set the alpha level in this study to 0.1. If the Type 1

error rate is 0.1, the statistical power (given the same sample size) would be 0.88, which means the combined error rate is $(0.5 \times 0.1 + 0.5 \times 0.12) = 0.11$. In other words, increasing the Type 1 error rate from 0.05 to 0.1 reduced the Type 1 error rate from 0.2 to 0.12, and the combined error rate from 0.125 to 0.11. In the latter scenario, our total probability of making an erroneous decision has become 0.015 smaller. As shown below, this approach can be extended to incorporate scenarios where the prior probability of H_0 and H_1 differ. Mudge, Baker, Edge, and Houlahan (2012) show that by choosing an alpha level based on the relative weight of Type 1 errors and Type 2 errors, and assuming beliefs about the prior probability that H_0 and H_1 are correct, decisions can be made more efficiently than when the default alpha level of 0.05 is used.

Winer (1962) writes: “The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type 1 and Type 2 errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels.” The reasoning here is that a design that has 70% power for the smallest effect size of interest would not balance the Type 1 and Type 2 error rates in a sensible manner. Similarly, in huge datasets it might be possible to achieve very high levels of power for all effect sizes that are still considered meaningful. If such a study has 99% power for effect sizes of interest, and thus a 1% Type 2 error rate, but uses a 5% alpha level, it also suffers from a lack of balance. This latter scenario is quite common in meta-analyses, where researchers by default use a 0.05 alpha level, while the meta-analysis typically have much higher power than 0.95 for all effect sizes of interest. In such cases, it seems sensible to lower the alpha level considerably.

Researchers can decide to either balance Type 1 and Type 2 error rates (e.g., setting both at 5%), or minimize the combined error rate. For any given sample size and effect size of interest there is an alpha level that minimizes the combined error rates. Because the chosen alpha level also influences the statistical power, and the Type 2 error rate is therefore dependent upon the Type 1 error rate, minimizing or balancing error rates requires an iterative procedure.

As an example, imagine researchers plan to perform a study which will be analyzed with an independent two-sided *t*-test. They initially plan to collect 50 participants per condition, and set their smallest effect size of interest to Cohen’s $d = 0.5$. They think a Type 1 error is just as costly as a Type 2 error, and believe H_0 is just

as likely to be true as H1. The combined error rate is minimized when they set alpha to 0.13 (see Figure 2, dotted line), which will give the study a Type 2 error rate of $\beta = 0.166$ to detect effects of $d = 0.5$. The combined error rate is 0.148, while it would have been 0.177 if the alpha level was set at 5%¹.

Instead of choosing an example that is more extreme, I think it is important to point out that a better justified alpha level not necessarily means a hugely different alpha level. The absolute benefits are more noticeable for study designs where Type 2 error rates are high when a 5% alpha level would be used, or where the Type 2 error is very small and the alpha level can be reduced from 0.05. Similarly, Figure 2 shows that if one were to use an alpha level of 0.005 following recommendations of Benjamin et al. (2018) when collecting a sample size of 100 per group to detect an effect of $d = 0.5$, the study would be much less efficient than it could be if a higher alpha level is chosen. The more important take-home message is that, perhaps counter-intuitively, decision making is sometimes slightly more efficient after *increasing* the alpha level from the default of 0.05.

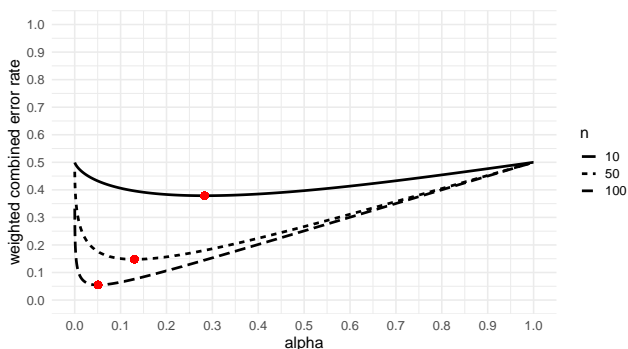


Figure 2. Weighted combined error rate (y-axis) for an independent t -test with $n = 10$, $n = 50$, and $n = 100$ per group and a smallest effect of interest of $d = 0.5$, for all possible alpha levels (x-axis).

Weighing the Relative Cost of Errors. Cohen (1988) considered a study design with a 5% Type 1 error rate and a 20% Type 2 error rate balanced. The reason for this was that instead of weighing both types of errors equally, he felt “Type I errors are of the order of four times as serious as Type II errors.” To determine the relative costs of Type 1 and Type 2 errors researchers should perform a cost-benefit analysis. As an example from another discipline, Field, Tyre, Jonzén, Rhodes, and Possingham (2004) quantify the relative costs of Type 1 errors when testing whether native species in Australia are declining. They come to the conclusion that when it comes to the Koala population, given its great economic value, a cost-benefit analysis indicates

the alpha level should be set to 1. In other words, one should always act as if the population is declining, because the relative cost of a Type 2 error compared to a Type 1 error is extremely large.

Although it can be difficult to formally quantify all factors that determine the costs of Type 1 and Type 2 errors, there is no reason to let the perfect be the enemy of the good. In research questions where there are no direct applications of the research findings, relative costs might be largely subjective. If you have a research strategy where you always follow up on an initial study testing a theoretical prediction with a replication and extension study, and want to innovate, you might want to weigh a Type 2 error more strongly than a Type 1 error. If you are relatively sure policies will be changed based on the outcome of your single study, or it is unlikely many researchers will have the resources to replicate your study, a Type 1 error rate might be weighed more strongly than a Type 2 error. There are no right or wrong answers, but you need to think through this question when you design a study.

If we adapt our calculations for the example above where researchers who plan to collect 50 participants per condition to detect a $d = 0.5$ effect, but now weigh the cost of Type 1 errors 4 times as much as Type 2 errors, balanced error rates are $\alpha = 0.0654$ and $\beta = 0.262$. You will notice this is exactly the scenario Cohen describes, where a 5% Type 1 error rate and 20% Type 2 error rate is deemed a balanced design because Type 1 errors are weighed 4 times as much as Type 2 errors.

One could also aim to minimize error rates in the latter scenario by setting alpha to 0.0387 and beta to 0.343. Remember the cost of a Type 1 error is twice as large as the cost of a Type 2 error. Imagine a Type 1 error has a cost of 400 and a Type 2 error has a cost of 100. We perform 100 studies, 50 where H0 is true and 50 where H1 is true. We will make 1.93 Type 1 errors and 17.14 Type 2 errors, with a respective cost of 773.6 and 1714, respectively. Increasing or decreasing the alpha level for this design will only increase the total cost of errors, assuming H0 and H1 are equally probable. Figure 3 visualizes the weighted combined error rate for this study design across the all possible alpha levels.

Incorporating prior probabilities. Miller and Ulrich (2019) explain how the choice for an optimal alpha level depends not just on the relative costs of Type 1 and Type 2 errors, but also on the base rate of true effects. In the extreme case, all studies a researcher designs test true hypotheses. In this case, there is no reason to worry

¹For the same scenario, balanced error rates are $\alpha = 0.149$ and $\beta = 0.149$.

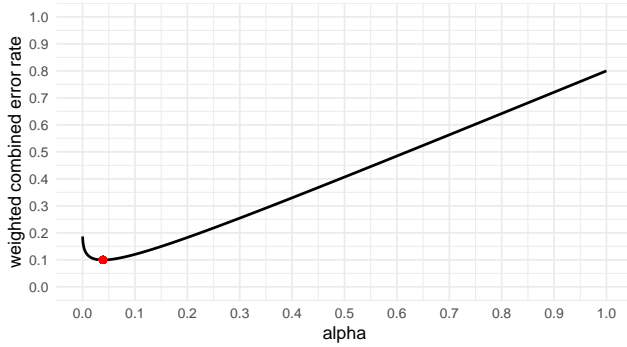


Figure 3. Weighted combined error rate (y-axis) for an independent t -test with $n = 50$ per group and a smallest effect of interest of $d = 0.5$, where Type 1 errors are weighed 4 times as much as Type 2 errors, for all possible alpha levels (x-axis).

about Type 1 errors, because a Type 1 error can only happen when the null hypothesis is true. Therefore, you can set the alpha level to 1 without any negative consequences. If the base rate of true hypotheses is very low, in the long run you will make many more Type 1 errors than Type 2 errors. For example, if you perform 1000 studies, and the base rate of true effects is 10%, with 5% Type 1 error rate and a 5% Type 2 error rate you will observe $1000 \times 0.9 \times 0.05 = 45$ Type 1 errors, and $1000 \times 0.1 \times 0.05 = 5$ Type 2 errors. If possible, researchers can take their expectations about the long run frequency of Type 1 and Type 2 errors into account when choosing an alpha level that either balances or minimizes the error rates.

If you want to balance or minimize error rates in the long run, you should lower the alpha level as the prior probability of a null effect increases, or increase the alpha level as the prior probability of a true effect increases. Because the base rate of true hypotheses is unknown, this step requires subjective judgment. This can not be avoided, because one always makes assumptions about base rates, even if the assumption is that a hypothesis is equally likely to be true as false (with both H_1 and H_0 having a 50% probability). Assuming equal prior probabilities for H_1 and H_0 , we saw above that balanced error rates assuming $d = 0.5$ and collection $n = 50$ per condition in a t -test would imply $\alpha = 0.149$ and $\beta = 0.149$. If you assume the null hypothesis is 10 times more likely than the alternative hypothesis (or the alternative hypothesis is 0.1 times as likely as the null hypothesis) then balanced error rates would require setting alpha to 0.0356 and beta to 0.355. If you believe the alternative hypothesis is twice as likely to be true as the null hypothesis, balancing error rates in

the long run would mean increasing the alpha level and increasing the power by choosing $\alpha = 0.213$ and $\beta = 0.107$.

The two approaches (balancing error rates or minimizing error rates) typically yield quite similar results. Where minimizing error rates might be slightly more efficient, balancing error rates might be slightly more intuitive (especially when the prior probability of H_0 and H_1 are equal). Note that although there is always an optimal choice of the alpha level, there is always a range of values that yield quite similar weighted error rates. It is useful to plot weighted error rates as in Figure 3. To balance or minimize error rates, researchers need to carefully consider the relative cost of Type 1 and Type 2 errors, the prior probability the null hypothesis is true, and the effect size they want to detect (Mudge, Baker, Edge, & Houlahan, 2012), because these factors are used to calculate the weighted combined error rates w :

$$w = \frac{(cost_{T1T2} \times \alpha + prior_{H1H0} \times \beta)}{prior_{H1H0} + cost_{T1T2}} \quad (1)$$

For example, imagine Type 1 errors are weighted 4 times as much as Type 2 errors ($cost_{T1T2} = 4$) and the alternative hypothesis is believed to be 2 times as likely as the null hypothesis ($prior_{H1H0} = 10$). With $\alpha = 0.213$ and $\beta = 0.107$ the weighted combined error rate is 0.142. **Sample Size Justification when Minimizing/Balancing Error Rates.** So far we have only considered cases, where researchers wanted to justify their alpha level on a fixed sample size. However, in practice researchers should justify their sample size before running a study. The approach of balancing or minimizing error rates can straightforwardly be generalized to a sample size justification. To do so researchers need to specify the weighted combined error rate they are aiming for. The required sample size, significance level, and power can be determined accordingly.

Let us illustrate this using the example of researchers studying an effect of $d = 0.5$ with a two sample t -test. We can estimate the sample size as well as the alpha and power they would require to achieve a weighted combined error rate of 5% when weighting Type 1 errors and Type 2 errors equally strong. Figure 4 shows alpha, beta, and the weighted combined error rate as a function of sample size when the alpha level is minimized. We can see that a weighted combined error rate of 5% (red line) is achieved at a sample size of 105. By iteratively optimizing the alpha level for different sample sizes and stopping when the desired error rate is reached, we can justify our sample size and the alpha level at the same

time.

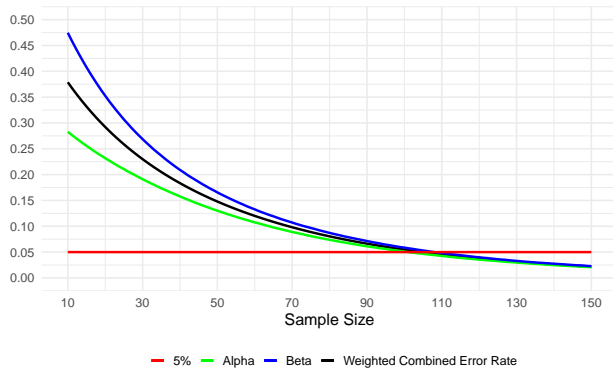


Figure 4. Weighted combined error rate (black), alpha (green), and beta (blue) for an independent t -test as a function of sample size when the alpha level is justified to achieve lowest errors at each sample size.

Lowering the Alpha Level as a Function of the Sample Size

Sometimes a researcher might feel it is too challenging to specify relative costs, prior probabilities, and the effect size of interest, and be tempted to fall back to the use of an alpha level of 0.05. For such cases an approach exists where the alpha level is lowered as a function of the sample size. The need to do so is discussed extensively by Leamer (1978). He writes “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.” This was already recognized by Harold Jeffreys (1939), who discusses ways to set the alpha level in the Neyman-Pearson approach to statistics: “We should therefore get the best result, with any distribution of alpha, by some form that makes the ratio of the critical value to the standard error increase with n . It appears then that whatever the distribution may be, the use of a fixed P limit cannot be the one that will make the smallest number of mistakes.” Good (1992) notes: “we have empirical evidence that sensible P values are related to weights of evidence and, therefore, that P values are not entirely without merit. 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.”

To understand these recommendations it is important to distinguish between statistical inferences based on error control and inferences based on likelihoods. An alpha level of 5% will limit incorrect decisions to a desired maximum (in the long run, and when all test assump-

tions are met). However, from a likelihood perspective it is possible that the observed data is much more likely when the null hypothesis is true, than when the alternative hypothesis is true, even when the observed p -value is smaller than 0.05. This situation, known as Lindley’s paradox, is visualized in Figure 1. It emerges because the critical value of a frequentist test approaches a limit as the sample size increases (e.g., $t = 1.96$ for a two-sided t -test). In Bayesian statistics the same Bayes factor requires a larger test statistic when the sample size is larger (see Jeffrey N. Rouder, Speckman, Sun, Morey, and Iverson (2009a) and Zellner (1971)).

To prevent Lindley’s paradox when using frequentist statistics one would need to adjust the alpha level in a way that the likelihood ratio (also called the Bayes factor) at the critical test statistic is always one. With such an alpha level, a significant p -value will always be as least as likely if the alternative is true than if the null is true, successfully avoiding the Lindley paradox. Faulkenberry (2019) and Jeffrey N. Rouder, Speckman, Sun, Morey, and Iverson (2009b) developed Bayes factors for t -tests and ANOVAs which can calculate the Bayes factor from the test statistic and degrees of freedom. Based on these papers we developed a shiny app (link) to reduce the alpha level proportional to the sample size and avoid Lindley’s paradox. However, it can be considered even paradoxical to reject the null hypothesis even so the evidence for the alternative from a Bayesian point of view is only weak as the case with a likelihood ratio of one. To reliably avoid false positives one would need a tool in which a significant p -value corresponds to moderate or strong evidence from a likelihoodist or Bayesian perspective. As a rule of thumb, if the data is 3 to 10 times more likely under the alternative than under the null this is considered moderate evidence, and data more than 10 times more likely under the alternative is considered strong evidence (H. Jeffreys, 1939, Appendix 1; Lee & Wagenmakers, 2013, p. 105). Therefore, we extended the shiny app to also allow researchers to adjust the alpha level in a way that a significant p -value will always provide moderate or strong evidence against the null hypothesis.

Our calculations are based on a unit information prior.² This prior is relatively wide which sometimes results in overstated evidence for H_0 . However, while it is usually wise to use more informed priors (e.g., McElreath, 2020), we believe that to avoid the Lindley paradox a wide prior is the more conservative choice. In other words, it will prevent overstating the evidence for H_1 and consequently be more likely to use an alpha level

²For t -tests the shiny app also allows to use the more common Cauchy prior.

that is too low, rather than an alpha level that is too high. In addition, in contrast to previous approaches that aimed to calculate a Bayes factor for every p -value (Colquhoun, 2017; e.g., D. Colquhoun, 2019) our application does not require researchers to specify the effect size under the alternative. Although the unit information prior is still an expectation for H1, not needing to specify the prior directly will likely increase the adoption among empirical researchers who sometimes find it difficult to specify a smallest effect size of interest or an expected effect size.

Let us illustrate this using one example study from the Deryl Bem's infamous paper on "feeling the future" (Bem, 2011). In this paper Deryl Bem claimed in a series of 9 studies that he found evidence for a phenomenon called *psi* which allows humans to predict the future. The publication of this implausible result in a well-known psychological journal, has sparked wide spread discussion about the replicability and data analysis practices in psychology Galak, LeBoeuf, Nelson, & Simmons (2012).

In study six of his paper, Bem reports the following one sample t -tests $t(149) = 1.80$, $p = .037$, $d = 0.15$, $p = .041$ and $t(149) = 1.77$, $p = .039$, $p = .041$, as evidence for "retroactive habituation". It seems like the p -values in this study are remarkably close to the 5% threshold, especially given the high sample size. Therefore, let us consider which alpha level Bem would have needed to use to avoid Lindley's Paradox, and to achieve moderate or strong evidence.

To avoid Lindley's paradox, Bem would need to use an alpha level of 0.0274. In other words, the observed p -value is more likely under the null hypothesis than under the alternative. However, it seems desirable to require at least moderate evidence for the alternative before claiming a discovery. To achieve this Bem would have needed to use a alpha level of 0.0077. To provide strong evidence given a significant p -value, he would have needed to use an even lower alpha level of 0.0020. We can see from this example, that $p = 0.05$ is too low of an evidential standard under large sample sizes.

Although in principle for small sample sizes the function could return an alpha level *larger* than 0.05, this is not the intended use. The reason for this is that we do not aim to replace the frequentist error rates with likelihood ratios but merely to reliably identify the cases where both approaches to statistical inference come to the same conclusions.

Comparison of The Two Approaches

When should we minimize or balance error rates and when should we avoid the Lindley paradox? In practice, it might be most convenient to use the first approach, whenever there is enough information to conduct power analysis, and if not to fall back on the second approach. Alternatively, one can also apply both approaches and choose the one that is more conservative (i.e., results in a lower alpha level). In addition, researchers aiming to find a compromise between Frequentist and Bayesian inference might be drawn more strongly towards the second approach (e.g., Good, 1992).

Figure 5 shows which alpha level is larger as a function of sample size (per group) and expected effect size for a two-sample t -test. For green values minimizing error rates resulted in higher alpha values, whereas red areas mean that avoiding the Lindley paradox resulted in higher alpha levels. We can see that when the expected effect size increases (x-axis) the alpha level will be smaller when using the method that minimizes error rates. This is mostly a result of how we set up the likelihood ratio calculations. Due to the unit information prior the method trying to avoid the Lindley paradox is not sensitive to the expected effect size. If the sample size increases on the other hand (y-axis) the alpha level reduces faster when avoiding the Lindley paradox then when minimizing error rates.

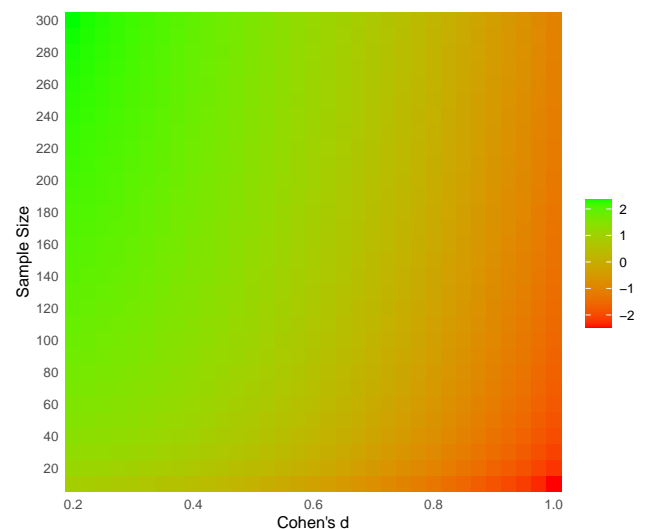


Figure 5. Heatmap of $\log(\alpha_{\text{minimize}}/\alpha_{\text{lindley}})$ as a function of sample size and expected effect size.

Discussion

Editors and reviewers should *always* ask authors to justify their choice of error rates, whenever researchers use data to make decisions about the presence or absence of an effect. As Skipper, Guenther, and Nass (1967) remarks: "If, in contrast with present policy, it were conventional that editorial readers for professional journals routinely asked: 'What justification is there for this level of significance?' authors might be less likely to indiscriminately select an alpha level from the field of popular eligibles." In a Neyman-Pearson approach to statistics, the alpha level should be set before the data is collected. When reviewers of for example a Registered Report would ask authors to justify their alpha level, it would be convenient if they can recommend some approaches to do so. The current manuscript hopefully helps to fill this gap.

If a power analysis can be performed (i.e., whenever a desired or expected effect size of interest can be specified) and relative costs and priors can be specified, researchers can design efficient by minimizing or balancing error rates, optimizing informational value, or aiming for specific posterior probabilities. If it is difficult to specify these aspects of the study, researchers can fall back to the approach where the alpha level is reduced as a function of the sample size.

A Shiny app is available that allows users to perform the calculations recommended in this article. It can be used to minimize or balance alpha and beta, which works by specifying the effect size of interest and the sample size, as well as an analytic power calculation. The effect size should be determined as in a normal a-priori power analysis (preferably based on the smallest effect size of interest, for recommendations, see Albers and Lakens (2018)). Alternatively, researchers lower the alpha level as a function of the sample size by specifying only their sample size. Whichever approach is used, it is strongly recommended to preregister the alpha level that researchers plan to use before the data is collected.

Because of the strong overreliance on a 5% error rate when designing studies, we have seen relatively few people attempt to justify their alpha level. As researchers start to justify their alpha, we will hopefully see the development of good examples and best practices for psychological science. It might be a challenge to get started, but the two approaches presented in the current article are one way to move away from the mindless use of a 5% alpha level. There is a lot to gain, and justifying alpha levels should improve our statistical inferences and increase the efficiency of the research we perform.

References

- Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, 74, 187–195. <https://doi.org/10.1016/j.jesp.2017.09.004>
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423–437.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. <https://doi.org/10.1038/S41562-017-0189-Z>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.
- Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12), 171085.
- Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, 73(sup1), 192–201.
- Cousins, R. D. (2017). The JeffreysLindley paradox and discovery criteria in high energy physics. *Synthese*, 194(2), 395–432. <https://doi.org/10.1007/s11229-014-0525-z>
- Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5), 553.
- Cumming, G. (2008). Replication and *p* Intervals: *P* Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, 3(4), 286–300. <https://doi.org/10.1111/j.1745-6924.2008.00079.x>
- 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.
- Field, S. A., Tyre, A. J., Jonzén, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, 7(8), 669–675.

- Fisher, R. A. (1935). *The design of experiments*. Oliver And Boyd; Edinburgh; London.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103(6), 933. Retrieved from <https://doi.org/10.1037/a0029709>
- Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, 2515245918771329. <https://doi.org/10.1177/2515245918771329>
- Good, I. J. (1992). The Bayes-Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, 87(419), 597. <https://doi.org/10.2307/2290192>
- Jeffreys, Harold. (1939). *Theory of probability* (1st ed). Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press.
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.
- Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. *The American Statistician*, 73(sup1), 82–90. <https://doi.org/10.1080/00031305.2018.1537891>
- Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1 edition). New York usw.: Wiley.
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187–192.
- McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC press.
- Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, 14(1), e0208631. <https://doi.org/10.1371/journal.pone.0208631>
- Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, 7(2), e32734. <https://doi.org/10.1371/journal.pone.0032734>
- Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706), 289–337. <https://doi.org/10.1098/rsta.1933.0009>
- Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive facilitation of recall' effect. *PloS One*, 7(3), e33423. Retrieved from <https://doi.org/10.1371/journal.pone.0033423>
- Rouder, Jeffrey N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009a). 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>
- Rouder, Jeffrey N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009b). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237.
- Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The Sacredness of .05: A Note concerning the Uses of Statistical Levels of Significance in Social Science. *The American Sociologist*, 2(1), 16–18.
- Tunç, D. U., Tunç, M. N., & Lakens, D. (2021). *The Epistemic and Pragmatic Function of Dichotomous Claims Based on Statistical Hypothesis Tests*. PsyArXiv. <https://doi.org/10.31234/osf.io/af9by>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). *Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011)*. Retrieved from <http://dx.doi.org/10.2139/ssrn.2001721>
- Winer, B. J. (1962). *Statistical principles in experimental design*. New York : McGraw-Hill.
- Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.