Many *Brain & Language* authors, editors and reviewers confuse significance and *statistical* significance

Abstract

In (xx%) of articles of the most recent (XX) volumes of *B&L*, statistical tests were used in a way that strongly implies that many authors, editors and reviewers falsely understood *statistical* significance as absolute significance. Specifically, completely known populations (such as word length of preselected, closed stimulus lists) are subjected to parameter estimation tests such as the *t*-test; thus, tests *estimating* a population parameter based on a sample are used on an exhaustively *measured* quantity, a population parameter that did not have to be estimated since it is precisely known. Such usage of inference tests indicates that researchers use them to quantify the practical relevance, or (common-language) significance, of a finding via *p*-value calculation - a crass misunderstanding. Researchers, including editors, reviewers and authors, must deepen their understanding of statistics.

Introduction

Consider the following quotes (neither from *Brain & Language*):

> Demographic and clinical characteristics of the two groups of participants are presented in Table 1. The two groups differed significantly in age, $t(50) = 2.87$, $p < .05$ (Joormann, Levens, and Gotlib 2011)

> The CWs across conditions were matched on word length and frequency: log frequency in CELEX (Baayen et al., 1993): HC = 2.71, LC = 2.70, SV = 2.69, F (2,646) = 2.35, p = 0.10.
> (Wang, Zhu, and Bastiaansen 2012)

> For HF words, the mean length was 4.73 letters (SD = .96), and the mean orthographic neighborhood size (N) (Coltheart, Davelaar, Jonasson, & Besner, 1977) was 4.77 (SD = 4.47). LF words had a mean length of 4.78 letters (SD = .85), and a mean N of 4.82 (SD = 4.37). HF and LF words did not differ in length or orthographic neighborhood size (all $t < 1$). The non-word mean N was 3.16 (SD = 3.67).
> (Bangert, Abrams, and Balota 2012)

The t and F statistics and the following p-values are meaningless, less than redundant, and imply that the researchers in question have false intuitions regarding statistical inference tests. Specifically, they imply that *statistical significance* is confused with *practical relevance*; this can be deduced from the observation that a crude parameter *estimation* test was applied to a *known* parameter, which indicates that the researchers misunderstand p-values as *magnitude quantifications*.

I do not wish to single out and ridicule any specific researcher. I freely admit that in the past, I have conducted similar tests, or worse, avoided them out of laziness even though I (lazily) believed them to be appropriate. However, the scientific literature is full of such statements, implying that many researchers, editors and reviewers make the same mistake. Even the volumes of high-quality journals such as *Brain & Language* are full of such misuses of statistics. In the following, I will discuss the appropriate interpretation of such tests, attempt to explain what they are (wrongly) used for instead, and show how prevalent such mistakes are by presenting the number of these mistakes in the last two years of *B&L*.

What's in a p?

Many statistical tests yield unintuitive results. In common academic discourse, the results of

hypothesis tests are referred to as *significance* or *not significant*. Likely, many, possibly most researchers know the proper definition of the terms and the related statistical concepts, such as *p*-values. They are repeated here for the sake of completeness for the example of a two-sample *t*-test:

> Assuming the hypothesis under test (which is that the difference between the two populations is X) is true, the *p*-value gives the probability of obtaining a sample as far or further from X as the present data.

At its heart, the NHST is a crude form of parameter estimation. It presents how surprising a given sample drawn from some population(s) is to those who have assumed a certain value for the population parameter. Although this is not what the null in null-hypothesis significance tests (NHST) refers to, in brain and behavioral sciences, X is near-universally set at 0, so *p* gives the surprise value of the obtained data for anybody who assumes the two populations to be identical. A low p value means that the data is highly surprising to those who assume no difference between the sampled populations. This is typically interpreted as allowing to say with some confidence (often 95%) that the parameter (the difference between the two populations) is not 0. No other entailment is afforded by the *p* value beyond providing some confidence in a crude estimate of the population parameter, and (at least in the Fischerian tradition[1]) only for the case where the test results in a low *p* value, which affords the estimation that the parameter is likely not 0.

Trivially, statistical significance does not entail practical significance (Goodman 2008); a near-zero effect maybe result in extremely small *p*-values if the sample size is high and/or the sample variance is low, and small samples may produce low *p*-values that misrepresent a small, nonexistent or opposing-sign effect (Gelman and Weakliem 2009). The nonequivalence of statistical and practical significance is even emphasised in many textbooks (Gliner, Leech, and Morgan 2002), and might be expected to be common knowledge.

However, it has been shown repeatedly that many, even most researchers, gravely misunderstand statistical concepts such as the meaning of p values (Haller and Krauss 2002)(M.-P. Lecoutre, Poitevineau, and Lecoutre 2003)(Oakes 1986)(Falk and Greenbaum 1995). Typically, it is discussed (Gigerenzer 2004)(Cohen 1994) that researchers intuitively interpret *p*-values as if they reflect the probabilities of hypotheses, that is, as p(Hypothesis | Data) - a Bayesian interpretation (Wagenmakers et al. 2008). *P*-values do not inform us about the probability of a hypothesis given the data, but about the probability of the data given some hypothesis, and as Cohen (1994) discusses, the *modus tollens* leading from p(Data | Hypothesis) to p(Hypothesis | Data) fails due to the probabilistic nature of the statement.
This is not the phenomenon to be discussed here. Instead, the present essay investigates the following problem: researchers use *parameter estimators* on known parameters, likely because they are misinterpreting these parameter estimators as *magnitude quantificators*. Such an interpretation is indefensible, wrong, and must be discontinued.
A related problem is that researchers are also known to infer from a test that fails to reject the null hypothesis that the null hypothesis is true. This is wrong (Cohen 1994), but a precondition of the misuse of statistics in the way discussed here.

Consider the following situation[2]: I wish to infer if I am significantly, or substantially, taller than my sister. I measure my height (6') and my sister's height (5'7). I now know that I am 5 inches taller than my sister. It

[1] In a Neyman-Pearson framework, entirely different problems become relevant (Gigerenzer 2004). I acknowledge that the present approach may at times not treat the fundamental differences between Neyman-Pearson and Fischerian concepts with enough care.

[2] A similar example has been used by Lindquist et al. (2013)

would make no sense at all to apply a *t*-test to these two values. Even if a low *p* resulted, all that this would allow me to say is that I am probably not exactly the same height as my sister - which is not very useful since I already have much more precise knowledge of the parameter in question (5“). Neither would an extremely low *p*-value give me any confidence in the statement that I am”extremely" or “a little” taller than my sister; this is not afforded by the *p*-value, which simply allows me to say that likely, I am not exactly the same height as my sister given that the obtained measures would be highly improbable if we were of equal height.
If I wanted to infer if me and my brother are taller than my mother and my sister, it would likewise make little sense to follow the measurement of all of my siblings with a *t*-test. Neither would the *t*-test make sense if I had 14 brothers and 15 sisters and wanted to see which group is, on average, taller, assuming I precisely knew all their heights, since I could simply calculate the two mean heights and precisely measure the difference.
Contrast with the following two situations, where statistical inference may well make sense: I wish to know if I am taller than my sister, yet my ruler is somehow magically inaccurate and adds or subtracts a variable number of inches from every measurement. Here it might make sense to take multiple measurements of my sister and I and calculate a 95% Confidence Interval; this would allow me to roughly estimate the difference in inches (ignoring the superior option of simply standing next to my sister and looking in the mirror). Alternatively, I may wish to infer if men are generally taller than women. While it is possible for me to measure all of my (hypothetical) 29 siblings, it would be practically impossible to measure all men and women; however, a reliable estimate of the parameter might result from measuring a lot of men and women and conducting a statistical test. Cases such as the latter two are where parameter estimations such as the NHST make sense; cases such as the first are not meaningfully solved by ANOVAs.

Researchers estimate known quantities

A *t*-test comparing the length of two lists of 80 word stimuli each is equivalent to the former case in the above example, a *t*-test of 2x80 subject mean reaction times to these words in a classification task is equivalent to the latter.
Some researchers, under compliance by editors and reviewers, use the crude parameter estimator that is the *p*-value in both kinds of situations (see below for the prevalence of this practice). They calculate *p* for the mean difference in ERP amplitude between two sentence types, between certain linguistic abilities for populations differing in some gene, or on reaction times following various stimuli. Here, parameter estimation makes sense because the true parameter is unknown.
However, researchers also calculate *p*-values using parameter estimation tests such as the *t*-Test or the ANOVA for populations that are exhaustively measured (as are my sister and I’s height), such as mean word length for two stimulus lists, or such as certain known parameters in two small, selected populations, like the age of two experimental groups.

So researchers already *know* the parameter (since they in fact must know e.g. the length of all their stimulus words to conduct the test of mean word length difference in the first place), and yet also *crudely estimate* it, with a tool allowing nothing but to say, with limited confidence, that e.g. two groups are not exactly identical with regards to the parameter, without any standardised quantification of the magnitude of the (possibly non-zero) difference.
Statistical inference tests are validly used to infer a population parameter from a limited sample of the population. It is pointless to apply them to an *exhaustive* sample of the total population, where the population parameter must not be estimated since it can be measured.

Interpreting low and high *p*'s

It is probably well-known that a researcher who declares his two groups identical based on a failed test is committing a grave mistake. Failed NHSTs do not at all allow the conclusion that the true parameter is that one under test (typically 0). The most pressing problem in this regard is that the power of tests, the likelihood of correctly rejecting a null hypothesis, is worryingly low in brain and behavioral sciences (Button et al. 2013)(Yarkoni 2009). Even if the difference between two populations is substantial, a NHST may still fail to reject the hypothesis of no difference if only a small number of samples are investigated. However, even beyond that, failed NHSTs do not afford accepting the null with great confidence (Cohen 1994)(Wagenmakers et al. 2008). A very simple demonstration of the fallacy of the test can be seen when considering that the actual status of the null hypothesis is *known*. The researcher estimating word length differences *knows* by what amount the mean word length between groups differs. If the words in one group measure on average 4.1 letters and the words in the other group 4.2, the population parameter is known to be 0.1 letter. For even such a small difference, a non-significant NHST, a test that fails to reject the zero-null hypothesis, is failing to reject a wrong null; in a two-sample *t*-test of the population difference, the null is that the difference is 0, so if the difference is anything but 0, such as 0.1, the null is wrong. Of course, many tests will fail to reject a null wrong by only such a small amount, for no other reason than insufficient power.

A test that may only allows to make with confidence the statement that a difference *does* exist, and never allows to say with confidence that no difference exists, is also most likely fully irrelevant to what the researchers have in mind in such situations, since usually, they want to show the equivalence of two groups. A *p*-value above 0.05 does not give confidence in the conclusion that two groups are equivalent; consider for example the situation of $p = 0.1$, a result clearly non-significant by current conventions. However, p in this case means that only in one out of 10 cases, a zero effect would result in data as extreme as the present observation, which presents more evidence against the null than for it. Of course, concluding that two populations are identical based on any other, even very high, *p*-value is similarly unjustified, for example since the test may simply be underpowered to detect a difference of the magnitude in the target population. So even if it would make sense to estimate known quantities, *t*- and *F*-family tests would still not be the appropriate tool for any researcher aiming to establish the equivalence of two groups regarding a parameter since it can never actually prove equivalence and is, at least with conventional type I thresholds, biased against non-equivalence.

Statistical methods for estimating population equivalence from samples do exist. Necessarily, they are not useful for *known* quantities; however, I will discuss them here briefly.

In the frequentist framework, researchers interested in the equivalence between two treatments may pre-define a Region of (practical) Equivalence (Lesaffre 2001) around the null and test if a CI falls entirely within this region; for example, a researcher investigating mean utterance length in men and women may assume that any difference in mean utterance length smaller than 2% is meaningless, and would infer from a 95% CI that falls entirely within [-0.02, 0.02] that likely, no meaningful difference exists and the populations are equivalent in this regard.

Alternatively, researchers may employ methods that explicitly allow to quantify the evidence in favour of one hypothesis, including a null hypothesis versus an alternative hypothesis, leading to the justified acceptance of the null hypothesis; such methods are provided by Bayesian statistics (Wagenmakers et al. 2008).

A researcher who finds a significant difference in word length between two groups and then says, with confidence, that the two groups differ is not committing as grave of

a sign as one who infers equivalence from a high *p*; however, he is still wasting time, print space and computing power to crudely estimate what he already precisely knows. I assume that most researchers are smart people. Why, then, are they practicing such a meaningless ritual?

One answer is that the NHST is indeed ritualistic (Gigerenzer 2004), that it is done with much respect, but little consideration. However, what is the meaning of this ritual? I argue that the most likely explanation is that researcher use the NHST ritual because they somehow believe it to be a measure of effect size - a magnitude estimator.[3]

When mean word length differences of known populations (in contrast to, e.g., mean word length in unknown populations, such as when comparing mean word length in men vs. women) are statistically estimated, the likely common rationale is excluding the relevance of the factor word length for the actual outcomes, such as reaction times or ERP amplitudes in response to these stimuli.

This is not afforded by the test, regardless of its outcome. *P* informs us how unlikely the data are under the null, which at best can be interpreted to give us some confidence in saying the H_0 is false; it entails no qualitative interpretation of the value, and it especially does not in any way help us decide how relevant the difference is. For a quantification of word length, the simple mean word length per group and their standard deviation are vastly superior to the crude estimation of if the parameter is or is not 0. To actually test if this difference in e.g. word length influences the results, I see no other, but also a simple possibility in computing the correlation between e.g. word length and the main outcome.

The direct implications of this error may admittedly be quite trivial.[4] Everybody is free to simply ignore a gratuitous *t* and *p*; if researchers argue that a non-significant inferential test establishes the equivalence of two groups regarding a parameter, it is easy to demonstrate the fallacy of this argument. In the worst case, a study will be cofounded because researchers used stimulus lists not well-controlled because an underpowered test failed to reject a false null.

More problematic might be the entailment that many researchers, editors and reviewers sometimes see *p*-values as measures of the relevance and practical significance of an effect. Since *p*-values are still the primary quantification of research outcomes in brain and behavioral sciences, this means that many researchers, editors and reviewers substantially misunderstand the main tool that is used to evaluate findings.

Regarding quantifications of effect sizes, others (Hentschke and Stüttgen 2011)(Cohen 1994)(Kruschke 2013)(Kline 2004), including the APA (Wilkinson 1999), have convincingly argued that such measures are both critically called for, and readily available in the form of Confidence (or Credible) Intervals, standardised effect sizes, and CIs of standardised effect sizes. However, regardless of the alternatives, researchers must understand, both intellectually and intuitively, the difference between practical and statistical significance.

In no way am I saying that all statistical tests of e.g. word length are wrong. Neither are all statistical tests of stimulus parameters wrong; for example, acceptability ratings of stimuli can be meaningfully subjected to statistical tests that result in parameter estimates (though a test that allows to conclude equivalence, such as Bayesian tests or checking if a 95% CI falls within a pre-defined region of practical equivalence, would be preferable to a test that can only ever reject a null), since acceptability ratings are random samples non-exhaustively drawn from a large population.

[3]There is also a legitimate interpretation of the estimate of e.g. stimulus parameters; it reflects the capabilities of the stimulus selectors to provide adequate stimuli *as a probabilistic measure*. Of course, it is typically of little interest if the people in charge of stimulus creation were *likely* to obtain adequate stimuli; rather, for the interpretation of a study's main outcome, it might be relevant if they did in fact do so.

[4]Though there are also reports of similar misuses of significance tests resulting in the loss of life, for example, in accident statistics influencing public policies (Hauer 2004).

However, parameter *estimation* of *known* parameters (such as a self-created stimulus list) reveal false intuitions regarding conventional statistics. Researchers must understand the difference between p, which is the surprise value of the observed data under the null, and actual measures of effect size which may take a shape such as Cohen's d (Hentschke and Stüttgen 2011). Note that standardised effect sizes (such as Cohen's rule of thumb of "small", "moderate" and "large" effects) do not straight-forwardly allow an assessment of the *relevance* of an effect either (since they are nothing but Cohen's rules of thumb). t-tests inform us how unlikely the parameter is to be zero, CIs inform us about a range of possible values for the effect, standardised effect sizes such as Cohen's d inform us how large a population effect is compared to the population variance; it is up to the researcher to understand and demonstrate how relevant any of these values are for his research questions. Instead of offering clear-cut solutions, one can only recommend researchers to abstain from ritualistic testing of any form, and consider each method's applicability and meaning on a case-by-case basis.

Prevalence of the problem in B&L

The initial three examples for the problem where collected nonsystematically. To estimate how common the problem is in neurolinguistics, a high-quality neurolinguistic journal, *Brain & Language*, was investigated.
Instances of the error can be easily found, not only in recent, but also in older publications:

> the two prime categories were equivalent in text frequency (Carroll et al., 1971), and in length (both t's < 1.1
> (Chiarello, Senehi, and Nuding 1987)

Here, the authors commit the "double sin" of both estimating a known quantity, and deducing equivalence (acceptance of the null) from a failed test - in this case, a test that leads the authors to accept a wrong null hypothesis; the difference in word length in this study was 0.16, a small quantity, but certainly not exactly zero.

Methods

The analysis was restricted to current volumes. For all articles published by B&L in the years 2012 and 2013 as of yet, two independent raters, neither of which was blind to the purpose of the experiment, investigated all published experimental papers (excluding reviews, simulation studies, editorials etc.). For each experiment reported in a study, the stimulus/materials sections were investigated for descriptive and inferential statistics of known quantities. If a descriptive and/or inferential statistic (such as mean and standard deviation) were reported, the study was coded as one where the researchers were interested in a known quantity, otherwise it was discarded. If an inferential statistic (such as a p-value) was reported, the study was coded as one where researchers answered that interest with an erroneous parameter estimate, otherwise as one where researchers did not commit the error. Rater agreement was generally good (yy%); in cases of disagreement, the author made the final call. Representative statements from every study committing the error are presented in the appendix.

Results

In total, N studies where found where researchers reported known quantities in their stimulus/materials section, and M (xx%) of these reported inferential statistics of these known values.
I abstain from computing the statistical significance of this finding; the evaluation of the significance of the findings is left to the reader.

References

Bangert, Ashley S., Richard A. Abrams, and David a Balota. 2012. "Reaching for words

and nonwords: Interactive effects of word frequency and stimulus quality on the characteristics of reaching movements." *Psychonomic bulletin & review* 19 (3) (mar): 513–520.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature reviews. Neuroscience* (apr).

Chiarello, C., J. Senehi, and S. Nuding. 1987. "Semantic priming with abstract and concrete words: differential asymmetry may be postlexical." *Brain and language* 31 (1) (may): 43–60.

Cohen, Jacob. 1994. "The earth is round (p<.05)." *American Psychologist* 49 (12): 997.

Falk, Ruma, and Charles W. Greenbaum. 1995. "Significance Tests Die Hard The Amazing Persistence of a Probabilistic Misconception." *Theory & Psychology* 5 (1): 75–98.

Gelman, Andrew, and David Weakliem. 2009. "Of beauty, sex, and power: statistical challenges in estimating small effects." *American Scientist* 97: 310–316.

Gigerenzer, Gerd. 2004. "Mindless statistics." *Journal of Socio-Economics* 33 (5) (nov): 587–606.

Gliner, Jeffrey A., Nancy L. Leech, and George A. Morgan. 2002. "Problems with null hypothesis significance testing (NHST): What do the textbooks say?" *The Journal of Experimental Education* 71 (1): 83–92.

Goodman, Steven. 2008. "A dirty dozen: twelve p-value misconceptions." *Seminars in hematology* 45 (3) (jul): 135–140.

Haller, Heiko, and Stefan Krauss. 2002. "Misinterpretations of significance: A problem students share with their teachers." *Methods of Psychological Research* 7 (1): 1–20.

Hauer, Ezra. 2004. "The harm done by tests of significance." *Accident Analysis & Prevention* 36 (3) (may): 495–500.

Hentschke, Harald, and Maik C. Stüttgen. 2011. "Computation of measures of effect size for neuroscience data sets." *Eur J Neurosci* 34 (12) (dec): 1887–1894.

Joormann, J., S. M. Levens, and I. H. Gotlib. 2011. "Sticky Thoughts: Depression and Rumination Are Associated With Difficulties Manipulating Emotional Material in Working Memory." *Psychological science* 22 (8) (aug): 979–983.

Kline, Rex B. 2004. "Beyond significance testing: Reforming data analysis methods in behavioral research."

Kruschke, John K. 2013. "Bayesian estimation supersedes the t test." *Journal of experimental psychology. General* 142 (2): 573–603.

Lecoutre, Marie-Paule, Jacques Poitevineau, and Bruno Lecoutre. 2003. "Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests." *International Journal of Psychology* 38 (1): 37–45.

Lesaffre, E. 2001. "The general concepts of an equivalence trial, applied to ASSENT-2, a large-scale mortality study comparing two fibrinolytic agents in acute myocardial infarction." *European Heart Journal* 22 (11) (jun): 898–902.

Lindquist, Martin A., Brian Caffo, and Ciprian Crainiceanu. 2013. "Ironing out the statistical wrinkles in 'ten ironic rules."' *Neuroimage* 81 (nov): 499–502.

Oakes, Michael W. 1986. *Statistical inference: A commentary for the social and behavioural sciences*. New York: Wiley.

Wagenmakers, Eric Jan, Michael Lee, Tom Lodewyckx, and Geoffrey J. Iverson. 2008. "Bayesian versus frequentist inference:" 181–207.

Wang, Lin, Zude Zhu, and Marcel Bastiaansen. 2012. "Integration or predictability? A further specification of the functional role of gamma oscillations in language comprehension." *Front Psychol* 3: 187.

Wilkinson, Leland. 1999. "Statistical methods in psychology journals: guidelines and explanations." *American Psychologist* 54 (8): 594.

Yarkoni, Tal. 2009. "Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al.(2009)." *Perspectives on Psychological Science* 4 (3): 294–298.