

Rescuing null hypothesis significance testing (NHST) using optimal alpha.

TRADITIONAL NULL HYPOTHESIS SIGNIFICANCE TESTING

Null hypothesis significance testing (NHST) has had a long and conflicted history from its introduction in the early 1900's to its recent more disputed role (Anderson et al. 2000, Wasserstein and Lazar 2016, Benjamin et al. 2018). For decades, NHST was the dominant approach to inferential statistics across a variety of disciplines including physiology (Curran-Everett and Benos 2004, Ludbrook 2013), ecology (Yoccoz 1991, Stephens et al 2007), psychology (Nickerson 2000), and human health (Stang et al. 2010) and, despite the increasing popularity of alternative approaches such as, information-theoretic techniques (Burnham and Anderson 2002), the use of effect sizes with associated confidence intervals (Gardner and Altman 1986) and Bayesian techniques (Hobbs and Hooten 2015), it continues to be widely used. However, NHST has been criticized and identified as contributing to the reproducibility crises in several disciplines (Aarts 2015, Baker 2016). Most recently, 800 scientists and statisticians signed a 'manifesto' calling for the concept of 'statistical significance' to be abandoned (Amrhein et al 2019).

The list of criticisms of NHST is long and varied (Johnson 1999, Anderson et al. 2000, Burnham and Anderson 2014, Levine et al.

2008, Dushoff et al. 2019, McShane et al 2019). Most criticisms rest on three fundamental properties of NHST and p-values – (1) that effect size is not an explicit consideration so that meaningful effects can be overlooked at small sample sizes and trivial effects can be determined to be statistically significant at large samples sizes, (2) that Type II errors are not considered and (3) that p-values don't provide what researchers want - error probability statistics (i.e. estimates of the probability of making a mistake when the null is rejected or not).

The ongoing debate about the use and relevance of p-values has been driven, in part, by the two competing schools of thought (i.e. Fisher versus Neyman-Pearson) on p-values – using them as inductive inference tools that provide evidence for or against a hypothesis or to set a threshold (usually $p < 0.05$) as a behavioral, decision-making tool (Mayo 2018) to reject the null or not (i.e. identify the evidence as 'statistically significant' or not). When scientists recommend abandoning 'statistical significance' they are opposed to the concept of a rejection threshold. But, the severest critics also oppose using p-values as evidence for or against hypotheses. However, practitioners have used and

continue to use p-values in a way that combines both philosophies (Kennedy-Shaffer 2019) – using p-values to make decisions about whether to reject or fail to reject a null hypothesis and as a measure of how convincing a rejection is (i.e. if the p-value is small then the null can be rejected with increased confidence).

Why 0.05?: There are two possible types of errors in NHST, rejecting the null when the null is true (i. e. Type I) or accepting the null hypothesis when it is false (i.e. Type II). The objective of setting a statistical threshold in NHST is to control the long-run probability of Type I errors. Choosing 0.05 as the threshold causes many, if not most, of the problems associated with traditional NHST. Fisher, Neyman, Pearson (and most of the rest of the world) considered an arbitrary statistical threshold to be a bad idea but they did not reject a considered significance threshold (Mayo 2018). So, while not all researchers are prepared to abandon the concept of statistical significance, there is a consensus that a ‘one-size-fits-all’ rejection threshold should be abandoned.

There have been many attempts to address criticisms of NHST and p-values. In fact, Neyman and Pearson (1933), in their seminal paper, recommended that the critical rejection regions should consider Type II error rates by only setting a rejection threshold for the ‘most powerful test’ (i.e. the alternative hypothesis that was most likely under the observed data). Savage (1961) noted the relationship between alpha and beta and suggested that if a researcher could identify a combination of alpha and

beta that they preferred, then one could select an optimal alpha for a particular sampling design. However, Savage did not discuss the grounds on which a preference might be based. Cornfield (1966) expanded on this and asserted that minimizing a linear function of alpha +beta was a logical consequence of adhering to the likelihood principle and noted that using a linear function meant that differences in cost of Type I and II errors could be incorporated. More recently, Gannon et al (2019) have suggested a similar goal of minimizing the combination of Type I and II error probabilities by controlling for effects of sample size on Type II error probabilities. In addition, there have been attempts to incorporate Type I and Type II errors and their relative costs into environmental decision making (Mapstone 1995, Munkittrick 2002, Field et al. 2005). There have also been suggestions for using a combination of p-value and effect size thresholds to decide when to reject the null (Goodman et al. 2019). Further, there have been other attempts to place p-values in the context of sample size and effect size (Betensky 2019, Fraser 2019) including ‘second generation’ null hypotheses, which set composite nulls that include no and trivial effects sizes (Blume et al. 2019).

However, none of these proposed an analytical method for combining Type I and II error probabilities to identify an ‘optimal’ alpha. Mudge et al.(2012) developed an approach to setting the statistical threshold for NHST – optimal alpha - that rather than controlling long-run probability of Type I errors, minimizes the cost or probability of making a Type I or II error for a specific test

(see explanation below). Optimal alpha is the rejection threshold at which researchers will make the fewest mistakes when they reject or fail to reject the null. So, optimal alpha appears solidly situated in the Neyman-Pearson school – designed to minimize error probability, albeit the error probability that researchers are most concerned about, the probability of making a mistake for a particular test. However, optimal alpha is actually a combination of the Fisherian and N-P philosophies. It is explicitly designed to address how to make the best decisions (behavior) but also provides a measure of how strongly you should favour the best decision (inference). ‘Optimal’ alpha provides a measure of the “severity” of the test - sensu Mayo - because the rejection threshold provides an estimate of the probability that the researcher is making an error when they reject or fail to reject the null. When the probability of error is small, the test is severe and when it is not, the test is weak.

OPTIMAL ALPHA

Understanding how optimal alpha is estimated is easier if we start with two simplifying assumptions, (1) that the costs of Type I and II errors are equal and (2) that the prior probabilities of the null and alternate hypotheses are equal. These assumptions are not necessary for estimating optimal alpha but they make the initial explanation easier to understand.

The objective of optimal alpha is to minimize the probability of making an error (either Type I or II). This implies that we are trying to identify the statistical threshold

where the probability of making an error is smallest. If we assume that the prior probability of $H_0 = H_A = 0.5$, then we want to minimize $(0.5 * \alpha + 0.5 * \beta) = (\alpha + \beta)/2$, where α is the probability of making a Type I error and β is the probability of making a Type II error. We multiply each of α and β by 0.5 because, in the simple example, we are assuming that the prior probability of each is 0.5.

There is an inevitable negative relationship between α and β (Figure 1). However, the shape of that relationship depends on the sampling design, the critical effect size and the type of statistical test used.

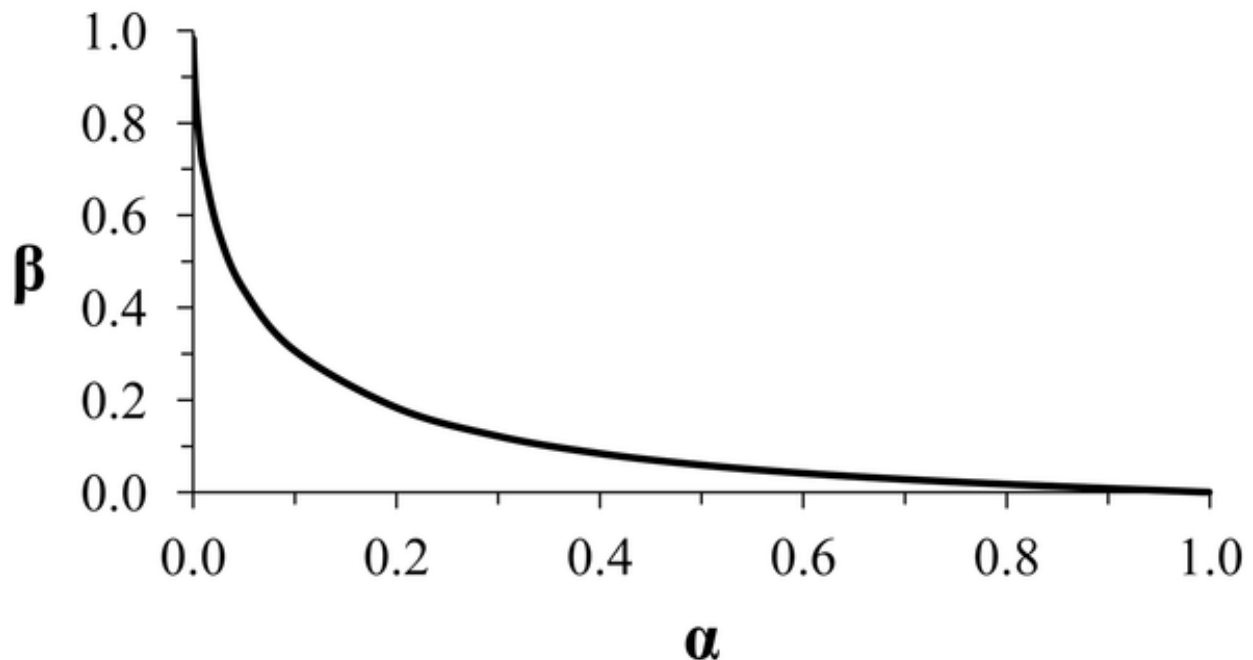


Figure 1: The non-linear relationship between α and β . The relationship between α and β for an independent 2-sample, 2-tailed t-test with $n_1 = n_2 = 10$ and critical effect size = 1σ

Estimating power or beta (i.e. $1 - \text{power}$) requires knowledge of the 1. statistical test, 2. sample size, 3. critical effect size, 4. variability unexplained by the alternative hypothesis model, and 5. alpha (Cohen 1988). However, if the statistical test, sample size, critical effect size and variability are known, it is a simple process to vary alpha and calculate the associated beta. Optimal alpha uses the proposed sample size, variability estimate and critical

effect size to build a curve describing the relationship between α and $(\alpha + \beta)/2$. We then identify the α value that minimizes $(\alpha + \beta)/2$ (Figure 2).

The result is that in using optimal alpha, we set a threshold that will minimize the probability of making a mistake when we choose to accept or reject the null. However, one key difficulty lies in identifying the critical effect size.

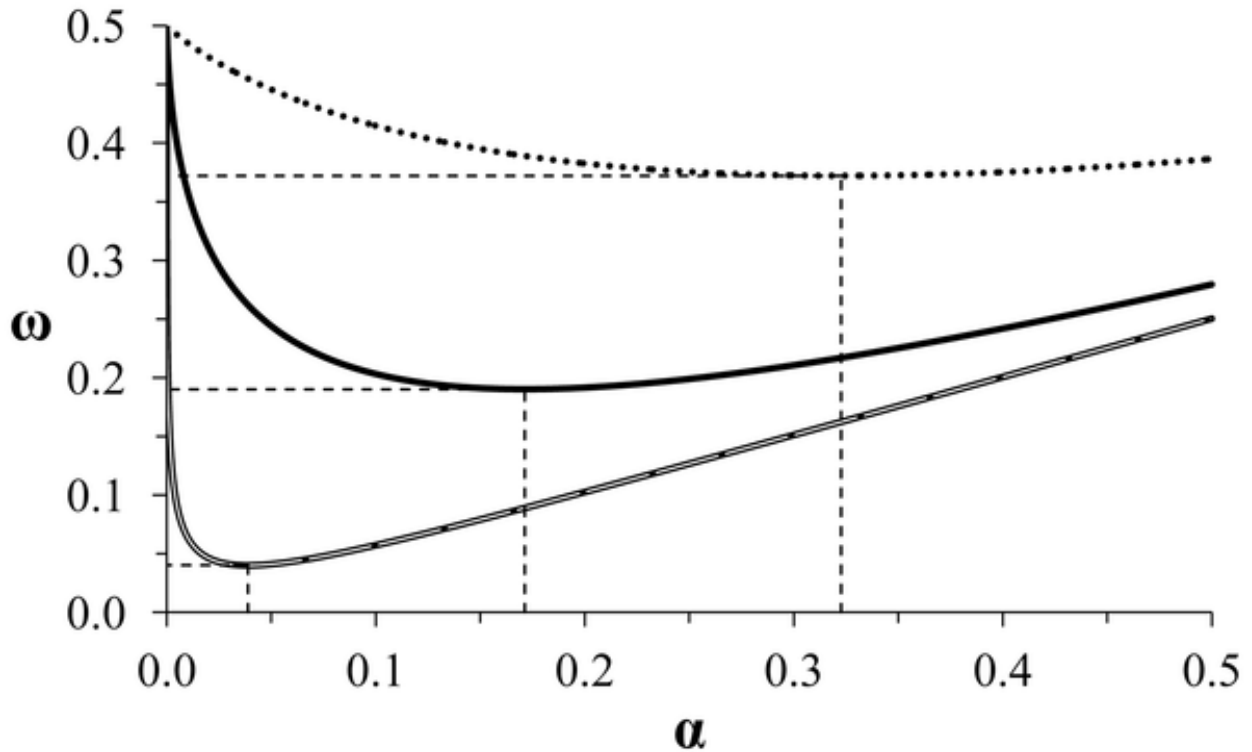


Figure 2: Determination of optimal α from the a priori combined probabilities of Type I and II error. α and ω for independent 2-tailed, 2 sample t-tests ($n_1 = n_2$). Data are for 3 (dotted line), 10 (solid line) and 30 (double line) samples per group, with critical effects sizes of 1σ where, $\omega = (\alpha + \beta)/2$ (i.e. average of Type I and Type II error). Drop lines indicate the minimum ω and its associated value for α .

Critical effect size: A t-test tests for a difference in mean values between two groups - the effect size is the difference in mean values between the two groups. An ANOVA tests for the ratio of among-group to within-group variability. Critical effect size is a more difficult concept than 'effect size' because it assumes that researchers can identify precisely the size of an effect that 'matters' (Munkittrick et al. 2009). That is, the size of an effect below which researchers are content to act as if there is no effect and, at or above which, they will act as if the effect is 'real'. The critical effect size is subjective and often in the eye of the beholder – the size of the effect of effluent on river fish populations that “matters” may be very different for recreational fisherman who use the river, operators of the company that release the effluent and people who use the product produced by the company that releases the effluent.

However, we should all consider critical effect sizes when null hypothesis significance testing. Traditional NHST allows us to ignore effect size and what we consider to be 'critical' effect sizes. By contrast, to estimate optimal alpha researchers must identify at least one critical effect size.

Our original advice was, if there is no objective rationale for a single critical effect size (and there rarely is) researchers should identify a small, a medium and a large effect size. The 'small' effect should be the effect below which it is difficult to imagine that any stakeholder would care. The 'large' effect size is the effect at which it is difficult to imagine that any stakeholder would

conclude it does not matter and the 'medium' effect size would be intermediate to the 'small' and 'large' effect size. Three critical effects sizes implies three optimal alpha values. This is still a viable approach but we now propose a dramatically different approach, which identifies the critical effect size for which the observed p-value would be the optimal alpha.

OPTIMAL ALPHA TO ESTIMATE EFFECT SIZE

Using optimal alpha in this way, researchers would no longer simply reject or fail to reject the null hypothesis, but instead would identify the critical effect size, based on an observed p-value, the test and the sample size, for which the best decision would be to reject the null hypothesis. Thus, the observed p-value would be interpreted as ‘evidence of an effect as large or smaller than the effect size for which the observed p-value was the optimal alpha’. This would be used to reject or fail to reject the null based on the effect size the researcher or reader believes is important.

For example, 0.79 SD is the critical effect size for a 2-tailed, 2-sample t-test with a sample size of 20 in each group if the observed p-value of 0.136 was the optimal alpha.

Thus, there is more evidence for an effect size of 0.79 SD’s or smaller than ‘no effect’. If the researcher or reader believes that 0.79 SD’s is biologically significant then they should reject the null hypothesis of no effect. If they believe the effect needs to be larger than 0.79 SD’s to be considered biologically significant, then they would fail to reject the null of no effect. So, optimal alpha could be used as Neyman and Pearson recommend – to make decisions, and meets the explicit objective of minimizing the probability/cost of making a mistake. More informally, it would be interpreted as “there is more evidence of an effect as large as 0.79 SD’s or smaller, than of no effect and more

evidence of no effect, than an effect larger than 0.79 SD’s.

For a second example, 2.2 SD’s as the critical effect size for a 2-tailed, 2-sample t-test with a sample size of 20 in each group if the observed p-value of 0.001 was the optimal alpha. This would be interpreted as “there is more evidence of an effect as large as 2.2 SD’s or smaller than of no effect and more evidence of no effect than an effect larger than 2.2 SD’s.

However, we can do better than simply identify the effect size at which to reject the null given the observed p-value – we can also make inferences about the strength of that evidence.

Characterizing evidence strength using optimal alpha: So, optimal alpha does more than simply provide the best decision-making approach because the absolute value of the optimal alpha can also be used to make inferences about the ‘severity’ of the test (Mayo 2018). Optimal alpha is estimated by identifying the statistical rejection threshold (i.e. alpha) at which the probability of making either a Type I or II error is minimized for a given critical effect size. This implies that when optimal alpha is estimated there is also an optimal beta (i.e. the probability of making a Type II error when the rejection threshold is set at the optimal alpha) and an optimal overall probability of error (i.e. (optimal α + optimal β)/2). Thus, when the observed p-value is \leq optimal α , the magnitude of optimal alpha provides an estimate of evidence strength. If

optimal alpha is very small and the observed p-value is less than optimal alpha then the evidence against the null is very strong. By contrast, if optimal alpha is very large and the observed p-value is less than optimal alpha the evidence is weak. Similarly, if the null is accepted and optimal beta is small, then the evidence for the null is strong, while if optimal beta is large, the evidence in support of the null is weak.

Thus, in the first example, we conclude that the evidence for an effect of 0.79 SD's or less is weak because optimal alpha = 0.136 and the evidence against an effect larger than 0.79 SD's is weak because optimal beta = 0.166. That is, if the null is true, the probability of rejecting the null is 0.136, which implies a relatively high probability of making a mistake, and if the alternate is larger than 0.79 SD's, the probability of accepting the null is 0.166, again a relatively high probability of making a mistake. By contrast, for the second example we have strong evidence of an effect of up to 2.2 SD's because optimal alpha = 0.001, which implies a small probability of rejecting the null if it is true and strong evidence that the effect is not larger than 2.2 because optimal beta = 0.0008, a small probability of accepting the null if the effect is larger than 2.2 SD's.

Relative cost assumptions: It isn't necessary to assume that Type I and II errors have equal costs. If users believe the costs of Type I and II errors are different and they are able to estimate the costs of Type I and II errors or, at least, the relative costs, the optimal alpha function can incorporate that information so as to minimize the cost rather than the probability of making an error. The optimal alpha function does this by weighting the errors. For example, if Type II errors are twice as costly as Type I errors the algorithm finds the statistical threshold that minimizes the function $(\alpha + 2\beta)/2$. If Type I errors are twice as costly as Type II errors the algorithm minimizes the function $(2\alpha + \beta)/2$.

Prior probability assumptions: Nor is it necessary to assume prior probabilities of $H_0 = H_A = 0.5$. Choosing prior probabilities of $H_0 = H_A = 0.5$ is analogous to choosing a vague or uninformative prior in Bayesian analysis. However, while optimal alpha is not explicitly Bayesian, it can incorporate Bayesian principles by making minor adjustments to the minimizing function. For example, if the prior probabilities of H_0 and H_A are estimated to be 0.3 and 0.7, respectively, the algorithm finds the statistical threshold that minimizes the function equal to $0.3\alpha + 0.7\beta$.

WHY OPTIMAL ALPHA?

Because optimal alpha is set in a simple Bayesian context – that is, estimating optimal alpha requires selecting prior probabilities of the null and alternate hypotheses - rejection thresholds are treated as error probabilities (and ARE error probabilities, if the priors are correct). Thus, researchers make decisions about rejecting or failing to reject the null based on a statistic that is directly relevant to the decision – the probability of making a mistake when they choose. Further, optimal alpha explicitly incorporates Type II errors because the objective of optimal alpha is to minimize the probability or cost of both errors. Optimal alpha is a function of sample size and will always be lower for strong than weak tests because the minimum probability of the combined errors will at a higher alpha for weak than strong tests. Optimal alpha also explicitly incorporates effects sizes because estimating the Type II error probability requires defining a critical effect size or sizes. So, when we fail to reject the null for a particular optimal alpha that optimal alpha is associated with a defined effect size. Similarly, when we reject the null for a particular optimal alpha the null is an implied composite null. By minimizing the probability of making a Type I or II error for a defined effect size we are implying that it is not a mistake to fail to reject the null when the effect is smaller than the defined critical effect size. We are in essence, saying that we should set the rejection threshold as if the null includes every effect from 0 up to, but not including,

the critical effect size. But, we must be cautious, here – the formal null hypothesis has not changed. We are still testing a null of zero effect against some alternate but the rejection threshold is being set to minimize the probability of making either error and failing to reject the null if the effect is smaller than the critical effect size will not be treated as an error.:

Would using optimal alpha change anything?: Optimal alpha addresses key criticisms of NHST but – does it matter? Is there any evidence that using optimal alpha would regularly result in different conclusions that using $\alpha=0.05$? This is really the key question – would using optimal alpha instead of traditional NHST change what ecologists believe about the world. Early pilot studies suggests using optimal alpha would change conclusions.

Pilot Study 1: Houlahan et al (2018) used linear regression to model the relationship between variability in community abundance and species richness (after controlling for mean abundance) for 91 different data sets. That is, $\log VS = \log SR + \log CA$. where VS is variance in community abundance, SR is species richness and CA is mean community abundance. We then estimated p-values for each of those 91 relationships using the statistical threshold of $\alpha=0.05$ and used an asterisk to denote which of the 91 studies achieved statistical significance. Here, we have returned to those studies, calculated optimal alpha for each of the 91 linear regressions and

identified how many times we would have reached a different conclusion than we reached in that paper.

We were unable to identify a single critical effect size and selected 3 effect sizes, small, medium and large, $r = 0.2$, $r=0.5$ and $r = 0.8$, respectively. For all optimal alpha calculations we assumed prior probabilities where $H_0 = H_A = 0.5$ and that Type I and II errors had equal costs.

We found that we would have reached a different conclusion using optimal alpha in 28.8% of tests overall – 30.7% for the small effect size, 29.6% for the medium and 26.3% for the large effect size. When we were trying to detect a small effect size optimal alpha rejected the null 24 times when $\alpha= 0.05$ would have accepted and accepted the null four times when $p<0.05$ would have rejected. By contrast, when we were trying to detect a large effect size, optimal alpha rejected the null when $\alpha= 0.05$

would have accepted three times and accepted the null when $p<0.05$ would have rejected, 21 times.

Pilot Study 2: Journal articles published between 2013 and 2018 were selected haphazardly from Ecology, the flagship journal of the Ecological Society of America. We retained only articles that tested null hypotheses using one of t-test, regression, and ANOVA analytical techniques, at a significance level of $\alpha=0.05$. To calculate optimal alpha we extracted sample size (N) for regression analyses, sample sizes (N_1 and N_2) and type of t-test (i.e. one, sample, two sample or paired and one or two-tailed) for t-tests, and the numerator and denominator degrees of freedom for ANOVA's. For all tests, we extracted the reported p-value. We calculated optimal alpha for one test from each of the 100 articles. We used different effects sizes for each of the tests (Table 1).

Table 1. Critical effect sizes for t-test, Regression, and ANOVA

Test	Critical effect size		
	Small	Medium	Large
Regression	$r=0.2$	$r=0.5$	$r=0.8$
t-test	$d=0.2$	$d=0.5$	$d=1.0$
ANOVA	$f^2=0.04$	$f^2=0.25$	$f^2=0.65$

For all optimal alpha calculations we assumed prior probabilities where $H_0 = H_A = 0.5$ and that Type I and II errors had equal costs. Conclusions reached using optimal alpha versus $p \leq 0.05$ conflicted in slightly more than 22% of tests (Table 2). When we were trying to detect a small effect size

optimal alpha rejected the null 21 times when $\alpha= 0.05$ would have accepted and accepted the null once when $\alpha= 0.05$ would have rejected. By contrast, when we were trying to detect a large effect size, optimal alpha rejected the null when $\alpha= 0.05$ would have accepted, three times and accepted the

null when $\alpha= 0.05$ would have rejected, 22 times.

Table 2. Number of times optimal α with its three critical effect sizes reached consistent and inconsistent conclusions compared to $\alpha =0.05$ ($N=100$).

Optimal α	Differences in decisions made between standard $\alpha =0.05$ and optimal α in percentages	
	Disagree	Agree
α_1 small effect size	22%	78%
α_2 medium effect size	20%	80%
α_3 large effect size	25%	75%

Summary: These preliminary pilot studies suggest that using optimal alpha rather than $\alpha= 0.05$ would result in different conclusions in about 22-28.5 % of tests of null hypotheses. We interpret this as meaning that in 22-28% of tests researchers made the wrong decision in accepting or rejecting the null. We are not suggesting that in 22 – 28% of tests the researchers did not identify the true hypothesis, because it is impossible to be certain at the time that the test is run which of the two hypotheses is true. However, we are suggesting that when there are two different approaches to decision-making and one has been clearly demonstrated to be superior, as is the case of optimal alpha versus NHST using $\alpha= 0.05$, then when the decision made using the inferior method is different than the decision that would have been made using the superior method, the researcher is making a mistake. They are making a mistake because they are not minimizing the probability of choosing the wrong hypothesis. Neither approach guarantees that we will make the right choice but it is always a mistake not to use

the approach that will give us the greatest probability of making the right choice.

Optimal alpha rescues NHST, but only when NHST is appropriate: Optimal alpha is useful in some contexts but barely mitigates the problem that NHST is a blunt tool. NHST makes sense in the early stages of attempting to explain a pattern or set of observations. As the science around a question matures, we should no longer be comparing and choosing between naïve, low information models of the world. We should be choosing among more plausible, high information, better tested models. As a discipline matures, fewer questions should be addressed with NHST because there are fewer questions in the very early stages of development. Using NHST for any particular question is a sign of an immature question and its common use on many questions is a sign of an immature discipline.

Optimal alpha is an index of sampling design quality: Each estimate of optimal alpha comes with an estimate of the optimal overall probability of making an error. Optimal alpha estimates the

statistical rejection threshold that will minimize the overall probability of making an error but simply because the minimum probability of error has been identified, doesn't mean that the probability of making an error is acceptably low. If the optimal overall probability of making an error is 0.47 it means that, for a given sampling design and a given critical effect size there is no way to reduce the probability of making an error below 0.47. This implies that the sampling design is not very good because the researcher will make the wrong decision 47% of the time. By contrast, if the optimal overall probability of making an error is 0.002 it means that the sampling design and chosen critical effect size allow us to choose an alpha that will reduce our probability of making a mistake to 0.2%. This implies a strong sampling design. So, the overall probability of making an error is an index of sampling design quality. There are two caveats. First, the prior probabilities must be approximately correct and second, critical effect sizes must be carefully considered.

Multiple comparisons corrections: Post-hoc multiple comparisons corrections are designed to control the probability of making Type I errors. However, they are ill-advised in all cases where beta and the relative costs of Type I and II error have not been explicitly considered. Multiple comparisons corrections from the most conservative like Bonferroni (Dunn 1961) to the most liberal, like the false discovery rate (Benjamini-Hochberg 1995) rarely, if ever, consider the effect that reductions in the test-wise alpha (say 0.05) to control for

the effects of multiple comparisons on the experiment-wide probability of Type I error, have on Type II error probabilities. However, we know that any reduction in Type I error probability will result in an increase in Type II error probability. It is indefensible to reduce Type I error probabilities without addressing the effects the reduction has on Type II error probabilities and the relative costs of Type I and II errors. Optimal alpha makes multiple comparison corrections ill-advised in all cases because optimal alpha is designed to minimize the probability of making an error on any kind and any change to alpha increases the probability of making an error. If the argument is that the researcher is more worried about Type I than Type II errors then quantify the relative costs of Type I and II errors and use optimal alpha to minimize the costs of errors rather than the probability of errors.

Advantages of multiple critical effect sizes: This is a fundamental difference between traditional NHST and optimal alpha – that the statistical rejection threshold can only be reasonably set if a critical effect size is set. When there is not a single critical effect size we have suggested using three critical effect sizes – small, medium and large, but three is an arbitrary choice. There is no reason not to use more critical effect sizes and get a more precise estimate of the likely size of the effect and the strength of the evidence. Setting multiple critical effects sizes allows researchers to reach much more nuanced conclusions (see example above). . For any observed p-value in a particular analysis it would be possible to reject the null for

some small enough effect size and accept the null for some large enough effect size. Using many critical effect sizes would allow us to identify the effect size below which the observed p-value would warrant rejection of the null and above which acceptance of the null was warranted. We would only reach a definitive conclusion when there was convincing evidence for setting a single critical effect size. In the large majority of studies there would be multiple critical effect sizes and researchers would always reach conclusions about the strength of evidence for effects less than or equal to a specific size and against effects greater than a specific size.

Expanding range of tests: One, to develop code for a larger variety of statistical tests. This would include mixed models and randomization tests. Power can be estimated through simulation so optimal alpha could also be estimated with simulation.

Assessing implications of using traditional NHST versus optimal alpha: Our pilot studies suggest that the gap between the decisions ecologists make when using NHST and the decisions they should make is relatively large. That is, our early results suggest that, if optimal alpha had been used, ecologists would have reached a different conclusion in 20 – 30% of tests. Further, in all cases, the results and conclusions would

have been more precise, more nuanced and more representative of the data if optimal alpha had been used rather than traditional NHST. This may have consequences for the rate of progress in ecology but almost certainly has more practical implications in areas such as medical research (Blackwelder 1982, Kumbhare et al. 2019).

Conclusions: If the objective of null hypothesis significance testing is to maximize the probability of making the correct decision when choosing between the null and the alternative then not only is optimal alpha superior to traditional NHST approaches, it addresses most of the concerns that have been raised about traditional NHST. For that reason alone, optimal alpha should replace traditional NHST. But, we can use optimal alpha and get more than simply a better decision – we can fundamentally change how we approach NHST so that all decisions about accepting or rejecting the null identify the inflection effect size where the evidence changes from support for the null to support against the null. Further, inferences based on optimal alpha will include explicit comment about the strength of the evidence for or against the null at a specific effect size. The next steps are to develop code for a wider variety of statistical tests and provide easier access to optimal alpha through a user-friendly website.

REFERENCES

- Aarts AA et al. 2015. Estimating the reproducibility of psychological science. *Science* 349: 943-950.
- Amrhein V, Greenland S, McShane B. 2019. Scientists rise up against statistical significance. *Nature* 567: 305-307.
- Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: Problems, prevalence and an alternative. *Journal of Wildlife Management* 64: 912-923.
- Baker M. 2016. 1,500 scientists lift the lid of reproducibility. *Nature* 533: 452-454.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B.* 57:289-300.
- Benjamin et al. 2018. Redefine Statistical Significance. *Nature Human Behavior* 2: 6-10.
- Betensky RA. 2019. The p-value requires context, not a threshold. *The American Statistician* 73: 115-117.
- Blackwelder WC. 1982. "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials* 3: 345-353.
- Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. 2019. An introduction to second-generation p-values. *The American Statistician* 73: 157-167.
- Burnham KP, Anderson DR. 2002. *Model Selection and multi-model Inference. A practical information-theoretic approach.* Springer-Verlag New York.
- Burnham KP, Anderson DR. 2014. P values are only an index to evidence: 29th- vs. 21st century statistical science. *Ecology* 95: 627-630.
- Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Publishers.
- Cornfield, J. 1966. Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* 20: 18-23.
- Curran-Everett D, Benos DJ. 2004. Guidelines for reporting statistics in journals published by the American Physiological Society. *American Journal of Physiology Regulatory, Integrative and Comparative Physiology* 287: R247-R249.

- Dunn OJ. 1961. Multiple Comparisons among means. *Journal of the American Statistical Association*. 56: 52-64.
- Dushoff et al. 2019. I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*. 00: 1-4.
- Field SA, Tyre AJ, Possingham HP. 2005. Optimizing allocation of monitoring effort under economic and observational constraints. *Journal of Wildlife Management* 69: 473-482.
- Fraser DAS. 2019. The p-value function and statistical inference. *The American Statistician* 73: 135-147.
- Gannon MA, de Braganca Pereira CA, Polpo A. 2019. Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels. *The American Statistician* 73: 213-222.
- Gardner MJ, Altman DG. 1986. Confidence intervals and p-values: estimation rather than hypothesis testing. *British Medical Journal* 292: 746-750.
- Goodman WM, Spruill SE, Komaroff E. 2019. A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use. *The American Statistician* 73: 168-185.
- Hobbs NT, Hooten MB. 2015. *Bayesian Analysis: A Statistical Primer for Ecologists* Princeton University Press NJ.
- Houlahan et al. 2018. Negative relationships between species richness and temporal variability are common but weak in natural systems. *Ecology* 99: 2592-2604.
- Johnson DH. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63: 763-772.
- 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: 82-90.
- Kumbhare D, Alavinia M, Furlan J. 2019. Hypothesis testing in superiority, noninferiority, and equivalence clinical trials. *American Journal of Physical Medicine & Rehabilitation* 98: 226-230.
- Levine TR, Weber R, Hullett C, Park HS, Massi Lindsay LL. 2008. A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research* 34: 171-187.

Ludbrook J. 2013. Should we use one-sided or two-sided P values in tests of significance. *Clinical and Experimental Pharmacology and Physiology*. 40: 357-361.

Mapstone BD. 1995. Scalable decision rules for environmental impact studies: Effect size, Type I, and Type II errors. *Ecological Applications* 5: 401-410.

Mayo DM. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. 2019. Abandon Statistical Significance. *The American Statistician* 73: 235-245.

Mudge JF, Baker LF, Edge CB, Houlahan JE. 2012. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE* 7: e32734.

Munkittrick KR, McGeachy SA, McMaster ME, Courtenay SC. 2002. Overview of freshwater fish studies from the pulp and paper environmental effects monitoring program. *Water Quality Research Journal Canada* 37: 49-77.

Munkittrick KR, Arens CJ, Lowell RB, Kaminski GP. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. *Environmental Toxicology and Chemistry* 28: 1361-1371.

Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*. 231: 289-337.

Nickerson RS. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5: 241-301.

Savage LJ. 1961. The foundations of statistics reconsidered. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1*, University of California Press 575-586.

Stang A, Poole C, Kuss O. 2010. The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology* 25: 225-230.

Stephens PA, Buskirk SW, Martinez del Rio C. 2007. Inference in ecology and evolution. *Trends In Ecology and Evolution* 22: 1192-197.

Wasserstein RL, Lazar NA. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70: 129-133.

Yoccoz NG. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*. 72: 106-111