Statistics in publishing: the (mis)use of the p-value (Part 2)

Stunt, Jonáh J.; Broekstra, Dieuwke C.; de Boer, Michiel R.

published in
Journal of Hand Surgery: European Volume
2022

DOI (link to publisher)
10.1177/17531934221115968

document version
Publisher's PDF, also known as Version of record

JHS(E)

Journal of Hand Surgery
(European Volume)
2022, Vol. 47(10) 1092–1095

DOI: 10.1177/17531934221115968
journals.sagepub.com/home/jhs
SAGE

Statistics in publishing: the (mis)use of the *p*-value (Part 2)

Introduction

This is the second of a two-part article, in which we discuss problems associated with null hypothesis significance testing (NHST) and the use of the *p*-value. In the first part, we described how NHST, and the *p*-value can be misused, the potential consequences and offered some practical solutions (Broekstra et al., 2022a). To illustrate this, we used a simulated dataset based on a randomized controlled trial (RCT) that investigated women with a distal radial fracture, randomized either to an intervention (cast + rehabilitation programme) or control (cast only) group (Broekstra et al., 2022b). In this second part, we will focus on some important conceptual problems of NHST and the *p*-value and propose an alternative solution to help address these problems.

Interpretation of *p*-values and significance

In medical sciences, we conduct empirical studies hoping that it will enable us to make statements about the effect or effectiveness of a certain treatment or therapy. As mentioned in Part 1, we often start from the idea that one (new) treatment is more effective or has less side effects than another (existing) treatment. We refer to this idea as the alternative hypothesis (H_1). This alternative hypothesis (H_1) directly challenges the null hypothesis (H_0), which assumes that there is no difference between the effect of two treatments.

After formulating our hypotheses, we gather the data, look at the extent to which the data support our alternative hypothesis and consequently draw conclusions about our hypothesis with the calculated *p*-value (Trafimow, 2019a). In the study example shown in part 1, the alternative and null hypotheses were:

- H_1: There is a difference in functional outcome between women with a distal radial fracture that receive cast immobilization only and women with a distal radial fracture that receive cast immobilization plus a rehabilitation programme.
- H_0: There is no difference in functional outcome between women with a distal radial fracture that receive cast immobilization only and women with a distal radial fracture that receive cast immobilization plus a rehabilitation programme.

In our example, we then conducted the study using a sample from the population of patients and analysed the gathered data. This resulted in a *p*-value of 0.07 as shown in Table 1 of Part 1 of Broekstra et al. (2022a), and most researchers would then conclude that there was no significant difference in functional outcome between both interventions. In our experience, many would follow this procedure routinely, without understanding exactly the underlying line of reason. However, this may lead to wrong conclusions with obvious erroneous impact on patient care. To illustrate this further, we will use a simple coin-flipping experiment.

Flipping a coin experiment

Imagine we want to know if a coin is 'fair' or unbiased. In order to determine that, we can conduct a simple experiment; we toss the coin a number of times, and the null hypothesis is that the probability (*p*) of obtaining a heads or tails is equal: $p(\text{heads}) = p(\text{tails}) = 0.5$. This is the 50:50 heads/tail percentage one would expect to find, if a fair coin is thrown numerous times. The alternative hypothesis assumes that the coin is 'unfair' or biased in some ways, and that the percentage would not be 50:50, namely, that $p(\text{heads}) \neq p(\text{tails})$ or $p(\text{heads}) \neq 0.5$. Suppose you toss the coin 5 times, and you throw 'heads' 5 times.

When we have formulated the hypothesis about the fairness of the coin and gathered our data, the next step is to derive the *p*-value, which should be straightforward in this case. As we had assumed that the coin is fair (so that the null hypothesis is true), which means that the probability of heads should be 0.5 every time we throw, the combined probability of throwing five heads is then $(0.5)^5 = 0.03$. In this case, throwing five heads out of five throws is the most

extreme outcome we could achieve. However, if we would have thrown four heads out of five, then we would be interested in the probability of throwing four or five heads, assuming that the coin is fair or unbiased. So the *p*-value calculated in our studies is in fact the probability of the observed, or more extreme observations, assuming that the null hypothesis is true. The *p*-value demonstrates the compatibility between the *observed* data and the *expected* data under the assumption that the null hypothesis is true, where 0 is complete incompatibility and 1 is perfect compatibility.

In our example study with the distal radial fracture patients (Broekstra et al., 2022a), the *p*-value of 0.07 gives the probability of observing a difference between the two groups of 1.21 points in terms of the Patient-Rated Wrist Evaluation (PRWE) score between both study groups, when we assumed that there was no difference in efficacy of both treatments in the entire population of patients. When the probability of an observed, or more extreme observation under the assumption that H_0 is true (i.e. the *p*-value), is very small, we state that this value does not support our null hypothesis and consequently reject it. As mentioned, the threshold for rejecting the null hypothesis is called the *significance level* and is labelled as alpha (α) and usually set at 5% (0.05) (Greenland et al., 2016).

Problems with NHST and the *p*-value

Inverse inference

NHST is the most commonly used statistical procedure in empirical science (Szucs and Ioannidis, 2017) and reaching statistically significance seems to have become an end in itself when conducting research. Despite its widespread use and popularity, NHST and the *p*-value have been criticized since they became universal. Methodologists and statisticians have addressed substantial problems associated with NHST and the *p*-value.

Probably the most important problem is the fact that NHST and the *p*-value answer the wrong question, namely: what is the probability of the observed (or more extreme) data given H_0 ($p(D|H)$), whereas we would like to know the probability of the hypothesis being true, given the data ($p(H|D)$) (Cohen, 1994). This is illustrated in Figure 1, where the curve indicates the probability distribution under the assumption that H_0 is true. The area under the curve represents possible *p*-values; more specifically, the *p*-value is the probability where a datapoint or more extreme data can be observed under this curve. If a datapoint falls within the green areas, the probability

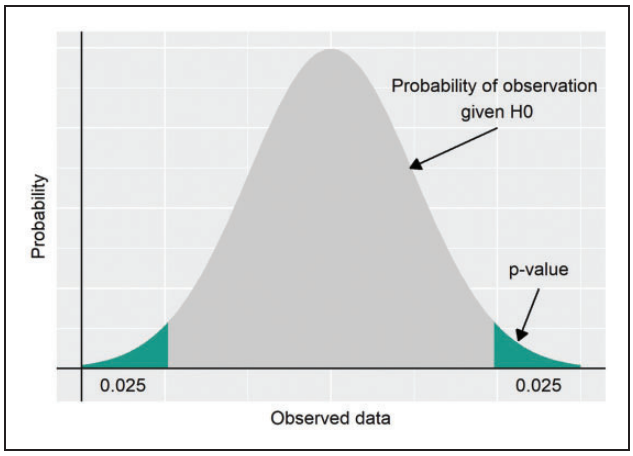

Figure 1. The curve is the probability of a certain observation, given the assumption that H_0 is true. The probability of finding a datapoint that falls within the green areas, given H_0, is so small that the H_0 will be rejected.

to find this datapoint, given the assumption that H_0 is true, would be small. This indicates that the results of the study are unlikely when the H_0 is true. This is not equivalent to the conclusion that the H_0 itself is unlikely and that therefore we have to reject it. However, the latter is exactly what we usually conclude from a small *p*-value. This problem in traversing from *p*-values to statements on hypothesis has been coined the problem of 'inverse inference' (Trafimow, 2017).

Returning to the experiment with the coin: we calculated that the probability of five heads out of five throws as $(0.5)^5 = 0.03$, and this is smaller than 5%. Within the framework of NHST, we would call this result statistically significant and reject the hypothesis that the coin is fair (H_0) and accept the hypothesis that there is more going on than merely coincidence (H_1): for example, a manipulated coin that only/mostly throws heads (Rosendaal, 2016). The problem is that in this case, we probably want to draw a conclusion about the *fairness of the coin*, but we have calculated the probability *of the findings of our experiment under the assumption that the coin is fair*. Again, this inverse way of reasoning calls upon a logical fallacy when making inferences about empirical findings (Trafimow, 2019b).

Dichotomous thinking

Another drawback of NHST is that it promotes a dichotomous way of thinking: researchers often use the outcome of a significance test as a dichotomous indicator for an effect, reducing empirical findings to two categories ($p < 0.05$: effect, $p > 0.05$: no effect). First of all, the threshold of 0.05 is an arbitrary one.

A *p*-value in a study might flip from 0.04 to 0.06 with one extra event in one of the study arms, leaving the data virtually the same, but leading to a completely different conclusion. Second, the *p*-value conflates the precision of the study result with the magnitude of the effect (effect size). Figure 2 shows how using the *p*-value as a dichotomous indicator can be misleading. Studies with different effect sizes can have the same *p*-value (A1 and A2). If we would only look at the *p*-value, we would conclude that the studies have the same outcome: results are equally compatible with the null hypothesis. However, looking at the effect size, the results of study A2 are, in terms of treatment effect, much more pronounced than the results of the study A1 (19% versus 3%). So, if we would quantify results in a dichotomous way, results of both studies would be the same, but if we would quantify results in a more comprehensive way, the first study (A1) would be far less convincing than the second study (A2) (Goodman, 2008). On the other hand, studies can have different degrees of significance, but at the same time identical estimates of effect. This is a result of varying precision of the estimate; in Figure 2, the estimate of study B1 is more precise (it has a narrow confidence interval) than the estimate of study B2 (which has a broad confidence interval). This is typically caused by differences in the sample sizes. These figures illustrate how important information is missed by looking only at the *p*-value.

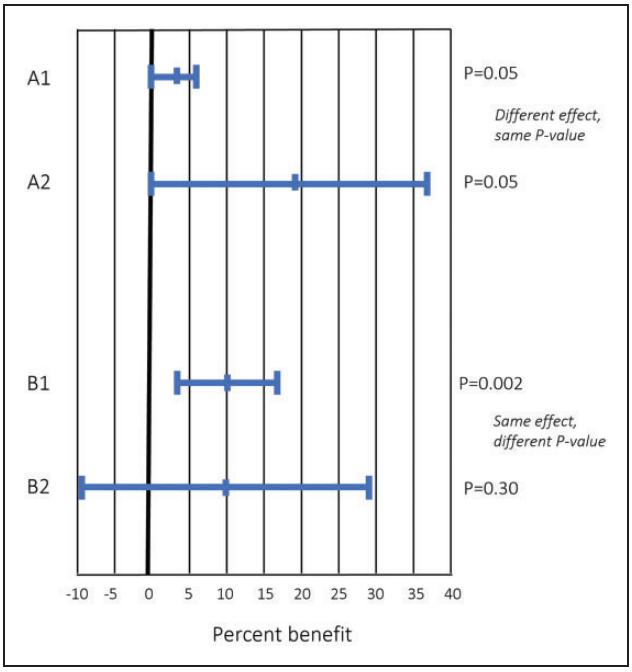

Figure 2. studies can have the same *p*-value while having different effect sizes (a). Studies can also have different *p*-values, but identical estimates of effect with different precision (b) (Goodman, 2008).

A useful alternative to NHST for the critical clinician when used properly: the confidence interval

Given the drawbacks of NHST, there is a need for a simple, alternative way of drawing inferences from the data of our studies. One that we have already alluded to in Part 1 of Broekstra et al. (2022a), and that is known to most researchers, is the confidence interval (CI). This is a measure of precision of the estimated effect: the narrower the CI, the more precise the estimated effect. In addition, the values within the CIs reflect treatment effects in the same units as the estimated treatment effect, namely that the values reflect directly interpretable effects. Notwithstanding these positive attributes, it is important to gain a more complete understanding of the interpretation of CIs and to prevent their misuse.

To start with misuse, many researchers just use CIs to conduct NHST: when the CI does not include the value of no effect, the null hypothesis is rejected, and the effect is significant. Using CIs in such a way, is exactly the same as using *p*-values to conduct NHST with the same inherent pitfalls. Additionally, many researchers, including one of the authors (Broekstra et al., 2022b), interpreted CIs as meaning 'there is a (95%) probability that the true population effect (effect in the entire population of patients) lies between the lower and upper limit of the CI'. This is a flawed interpretation, which results from the fact that CIs, like *p*-values, do not provide information about the probability of (hypotheses on) the true population effects. Instead, CIs, like *p*-values, are compatibility measures. For that reason, some authors have proposed the term 'compatibility intervals' (Amrhein et al., 2019). Specifically, the CI contains a range of population effects that are more compatible with the data than values outside the CI (Rafi and Greenland, 2020). It is important to acknowledge two things in the interpretation. First, the interval reflects a continuous measure; values just outside the interval are almost as compatible as values just inside the interval. Second, the most compatible effect is the effect that was calculated from the study data and effects near to that (Amrhein et al., 2019).

In our example we found an adjusted mean difference in PRWE score between both treatment groups of 1.21 with a 95% CI ranging from –0.12 to 2.54.

This (naturally) means that a difference of 1.21 points in PRWE score is most compatible with the study data. Differences between –0.12 and 2.54 are most compatible with the study data, and a value of 2.55 is almost as compatible as a value of 2.54. Moreover, a difference of 2 points is more compatible with the data than a difference of 0 points, which illustrates that just dismissing the intervention because '0' is in the interval is nonsensical. Finally, as discussed in Part 1 (Broekstra et al., 2022a), it does make sense to take into account the minimal clinically important change (MCIC). Now, we can do that using the CI. In our example, the MCIC was 20 points. This value is highly incompatible with our data and that would be a reason to dismiss (at least until we have other evidence) the intervention that we studied.

Conclusion

To summarize, the *p*-value is a hard-to-interpret parameter, with limited utility. It is a magic word for many researchers, but it does not tell us what we want to know and often leads to confusion about the relevancy and certainty of our empirical findings. Most importantly, using NHST based on the *p*-value can easily lead to erroneous conclusions. Confidence or compatibility intervals provide a useful alternative if interpreted correctly. The interval contains values of the population effect more compatible with the data, than values that are outside the interval. That is, it should be interpreted as a continuous measure. The narrower the CI, the more precise our estimate of the effect.

ORCID iD Dieuwke C. Broekstra https://orcid.org/0000-0002-7134-7007

References

Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature. 2019, 567: 305–7.

Broekstra DC, de Boer MR, Stunt JJ. Statistics in publishing: the (mis) use of the *p*-value (part 1). J Hand Surg Eur. 2022a, 47: 677–80.

Broekstra DC, Mouton LJ, van der Sluis CK, IJpma FFA, Stenekes MW. Hand function in patients with distal radius factures after home-based kinaesthetic motor imagery training. J Hand Surg Eur. 2022, 47: 656–58.

Cohen J. The earth is round ($p < 0.05$). Am Psychol. 1994, 49: 997–1003.

Goodman S. A dirty dozen: twelve *p*-value misconceptions. Semin Hematol. 2008, 45: 135–40.

Greenland S, Senn S, Rothman K et al. Statistical tests, *p*-values confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016, 31: 337–50.

Rosendaal FR. The *p*-value: a clinician's disease? Eur J Intern Med. 2016, 35: 20–3.

Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Medical Res Methodol. 2020, 20: 244.

Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci. 2017, 11: 390.

Trafimow D. A taxonomy of model assumptions on which *p* is based and implications for added benefit in the sciences. Int J Soc Res Methodol. 2019a. 22: 571–83.

Trafimow D. A frequentist alternative to significance testing, *p*-values and confidence intervals. Econometrics. 2019b, 7: https://doi.org/10.3390/econometrics7020026

Trafimow D. Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. Educ Psychol Meas. 2017, 77: 831–54.

Jonáh J. Stunt[1], Dieuwke C. Broekstra[2,*] and Michiel R. de Boer[3]

[1]Department of Health Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2]Department of Plastic Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[3]Department of General Practice and Elderly Care Medicine, University Medical Center, Groningen, University of Groningen, Groningen, The Netherlands

*Corresponding author: d.c.broekstra@umcg.nl

Twitter: @DCBroekstra