Null Hypothesis Significance Testing: a short tutorial

Although thoroughly criticized, null hypothesis significance testing 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, effect size, and Bayesian factor, 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 highlight good practices.

Null Hypothesis Significance Testing:
a short tutorial

Cyril R. Pernet

Centre for Clinical Brain Sciences, NeuroImaging Sciences,
The University of Edinburgh

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 highlight good practices.

The Null Hypothesis Significance Testing framework

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 and Pearson (1933) and p-value developed by Fisher (1959).

Fisher, significance testing, and the p-value

The method developed by Fisher (1959) allows to compute 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(Obs|H0), and is equal to the area under the null probability distribution curve in e.g. [- ∞ -t] and [+t +∞] for a two-tailed t-test (Turkheimer, Aston, & Cunningham, 2004). Following Fisher, the smaller is the p-value, the greater is the likelihood that the null hypothesis is false. The p-value however only allows to judge whether the evidence is significant in the sense of worth further investigation. The reason for this is that only H0 is tested whilst the effect under study has not itself been investigated.

What is not a p-value?

The p-value *is not the probability of the null hypothesis of being true, p(H0)* (Krzywinski & Altman, 2013). This common misconception arises from a confusion between the probability of an observation given the null p(Obs|H0) and the probability of the null given an observation p(H0|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 H0. 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 obtained because the p-value 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 positive can also be obtained by aggregating results (Ioannidis, 2005). If there is an effect however, the probability to replicate is function of the (unknown) population effect size with no good ways 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). 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 given F value to be in the critical interval [+2 +∞] is less than 5%. Because the space of results is dichotomized, we can 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 and not rejecting H0 when there is an effect). The erroneous rejection of H0 when there is no effect is known as type I error and corresponds to the p-value.

Acceptance or rejection of H0?

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). 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, and again one cannot conclude that the null hypothesis is true. This distinction matters because 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, 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) provide estimates of effect size. CI are builds that fail to cover the true value at a rate of alpha, the Type I error rate (Morey & Rouder, 2011) and therefore indicate if values can be rejected by a two tailed test with a given alpha. 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% of replication success (Lakens & Evers, 2014). 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.

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 in 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 as 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 H0, 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, Morey, Rouder, & Wagenmakers, 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, null hypothesis testing is not sufficient for establishing beliefs or estimating the probability of an effect. To estimate the probability that a claim is correct, a Bayesian analysis is a better alternative to null hypothesis testing. 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. A simple solution is to have 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, p values) because of the increase type one 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.

References

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. http://doi.org/10.3389/fpsyg.2014.00781

Fisher, R. A. (1959). *Statistical methods and scientific inference.* (2nd ed.). New-York: Hafner Publishing.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379.

Gelman, A. (2013). P Values and Statistical Practice: *Epidemiology*, *24*(1), 69–72. http://doi.org/10.1097/EDE.0b013e31827886f7

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. http://doi.org/10.3758/s13423-013-0572-3

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Med*, *2*(8), e124. http://doi.org/10.1371/journal.pmed.0020124

Johnson, V. E. (2013). Revised Standards for Statistical Evidence. *Proceedings of the National Academy of Sciences USA*, *110*(48), 19313–19317. http://doi.org/10.1073/pnas.1313476110

Killeen, P. R. (2005). An Alternative to Null-Hypothesis Significance Tests. *Psychological Science*, *16*(5), 345–353. http://doi.org/10.1111/j.0956-7976.2005.01538.x

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. http://doi.org/10.1177/1745691611406925

Krzywinski, M., & Altman, N. (2013). Points of significance: Significance, P values and t-tests. *Nature Methods*, *10*(11), 1041–1042.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00863

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278–292.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617–640. http://doi.org/10.3758/PBR.16.4.617

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 416–419.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A.*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. http://doi.org/10.1037//1082-989X.5.2.241

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, in press*.

Rosenthal, R. (1991). Cumulating psychology: an appreciation of Donald T. Campbell. *Psychological Science*, *2*(4), 213–221.

Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*(245). http://doi.org/10.3389/fpsyg.2015.00245

Tan, S. H., & Tan, S. B. (2010). The Correct Interpretation of Confidence Intervals. *Proceedings of Singapore Healthcare▪ Volume*, *19*(3), 276.

Turkheimer, F. E., Aston, J. A. D., & Cunningham, V. J. (2004). On the logic of hypothesis testing in functional imaging. *European Journal of Nuclear Medicine and Molecular Imaging*, *31*(5), 725–732. http://doi.org/10.1007/s00259-003-1387-7

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