OPINION ARTICLE

REVISED Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice [version 4; peer review: 2 approved, 2 not approved]

Null hypothesis significance testing: a short tutorial

Cyril Pernet

Centre for Clinical Brain Sciences (CCBS), Neuroimaging Sciences, The University of Edinburgh, Edinburgh, UK

v4 **First published:** 25 Aug 2015, **4**:621
https://doi.org/10.12688/f1000research.6963.1
Second version: 13 Jul 2016, **4**:621
https://doi.org/10.12688/f1000research.6963.2
Third version: 10 Oct 2016, **4**:621
https://doi.org/10.12688/f1000research.6963.3
Fourth version: 26 Sep 2017, **4**:621
https://doi.org/10.12688/f1000research.6963.4
Latest published: 12 Oct 2017, **4**:621
https://doi.org/10.12688/f1000research.6963.5

Abstract
Although thoroughly criticized, null hypothesis significance testing (NHST) remains the statistical method of choice used to provide evidence for an effect, in biological, biomedical and social sciences. In this short guide, I first summarize the concepts behind the method, distinguishing test of significance (Fisher) and test of acceptance (Newman-Pearson) and point to common interpretation errors regarding the p-value. I then present the related concepts of confidence intervals and again point to common interpretation errors. Finally, I discuss what should be reported in which context. The goal is to clarify concepts to avoid interpretation errors and propose simple reporting practices.

Keywords
null hypothesis significance testing , tutorial , p-value , reporting , confidence intervals

Open Peer Review

Approval Status ✗ ✗ ✓ ✓

	1	2	3	4
version 5 (revision) 12 Oct 2017				
version 4 (revision) 26 Sep 2017				✓ view
version 3 (revision) 10 Oct 2016			✓ view	? view
version 2 (revision) 13 Jul 2016			? view	
version 1 25 Aug 2015	✗ view	✗ view		

1. **Daniel Lakens**, Eindhoven University of Technology, Eindhoven, The Netherlands
2. **Marcel ALM van Assen**, Tilburgh University, Tilburg, The Netherlands
3. **Stephen J. Senn**, Luxembourg Institute of

Health, Strassen, Luxembourg

4. **Dorothy Vera Margaret Bishop**, University of Oxford, Oxford, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Cyril Pernet (cyril.pernet@ed.ac.uk)

Author roles: Pernet C: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Pernet C. **Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice [version 4; peer review: 2 approved, 2 not approved]** F1000Research 2017, **4**:621 https://doi.org/10.12688/f1000research.6963.4

First published: 25 Aug 2015, **4**:621 https://doi.org/10.12688/f1000research.6963.1

REVISED **Amendments from Version 3**

The manuscript has been changed according to the reviewer recommendations: title change, typos, a few other edits, along with 1 major update: the use of a concrete example. The data for that example are given (new Table 1) and used for the Figure 1 (updated), and testing: Fisher, Neyman-Pearson, Equivalence testing, and Bayes Factor (the last two being new as well as recommended). A few other references were also added.

See referee reports

The Null Hypothesis Significance Testing framework

NHST is a method of statistical inference by which an experimental factor is tested against a hypothesis of no effect or no relationship based on a given observation. The method is a combination of the concepts of significance testing developed by Fisher in 1925 and of acceptance based on critical rejection regions developed by Neyman & Pearson in 1928. In the following I am first presenting each approach, highlighting the key differences and common misconceptions that result from their combination into the NHST framework (for a more mathematical comparison, along with the Bayesian method, see Christensen, 2005). I next present the related concept of confidence intervals. I finish by discussing practical aspects in using NHST and reporting practice.

Fisher, significance testing, and the p-value

The method developed by (Fisher, 1934; Fisher, 1955; Fisher, 1959) allows us to compute the probability of observing a result at least as extreme as a test statistic (e.g. t value), assuming the null hypothesis of no effect is true. This probability or p-value reflects (1) the conditional probability of achieving the observed outcome or larger: $p(Obs \geq t|H0)$, and (2) is therefore a cumulative probability rather than a point estimate. It is equal to the area under the null probability distribution curve from the observed test statistic to the tail of the null distribution (Turkheimer *et al.*, 2004 – Figure 1). The approach proposed is of 'proof by contradiction' (Christensen, 2005), we pose the null model and test if data conform to it.

In practice, it is recommended to set a *level of significance* (a theoretical p-value) that acts as a reference point to identify significant results, that is to identify results that differ from the null-hypothesis of no effect. Fisher recommended using p=0.05 to judge whether an effect is significant or not as it is roughly two standard deviations away from the mean for the normal distribution (Fisher, 1934 page 45: 'The value for which p=.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not'). It is important to appreciate that this threshold is partly subjective: 2 standard deviations seems reasonable in biological, human and social sciences but in particle physics, this threshold

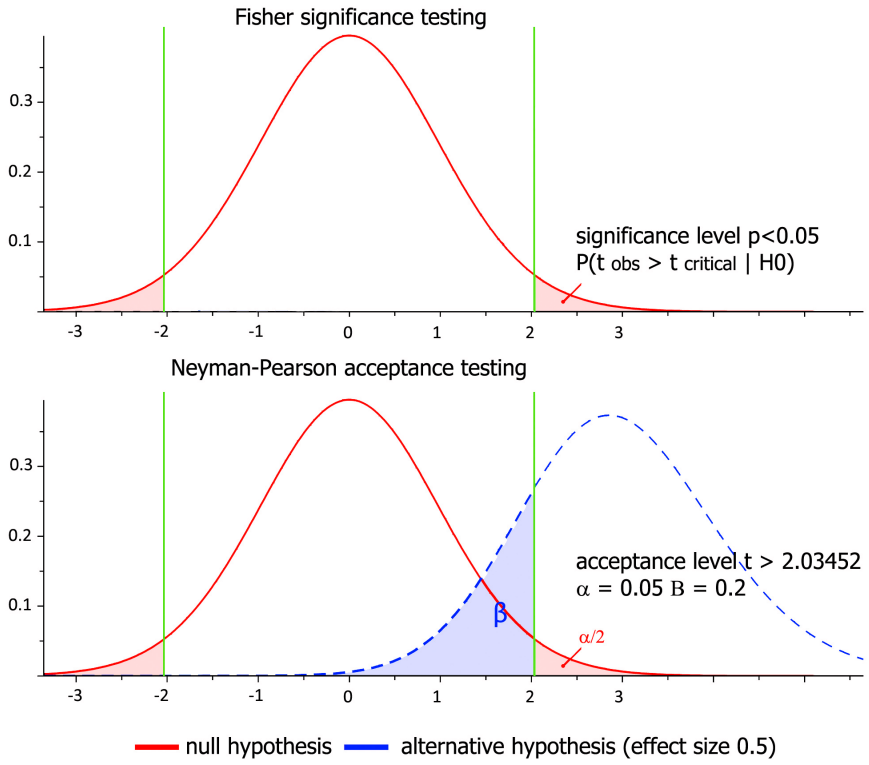

Figure 1. Illustration of the difference between the Fisher and Neyman-Pearson procedures. The figure was prepared with G-power for a two-sided one-sample t-test, an effect size of 0.5. In Fisher's procedure, only the null-hypothesis is posed, and the observed p-value is compared to an a priori level of significance. If the observed p-value is below this level (here p=0.05), one rejects H0. In Neyman-Pearson's procedure, the null and alternative hypotheses are specified along with an a priori level of acceptance (here alpha 0.05 beta 0.8). If the observed statistical value t value is outside the critical region (here [-∞ -2.57] [2.57+∞]), one rejects H0.

is set at 5 standard deviations. That choice of two standard deviations is also contested, for instance in psychology calls for p<.001 (Colquhoun, 2014) or p<.005 (Benjamin *et al.*, 2017) have been made. A key aspect of Fishers' theory is that the significance threshold is only part of the process to accept or reject a hypothesis. Since only the null-hypothesis is tested, p-values are meant to be used in a graded manner to decide whether the evidence is worth additional investigation and/or replication (Fisher, 1971 page 13: 'it is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require […]' and 'no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon'). How small the level of significance is, is thus left to researchers.

What is not a p-value? Common mistakes

The p-value *is not an indication of the strength or magnitude of an effect*. Any interpretation of the p-value in relation to the effect under study (strength, reliability, probability) is wrong, since p-values are based on H0. In addition, while p-values are randomly distributed (if all the assumptions of the test are met) when there is no effect, their distribution depends of both the population effect size and the number of participants, making impossible to infer strength of effect from them.

Similarly, 1-p *is not the probability of replicating an effect*. Often, a small value of p is considered to mean a strong likelihood of getting the same results on another try, but again this cannot be obtained because the p-value is not informative about the effect itself (Miller, 2009). Because the p-value depends on the number of subjects, it can only be used in high powered studies to interpret results. In low powered studies (typically small number of subjects), the p-value has a large variance across repeated samples, making it unreliable to estimate replication (Halsey *et al.*, 2015).

A (small) p-value *is not an indication favouring a given hypothesis*. Because a low p-value only indicates a misfit of the null hypothesis to the data, it cannot be taken as evidence in favour of a specific alternative hypothesis more than any other possible alternatives such as measurement error and selection bias (Gelman, 2013). Some authors have even argued that the more (a priori) implausible the alternative hypothesis, the greater the chance that a finding is a false alarm (Krzywinski & Altman, 2013; Nuzzo, 2014).

The p-value *is not the probability of the null hypothesis p(H0) being true,* (Krzywinski & Altman, 2013). This common misconception arises from a confusion between the probability of an observation given the null p(Obs≥t|H0) and the probability of the null given an observation p(H0|Obs≥t) that is then taken as an indication for p(H0) (see Nickerson, 2000).

Neyman-Pearson, hypothesis testing, and the α-value

Neyman & Pearson (1933) proposed a framework of statistical inference for applied decision making and quality control. In such framework, two hypotheses are proposed: the null hypothesis of no effect and the alternative hypothesis of an effect, along with a control of the long run probabilities of making errors. The first key concept in this approach, is the establishment of an *alternative hypothesis* along with an a priori effect size. This differs markedly from Fisher who proposed a general approach for scientific inference conditioned on the null hypothesis only. The second key concept is the *control of error rates*. Neyman & Pearson (1928) introduced the notion of critical intervals, therefore dichotomising the space of possible observations into correct vs. incorrect zones. This dichotomisation allows one to distinguish correct results (rejecting H0 when there is an effect and not rejecting H0 when there is no effect) from errors (rejecting H0 when there is no effect, the type I error, and not rejecting H0 when there is an effect, the type II error). In this context, alpha is the probability of committing a Type I error in the long run and Beta is the probability of committing a Type II error in the long run.

The (theoretical) difference in terms of hypothesis testing between Fisher and Neyman-Pearson is illustrated on Figure 1. In the 1st case, we choose a level of significance for observed data of 5%, and compute the p-value. If the p-value is below the level of significance, it is used to reject H0. Importantly, the exact p-value is then taken as measure of evidence. In the 2nd case, we set a critical interval based on the a priori effect size and error rates (alpha and beta). If an observed statistic value is below and above the critical values (the bounds of the confidence region), it is deemed significantly different from H0. Note that in this framework, the p-value is irrelevant (Szucs & Ioannidis, 2017). In the NHST framework, the level of significance is (in practice) assimilated to the alpha level (the level of acceptance), which appears as a simple decision rule: if the p-value is less or equal to alpha, the null is rejected. It is however a common mistake to assimilate these two concepts. The level of significance set for a given sample is not the same as the frequency of acceptance alpha found on repeated sampling because alpha (a point estimate) is meant to reflect the long run probability whilst the p-value (a cumulative estimate) reflects the current probability (Fisher, 1955; Hubbard & Bayarri, 2003).

Imagine you want to test that median reaction times between two experimental conditions differ. We first compute the difference per participant, given the mean difference and associated standard deviation (Table 1). The null hypothesis is that the mean reaction time difference is 0, and a one sample Student t-test gives t(34)=0.3037 p=0.7632. Following Fisher hypothesis testing, setting the level of significance a p=0.05, we cannot reject the null and thus continue to assume conditions do not differ (see below *Acceptance or rejection of H0?)*. If we follow Neyman-Pearson, we must specify an alternative hypothesis along with our alpha and beta rates. The simplest alternative hypothesis is to state that condition differ, i.e. mean reaction time differences are not equal to 0 and we chose our acceptance level with alpha 0.05. We are also compelled to define beta (which is not the case for Fisher hypothesis testing). To compute the prior probability of type II error, we need to define an a-priori effect size as well. Assuming reaction times differences cannot be less than 10 ms (± 20 ms), which correspond to a medium effect size (d=0.5), then 34 subjects are needed to achieve 80% power (1-beta). The results (t and p values) are the same but we gain in certainty regarding the type

Table 1. Data from 35 subjects (S1 to S35) showing a difference between two conditions in median reaction times, used for Figure 1 and testing (Fisher, Neyman-Pearson, Equivalence testing, Bayes Factor).

S1	26.00	S13	-19.41	S25	66.78
S2	23.62	S14	-26.58	S26	-15.09
S3	-3.38	S15	23.18	S27	-12.07
S4	9.80	S16	53.61	S28	-16.85
S5	-13.38	S17	-17.49	S29	-10.34
S6	-28.99	S18	-7.02	S30	40.94
S7	-9.13	S19	-56.22	S31	-0.31
S8	8.60	S20	35.68	S32	5.16
S9	-9.89	S21	-25.53	S33	-18.64
S10	38.27	S22	2.30	S64	-16.47
S11	-31.21	S23	49.71	S35	-11.41
S12	-15.15	S24	30.79		

II error (here less than 20% chance to see this result if there was an effect). One typical issue in NHST is defining H0, defining H1 as a simple difference, taking a conventional alpha say 0.05, and not defining beta (and thus a priori power). By doing so, we cannot go into the acceptance / rejection inference mode (i.e. take a binary decision) because we have not dichotomized the decision space. That case is more akin to significance hypothesis testing (Fisher) and inference must thus be graded rather than binary.

Acceptance or Rejection of H0?

The acceptance level α can also be viewed as the maximum probability that a test statistic falls into the rejection region when the null hypothesis is true (Johnson, 2013). Therefore, one can only reject the null hypothesis if the test statistics falls into the critical region(s), or fail to reject this hypothesis. In the latter case, all we can say is that no significant effect was observed, but one cannot conclude that the null hypothesis is true. This is another common mistake in using NHST: there is a profound difference between accepting the null hypothesis and simply failing to reject it (Killeen, 2005). By failing to reject, we simply continue to assume that H0 is true, which implies that one cannot argue against a theory from a non-significant result (absence of evidence is not evidence of absence). To accept the null hypothesis, tests of equivalence (Walker & Nowacki, 2011) or Bayesian approaches (Dienes, 2014; Kruschke, 2011) must be used.

Taking again the data from Table 1, The NHST tells us the 95% CI of the mean reaction time difference is [-8.11 10.97]. Since the CI includes 0, we cannot reject H0, and we continue to assume that conditions do not differ. A test of equivalence considers instead equivalence bounds. Here the bounds are taken to be -10ms and +10ms, i.e. we consider that any difference lower than 10 ms is meaningless (same as the effect size considered above). The computation is simply two one sample t-tests at alpha = 10%, one for the lower bound (-10ms) and one for the upper bound (+10 ms) and rejecting for differences (inverting H0 and H1 - Lakens, 2017). In this case, because results indicate that the difference is within the equivalence bounds, i.e. significantly above the lower bound ($t(34)=-1.82$ $p=.038$) and significantly below the upper bound ($t(34)=2.43$ $p=.001$), we can conclude that conditions do not differ, i.e. we can accept H0. Using a Bayesian t-test with normal priors (0 mean, 1 standard deviation, Dienes, 2014), the Bayes factor is 0.175 for H1 (BF_{10}) and 5.73 for H0 (BF_{01}) which gives moderate evidence for H0. Being able to confirm H0, we can argue against a theory that proposed differences – this is not possible using the NHST framework.

Confidence intervals

Confidence intervals (CI) are constructs that fail to cover the true value at a rate of alpha, the Type I error rate (Morey & Rouder, 2011) and therefore indicate if observed values can be rejected by a (two tailed) test with a given alpha. CI have been advocated as alternatives to p-values because (i) they allow judging the statistical significance and (ii) provide estimates of effect size. Assuming the CI (a)symmetry and width are correct (but see Wilcox, 2012), they also give some indication about the likelihood that a similar value can be observed in future studies. For future studies of the same sample size, 95% CI give about 83% chance of replication success (Cumming & Maillardet, 2006). If sample sizes however differ between studies, there is no warranty that a CI from one study will be true at the rate alpha in a different study, which implies that CI cannot be compared across studies at this is rarely the same sample sizes.

Although CI provide more information, they are not less subject to interpretation errors (see Savalei & Dunn, 2015 for a review). The most common mistake is to interpret CI as the probability that a parameter (e.g. the population mean) will fall in that interval X% of the time. The correct interpretation is that, for repeated measurements with the same sample sizes, taken from the same population, X% of times the CI obtained will contain the true parameter value (Tan & Tan, 2010). The alpha value has the same interpretation as testing against H0, e.g. a 95% CI is wrong in 5% of the times in the long run (i.e. if we repeat the experiment many times). This implies that CI do not allow to make strong statements about the parameter of interest (e.g. the mean difference) or about H1 (Hoekstra *et al.*, 2014). To make a statement about the probability of a parameter of interest (e.g. the probability of the mean), Bayesian intervals must be used (Morey & Rouder, 2011).

From the Table 1, the 95% confidence interval is [-8.11 10.97] and this means that this interval will be wrong 5% of the times when we repeat the experiment. The 95% Bayesian interval of the mean is in this case the same. Its meaning is however completely different: there is 95% probability that the average is in this interval.

The (correct) use of NHST

NHST has always been criticized, and yet is still used every day in scientific reports (Nickerson, 2000). One question to ask oneself is what is the goal of a scientific experiment at hand? If the goal is to establish a discrepancy with the null hypothesis and/or establish a pattern of order (i.e. establish that A > B),

because both requires ruling out equivalence (i.e. ruling out A=B), then NHST is a good tool (Frick, 1996; Walker & Nowacki, 2011). If the goal is to test the presence of an effect (i.e. compute its probability) and/or establish some quantitative values related to an effect, then NHST is not the method of choice since testing can only reject the null hypothesis.

While a Bayesian analysis is suited to estimate the probability that a hypothesis is correct, like NHST, it does not prove a theory by itself, but adds to its plausibility (Lindley, 2000). It has however another advantage: it allows to choose between competing hypotheses while NHST cannot prove any specific hypothesis (Szucs & Ioannidis, 2017). No matter what testing procedure is used and how strong results are, Fisher (1959, p13) reminds us that ' […] no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon'. Similarly, the recent statement of the American Statistical Association (Wasserstein & Lazar, 2016) makes it clear that conclusions should be based on the researchers understanding of the problem in context, along with all summary data and tests, and that no single value (being p-values, Bayesian factor or else) can be used to support or invalidate a theory.

What to report and how?

Considering that quantitative reports will always have more information content than binary (significant or not) reports, we can always argue that raw and/or normalized effect size, confidence intervals, or Bayes factor must be reported. Reporting everything can however hinder the communication of the main result(s), and we should aim at giving only the information needed, at least in the core of a manuscript. I recommend adopting 'optimal reporting' in the result section to keep the message clear, but have detailed supplementary material. When the hypothesis is about the presence/absence (two-sided test) or order of an effect (one-sided test), and providing that a study has sufficient power, NHST is appropriate and it is sufficient to report in the text the actual p-value since it conveys the information needed to rule out equivalence. When the hypothesis and/or the discussion involve some quantitative value, and because p-values do not inform on the effect, it is essential to report on effect sizes (Lakens, 2013), preferably accompanied by confidence or credible intervals. The reasoning is simply that one cannot predict and/or discuss quantities without accounting for variability.

Because science progress is obtained by cumulating evidence (Rosenthal, 1991), scientists should also anticipate the secondary use of the data. With today's electronic articles, there are no reasons for not including all derived data (mean, standard deviations, effect size, CI, Bayes factor) as supplementary tables (or even better also share raw data). It is also essential to report the context in which tests were performed – that is to report all tests performed (all t, F, p values) because of the increase type I error rate due to selective reporting (multiple comparisons and p-hacking problems - Ioannidis, 2005). Providing all of this information allows (i) other researchers to directly and effectively compare their results in quantitative terms (replication of effects beyond significance, Open Science Collaboration, 2015), (ii) to compute power to future studies (Lakens & Evers, 2014), and (iii) to aggregate results for meta-analyses whilst minimizing publication bias (van Assen *et al.*, 2014).

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

Benjamin DJ, Berger J, Johannesson M, *et al.*: **Redefine statistical significance**. 2017.
Reference Source

Christensen R: **Testing Fisher, Neyman, Pearson, and Bayes.** *Am Stat.* 2005; **59**(2): 121–126.
Publisher Full Text

Colquhoun D: **An investigation of the false discovery rate and the misinterpretation of *p*-values.** *R Soc open sci.* 2014; **1**(3): 140216.
PubMed Abstract | Publisher Full Text | Free Full Text

Cumming G, Maillardet R: **Confidence intervals and replication: where will the next mean fall?** *Psychol Methods.* 2006; **11**(3): 217–227.
PubMed Abstract | Publisher Full Text

Dienes Z: **Using Bayes to get the most out of non-significant results.** *Front Psychol.* 2014; **5**: 781.
PubMed Abstract | Publisher Full Text | Free Full Text

Fisher RA: **Statistical Methods for Research Workers**. (Vol. 5th Edition). Edinburgh, UK: Oliver and Boyd. 1934.
Reference Source

Fisher RA: **Statistical Methods and Scientific Induction.** *Journal of the Royal Statistical Society, Series B.* 1955; **17**(1): 69–78.
Reference Source

Fisher RA: **Statistical methods and scientific inference**. (2nd ed.). NewYork: Hafner Publishing, 1959.
Reference Source

Fisher RA: **The Design of Experiments**. Hafner Publishing Company, New-York. 1971.
Reference Source

Frick RW: **The appropriate use of null hypothesis testing.** *Psychol Methods.* 1996; **1**(4): 379–390.
Publisher Full Text

Gelman A: ***P* values and statistical practice.** *Epidemiology.* 2013; **24**(1): 69–72.
PubMed Abstract | Publisher Full Text

Halsey LG, Curran-Everett D, Vowler SL, *et al.*: **The fickle *P* value generates irreproducible results.** *Nat Methods.* 2015; **12**(3): 179–85.
PubMed Abstract | Publisher Full Text

Hoekstra R, Morey RD, Rouder JN, *et al.*: **Robust misinterpretation of confidence intervals.** *Psychon Bull Rev.* 2014; **21**(5): 1157–1164.
PubMed Abstract | Publisher Full Text

Hubbard R, Bayarri MJ: **Confusion over measures of evidence (*p*'s) versus errors ([alpha]'s) in classical statistical testing.** *Am Stat.* 2003; **57**(3): 171–182.
Publisher Full Text

Ioannidis JP: **Why most published research findings are false.** *PLoS Med.* 2005; **2**(8): e124.
PubMed Abstract | Publisher Full Text | Free Full Text

Johnson VE: **Revised standards for statistical evidence.** *Proc Natl Acad Sci U S A.* 2013; **110**(48): 19313–19317.
PubMed Abstract | Publisher Full Text | Free Full Text

Killeen PR: **An alternative to null-hypothesis significance tests.** *Psychol Sci.* 2005; **16**(5): 345–353.
PubMed Abstract | Publisher Full Text | Free Full Text

Kruschke JK: **Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison.** *Perspect Psychol Sci.* 2011; **6**(3): 299–312.
PubMed Abstract | Publisher Full Text

Krzywinski M, Altman N: **Points of significance: Significance, *P* values and *t*-tests.** *Nat Methods.* 2013; **10**(11): 1041–1042.
PubMed Abstract | Publisher Full Text

Lakens D: **Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs.** *Front Psychol.* 2013; **4**: 863.
PubMed Abstract | Publisher Full Text | Free Full Text

Lakens D: **Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses.** *Soc Psychol Personal Sci.* 2017; **8**(4): 355–362.
PubMed Abstract | Publisher Full Text | Free Full Text

Lakens D, Evers ER: **Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies.** *Perspect Psychol Sci.* 2014; **9**(3): 278–292.
PubMed Abstract | Publisher Full Text

Lindley D: **The philosophy of statistics.** *J R Stat Soc.* 2000; **49**(3): 293–337.
Publisher Full Text

Miller J: **What is the probability of replicating a statistically significant effect?** *Psychon Bull Rev.* 2009; **16**(4): 617–640.
PubMed Abstract | Publisher Full Text

Morey RD, Rouder JN: **Bayes factor approaches for testing interval null hypotheses.** *Psychol Methods.* 2011; **16**(4): 406–419.
PubMed Abstract | Publisher Full Text

Neyman J, Pearson ES: **On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I.** *Biometrika.* 1928; **20A**(1/2): 175–240.
Publisher Full Text

Neyman J, Pearson ES: **On the problem of the most efficient tests of statistical hypotheses.** *Philos Trans R Soc Lond Ser A.* 1933; **231**(694–706): 289–337.
Publisher Full Text

Nickerson RS: **Null hypothesis significance testing: a review of an old and continuing controversy.** *Psychol Methods.* 2000; **5**(2): 241–301.
PubMed Abstract | Publisher Full Text

Nuzzo R: **Scientific method: statistical errors.** *Nature.* 2014; **506**(7487): 150–152.
PubMed Abstract | Publisher Full Text

Open Science Collaboration: **PSYCHOLOGY. Estimating the reproducibility of psychological science.** *Science.* 2015; **349**(6251): aac4716.
PubMed Abstract | Publisher Full Text

Rosenthal R: **Cumulating psychology: an appreciation of Donald T. Campbell.** *Psychol Sci.* 1991; **2**(4): 213–221.
Publisher Full Text

Savalei V, Dunn E: **Is the call to abandon *p*-values the red herring of the replicability crisis?** *Front Psychol.* 2015; **6**: 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Szucs D, Ioannidis JPA: **When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment.** *Front Hum Neurosci.* 2017; **11**: 390.
PubMed Abstract | Publisher Full Text | Free Full Text

Tan SH, Tan SB: **The Correct Interpretation of Confidence Intervals.** *Proceedings of Singapore Healthcare.* 2010; **19**(3): 276–278.
Publisher Full Text

Turkheimer FE, Aston JA, Cunningham VJ: **On the logic of hypothesis testing in functional imaging.** *Eur J Nucl Med Mol Imaging.* 2004; **31**(5): 725–732.
PubMed Abstract | Publisher Full Text

van Assen MA, van Aert RC, Nuijten MB, *et al.*: **Why Publishing Everything Is More Effective than Selective Publishing of Statistically Significant Results.** *PLoS One.* 2014; **9**(1): e84896.
PubMed Abstract | Publisher Full Text | Free Full Text

Walker E, Nowacki AS: **Understanding equivalence and noninferiority testing.** *J Gen Intern Med.* 2011; **26**(2): 192–196.
PubMed Abstract | Publisher Full Text | Free Full Text

Wasserstein RL, Lazar NA: **The ASA's Statement on *p*-Values: Context, Process, and Purpose.** *Am Stat.* 2016; **70**(2): 129–133.
Publisher Full Text

Wilcox R: **Introduction to Robust Estimation and Hypothesis Testing**. Edition 3, Academic Press, Elsevier: Oxford, UK, ISBN: 978-0-12-386983-8. 2012.
Reference Source

Open Peer Review

Current Peer Review Status: ✗ ✗ ✓ ✓

Version 4

Reviewer Report 09 October 2017

https://doi.org/10.5256/f1000research.13792.r26375

✓ **Dorothy Vera Margaret Bishop**
Department of Experimental Psychology, University of Oxford, Oxford, UK

The Pernet paper is much improved and I think will be a useful addition to the literature. There are a handful of minor typos to be corrected, but once this is done, I am happy to recommend acceptance.

p 4, col 1, para 2, line 6; 'depends on'
p 4, col 2, para 3, line 12 'conditions'
p 5, col 2, para 2, last line 'studies, as these rarely use the same sample sizes'
p 6, para 2, line 12 need apostrophe - researcher's
p 6, para 2 line 13 - material in brackets rephrase as 'whether p-values, Bayes factors or something else)'
p 7 - I don't think the word PSYCHOLOGY is part of the article title for the Open Science Collaboration paper

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 09 Oct 2017

Cyril Pernet, The University of Edinburgh, Edinburgh, UK

Thank you again for your revision - using the example did improve a lot.
I have fixed the last typos, thank you for that.

Competing Interests: No competing interests were disclosed.

Version 3

Reviewer Report 03 February 2017

https://doi.org/10.5256/f1000research.10487.r19282

?

Dorothy Vera Margaret Bishop
Department of Experimental Psychology, University of Oxford, Oxford, UK

I can see from the history of this paper that the author has already been very responsive to reviewer comments, and that the process of revising has now been quite protracted.

That makes me reluctant to suggest much more, but I do see potential here for making the paper more impactful. So my overall view is that, once a few typos are fixed (see below), this could be published as is, but I think there is an issue with the potential readership and that further revision could overcome this.

I suspect my take on this is rather different from other reviewers, as I do not regard myself as a statistics expert, though I am on the more quantitative end of the continuum of psychologists and I try to keep up to date. I think I am quite close to the target readership , insofar as I am someone who was taught about statistics ages ago and uses stats a lot, but never got adequate training in the kinds of topic covered by this paper. The fact that I am aware of controversies around the interpretation of confidence intervals etc is simply because I follow some discussions of this on social media. I am therefore very interested to have a clear account of these issues.

This paper contains helpful information for someone in this position, but it is not always clear, and I felt the relevance of some of the content was uncertain. So here are some recommendations:

1. I wondered about changing the focus slightly and modifying the title to reflect this to say something like: Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice

   As one previous reviewer noted, it's questionable that there is a need for a tutorial introduction, and the limited length of this article does not lend itself to a full explanation. So it might be better to just focus on explaining as clearly as possible the problems people have had in interpreting key concepts. I think a title that made it clear this was the content would be more appealing than the current one.

2. P 3, col 1, para 3, last sentence. Although statisticians always emphasise the arbitrary nature of $p < .05$, we all know that in practice authors who use other values are likely to have their analyses queried. I wondered whether it would be useful here to note that in some disciplines different cutoffs are traditional, e.g. particle physics. Or you could cite David Colquhoun's paper in which he recommends using $p < .001$ (

http://rsos.royalsocietypublishing.org/content/1/3/140216) - just to be clear that the traditional p < .05 has been challenged.

3. Having read the section on the Fisher approach and Neyman-Pearson approach I felt confused. I have to confess that despite years of doing stats, this distinction had eluded me (which is why I am a good target reader), but I wasn't really entirely enlightened after reading this. As I understand it, I have been brought up doing null hypothesis testing, so am adopting a Fisher approach. But I also talk about setting alpha to .05, and understand that to come from the Neyman-Pearson approach. If I have understood this correctly, these do amount to the same thing (as the author states, they are assimilated in practice), but we are then told this is a 'common mistake'. But the explanation of the difference was hard to follow and I found myself wondering whether it would actually make any difference to what I did in practice. In order to understand the last sentence before 'Acceptance or rejection of H0' I would need some good analogy. Maybe it would be possible to explain this better with the tried-and-tested example of tossing a coin. So in Fisher approach you do a number of coin tosses to test whether the coin is unbiased (Null hypothesis); you can then work out p as the probability of the null given a specific set of observations, which is the p—value.

   What I can't work out is how you would explain the alpha from Neyman-Pearson in the same way (though I can see from Figure 1 that with N-P you could test an alternative hypothesis, such as the idea that the coin would be heads 75% of the time).

4. The section on acceptance or rejection of H0 was good, though I found the first sentence a bit opaque and wondered if it could be made clearer. Also I wondered if this rewording would be accurate (as it is clearer to me): instead of:

   'By failing to reject, we simply continue to assume that H0 is true, which implies that one cannot….' have 'In failing to reject, we do not assume that H0 is true; one cannot argue against a theory from a non-significant result.'

   I felt most readers would be interested to read about tests of equivalence and Bayesian approaches, but many would be unfamiliar with these and might like to see an example of how they work in practice – if space permitted.

5. Confidence intervals: I simply could not understand the first sentence – I wondered what was meant by 'builds' here. I understand about difficulties in comparing CI across studies when sample sizes differ, but I did not find the last sentence on p 4 easy to understand.

6. P 5: The sentence starting: 'The alpha value has the same interpretation' was also hard to understand, especially the term '1-alpha CI'. Here too I felt some concrete illustration might be helpful to the reader. And again, I also found the reference to Bayesian intervals tantalising – I think many readers won't know how to compute these and something like a figure comparing a traditional CI with a Bayesian interval and giving a source for those who want to read on would be very helpful. The reference to 'credible intervals' in the penultimate paragraph is very unclear and needs a supporting reference – most readers will not be familiar with this concept.

Typos etc:

P 3, col 1, para 2, line 2; “allows us to compute”
P 3, col 2, para 2, ‘probability of replicating’
P 3, col 2, para 2, line 4 ‘informative about’
P 3, col 2, para 4, line 2 delete ‘of’
P 3, col 2, para 5, line 9 – ‘conditioned’ is either wrong or too technical here: would ‘based’ be acceptable as alternative wording
P 3, col 2, para 5, line 13 ‘This dichotomisation allows one to distinguish’
P 3, col 2, para 5, last sentence, delete ‘Alternatively’.
P 3, col 2, last para line 2 ‘first’
P 4, col 2, para 2, last sentence is hard to understand; not sure if this is better: ‘If sample sizes differ between studies, the distribution of CIs cannot be specified a priori’
P 5, col 1, para 2, ‘a pattern of order’ – I did not understand what was meant by this
P 5, col 1, para 2, last sentence unclear: possible rewording: “If the goal is to test the size of an effect then NHST is not the method of choice, since testing can only reject the null hypothesis.’ (??)
P 5, col 1, para 3, line 1 delete ‘that’
P 5, col 1, para 3, line 3 ‘on’ -> ‘by’
P 5, col 2, para 1, line 4 , rather than ‘Here I propose to adopt’ I suggest ‘I recommend adopting’
P 5, col 2, para 1, line 13 ‘with’ -> ‘by’
P 5, col 2, para 1 – recommend deleting last sentence
P 5, col 2, para 2, line 2 ‘consider’ -> ‘anticipate’
P 5, col 2, para 2, delete ‘should always be included’
P 5, col 2, para 2, ‘type one’ -> ‘Type I’

References

1. Colquhoun D: An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*. 2014; **1** (3): 140216 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Sep 2017

Cyril Pernet, The University of Edinburgh, Edinburgh, UK

I wondered about changing the focus slightly and modifying the title to reflect this to say something like: Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice

1. As one previous reviewer noted, it’s questionable that there is a need for a tutorial introduction, and the limited length of this article does not lend itself to a full explanation. So it might be better to just focus on explaining as clearly as possible the problems people have had in interpreting key concepts. I think a title that made it clear this was the content would be more appealing than the current one.

Thank you for the suggestion – you indeed saw the intention behind the 'tutorial' style of the paper.

1. P 3, col 1, para 3, last sentence. Although statisticians always emphasise the arbitrary nature of p < .05, we all know that in practice authors who use other values are likely to have their analyses queried. I wondered whether it would be useful here to note that in some disciplines different cutoffs are traditional, e.g. particle physics. Or you could cite David Colquhoun's paper in which he recommends using p < .001 (http://rsos.royalsocietypublishing.org/content/1/3/140216) - just to be clear that the traditional p < .05 has been challenged.

I have added a sentence on this citing Colquhoun 2014 and the new Benjamin 2017 on using .005.

1. Having read the section on the Fisher approach and Neyman-Pearson approach I felt confused. I have to confess that despite years of doing stats, this distinction had eluded me (which is why I am a good target reader), but I wasn't really entirely enlightened after reading this. As I understand it, I have been brought up doing null hypothesis testing, so am adopting a Fisher approach. But I also talk about setting alpha to .05, and understand that to come from the Neyman-Pearson approach. If I have understood this correctly, these do amount to the same thing (as the author states, they are assimilated in practice), but we are then told this is a 'common mistake'. But the explanation of the difference was hard to follow and I found myself wondering whether it would actually make any difference to what I did in practice. In order to understand the last sentence before 'Acceptance or rejection of H0' I would need some good analogy. Maybe it would be possible to explain this better with the tried-and-tested example of tossing a coin. So in Fisher approach, you do a number of coin tosses to test whether the coin is unbiased (Null hypothesis); you can then work out p as the probability of the null given a specific set of observations, which is the p—value.

   What I can't work out is how you would explain the alpha from Neyman-Pearson in the same way (though I can see from Figure 1 that with N-P you could test an alternative hypothesis, such as the idea that the coin would be heads 75% of the time).

I agree that this point is always hard to appreciate, especially because it seems like in practice it makes little difference. I added a paragraph but using reaction times rather than a coin toss – thanks for the suggestion.

1. The section on acceptance or rejection of H0 was good, though I found the first sentence a bit opaque and wondered if it could be made clearer. Also I wondered if this rewording would be accurate (as it is clearer to me): instead of:

   'By failing to reject, we simply continue to assume that H0 is true, which implies that one cannot….' have 'In failing to reject, we do not assume that H0 is true; one cannot argue against a theory from a non-significant result.'

   I felt most readers would be interested to read about tests of equivalence and

Bayesian approaches, but many would be unfamiliar with these and might like to see an example of how they work in practice – if space permitted.

Added an example based on new table 1, following figure 1 – giving CI, equivalence tests and Bayes Factor (with refs to easy to use tools)

1. Confidence intervals: I simply could not understand the first sentence – I wondered what was meant by 'builds' here. I understand about difficulties in comparing CI across studies when sample sizes differ, but I did not find the last sentence on p 4 easy to understand.

Changed builds to constructs (this simply means they are something we build) and added that the implication that probability coverage is not warranty when sample size change, is that we cannot compare CI.

1. P 5: The sentence starting: 'The alpha value has the same interpretation' was also hard to understand, especially the term '1-alpha CI'. Here too I felt some concrete illustration might be helpful to the reader. And again, I also found the reference to Bayesian intervals tantalising – I think many readers won't know how to compute these and something like a figure comparing a traditional CI with a Bayesian interval and giving a source for those who want to read on would be very helpful. The reference to 'credible intervals' in the penultimate paragraph is very unclear and needs a supporting reference – most readers will not be familiar with this concept.

I changed 'i.e. we accept that 1-alpha CI are wrong in alpha percent of the times in the long run' to ', 'e.g. a 95% CI is wrong in 5% of the times in the long run (i.e. if we repeat the experiment many times).' – for Bayesian intervals I simply re-cited Morey & Rouder, 2011.

Typos

P 3, col 1, para 2, line 2; "allows us to compute"
P 3, col 2, para 2, 'probability of replicating'
P 3, col 2, para 2, line 4 'informative about'
P 3, col 2, para 4, line 2 delete 'of'
P 3, col 2, para 5, line 9 – 'conditioned' is either wrong or too technical here: would 'based' be acceptable as alternative wording
P 3, col 2, para 5, line 13 'This dichotomisation allows one to distinguish'
P 3, col 2, para 5, last sentence, delete 'Alternatively'.
P 3, col 2, last para line 2 'first'
P 4, col 2, para 2, last sentence is hard to understand; not sure if this is better: 'If sample sizes differ between studies, the distribution of CIs cannot be specified a priori'

It is not the CI cannot be specified, it's that the interval is not predictive of anything anymore! I changed it to 'If sample sizes, however, differ between studies, there is no warranty that a CI from one study will be true at the rate alpha in a different study, which implies that CI cannot be compared across studies at this is rarely the same sample sizes'

P 5, col 1, para 2, 'a pattern of order' – I did not understand what was meant by this

I added (i.e. establish that A > B) – we test that conditions are ordered, but without further

specification of the probability of that effect nor its size

P 5, col 1, para 2, last sentence unclear: possible rewording: "If the goal is to test the size of an effect then NHST is not the method of choice, since testing can only reject the null hypothesis.' (??)

Yes it works – thx

P 5, col 1, para 3, line 1 delete 'that'
P 5, col 1, para 3, line 3 'on' -> 'by'
P 5, col 2, para 1, line 4 , rather than 'Here I propose to adopt' I suggest 'I recommend adopting'
P 5, col 2, para 1, line 13 'with' -> 'by'
P 5, col 2, para 1 – recommend deleting last sentence
P 5, col 2, para 2, line 2 'consider' -> 'anticipate'
P 5, col 2, para 2, delete 'should always be included'
P 5, col 2, para 2, 'type one' -> 'Type I'

Typos fixed, and suggestions accepted – thanks for that.

Competing Interests: No competing interests were disclosed.

Reviewer Report 03 November 2016

https://doi.org/10.5256/f1000research.10487.r17400

Stephen J. Senn
Luxembourg Institute of Health, Strassen, L-1445, Luxembourg

The revisions are OK for me, and I have changed my status to Approved.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 28 September 2016

https://doi.org/10.5256/f1000research.9903.r16257

?

Stephen J. Senn
Luxembourg Institute of Health, Strassen, L-1445, Luxembourg

On the whole I think that this article is reasonable, my main reservation being that I have my doubts on whether the literature needs yet another tutorial on this subject.

A further reservation I have is that the author, following others, stresses what in my mind is a relatively unimportant distinction between the Fisherian and Neyman-Pearson (NP) approaches. The distinction stressed by many is that the NP approach leads to a dichotomy accept/reject based on probabilities established in advance, whereas the Fisherian approach uses tail area probabilities calculated from the observed statistic. I see this as being unimportant and not even true. Unless one considers that the person carrying out a hypothesis test (original tester) is mandated to come to a conclusion on behalf of all scientific posterity, then one must accept that any remote scientist can come to his or her conclusion depending on the personal type I error favoured. To operate the results of an NP test carried out by the original tester, the remote scientist then needs to know the p-value. The type I error rate is then compared to this to come to a personal accept or reject decision (1). In fact Lehmann (2), who was an important developer of and proponent of the NP system, describes exactly this approach as being good practice. (See Testing Statistical Hypotheses, 2nd edition P70). Thus using tail-area probabilities calculated from the observed statistics does not constitute an operational difference between the two systems.

A more important distinction between the Fisherian and NP systems is that the former does not use alternative hypotheses(3). Fisher's opinion was that the null hypothesis was more primitive than the test statistic but that the test statistic was more primitive than the alternative hypothesis. Thus, alternative hypotheses could not be used to justify choice of test statistic. Only experience could do that.

Further distinctions between the NP and Fisherian approach are to do with conditioning and whether a null hypothesis can ever be accepted.

I have one minor quibble about terminology. As far as I can see, the author uses the usual term 'null hypothesis' and the eccentric term 'nil hypothesis' interchangeably. It would be simpler if the latter were abandoned.

References

1. Senn S: A comment on replication,p-values and evidence S.N.Goodman,Statistics in Medicine 1992;11:875-879. *Statistics in Medicine*. 2002; **21** (16): 2437-2444 Publisher Full Text

2. Lehmann E L: Testing Statistical Hypotheses, 2nd edition. *Chapman and Hall*. 1993.
3. Senn S: You may believe you are a Bayesian but you are probably wrong. *RMM*. 2011; **2**: 41-66
Reference Source

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 10 November 2015

https://doi.org/10.5256/f1000research.7499.r11036

Marcel ALM van Assen
Department of Methodology and Statistics, Tilburgh University, Tilburg, The Netherlands

Null hypothesis significance testing (NHST) is a difficult topic, with misunderstandings arising easily. Many texts, including basic statistics books, deal with the topic, and attempt to explain it to students and anyone else interested. I would refer to a good basic text book, for a detailed explanation of NHST, or to a specialized article when wishing an explaining the background of NHST. So, what is the added value of a new text on NHST? In any case, the added value should be described at the start of this text. Moreover, the topic is so delicate and difficult that errors, misinterpretations, and disagreements are easy. I attempted to show this by giving comments to many sentences in the text.

Abstract: “null hypothesis significance testing is the statistical method of choice in biological, biomedical and social sciences to investigate if an effect is likely”. No, NHST is the method to test the hypothesis of no effect.

Intro: “Null hypothesis significance testing (NHST) is a method of statistical inference by which an observation is tested against a hypothesis of no effect or no relationship.” What is an ‘observation’? NHST is difficult to describe in one sentence, particularly here. I would skip this sentence entirely, here.

Section on Fisher; also explain the one-tailed test.

Section on Fisher; p(Obs|H0) does not reflect the verbal definition (the ‘or more extreme’ part).

Section on Fisher; use a reference and citation to Fisher's interpretation of the p-value

Section on Fisher; "This was however only intended to be used as an indication that there is something in the data that deserves further investigation. The reason for this is that only H0 is tested whilst the effect under study is not itself being investigated." First sentence, can you give a reference? Many people say a lot about Fisher's intentions, but the good man is dead and cannot reply... Second sentence is a bit awkward, because the effect is investigated in a way, by testing the H0.

Section on p-value; Layout and structure can be improved greatly, by first again stating what the p-value is, and then statement by statement, what it is not, using separate lines for each statement. Consider adding that the p-value is randomly distributed under H0 (if all the assumptions of the test are met), and that under H1 the p-value is a function of population effect size and N; the larger each is, the smaller the p-value generally is.

Skip the sentence "If there is no effect, we should replicate the absence of effect with a probability equal to 1-p". Not insightful, and you did not discuss the concept 'replicate' (and do not need to).

Skip the sentence "The total probability of false positives can also be obtained by aggregating results (Ioannidis, 2005)." Not strongly related to p-values, and introduces unnecessary concepts 'false positives' (perhaps later useful) and 'aggregation'.

Consider deleting; "If there is an effect however, the probability to replicate is a function of the (unknown) population effect size with no good way to know this from a single experiment (Killeen, 2005)."

The following sentence; " Finally, a (small) p-value *is not an indication favouring a hypothesis*. A low p-value indicates a misfit of the null hypothesis to the data and cannot be taken as evidence in favour of a specific alternative hypothesis more than any other possible alternatives such as measurement error and selection bias (Gelman, 2013)." is surely not mainstream thinking about NHST; I would surely delete that sentence. In NHST, a p-value is used for testing the H0. Why did you not yet discuss significance level? Yes, before discussing what is not a p-value, I would explain NHST (i.e., what it is and how it is used).

Also the next sentence "The more (a priori) implausible the alternative hypothesis, the greater the chance that a finding is a false alarm (Krzywinski & Altman, 2013; Nuzzo, 2014)." is not fully clear to me. This is a Bayesian statement. In NHST, no likelihoods are attributed to hypotheses; the reasoning is "IF H0 is true, then...".

Last sentence: "As Nickerson (2000) puts it 'theory corroboration requires the testing of multiple predictions because the chance of getting statistically significant results for the wrong reasons in any given case is high'." What is relation of this sentence to the contents of this section, precisely?

Next section: "For instance, we can estimate that the probability of a given F value to be in the critical interval [+2 +∞] is less than 5%" This depends on the degrees of freedom.

"When there is no effect (H0 is true), the erroneous rejection of H0 is known as type I error and is equal to the p-value." Strange sentence. The Type I error is the probability of erroneously rejecting

the H0 (so, when it is true). The p-value is … well, you explained it before; it surely does not equal the Type I error.

Consider adding a figure explaining the distinction between Fisher's logic and that of Neyman and Pearson.

"When the test statistics falls outside the critical region(s)" What is outside?

"There is a profound difference between accepting the null hypothesis and simply failing to reject it (Killeen, 2005)" I agree with you, but perhaps you may add that some statisticians simply define "accept H0"' as obtaining a p-value larger than the significance level. Did you already discuss the significance level, and it's mostly used values?

"To accept or reject equally the null hypothesis, Bayesian approaches (Dienes, 2014; Kruschke, 2011) or confidence intervals must be used." Is 'reject equally' appropriate English? Also using Cis, one cannot accept the H0.

Do you start discussing alpha only in the context of Cis?

"CI also indicates the precision of the estimate of effect size, but unless using a percentile bootstrap approach, they require assumptions about distributions which can lead to serious biases in particular regarding the symmetry and width of the intervals (Wilcox, 2012)." Too difficult, using new concepts. Consider deleting.

"Assuming the CI (a)symmetry and width are correct, this gives some indication about the likelihood that a similar value can be observed in future studies, with 95% CI giving about 83% chance of replication success (Lakens & Evers, 2014)." This statement is, in general, completely false. It very much depends on the sample sizes of both studies. If the replication study has a much, much, much larger N, then the probability that the original CI will contain the effect size of the replication approaches (1-alpha)*100%. If the original study has a much, much, much larger N, then the probability that the original Ci will contain the effect size of the replication study approaches 0%.

"Finally, contrary to p-values, CI can be used to accept H0. Typically, if a CI includes 0, we cannot reject H0. If a critical null region is specified rather than a single point estimate, for instance [-2 +2] and the CI is included within the critical null region, then H0 can be accepted. Importantly, the critical region must be specified a priori and cannot be determined from the data themselves." No. H0 cannot be accepted with Cis.

"The (posterior) probability of an effect can however not be obtained using a frequentist framework." Frequentist framework? You did not discuss that, yet.

"X% of times the CI obtained will contain the same parameter value". The same? True, you mean?

"e.g. X% of the times the CI contains the same mean" I do not understand; which mean?

"The alpha value has the same interpretation as when using H0, i.e. we accept that 1-alpha CI are wrong in alpha percent of the times. " What do you mean, CI are wrong? Consider rephrasing.

"To make a statement about the probability of a parameter of interest, likelihood intervals (maximum likelihood) and credibility intervals (Bayes) are better suited." ML gives the likelihood of the data given the parameter, not the other way around.

"Many of the disagreements are not on the method itself but on its use." Bayesians may disagree.

"If the goal is to establish the likelihood of an effect and/or establish a pattern of order, because both requires ruling out equivalence, then NHST is a good tool (Frick, 1996)" NHST does not provide evidence on the likelihood of an effect.

"If the goal is to establish some quantitative values, then NHST is not the method of choice." P-values are also quantitative... this is not a precise sentence. And NHST may be used in combination with effect size estimation (this is even recommended by, e.g., the American Psychological Association (APA)).

"Because results are conditioned on H0, NHST cannot be used to establish beliefs." It can reinforce some beliefs, e.g., if H0 or any other hypothesis, is true.

"To estimate the probability of a hypothesis, a Bayesian analysis is a better alternative." It is the only alternative?

"Note however that even when a specific quantitative prediction from a hypothesis is shown to be true (typically testing H1 using Bayes), it does not prove the hypothesis itself, it only adds to its plausibility." How can we *show* something is true?

I do not agree on the contents of the last section on 'minimal reporting'. I prefer 'optimal reporting' instead, i.e., the reporting the information that is essential to the interpretation of the result, to any ready, which may have other goals than the writer of the article. This reporting includes, for sure, an estimate of effect size, and preferably a confidence interval, which is in line with recommendations of the APA.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 06 Jul 2016

Cyril Pernet, The University of Edinburgh, Edinburgh, UK

- *Null hypothesis significance testing (NHST) is a difficult topic, with misunderstandings arising easily. Many texts, including basic statistics books, deal with the topic, and attempt to explain it to students and anyone else interested. I would refer to a good basic text book, for a detailed explanation of NHST, or to a specialized article when wishing an explaining the background of NHST. So, what is the added value of a new text on NHST? In any case, the added value should be described at the start of this text. Moreover, the topic*

is so delicate and difficult that errors, misinterpretations, and disagreements are easy. I attempted to show this by giving comments to many sentences in the text.

The idea of this short review was to point to common interpretation errors (stressing again and again that we are under H0) being in using p-values or CI, and also proposing reporting practices to avoid bias. This is now stated at the end of abstract.

Regarding text books, it is clear that many fail to clearly distinguish Fisher/Pearson/NHST, see Glinet et al (2012) J. Exp Education 71, 83-92. If you have 1 or 2 in mind that you know to be good, I'm happy to include them.

- *Abstract: "null hypothesis significance testing is the statistical method of choice in biological, biomedical and social sciences to investigate if an effect is likely". No, NHST is the method to test the hypothesis of no effect.*

  I agree – yet people use it to investigate (not test) if an effect is likely. The issue here is wording. What about adding this distinction at the end of the sentence?: 'null hypothesis significance testing is the statistical method of choice in biological, biomedical and social sciences used to investigate if an effect is likely, even though it actually tests for the hypothesis of no effect'.

- *Intro: "Null hypothesis significance testing (NHST) is a method of statistical inference by which an observation is tested against a hypothesis of no effect or no relationship." What is an 'observation'? NHST is difficult to describe in one sentence, particularly here. I would skip this sentence entirely, here.*

  I think a definition is needed, as it offers a starting point. What about the following: 'NHST is a method of statistical inference by which an experimental factor is tested against a hypothesis of no effect or no relationship based on a given observation'

- *Section on Fisher; also explain the one-tailed test.*
  Section on Fisher; p(Obs|H0) does not reflect the verbal definition (the 'or more extreme' part).
  Section on Fisher; use a reference and citation to Fisher's interpretation of the p-value
  Section on Fisher; "This was however only intended to be used as an indication that there is something in the data that deserves further investigation. The reason for this is that only H0 is tested whilst the effect under study is not itself being investigated." First sentence, can you give a reference? Many people say a lot about Fisher's intentions, but the good man is dead and cannot reply... Second sentence is a bit awkward, because the effect <u>is</u> investigated in a way, by testing the H0.

  The section on Fisher has been modified (more or less) as suggested: (1) avoiding talking about one or two tailed tests (2) updating for $p(Obs \geq t|H0)$ and (3) referring to Fisher more explicitly (ie pages from articles and book) ; I cannot tell his intentions but these quotes leave little space to alternative interpretations.

- *Section on p-value; Layout and structure can be improved greatly, by first again stating*

what the p-value is, and then statement by statement, what it is not, using separate lines for each statement. Consider adding that the p-value is randomly distributed under H0 (if all the assumptions of the test are met), and that under H1 the p-value is a function of population effect size and N; the larger each is, the smaller the p-value generally is.

Done

- *Skip the sentence "If there is no effect, we should replicate the absence of effect with a probability equal to 1-p". Not insightful, and you did not discuss the concept 'replicate' (and do not need to).*

  Done

- *Skip the sentence "The total probability of false positives can also be obtained by aggregating results (Ioannidis, 2005)." Not strongly related to p-values, and introduces unnecessary concepts 'false positives' (perhaps later useful) and 'aggregation'.*

  Done

- *Consider deleting; "If there is an effect however, the probability to replicate is a function of the (unknown) population effect size with no good way to know this from a single experiment (Killeen, 2005)."*

  Done

- *The following sentence; " Finally, a (small) p-value is not an indication favouring a hypothesis. A low p-value indicates a misfit of the null hypothesis to the data and cannot be taken as evidence in favour of a specific alternative hypothesis more than any other possible alternatives such as measurement error and selection bias (Gelman, 2013)." is surely not mainstream thinking about NHST; I would surely delete that sentence. In NHST, a p-value is used for testing the H0. Why did you not yet discuss significance level? Yes, before discussing what is not a p-value, I would explain NHST (i.e., what it is and how it is used).*

  Also the next sentence "The more (a priori) implausible the alternative hypothesis, the greater the chance that a finding is a false alarm (Krzywinski & Altman, 2013; Nuzzo, 2014)." is not fully clear to me. This is a Bayesian statement. In NHST, no likelihoods are attributed to hypotheses; the reasoning is "IF H0 is true, then…".

  The reasoning here is as you state yourself, part 1: 'a p-value is used for testing the H0; and part 2: 'no likelihoods are attributed to hypotheses' it follows we cannot favour a hypothesis. It might seems contentious but this is the case that all we can is to reject the null – how could we favour a specific alternative hypothesis from there? This is explored further down the manuscript (and I now point to that) – note that we do not need to be Bayesian to favour a specific H1, all I'm saying is this cannot be attained with a p-value.

- *Last sentence: "As Nickerson (2000) puts it 'theory corroboration requires the testing of multiple predictions because the chance of getting statistically significant results for the wrong reasons in any given case is high'." What is relation of this sentence to the contents of this section, precisely?*

  The point was to emphasise that a p value is not there to tell us a given H1 is true and can only be achieved through multiple predictions and experiments. I deleted it for clarity.

- *Next section: "For instance, we can estimate that the probability of a given F value to be in the critical interval [+2 +∞]* is less than 5%" This depends on the degrees of freedom.

  This sentence has been removed

- *"When there is no effect (H0 is true), the erroneous rejection of H0 is known as type I error and is equal to the p-value." Strange sentence. The Type I error is the probability of erroneously rejecting the H0 (so, when it is true). The p-value is … well, you explained it before; it surely does not equal the Type I error.*

  Indeed, you are right and I have modified the text accordingly. When there is no effect (H0 is true), the erroneous rejection of H0 is known as type 1 error. Importantly, the type 1 error rate, or alpha value is determined a priori. It is a common mistake but the level of significance (for a given sample) is not the same as the frequency of acceptance alpha found on repeated sampling (Fisher, 1955).

- *Consider adding a figure explaining the distinction between Fisher's logic and that of Neyman and Pearson.*

  A figure is now presented – with levels of acceptance, critical region, level of significance and p-value.

- "*When the test statistics falls outside the critical region(s)" What is outside?*

  "There is a profound difference between accepting the null hypothesis and simply failing to reject it (Killeen, 2005)" I agree with you, but perhaps you may add that some statisticians simply define "accept H0"' as obtaining a p-value larger than the significance level. Did you already discuss the significance level, and it's mostly used values?

  "To accept or reject equally the null hypothesis, Bayesian approaches (Dienes, 2014; Kruschke, 2011) or confidence intervals must be used." Is 'reject equally' appropriate English? Also using Cis, one cannot accept the H0.

  I should have clarified further here – as I was having in mind tests of equivalence. To clarify, I simply states now: 'To accept the null hypothesis, tests of equivalence or Bayesian approaches must be used.'

- *Do you start discussing alpha only in the context of Cis?*

It is now presented in the paragraph before.

- *"CI also indicates the precision of the estimate of effect size, but unless using a percentile bootstrap approach, they require assumptions about distributions which can lead to serious biases in particular regarding the symmetry and width of the intervals (Wilcox, 2012)." Too difficult, using new concepts. Consider deleting.*

  Done

- *"Assuming the CI (a)symmetry and width are correct, this gives some indication about the likelihood that a similar value can be observed in future studies, with 95% CI giving about 83% chance of replication success (Lakens & Evers, 2014)." This statement is, in general, completely false. It very much depends on the sample sizes of both studies. If the replication study has a much, much, much larger N, then the probability that the original CI will contain the effect size of the replication approaches (1-alpha)*100%. If the original study has a much, much, much larger N, then the probability that the original Ci will contain the effect size of the replication study approaches 0%.*

  Yes, you are right, I completely overlooked this problem. The corrected sentence (with more accurate ref) is now *"Assuming the CI (a)symmetry and width are correct, this gives some indication about the likelihood that a similar value can be observed in future studies. For future studies of the same sample size, 95% CI giving about 83% chance of replication success (Cumming and Mallardet, 2006). If sample sizes differ between studies, CI do not however warranty any a priori coverage".*

- *"Finally, contrary to p-values, CI can be used to accept H0. Typically, if a CI includes 0, we cannot reject H0. If a critical null region is specified rather than a single point estimate, for instance [-2 +2] and the CI is included within the critical null region, then H0 can be accepted. Importantly, the critical region must be specified a priori and cannot be determined from the data themselves." No. H0 cannot be accepted with Cis.*

  Again, I had in mind equivalence testing, but in both cases you are right we can only reject and I therefore removed that sentence.

- *"The (posterior) probability of an effect can however not be obtained using a frequentist framework." Frequentist framework? You did not discuss that, yet.*

  Removed

- *"X% of times the CI obtained will contain the same parameter value". The same? True, you mean?*

  "e.g. X% of the times the CI contains the same mean" I do not understand; which mean?

  "The alpha value has the same interpretation as when using H0, i.e. we accept that 1-alpha CI are wrong in alpha percent of the times. " What do you mean, CI are wrong? Consider

rephrasing.

"To make a statement about the probability of a parameter of interest, likelihood intervals (maximum likelihood) and credibility intervals (Bayes) are better suited." ML gives the likelihood of the data given the parameter, not the other way around.

corrected

- "*Many of the disagreements are not on the method itself but on its use." Bayesians may disagree.*

  removed

- *"If the goal is to establish the likelihood of an effect and/or establish a pattern of order, because both requires ruling out equivalence, then NHST is a good tool (Frick, 1996)" NHST does not provide evidence on the likelihood of an effect.*

  "If the goal is to establish some quantitative values, then NHST is not the method of choice." P-values are also quantitative… this is not a precise sentence. And NHST may be used in combination with effect size estimation (this is even recommended by, e.g., the American Psychological Association (APA)).

  Yes, p-values must be interpreted in context with effect size, but this is not what people do. The point here is to be pragmatic, does and don't. The sentence was changed.

- *"Because results are conditioned on H0, NHST cannot be used to establish beliefs." It can reinforce some beliefs, e.g., if H0 or any other hypothesis, is true.*

  "To estimate the probability of a hypothesis, a Bayesian analysis is a better alternative." It is the only alternative?

  Not for testing, but for probability, I am not aware of anything else.

- *"Note however that even when a specific quantitative prediction from a hypothesis is shown to be true (typically testing H1 using Bayes), it does not prove the hypothesis itself, it only adds to its plausibility." How can we show something is true?*

  Cumulative evidence is, in my opinion, the only way to show it. Even in hard science like physics multiple experiments. In the recent CERN study on finding Higgs bosons, 2 different and complementary experiments ran in parallel – and the cumulative evidence was taken as a proof of the true existence of Higgs bosons.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 October 2015

https://doi.org/10.5256/f1000research.7499.r10159

Daniel Lakens
School of Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

I appreciate the author's attempt to write a short tutorial on NHST. Many people don't know how to use it, so attempts to educate people are always worthwhile. However, I don't think the current article reaches it's aim. For one, I think it might be practically impossible to explain a lot in such an ultra short paper - every section would require more than 2 pages to explain, and there are many sections. Furthermore, there are some excellent overviews, which, although more extensive, are also much clearer (e.g., Nickerson, 2000). Finally, I found many statements to be unclear, and perhaps even incorrect (noted below). Because there is nothing worse than creating more confusion on such a topic, I have extremely high standards before I think such a short primer should be indexed. I note some examples of unclear or incorrect statements below. I'm sorry I can't make a more positive recommendation.

“investigate if an effect is likely” – ambiguous statement. I think you mean, whether the observed DATA is probable, assuming there is no effect?

The Fisher (1959) reference is not correct – Fischer developed his method much earlier.

“This p-value thus reflects the conditional probability of achieving the observed outcome or larger, p(Obs|H0)” – please add 'assuming the null-hypothesis is true'.

“p(Obs|H0)” – explain this notation for novices.

“Following Fisher, the smaller the p-value, the greater the likelihood that the null hypothesis is false.” This is wrong, and any statement about this needs to be much more precise. I would suggest direct quotes.

“there is something in the data that deserves further investigation” –unclear sentence.

“The reason for this” – unclear what ‘this’ refers to.

“*not the probability of the null hypothesis of being true, p(H0)”* – second of can be removed?

“Any interpretation of the p-value in relation to the effect under study (strength, reliability, probability) is indeed
wrong, since the p-value is conditioned on H0” - incorrect. A big problem is that it depends on the sample size, and that the probability of a theory depends on the prior.

"If there is no effect, we should replicate the absence of effect with a probability equal to 1-p." I don't understand this, but I think it is incorrect.

"The total probability of false positives can also be obtained by aggregating results (Ioannidis, 2005)." Unclear, and probably incorrect.

"By failing to reject, we simply continue to assume that H0 is true, which implies that one cannot, from a nonsignificant result, argue against a theory" – according to which theory? From a NP perspective, you can ACT as if the theory is false.

"(Lakens & Evers, 2014") – we are not the original source, which should be cited instead.

" Typically, if a CI includes 0, we cannot reject H0." - when would this not be the case? This assumes a CI of 1-alpha.

"If a critical null region is specified rather than a single point estimate, for instance [-2 +2] and the CI is included within the critical null region, then H0 can be accepted." – you mean practically, or formally? I'm pretty sure only the former.

The section on 'The (correct) use of NHST' seems to conclude only Bayesian statistics should be used. I don't really agree.

"we can always argue that effect size, power, etc. must be reported." – which power? Post-hoc power? Surely not? Other types are unknown. So what do you mean?

The recommendation on what to report remains vague, and it is unclear why what should be reported.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 06 Jul 2016

Cyril Pernet, The University of Edinburgh, Edinburgh, UK

- *"investigate if an effect is likely" – ambiguous statement. I think you mean, whether the observed DATA is probable, assuming there is no effect?*

  This sentence was changed, following as well the other reviewer, to 'null hypothesis significance testing is the statistical method of choice in biological, biomedical and social sciences to investigate if an effect is likely, even though it actually tests whether the observed data are probable, assuming there is no effect'

- *The Fisher (1959) reference is not correct – Fischer developed his method much earlier.*

Changed, refers to Fisher 1925

- *"This p-value thus reflects the conditional probability of achieving the observed outcome or larger, p(Obs|H0)" – please add 'assuming the null-hypothesis is true'. "p(Obs|H0)" – explain this notation for novices.*

  I changed a little the sentence structure, which should make explicit that this is the condition probability.

- *"Following Fisher, the smaller the p-value, the greater the likelihood that the null hypothesis is false." This is wrong, and any statement about this needs to be much more precise. I would suggest direct quotes.*

  This sentence has been removed

- *"there is something in the data that deserves further investigation" –unclear sentence. "The reason for this" – unclear what 'this' refers to.*

  This has been changed to '[...] to decide whether the evidence is worth additional investigation and/or replication (Fisher, 1971 p13)'

- *"not the probability of the null hypothesis of being true, p(H0)" – second of can be removed?*

  my mistake – the sentence structure is now '*not the probability of the null hypothesis p(H0), of being true,' ; hope this makes more sense (and this way refers back to p(Obs>t|H0)*

- *"Any interpretation of the p-value in relation to the effect under study (strength, reliability, probability) is indeed wrong, since the p-value is conditioned on H0" - incorrect. A big problem is that it depends on the sample size, and that the probability of a theory depends on the prior.*

  Fair enough – my point was to stress the fact that p value and effect size or H1 have very little in common, but yes that the part in common has to do with sample size. I left the conditioning on H0 but also point out the dependency on sample size.

- *"If there is no effect, we should replicate the absence of effect with a probability equal to 1-p." I don't understand this, but I think it is incorrect.*

  Removed

- *"The total probability of false positives can also be obtained by aggregating results (Ioannidis, 2005)." Unclear, and probably incorrect.*

  Removed

- *"By failing to reject, we simply continue to assume that H0 is true, which implies that one cannot, from a nonsignificant result, argue against a theory" – according to which theory? From a NP perspective, you can ACT as if the theory is false.*

  The whole paragraph was changed to reflect a more philosophical take on scientific induction/reasoning. I hope this is clearer.

- *"(Lakens & Evers, 2014") – we are not the original source, which should be cited instead.*

  done

- *" Typically, if a CI includes 0, we cannot reject H0." - when would this not be the case? This assumes a CI of 1-alpha. "If a critical null region is specified rather than a single point estimate, for instance [-2 +2] and the CI is included within the critical null region, then H0 can be accepted." – you mean practically, or formally? I'm pretty sure only the former.*

  Changed to refer to equivalence testing

- *The section on 'The (correct) use of NHST' seems to conclude only Bayesian statistics should be used. I don't really agree.*

  I rewrote this, as to show frequentist analysis can be used - I'm trying to sell Bayes more than any other approach.

- ***"**we can always argue that effect size, power, etc. must be reported." – which power? Post-hoc power? Surely not? Other types are unknown. So what do you mean? The recommendation on what to report remains vague, and it is unclear why what should be reported.*

  I'm arguing we should report it all, that's why there is no exhausting list – I can if needed.

Competing Interests: No competing interests were disclosed.