OPINION ARTICLE

Null hypothesis significance testing: a short tutorial [version 1; peer review: 2 not approved]

Cyril Pernet

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

V1 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) is the statistical method of choice in biological, biomedical and social sciences to investigate if an effect is likely. In this short tutorial, I first summarize the concepts behind the method while pointing to common interpretation errors. I then present the related concepts of confidence intervals, and discuss what should be reported in which context. The goal is to clarify concepts, present statistical issues that researchers face using the NHST framework and propose reporting practices.

Keywords

null hypothesis significance testing , tutorial , p-value , reporting

Open Peer Review**Approval Status**

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

1. **Daniel Lakens** , Eindhoven University of Technology, Eindhoven, The Netherlands
2. **Marcel ALM van Assen**, Tilburg 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)

Competing interests: No competing interests were disclosed.

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

Copyright: © 2015 Pernet C. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Pernet C. **Null hypothesis significance testing: a short tutorial [version 1; peer review: 2 not approved]** F1000Research 2015, 4:621 <https://doi.org/10.12688/f1000research.6963.1>

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

The Null Hypothesis Significance Testing framework

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. The method as practiced nowadays is a combination of the concepts of critical rejection regions developed by [Neyman & Pearson \(1933\)](#) and the p-value developed by [Fisher \(1959\)](#).

Fisher, significance testing, and the p-value

The method developed by [Fisher \(1959\)](#) allows computation of the probability of observing a result at least as extreme as a test statistic (e.g. t value), assuming the null hypothesis is true. This p-value thus reflects the conditional probability of achieving the observed outcome or larger, $p(\text{Obs}|\text{H}_0)$, and is equal to the area under the null probability distribution curve, for example $[-\infty -t]$ and $[+t +\infty]$ for a two-tailed t-test ([Turkheimer et al., 2004](#)). Following Fisher, the smaller the p-value, the greater the likelihood that the null hypothesis is false. 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 H_0 is tested whilst the effect under study is not itself being investigated.

What is not a p-value?

The p-value is *not* the probability of the null hypothesis of being true, $p(\text{H}_0)$ ([Krzywinski & Altman, 2013](#)). This common misconception arises from confusion between the probability of an observation given the null $p(\text{Obs}|\text{H}_0)$ and the probability of the null given an observation $p(\text{H}_0|\text{Obs})$ (see [Nickerson \(2000\)](#) for a detailed demonstration). 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 indeed wrong, since the p-value is conditioned on H_0 . Similarly, $1-p$ is *not* the probability to replicate 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 ascertained from the p-value because it is not informative on the effect itself ([Miller, 2009](#)). If there is no effect, we should replicate the absence of effect with a probability equal to $1-p$. The total probability of false positives can also be obtained by aggregating results ([Ioannidis, 2005](#)). 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](#)). 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](#)). The more (a priori) implausible the alternative hypothesis, the greater the chance that a finding is a false alarm ([Krzywinski & Altman, 2013](#); [Nuzzo, 2014](#)). 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’.

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

[Neyman & Pearson \(1933\)](#) introduced the notion of critical intervals over which the probability of observing a test statistic is less than a stipulated significance level, α . If the statistic value falls within those intervals, it is deemed significantly different from that

expected under the null hypothesis. For instance, we can estimate that the probability of a given F value to be in the critical interval $[+2 +\infty]$ is less than 5%. Because the space of results is dichotomized, we can distinguish correct results (rejecting H_0 when there is an effect and not rejecting H_0 when there is no effect) from errors (rejecting H_0 when there is no effect and not rejecting H_0 when there is an effect). When there is no effect (H_0 is true), the erroneous rejection of H_0 is known as type I error and is equal to the p-value.

Acceptance or failure to reject H_0 ?

The significance level α is defined to be the maximum probability that a test statistic falls into the rejection region when the null hypothesis is true ([Johnson, 2013](#)). When the test statistics falls outside the critical region(s), all we can say is that no significant effect is observed, but one cannot conclude that the null hypothesis is true, i.e. we cannot accept H_0 . 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 H_0 is true, which implies that one cannot, from a non-significant result, argue against a theory. We cannot accept the null hypothesis, because all we have done is not disprove it. To accept or reject equally the null hypothesis, Bayesian approaches ([Dienes, 2014](#); [Kruschke, 2011](#)) or confidence intervals must be used.

Confidence intervals

Confidence intervals (CI) have been advocated as alternatives to p-values because (i) they allow judging the statistical significance and (ii) they provide estimates of effect size. CI fail to cover the true value at a rate of α , the type I error rate ([Morey & Rouder, 2011](#)) and therefore indicate if values can be rejected by a two tailed test with a given α . 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](#)). 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](#)). Finally, contrary to p-values, CI can be used to accept H_0 . Typically, if a CI includes 0, we cannot reject H_0 . 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 H_0 can be accepted. Importantly, the critical region must be specified a priori and cannot be determined from the data themselves.

Although CI provide more information, they are not less subject to interpretation errors (see [Savalei & Dunn, 2015](#) for a review). People often interpret X% CI as the probability that a parameter (e.g. the mean) will fall into that interval X% of the time. The (posterior) probability of an effect can however not be obtained using a frequentist framework. The CI represents the bounds for which one has X% confidence. 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 same parameter value, e.g. X% of the times the CI contains the same mean ([Tan & Tan, 2010](#)). The alpha value has the same interpretation as when using H_0 , i.e. we accept that $1-\alpha$ CI are wrong in alpha percent of the

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, likelihood intervals (maximum likelihood) and credibility intervals (Bayes) are better suited.

The (correct) use of NHST

NHST has always been criticized, and yet is still used every day in scientific reports (Nickerson, 2000). Many of the disagreements are not on the method itself but on its use. The question one should ask is what is the goal of a scientific experiment at hand? 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). If the goal is to establish some quantitative values, then NHST is not the method of choice. Because results are conditioned on H0, NHST cannot be used to establish beliefs. To estimate the probability of a hypothesis, a Bayesian analysis is a better alternative. To estimate parameters (point estimates and variances), alternative approaches are also better suited. 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.

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 effect size, power, etc. 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. Here I propose to adopt minimal reporting in the result section to keep the message clear, but have detailed supplementary material. When the hypothesis is about the presence/absence or order of an effect, 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 with confidence, likelihood or credible intervals depending on the question at hand. The reasoning is simply that one cannot predict and/or discuss quantities without accounting for variability. For the reader to understand and fully appreciate the results, nothing else is needed.

Because science progress is obtained by cumulating evidence (Rosenthal, 1991), scientists should also consider the secondary use of the data. With today's electronic articles, there are no reasons for not including all of derived data: mean, standard deviations, effect size, CI, Bayes factor should always be included 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 of the tests performed (all t, F, and p values) because of the increase in type I error rate due to selective reporting (multiple comparisons problem - 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.

Competing interests

No competing interests were disclosed.

Grant information

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

References

- 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 and scientific inference.** (2nd ed.). New-York: Hafner Publishing, 1959.
[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](#)
- Hoekstra R, Morey RD, Roudier JN, *et al.*: **Robust misinterpretation of confidence intervals.** *Psychon Bull Rev.* 2014; 21(5): 1157–1164.
[PubMed Abstract](#) | [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, 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](#)
- 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, Roudier 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 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. **Estimating the reproducibility of psychological science.** *Science*. in press. 2015.

[Reference Source](#)

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](#)

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](#)

Wilcox R: **Introduction to Robust Estimation and Hypothesis Testing.** (3rd ed.). Oxford, UK: Academic Press, Elsevier, 2012.

[Reference Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 10 November 2015

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

© 2015 van Assen M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Marcel ALM van Assen

Department of Methodology and Statistics, Tilburg 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; $p(\text{Obs} | H_0)$ 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 H_0 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 $[+2, +\infty]$ 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 H_0 " 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 H_0 .

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 H_0 . Typically, if a CI includes 0, we cannot reject H_0 . 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 H_0 can be accepted. Importantly, the critical region must be specified a priori and cannot be determined from the data themselves." No. H_0 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 H_0 , i.e. we accept that 1-alpha CIs are wrong in alpha percent of the times." What do you mean, CIs 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 H_0 , NHST cannot be used to establish beliefs." It can reinforce some beliefs, e.g., if H_0 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 H_1 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 H_0) 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; $p(\text{Obs} | H_0)$ 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 H_0 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 H_0 .

The section on Fisher has been modified (more or less) as suggested: (1) avoiding talking about one or two tailed tests (2) updating for $p(\text{Obs} \geq t | H_0)$ 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 H_0 (if all the assumptions of the test are met), and that under H_1 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 H_0 . 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 H_0 is true, then..."

The reasoning here is as you state yourself, part 1: 'a p -value is used for testing the H_0 '; and part 2: 'no likelihoods are attributed to hypotheses' it follows we cannot favour a hypothesis. It might seem contentious but this is the case that all we can do 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 H_1 , 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 H_1 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 $[+2 +\infty]$ is less than 5%" This depends on the degrees of freedom.*

This sentence has been removed

- *"When there is no effect (H_0 is true), the erroneous rejection of H_0 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 H_0 (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 (H_0 is true), the erroneous rejection of H_0 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 H_0 " 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 H_0 .

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 Mallard, 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 H_0 . Typically, if a CI includes 0, we cannot reject H_0 . 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 H_0 can be accepted. Importantly, the critical region must be specified a priori and cannot be determined from the data themselves." No. H_0 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 H_0 , 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 H_0 , NHST cannot be used to establish beliefs." It can reinforce some beliefs, e.g., if H_0 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 H_1 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>

© 2015 Lakens D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 its 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, $p(\text{Obs} | H_0)$ " – please add 'assuming the null-hypothesis is true'.

" $p(\text{Obs} | H_0)$ " – 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, $p(H_0)$* " – 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 H_0 " – 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 H_0 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 H_0 .” – 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 H_0 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, $p(\text{Obs} | H_0)$ ” – please add ‘assuming the null-hypothesis is true’. “ $p(\text{Obs} | H_0)$ ” – 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, $p(H_0)$ " – second of can be removed?*

my mistake – the sentence structure is now '*not the probability of the null hypothesis $p(H_0)$, of being true,*' ; hope this makes more sense (and this way refers back to $p(\text{Obs} > t | H_0)$)

- *"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 H_0 " - 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 H_1 have very little in common, but yes that the part in common has to do with sample size. I left the conditioning on H_0 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 H_0 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 H_0 .” - 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 H_0 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.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research