Journal of Physiotherapy 65 (2019) 178–181

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Appraisal

Research Note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary

Significance tests and hypothesis tests (collectively referred to here as null hypothesis statistical tests)[1] have become widespread in health research. Over the last century or so, health researchers have published reports of hundreds of millions of such tests.

Many researchers think that 'statistical analysis' of quantitative research data necessarily involves conducting null hypothesis statistical tests, and many assume that null hypothesis statistical tests provide the most informative and rigorous ways to analyse data. So, it may come as a surprise that null hypothesis statistical tests have been subject to sustained and serious criticism since the earliest days of their existence.[2–5] Neither significance tests nor hypothesis tests provide informative tests of meaningful hypotheses. The products of null hypothesis statistical tests – *p*-values and claims of statistical significance – are widely misinterpreted, often in ways that make the tests appear more informative than they really are.[1,6]

In the last few years there has been renewed criticism of null hypothesis statistical testing, notably in a recent special issue of *The American Statistician* on statistical inference.[7] Leading statisticians have argued that the concept of statistical significance should be retired.[8] Maybe, finally, the demise of null hypothesis statistical tests is imminent.

This Research Note provides a brief overview and critique of null hypothesis statistical tests for non-statisticians. It is divided into three parts. Part one explains the reasoning behind null hypothesis statistical tests and distinguishes between significance testing and hypothesis testing. Part two outlines some problems with null hypothesis statistical tests. Part three considers how researchers could rigorously analyse data and report statistical analyses without conducting null hypothesis statistical tests.

Significance testing and hypothesis testing

When researchers conduct research, they are faced with a problem. They can only collect data from people (or animals, or cells) who comprise a small part – a sample – of a much larger population. The *sample* is of little intrinsic interest, not least because it makes up a small proportion of the population. Instead, it is the *population* from which the sample was drawn that is usually of interest. Unfortunately, even if a sample can be drawn at random from the population, the vagaries of chance mean that the sample can only ever represent its population imperfectly. Nonetheless, researchers must somehow make inferences about populations using data from samples. The process of making inferences about populations using data from samples is called statistical inference.

For the last 100 years, most statistical inference has been conducted within a 'frequentist' framework. Frequentist statistical inference involves imagining what would happen if a study was conducted in exactly the same way a very large number of times, each time on a new sample drawn randomly from the same population. Frequentists are concerned with how a study's findings (or, more precisely, how a statistic such as a mean, a difference between two means, a risk ratio, or a correlation) would vary across the many imagined replications of the study. In a frequentist analysis, data obtained from a study sample are used to infer how the study findings would vary across the imagined replications of the study. When researchers conduct frequentist analyses, they hope to find that there is little variability across imagined study replications. It is thought that little variation across imagined study replications implies that the findings from the study that was actually conducted are, in some sense, robust, replicable or believable.

Early in the last century, Sir Ronald Fisher developed a frequentist approach to statistical inference called significance testing. Significance testing starts by positing a 'null hypothesis'. The null hypothesis is that there are no effects on the population from which the study sample was drawn. Here, the term 'effects' is used in a statistical sense. Statistical effects need not be causal effects. Effects are usually quantified in terms of differences or ratios. Here are three examples: 1) In clinical trials, the null hypothesis is usually that there is no effect of the intervention on health outcomes, and the effect may be quantified in terms of the difference in mean outcomes with and without treatment. 2) In cohort studies of aetiology, the null hypothesis may be that there is no effect of exposure on risk of disease, and the effect may be quantified in terms of the ratio of risks in exposed and unexposed people. 3) In a cross-sectional study, the null hypothesis may be that there is no relationship between two variables, and the effect may be quantified in terms of the correlation between the variables. Note that the null hypothesis is a hypothesis about the population, not about a sample.

Fisher developed mathematical procedures that use data from the study sample to determine how an effect estimated using data from a study sample would vary across many imagined replicates of the study *if the null hypothesis was true*. He used these procedures to calculate the proportion of imagined study replicates that would find effects at least as large as the effect that was actually observed in the real study. That proportion can be thought of as a probability: it is the probability, when repeatedly conducting a study, of observing an effect that is at least as large as the actually observed effect if the null hypothesis is true. This probability is called '*p*'.

Fisher argued that if, in a particular study, it could be shown that effects of at least the size that was actually observed would be unlikely to be observed in imagined study replicates if the null hypothesis is true (ie, if '*p*' is small), this could be taken as evidence that the null hypothesis is false. The convention has become that a $p < 0.05$ (ie, $p < 5\%$) is considered unlikely. So, when a researcher analyses data from a study and observes an effect for which $p < 0.05$, the effect is declared to be 'statistically significant'. In significance testing, statistically significant findings are interpreted as evidence that the null hypothesis is false.

It was not long until mathematicians identified fundamental problems with significance testing. That motivated the development,

https://doi.org/10.1016/j.jphys.2019.05.001

by Jersey Neyman and Egon Pearson, of an alternative frequentist approach to statistical inference, referred to here as hypothesis testing.

There are *two* hypotheses in hypothesis testing: the *null* hypothesis, in which there is no difference in the population, and the *alternative* hypothesis, in which there is a difference in the population. In hypothesis testing, *p*-values serve only to tell the researcher how to choose between these two competing hypotheses. When $p < 0.05$, the researcher chooses to accept the alternative hypothesis, and when $p \geq 0.05$, the researcher chooses to accept the null hypothesis. In the long run (over many hypothesis tests across many studies), hypothesis testing is optimal, in a strictly mathematical sense, because it minimises the probability of making certain types of false decisions. In any single hypothesis test there is no guarantee that the choice is correct. But there is at least the assurance that, in the long run, choices made in this way will be correct as often as is possible.

While Fisherian significance testing and Neyman-Pearson's hypothesis testing both involve nominating a null hypothesis and calculating *p*-values, and the two are mathematically very similar, they greatly differ in their interpretation. The two parties argued the merits of their own approaches for decades. Despite a long and rancorous feud, the two positions were never reconciled. Perhaps as a result, modern researchers often combine elements of the two approaches. Modern statistical inference is often a logically inconsistent mish-mash of the two traditions.[5] For more extensive discussions of the history and rationale of significance testing and hypothesis testing the interested reader is referred to references 2, 4, and 5.

Problems

Some problems with null hypothesis statistical testing are listed here. Others are discussed elsewhere.[2–4] All of the problems discussed here apply both to significance testing and hypothesis testing.

P does not tell us the probability that a hypothesis is (or is not) true

P is the probability of observing the actually observed data, given that the null hypothesis is true. That probability is, or should be, of little or no intrinsic interest to researchers. There should be more interest in the probability that the null hypothesis is true, given the actually observed data.

If you think that these two probabilities (the probability of observing the actually observed data, given that the null hypothesis is true, and the probability that the null hypothesis is true, given the actually observed data) are the same thing, or that they should behave in the same way, think again. The two probabilities need be no more alike than, say, the probability that a person is prime minister of India, given that the person is a man, and the probability that a person is a man, given that the person is prime minister of India.

P is not evidence

A related problem is that the probability of an observation given a particular hypothesis (like the null hypothesis) does not, on its own, provide evidence for or against that hypothesis. After all, there is a low probability of drawing an ace of spades from a full pack of cards, given that it is a standard pack of cards, but drawing an ace of spades from a pack of cards would not make us think that the pack of cards is not a standard pack of cards. It is only possible to quantify the strength of evidence for a hypothesis *compared with* a competing hypothesis. Data provide evidence *for* hypothesis A and *against* hypothesis B if the data are more likely to be observed given hypothesis A than given hypothesis B.[3,4]

Significant findings are not replicable

If a study could be perfectly replicated many times, each time randomly drawing a new sample from the population, it would be found that the *p*-values, like other statistics, varied from sample to sample. Cumming illustrated this with a brilliant and entertaining video titled 'dance of the p values'.[9,10] Boos and Stefanski conducted a formal analysis of the reproducibility of *p*-values across hypothetical study replicates and showed that *p*-values are alarmingly volatile. They concluded that 'the probability of non-replication of published studies with *p*-values in the range 0.005 to 0.05 is roughly 0.33.' This means that even if a study with a statistically significant finding could be conducted again in exactly the same way on a new sample drawn randomly from the same population, there could be a high probability of failing to reproduce the statistically significant finding. Therefore, for purely statistical reasons, researchers' findings of statistical significance are inherently difficult to reproduce. (Incidentally, Boos and Stefanski's findings suggest that the current perception of a crisis of replicability in science may be due, at least in part, to the conceptualisation of replicability in terms of the replicability of claims of statistical significance.)

In most clinical research, the null hypothesis must be false

The null hypothesis posits that the effect of interest is *exactly* zero in the population. It does not posit that the effect is nearly zero, or too small to be interesting, or close enough to zero that it can be considered zero for all practical purposes.

In clinical research, it is generally implausible that the effect of interest is exactly zero. For example, it is usually inconceivable, in a clinical trial, that there is absolutely no effect whatsoever, on average, of the intervention on the trial outcomes. It could be expected that pretty much all interventions have some effect: some interventions have trivially small effects that are too small to be detected and other interventions have large and obvious effects, but it is hard to imagine any intervention that has *exactly* no effect in the population of interest. Moreover, even the best conducted studies are likely to have some bias. It is inconceivable that any study has absolutely no bias whatsoever.

These examples illustrate that generally, in clinical research, the null hypothesis cannot be true. This implies that all claims of statistical significance must be correct. It also implies that all findings of a lack of statistical significance represent failures to detect an effect that really does exist.

We need to know about the size of effects

The main way that clinical research can contribute to understanding disease and healthcare is by providing estimates of the size of effects. Information about the size of effects, rather than just the existence of effects, is needed if important causes of disease are to be identified, the primary mechanisms by which health interventions work are to be determined, or the effects of interventions are to be assessed.

There is, for example, little value is demonstrating that a health intervention 'has an effect'. (After all, almost all health interventions have some effect.) The scientific challenge is to distinguish between interventions whose effects are large enough to make them worth their costs, risks and inconveniences.[11] It can only be decided if an intervention is worthwhile if the size of the effects of the intervention can be estimated. Similar arguments can be applied to studies that investigate the mechanisms of disease or the mechanisms by which interventions have their effects. The primary challenge is to determine the extent to which a mechanism causes disease or modifies a disease pathway. Only when the size of such effects is known can it be decided whether a mechanism is important or not.

It is important to understand that null hypothesis statistical testing in general, and *p*-values specifically, tell us nothing about the size of effects. A small and statistically significant *p*-value could arise when the data provide clear evidence of a large and clinically important effect. However, the same *p*-value could also arise when the data do not provide clear evidence of an important effect. Even more alarming, the same *p*-value could arise when the data provide clear evidence that the effect is small and unimportant. Statistically non-significant findings are also uninformative about sizes of effects. A large, statistically non-

Box 1. An example of how study objectives (in this case, the objectives of a clinical trial) can be framed in terms of the estimation of effects rather than as a test of a hypothesis.

Instead of framing study objectives in terms of tests of hypotheses, like …	**frame study objectives in terms of estimation of effects, like** …
'The aim of this study is to test the hypothesis that, compared with [control intervention], [intervention] affects [outcome] in [population].' OR 'The aim of this study is to test the null hypothesis that, compared with [control intervention], [intervention] does not affect [outcome] in [population].'	'The aim of this study is to estimate the average effect of [intervention] compared with [control intervention] on [outcome] in [population].'

significant *p*-value could arise when the data provide clear evidence of little or no effect, but the same *p*-value could arise when the data provide evidence that there could be a large and clinically important effect. So *p*-values or null hypothesis significance testing cannot be relied on for information about the size of effects.

Quantitative health research *can* provide information about the size of effects. However, it is not possible to quantify the size of effects using null hypothesis significance tests. The next section argues that it is possible to make useful inferences about the sizes of effects without ever conducting null hypothesis significance tests.

Alternatives

The preceding section argued that there are serious problems with null hypothesis significance tests. Null hypothesis significance tests interrogate a hypothesis that cannot be true, they are widely misinterpreted, and they do not tell us what really needs to be known. That does not mean that the vast scientific literature that has reported null hypothesis significance tests is entirely wrong or useless. Fortunately, there are simple alternatives to null hypothesis significance testing that are built on the same frequentist foundations as null hypothesis significance testing and which are meaningful and relatively easily interpreted. By reading between the lines, it is often possible to apply better approaches to the interpretation of studies that have conducted and reported null hypothesis significance tests. Even though null hypothesis significance tests are very problematic, decades of research that have used null hypothesis significance tests do not need to be discarded.

A sensible way to make statistical inferences is to focus on the size of effects and the precision with which those effects can be estimated. This is sometimes called 'estimation'. Estimation can be conducted within a frequentist framework. With an estimation approach to data analysis, the central task is to use statistics derived from sample data to estimate parameters of populations. Those parameters represent the size of the effects of interest in the population. It is recognised that, because of the vagaries of chance, estimates of effect obtained from sample data may underestimate or overestimate the effect in the population. Therefore, estimation also involves quantification of the degree of uncertainty (or imprecision) of estimates of effects, conventionally with the use of confidence intervals. The mathematics underlying frequentist estimation is very similar to the mathematics that underlies null hypothesis significance tests, but estimation does not involve explicitly nominating a null hypothesis, interpreting *p*-values, or arbitrating on statistical significance. Cumming provides a clear, non-technical account of the theory and practice of an estimation approach to statistical inference.[10] Estimation provides a simple alternative to null hypothesis significance testing that is meaningful and relatively easily understood. Estimation provides a good alternative to null hypothesis significance testing. So it is not necessary to conduct null hypothesis significance tests.

What would quantitative health researchers have to do if they were to stop using null hypothesis significance tests statistical tests and start using an estimation approach to the analysis of their data? Here are three simple, easily implemented suggestions:

1. Define study objectives in terms of estimation of effects

Conventionally, researchers define study objectives by stating the hypotheses that they wish to test. Some researchers explicitly nominate the null hypotheses. This practice only makes sense if it is conceivable that the null hypothesis could be exactly true.

A better way to frame the objectives of quantitative clinical research is in terms of estimating the size of effects. Box 1 provides an example.

2. Do not report statistical significance

Late last year, the American Statistical Association convened a 2-day symposium on statistical inference. The proceedings were published as 43 papers in *The American Statistician*, the Association's official journal. Most of the papers were critical of the practice of null hypothesis significance testing.

In the lead editorial, Wasserstein and colleagues stated:

> *The ASA Statement on P-Values and Statistical Significance* [published in 2016] *stopped just short of recommending that declarations of 'statistical significance' be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term 'statistically significant' entirely. Nor should variants such as 'significantly different,' '$p < 0.05$,' and 'non-significant' survive, whether expressed in words, by asterisks in a table, or in some other way.*

Just a few weeks later, Amrhein and colleagues, wrote in the journal *Nature*: *we agree, and call for the entire concept of statistical significance to be abandoned.*[8]

In this author's opinion, these recommendations are a sensible response to the well-known problems with null hypothesis significance tests presented in the preceding section. It is acknowledged that not all experts agree (see, for example, reference 12).

3. Focus on the estimated size of the effect

Most clinical research should be concerned with estimation of the size of effects. For example, as has been argued above, clinical trials should be concerned with the size of the (beneficial and harmful) effects of intervention, aetiologic studies should be concerned with how much the risk of disease is increased by a particular exposure, and so on.

Interpretation of the size of effects can proceed in three steps. First, the researcher should consider the 'point estimate' of the effect (the point estimate is the best estimate of the effect in the population). Is the point estimate of the effect large enough to be, in some sense, important? Or is the effect trivially small? Second, the researcher should consider the accuracy of the estimate. Is the point estimate likely to be biased? (For example, if the estimate is an effect of intervention obtained in a clinical trial, how much could the estimate have been biased by loss to follow-up? If the estimate is of the causal effect of exposure in an observational study, how much could the estimate have been biased by uncontrolled confounding?) Lastly, the researcher should consider the precision of the estimate. How

precise is the estimate of effect? Can we be reasonably confident about the estimate?

The precision of an estimate is usually quantified with confidence intervals. A simple, though not strictly accurate, interpretation of confidence intervals is that they are the interval within which the true effect (the average effect in the population) probably lies. At issue is whether the two extremes of the confidence interval have the same substantive interpretation. If the lower and upper confidence limits have the same substantive interpretation (eg, both the lower and upper confidence limits of the effect of treatment obtained from a clinical trial are large enough to be clinically important), the study provides a clear answer. If, on the other hand, the lower and upper confidence have different interpretations (eg, the lower limit is of a trivially small effect whereas the upper limit is of a clinically important effect), the study leaves interpretative uncertainty.[8] Amrhein et al recommend that confidence intervals are interpreted by:

> *describ[ing] the practical implications of all values inside the [confidence] interval, especially the observed effect (or point estimate) and the limits. In doing so, [we] should remember that all the values between the interval's limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval.*[8]

Critics of the estimation approach recommended here might reasonably point out that frequentist estimation suffers from some of the same logical flaws as frequentist statistical tests. If rigorous estimates of effects in populations are to be obtained using data from samples, frequentism must be abandoned altogether and alternative ways of conducting statistical inference be adopted, such as Bayesian inference. A pragmatic defence of frequentist estimation against this criticism is that the point estimates and confidence intervals obtained with frequentist estimation are generally consistent with those obtained using Bayesian estimation in the presence of weak priors. (For a clear explanation of Bayesian inference, the meaning of 'priors' and the relationship between frequentist and Bayesian estimation, see reference 13.) For now, at least, frequentist point estimates and confidence intervals may be good enough to inform the progress of science. They are, at any rate, more informative than significance and null hypothesis tests.

Competing interests: Nil.

Sources of support: Nil.

Acknowledgements: The author is supported by the Australian NHMRC (APP1117192).

Provenance: Invited. Not peer reviewed.

Correspondence: Rob Herbert, Neuroscience Research Australia (NeuRA), Australia. Email: r.herbert@neura.edu.au

Rob Herbert
Neuroscience Research Australia (NeuRA), Australia

References

1. Nickerson RS. *Psychol Methods*. 2000;5(2):241–301.
2. Barnett V. *Comparative Statistical Inference*. London, New York: Wiley; 1973.
3. Goodman SN, Royall R. *Am J Public Health*. 1988;78(12):1568–1574.
4. Royall RM. *Statistical Evidence: A Likelihood Paradigm*. 1st ed. London, New York: Chapman & Hall; 1997.
5. Gigerenzer G. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge England, New York: Cambridge University Press; 1989.
6. Greenland S, et al. *Eur J Epidemiol*. 2016;31:337–350.
7. Wasserstein RL, et al. *Am Stat*. 2019;73(Suppl 1):1–19.
8. Amrhein V, et al. *Nature*. 2019;567:305–307.
9. Cumming G. Viewed 22 April 2019, from https://latrobe.figshare.com/articles/Introduction_to_Statistics_Dance_of_the_p_Values_Prof_Geoff_Cumming/7453169.
10. Cumming G. *Multivariate applications series*. New York: Routledge; 2012.
11. Herbert RD, et al. *Practical Evidence-Based Physiotherapy*. 2nd ed. Oxford: Elsevier; 2011.
12. Ioannidis JPA. *JAMA*. 2019.
13. Greenland S. *Int J Epidemiol*. 2006;35:765–775.