

Justify Your Alpha: A Practical Guide

Daniël Lakens¹

¹ Eindhoven University of Technology, The Netherlands

Justify Everything.

Keywords: power analysis, ANOVA, hypothesis test, sample size justification, repeated measures

Word count: 4654 words.

In a Neyman-Pearson approach to hypothesis testing researchers rely on data to decide how to act. Studies are designed such that the rate at which erroneous decisions about how to act are controlled in the long run. If resources were infinite, we could reduce error rates by collecting so much data that the chance of making a wrong decision is incredibly small. But resources are often limited, which means that researchers need to decide how to choose their error rates. 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). For any fixed sample size and true effect size 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 justify the factors in their study that determine their error rates, such as the alpha level, the desired statistical power, and their sample size. One reason researchers predominantly rely on norms instead of rationally chosen error rates is the absence of explanations how to justify a different alpha level than 0.05. In this article I will explain why error rates need to be justified, and will provide two practical guidelines that can be used to justify the alpha level based on the sample size, or to balance or minimize the type 1 and Type 2 error rate.

Why error rates should be justified

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.” Despite this good advice, social scientists developed the norm to always use an alpha level of 0.05 as a significance threshold

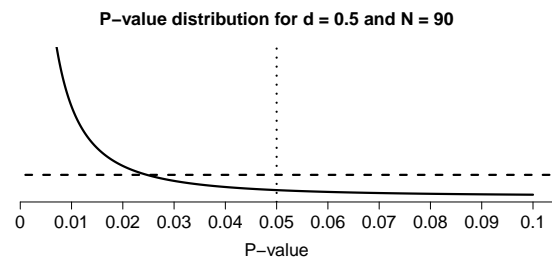


Figure 1. *P*-value distributions for a two-sided independent *t*-test with $N = 150$ and $d = 0.5$ (black curve) or $d = 0$ (horizontal dashed line).

when testing hypotheses. This mindless use of null hypothesis significance test has been criticized for more than half a century (Bakan, 1966; Gigerenzer, 2018), but there have been few attempts to provide researchers with practical guidelines to do anything else than blindly following norms. 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 is that as the statistical power increases, *p*-values below 0.05 can be more likely when there is no effect than when there is an effect. This is known as Lindley’s paradox (see Lindley (1957) Cousins (2017)). The distribution of *p*-values is a function of the statistical power. When the statistical power is extremely high, not only will most *p*-values fall below 0.05, most will also fall below 0.01. When there is no true effect *p*-values are uniformly distributed, and 1% of observed *p*-values fall between 0.04 and 0.05. In Figure 1 we see that with high power small *p*-values are more likely to be observed when there is an effect than when there is no effect (the black curve falls about the dashed horizontal line). But observing a *p*-value of 0.04 is more likely when the null hypothesis is true than when the alternative hypothesis is true and we have very high power (the horizontal dashed line falls above the black curve for *p*-values larger than ~0.025).

Researchers often want to interpret a significant test result as “support” for the alternative hypothesis. If so, it makes sense to choose the alpha level such that when a significant *p*-value is found, the *p*-value is actually more likely when the

All code used to create this manuscript is provided in an OSF repository at <https://osf.io/xxxxx/>.

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

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 large), the alpha level should be reduced. If we would lower the alpha level in Figure 1 to 0.02 then the for all observed significant p -values the alternative hypothesis is more likely than the null hypothesis.

The second reason to carefully choose an alpha level is that it makes decision making based on data more efficient. If researchers use hypothesis tests to make decisions about how to act while balancing error rates, then it is typically possible to make decisions more efficiently by choosing a different alpha level 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 costs of both types of errors.

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

As a young scholar 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) 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. Fischer 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 used, instead of the notion 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 idea that a deviation of two standard deviations (a z -score of 2) the early work of Gosset on the t -distribution (Cowles & Davis, 1982).

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 b 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 α of .05. The b 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.”

Ironically, we see that the norm to aim for 80% power is built on the norm to set the alpha level at 5%. If we ignore the slightly depressing practice of building conventions upon conventions, at the core the approach by Cohen highlights the real decision researchers need to make, namely to determine the relative seriousness of Type 1 to Type 2 errors, and then balancing both types of errors when designing a study. If a Type 1 error is weighted four times as strongly as a Type 2 error, the *weighted* error rates in the study are balanced. Again instead of imitating the value chosen by Cohen, researchers should aim to imitate the approach he used to determine the error rates.

Lowering the Alpha Level as a Function of the Sample Size

The idea behind this recommendation is discussed most 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 Jeffreys (1939), who discusses ways to set the alpha level in Neyman-Pearson 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.”

To understand this recommendation it is important to distinguish between statistical inferences based on error control and inferences based on likelihoods. When the alpha level is set to 5% a researcher will not conclude there is an effect when there is no true effect more than 5% of the time (in the long run, and when all test assumptions hold). 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. As explained above, this is known as Lindley's paradox. Compared to frequentist hypothesis tests, where the critical value approaches a limit as the sample size increases (i.e., a critical value of 1.96 for $p = 0.05$ in a two-sided test), the critical value for a Bayes factor increases as the sample size increases (see Rouder, Speckman, Sun, Morey, and Iverson (2009) and Zellner (1971)).

To prevent Lindley's paradox one would need to lower the alpha level as a function of the statistical power. 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." Based on the observation by Jeffrey's (1939) that, under specific circumstances, the Bayes factor against the null hypothesis is approximately inversely proportional to (\sqrt{N}) , Good (1982) suggests a standardized p -value to bring p -values in closer relationship with weights of evidence:

$$p_{stan} = p \sqrt{\frac{N}{100}} \quad (1)$$

where p is the observed p -value and N is the total sample size. This formula standardizes the p -value to the evidence against the null hypothesis that would be observed if p_{stan} was the tail area probability observed in a sample of 100 participants. It makes things easier to think about a standardized alpha level:

$$\alpha_{stan} = \frac{\alpha}{\sqrt{\frac{N}{100}}} \quad (2)$$

With 100 participants α and α_{stan} are the same, but as the sample size increases beyond 100 the alpha level becomes smaller. For example, an $\alpha = 0.05$ for a sample size of 500 would become an α_{stan} of 0.022.

As mentioned by Rouder et al. (2009): "NP testing can be made consistent by allowing Type I error rates to decrease toward zero as the sample size increases. How this rate should decrease with sample size, however, is neither obvious nor stipulated by the statistical theory underlying NP testing." There are many ways to calibrate p -values so that they are more in line with Bayes factors (see Sellke, Bayarri, and Berger (2001) and Benjamin et al. (2018)), but not everyone thinks such calibrations are sensible (Senn, 2001). Regardless, reducing the alpha level as a function off the sample size should be seen as a quick fix for researchers who are willing to prevent Lindley's paradox, but who for either philosophical or practical reasons do not want to use likelihoods or Bayes factors. However, there is a second approach to justifying alpha levels that is more consistent with the Neyman-Pearson philosophy, namely to minimize or balance Type 1 and Type 2 error rates.

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 would make decision making overall most efficient. Researchers can choose to design a study with a statistical power and alpha level that minimize the *combined error rate*. For example, with a 5% alpha level and a statistical power of 80% the combined error rate is $5 + 20 = 25\%$. If the statistical power is 99% and the alpha is 5% the combined error rate is $1 + 5 = 6\%$. 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 if desired beliefs about the prior probability that H_0 and H_1 are true, decision making can be more efficient than when relying on the default alpha level of 0.05.

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, and a study design which has 99% power but uses a 5% alpha level similarly suffers from a lack of balance. A common example is a meta-analysis, which often have extremely high power for any effect size that would be considered meaningful, but where researchers often use a 5% Type 1 error rate anyway.

Researchers can decide to either balance Type 1 and Type 2 error rates (e.g., setting both at 5%), or minimize error rates. For any given sample size and effect size one wants to detect, there is an alpha level that minimizes the combined error rates. Because the alpha level also influences the statistical power and thus the Type 2 error rate and iterative procedure is needed that determines the optimal value for the alpha level. For example, for a study which will be analyzed with an independent two-sided t-test with $n = 64$ per condition, and an effect size that the researchers wants to detect of $d = 0.5$, the combined error rate is minimized when $\alpha = 0.0997$ and the Type 2 error rate $\beta = 0.122$. For the same scenario, balanced error rates are $\alpha = 0.111$ and $\beta = 0.111$.

Weighing the relative cost of errors. Cohen (1988) considered a Type 1 error rate of 5% and a 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. 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, and in many research questions where there are no direct applications of the research findings, relative costs might be largely subjective.

If we adapt our calculations for the *t*-test above, with $d = 0.5$, $n = 64$, but now weigh the cost of Type 1 errors 4 times as much as Type 2 errors, balanced error rates are $\alpha = 0.0498$ and $\beta = 0.199$. You personally might believe the cost of a Type 1 error is only twice as severe as a Type 2 error, in which case balanced error rates would mean designing a study with $\alpha = 0.0754$ and $\beta = \text{r format(res4$beta, digits = 3, nsmall = 2)}$.

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, if all studies a researcher designs test true hypotheses there is no reason to worry about Type 1 errors (because the null hypothesis is never true, you can never conclude there is an effect when there is no effect) and you can set the alpha level to 1. 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 $1000 \times 0.1 \times 0.05 = 5$ Type 2 errors, and $1000 \times 0.9 \times 0.05 = 45$ Type 1 errors. The long run frequency of Type 1 errors and Type 2 errors is not balanced in this example.

If you want to balance or minimize error rates in the long run, you should lower the alpha level as a function of the base rate of true hypotheses. 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 this assumption is that a hypothesis is equally likely to be true as false (a ratio of $H1/H0$ of $1/1 = 1$). Assuming equal prior probabilities for $H1$ and $H0$, we saw above that balanced error rates assuming $d = 0.5$ and collection $n = 64$ per condition in a *t*-test would imply $\alpha = 0.111$ and $\beta = 0.111$. If you assume it is ten times as likely that the null hypothesis is true than that the alternative hypothesis is true (a ratio of $H1/H0$ of $1/10 = 0.1$) balanced error rates would require setting $\alpha = 0.0275$ and $\beta = 0.275$. If you believe the alternative hypothesis is twice as likely to be true than the null hypothesis, balancing error rates in the long run would mean increasing the alpha level and increasing the power by choosing $\alpha = 0.159$ and $\beta = 0.0795$.

To minimize the combined error rates Equation (3) needs to be minimize the α and β , which requires an iterative procedure since the power, or $1 - \beta$, also depends on the α level.

When balancing error rates the difference between the α and β is minimized in an iterative procedure. Both approaches 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 $H0$ and $H1$ are equal).

$$\frac{(cost_{T1T2} \times \alpha + prior_{H1H0} \times \beta)}{prior_{H1H0} + 1} \quad (3)$$

I have created an R function and shiny app to minimize α and β (or their difference) which works by specifying the effect size and sample size, as well as an analytic power calculation. The effect size should be determined as in a normal power analysis (preferably based on the smallest effect size of interest), and the sample size can be increased until the error rates are deemed acceptable given the available resources.

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." Reviewers and editors who embrace this recommendation should expect some annoyance. It is often confronting to be asked to justify a norm you have taken for granted your entire life, only to realize you don't have any justification.

If a power analysis can be performed (i.e., whenever a desired or expected effect size of interest can be specified) researchers could do much worse than to design a study such that the combined error rate is minimized or balanced. If it is difficult to specify an effect size of interest (and thus to perform an a-priori power analysis and calculate a Type 2 error rate) researchers can fall back to the approach where the alpha level is reduced as a function of the sample size.

A benefit of lowering the alpha level as a function of the sample size as suggested by Good (1982) is that it is easy and requires no assumptions about prior probabilities or relative costs of different types of errors, and it largely prevents Lindley's paradox, which is one serious concern about the normative use of a 5% alpha level. A downside of the approach is that the approach is not very principled, in that it only roughly tracks the evidence as provided by a Bayes factor with an informed prior, and the standardization against a sample size of 100 is rather random. The benefit of balancing

or minimizing error rates is that it fits perfectly with Neyman-Pearson approaches to statistical inferences. A downside of this approach is that it is difficult to choose the relative costs of Type 1 and Type 2 errors (although it no reasons can be given to weigh them differently, equal weights is a valid option) and that the prior probability of examining a real effect is unknown, but has to be chosen (and although Mudge and colleagues suggest equal prior for H_0 and H_1 are a valid default, this is more debatable).

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. Examples in other research domains exist. For example, 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, 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 1 error compared to a Type 1 error is extremely large. As researchers start to justify their alpha, we will hopefully see the development of similar examples and best practices in psychological science. It might be a challenge to get started, but there is a lot to gain, both in improving our statistical practices, as in the efficiency of the research we perform.

References

- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423–437.
- 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. doi:10.1038/S41562-017-0189-Z
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.
- Cousins, R. D. (2017). The JeffreysLindley paradox and discovery criteria in high energy physics. *Synthese*, 194(2), 395–432. doi: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.
- 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.
- Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, 2515245918771329. doi:10.1177/2515245918771329
- Good, I. J. (1982). C140. Standardized tail-area probabilities. *Journal of Statistical Computation and Simulation*, 16(1), 65–66. doi:10.1080/00949658208810607
- Good, I. J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, 87(419), 597. doi:10.2307/2290192
- Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford [Oxfordshire] : New York: Clarendon Press ; Oxford University Press.
- Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1 edition.). New York usw.: Wiley.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187–192.
- Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, 14(1), e0208631. doi: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. doi: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. doi:10.1098/rsta.1933.0009
- 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. doi:10.3758/PBR.16.2.225
- 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.
- Senn, S. (2001). Two cheers for P-values? *Journal of Epidemiology and Biostatistics*, 6(2), 193–204.
- 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.

- 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.