International Journal of Nursing Studies 99 (2019) 103384

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

Statistical significance testing and p-values: Defending the indefensible? A discussion paper and position statement

Peter Griffiths[a,*], Jack Needleman[b]

[a] University of Southampton, UK and Executive Editor, International Journal of Nursing Studies, United Kingdom
[b] Department of Health Policy and Management, University of California, Los Angeles School of Public Health, Los Angeles, USA

ARTICLE INFO

Keywords:
Confidence intervals
Probability
Data interpretation
Statistical

ABSTRACT

Much statistical teaching and many research reports focus on the 'null hypothesis significance test'. Yet the correct meaning and interpretation of statistical significance tests is elusive. Misinterpretations are both common and persistent, leading many to question whether significance tests should be used at all. While most take aim at the arbitrary declaration of $p < 0.05$ as a threshold for determining 'significance', others extend the critique to suggest the 'p-value' should be dispensed with entirely.

P-values and significance tests are still widely used as if they give a measure of the size and importance of relationships, even though this misunderstanding has been observed and discussed for many years. We argue that p-values and significance tests are intrinsically misleading. Point estimates of relationships and confidence intervals give direct information about the effect and the uncertainty of the estimate without recourse to interpreting how a particular p-value might have arisen or indeed referring to them at all.

In this paper we briefly outline some of the problems with significance testing, offer a number of examples selected from a recent issue of the International Journal of Nursing Studies and discuss some proposed responses to these problems. We conclude by offering some guidance to authors reporting statistical tests in journals and present a position statement that has been adopted by the International Journal of Nursing Studies to guide its' authors in reporting the results of statistical analyses.

While stopping short of calling for an outright ban on reporting p-values and significance tests we urge authors (and journals) to place more emphasis on measures of effect and estimates of precision/uncertainty and, following the position of the American Statistical Association, emphasise that authors (and readers) should avoid using 0.05 or any other cut off for a p-value as the basis for a decision about the meaningfulness/importance of an effect. If point estimates and confidence intervals are used, then the p-value may be redundant and can be omitted from reports. When authors talk about 'significance' they need to be explicit when referring to statistical significance and we recommend authors adopt the language of 'importance' when talking about effect sizes to avoid any confusion.

What is already known about the topic?

- Statistical significance tests and p-values are a widely used and reported in research papers
- Both are subject to widespread misinterpretation, because they are used as if they give information about the size or importance of effects

What this paper adds

- The language of 'significance' may be intrinsically misleading and 'importance' should be used when talking about effect sizes
- We encourage emphasis on effect sizes and confidence intervals when reporting or discussing results and drawing conclusions
- If point estimates and confidence intervals are used then the p-value may be redundant and can be omitted from reports.

1. Statistical significance testing and p-values

It can sometimes seem that statistical significance testing is the ultimate tool of quantitative data analysis. Much statistical teaching and many research reports focus on the so called 'null hypothesis

* Corresponding author at: University of Southampton, B67, University Road, Southampton, SO17 1BJ, United Kingdom.
E-mail address: peter.griffiths@soton.ac.uk (P. Griffiths).
@workforcesoton, @ijnsjournal (P. Griffiths)

https://doi.org/10.1016/j.ijnurstu.2019.07.001
0020-7489/

significance test', sometimes to the exclusion of almost any other consideration (Greenland et al., 2016). Yet the correct meaning and interpretation of statistical significance tests is elusive. Misinterpretations are both common and persistent, leading many to question whether significance tests should be used at all. While most take aim at the arbitrary declaration of $p < 0.05$ as a threshold for determining 'significance', others extend the critique to suggest the 'p-value' should be dispensed with entirely.

In this paper we briefly outline some of the problems with significance testing and discuss some proposed responses to these problems. Our paper concludes by offering some guidance to authors reporting statistical tests in journals and presents a position statement that has been adopted by the International Journal of Nursing Studies to guide its' authors in reporting the results of statistical analyses.

1.1. Significance tests as understood

These two explanations of statistical significance capture a common (mis)understanding:

> *"Statistical significance measures how likely that any apparent differences in outcome between treatment and control groups are real and not due to chance."* (Leung, 2001) p 201

Or

> *"Statistical significance is the probability that the observed difference between two groups is due to chance."* (Sullivan and Feinn, 2012) p 279

These descriptions are formulated around the sorts of questions that researchers ask – whether there is a difference, or, more importantly 'what is the difference?' In both cases it appears that the information contained in the significance tests relates directly to the difference (or effect) that has been observed in a given study and so appears to provide some measure of that relationship. So, if, in a trial comparing two antihypertensive agents, we observed that one provides a 10% improvement in the number of people who respond to treatment, these interpretations suggest that the significance test measures how likely it is that this observed effect is due to chance or is 'real'. The same explanation would generalise to all observed associations.

Under both these explanations, a low p value seems to confirm or provide evidence that the particular result observed result is 'true'. In our example a low p-value is evidence that the effect is 'really' 10%. Conversely a high p-value is widely treated as evidence that there is no difference or relationship (Alderson, 2004) or, more subtly, that effect is 'not' 10%.

It is not our intention to criticise the authors quoted above: these are honest attempts to explain a tricky concept in a simple way that can be grasped intuitively. These explanations and similar formulations are often repeated. However, most statisticians would likely agree that they are wrong. Indeed, simple explanations of the significance test and p-value are always more or less wrong, while most "correct" explanations make little intuitive sense and do not make it easy to understand the proper inferences that can be made from a test (Greenland et al., 2016).

1.2. A more technical definition

A more "correct" definition, based on that offered by Greenland et al. (2016) might go something like this:

- The value of p is the frequency probability of the observed data assuming a "null" hypothesis (typically 'no difference' or 'no relationship') and a set of assumptions about the population from which the data is drawn.

Other important assumptions such as normally distributed data are rarely the focus of interest for those performing the test (or those reading the results) but remain important and so the definition can be extended to a more general formulation. The p-value is:

> *". . . a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (all the assumptions used to compute the P value) were correct."* (Greenland et al., 2016) p 339

We suspect that at this point some readers may begin to glaze over. Please bear with us – we sympathise. Others may be busily spotting technical deficiencies in our summary definition although we trust that these readers would concede that we are somewhere near the mark. So, what are the consequences of these rather different descriptions of the significance test?

The 'correct' definition of the significance appears to be answering an entirely different question to the ones researchers ask. Indeed it is a remarkably odd question that is asked about an entirely hypothetical situation. The researcher's question "*what is the effect of treatment a compared to b?*" is answered with " *. . . if there is no difference in effect between a and b the probability of what you have seen is X*". When we consider the more general definition of the p-value above, the answer becomes even more obscure.

Those who glazed over above might reasonably ask – does that really matter? Doesn't the p-value tell us something about the effect really? Should we, as applied health researchers, get concerned about these statistical niceties? For many, quoting the 'significance' is one of the main ways of describing the results of research and, for researchers, undertaking a course in statistics to learn how to perform tests of significance is a rite of passage. What does it matter if the explanations offered are 'technically' incorrect, as long as there are no practical consequences?

1.3. Consequences of misunderstanding

The basic logic of the significance test is that if p-value (probability) is low, then the assumed 'null' hypothesis is unlikely to be correct and it is 'rejected'. However, because the resulting p-value is not answering a question that most people would think to ask, the path is set for one of the most common misinterpretations of the p value – that it is a measure of effect (Badenes-Ribera et al., 2016; Leung, 2001; Oakes, 1986). This is erroneous because under any given set of assumptions the obtained p value is a product of the effect, the inherent variability of what is being studied, and the sample size (linked to random variation due to sampling or measurement error). It cannot be used as a measure of effect size because it is confounded (Stang et al., 2010).

The true situation is worse because, for most people, even the 'correct' interpretation above focusses only on the null hypothesis, of no effect/relationship and not the other assumptions of the underlying statistical model. A low p-value may lead us to conclude that one or more of the assumptions of the model are incorrect because the data observed are not compatible with the model, so we reject the null, but we do not know which of the assumptions are incorrect (Greenland et al., 2016).

Furthermore, the use of the word 'significance' strongly implies importance, another common misinterpretation that has been shown to persist over many years, even among researchers who confidently use and interpret statistical tests (Badenes-Ribera et al., 2016; Oakes, 1986; Ziliak and McCloskey, 2008). The extent that statistical significance gives no information about importance can be illustrated by the simple fact that given a sufficiently large sample *any* effect or degree of association will be statistically significant, no matter how tiny (Demidenko, 2016).

So, to summarise, statistical significance tests and p-values tell us essentially nothing about the size or importance of an effect and may even fail to deliver on the slim promise of quantifying the probability of the observed data given the null hypothesis. A small p-value and 'significant' result can result from a tiny and unimportant effect. A 'non-significant' high p-value can be observed when the estimated effect is close to the null or when it is far from the null and potentially important, but estimated with a high degree of imprecision or uncertainty.

1.4. The p-problem: some recent examples

With this in mind we reviewed quantitative analyses and systematic reviews published in the most recent volume of the International Journal of Nursing Studies at the time of writing (2019, vol 93). In doing so we do not wish to single out particular authors for blame or praise, but rather show how pervasive problems associated with simplistic interpretations of significance tests are. We found examples of the problems that arise from the interpretation of significance papers in the first three papers we looked at.

In reviewing outcomes of person centred rehabilitation Yun and Choi (2019) concluded that *"there was strong evidence regarding the positive effects of person-centered care on occupational performance and rehabilitation satisfaction" (p 74).* It is hard to escape the implication that these are important findings – effects that will really matter. However, the conclusion was based on reporting whether or not results of studies on a number of outcomes were (statistically) 'significant' in the studies reviewed. The conclusion of 'strong' evidence was based on outcomes where all the studies they reviewed gave significant results. There is no mention of the size or importance of effects, and in some cases, the conclusions seem to be based on a single study. Now it is clear enough that the authors are referring to the strength of the evidence and so not directly referring to the size of the effect. Leaving aside whether their interpretation of 'strong' evidence is correct on purely methodological grounds, it is still hard to escape reading this without getting an impression of certainty that is not just about *some* effect, but about an important one.

A second example comes from a cohort study examining work-related psychosocial risk factors and risk of disability pension. In the abstract, the following results are reported:

> *"Among nursing professionals and care assistants, high quantitative job demands and low social support, but not job control, were associated with future disability pension"* (Leineweber et al., 2019) (p 12)

The authors appear to have taken the approach that because the null could not be rejected, the result should be interpreted as though the null was the correct value. However, looking in detail at the results, aspects of job control *were* associated with future disability pensions, especially for registered nurses. While the reported hazard ratios were not statistically significant, the estimated hazards of future disability pension were increased by 29% for nurses who reported low participation in decision making and 34% for those who reported too little influence. Both these varibales were measures of job control. In addition, while the reported hazard ratios were not 'statistically significant,' in the group of 2576 nursing professionals essentially the same relationships were observed in a much larger group of 66,252 'all other occupations'. Among this group the estimated hazards increased by 32% for low decision making and 28% for low influence. So had this group been the focus of interest the same observed relationship would likely have been presented as demonstrating that job control *was* associated with disability and that seems an appropriate conclusion – with no basis to claim that the relationships in registered nurses was any different to that in the general population. There is no suggestion that these associations are overlooked because they are small or unimportant, rather we conclude that they are overlooked because they are not statistically significant and the authors are implicitly taking this as evidence of 'no effect'.

A third example comes from a review of research on the relationship between specialty nurse certification and patient, nurse and organizational outcomes (Whitehead et al., 2019). Unlike the previous example, this review clearly reports and emphasises observed effects in addition to statistical significance. Despite this helpful emphasis, the authors still have to struggle with the apparent implications of the word 'significant'. One passage, describing the results of a study by Hughes et al. (2001)' illustrates this:

> *"*Hughes et al. (2001) *reported weak, but statistically significant correlations ($r = 0.07$, $p < 0.01$) between certification and job satisfaction. The magnitude of the correlation suggests that certification had limited practical significance in accounting for variances in nurses' job perceptions"* (Whitehead et al., 2019 p7).

The weak correlation, apparently elevated by the word 'significant', has to be talked down and yet cannot be fully dismissed because it still appears to be important in some way. The authors' go on to point out that this weak correlation has '*limited practical significance*' although, in truth, it is difficult to imagine *any* practical significance arising from such a small correlation in the context.

These three examples were selected more or less at random from the International Journal of Nursing Studies. It is a small sample but we have little doubt that we would not have to search long for similar examples in other journals. These papers were all selected for publication after peer review and we believe that they remain worthy of publication. However the deeply embedded conventions in how researchers talk about statistical significance have caused all these authors problems even where researchers properly place the emphasis on estimated effects, as in the case of Whitehead et al. (2019).

More widely, well-meaning attempts to emphasise effect size over the significance test tend to compound the issue by implying that p-values might, in fact, be a measure of effect. For example the title of the pieces "Using Effect Size-or Why the P Value Is Not Enough" (Sullivan and Feinn, 2012) and "Balancing statistical and clinical significance in evaluating treatment effects" (Leung, 2001) still, to us, seem to imply that statistical significance and effect sizes are closely linked and that p-values do have a role in interpreting effect sizes. In focussing on p-values such titles have the potential to retain an important role for p-values when other results from statistical analyses, in particular point estimates of relationships and confidence intervals, give direct information about the effect and the uncertainty of the estimate without recourse to interpreting how a particular p-value might have arisen, or indeed referring to it at all.

1.5. Responses to the problems – a ban on $p < 0.05$?

It seems as if significance tests may be intrinsically misleading to all but the most statistically literate (and perhaps even to many who would consider themselves as such). Because of these and other problems, a number of journals have banned the reporting of p-values in general (Cummings, 2018; Gill, 2018; Woolston, 2015) or, more specifically, the reporting of statistical significance, typically based on a threshold of $p < 0.05$ (Anon, 2019). Recently, in a guest editorial for the International Journal of Nursing Studies a group of researchers, academics and educators engaged with nursing research made a call for nursing journals to institute a

similar policy (Hayat et al., 2019). Following recent guidance from the American Statistical Association (Wasserstein and Lazar, 2016), Hyat and colleagues call for journals to require the reporting exact p-values (rather than simple thresholds of 'significance', such as $p < 0.05$), to require that p-values are reported only alongside measures of effect size and a confidence interval and, perhaps most significantly, that authors should "*avoid using 0.05 or any other cut off for a p-value as the basis for a decision about the meaningfulness/importance of an effect*" (Hayat et al., 2019).

While we endorse these suggestions, it is important to recognise that they do not, in themselves remove all problems. Any move that places more emphasis on estimating effects and direct consideration of the precision of the effect (indicated by the confidence interval) is to be welcomed. There remains a danger that presenting exact p-values still implies they carry important meaning and, in any case, all too often authors resort to interpreting confidence intervals and exact p-values as if they were nothing more than data used to perform the simple binary null hypothesis significance test (Coulson et al., 2010; Stang et al., 2010). Researchers clearly struggle to find an appropriate language to talk about their results without recourse to the significance test.

1.6. Effect sizes and clinical 'significance'

Furthermore, while authors increasingly talk about effect size, the standardised terms of 'small', 'medium' and 'large' are based on statistical distributions. Like statistical 'significance' they have no direct bearing on whether an effect is important or not, even though they appear to. A "small" effect on a population health outcome might be an important one because it is an important outcome and it effects many people. On the other hand a "large" effect may have little importance. A large (standardised) effect on practitioner knowledge resulting from a training course may have little practical importance, especially if there was little underlying variation or there is no effect on subsequent behaviours. In general, a "large" standardised effect on a measured variable might be rather small in absolute terms when population variation is low and so standard deviations are small. And, finally, in talking of effect sizes, it is easy to forget that in many cases what is being reported is not in fact an "effect" but an "association" in an observational study, thus inadvertently introducing a causal language into studies where causal inference may not be warranted.

Thus far we have only mentioned 'clinical significance' in passing. Those who have written about clinical significance have done a great service by emphasising the importance of presenting outcomes whose meaning and importance are easily interpretable. The discourse goes far beyond simply rejecting "statistical significance" as an effect measure. For example, clinical significance can be more easily considered by presenting the results of studies that might have previously given risk ratios as a 'number needed to treat' – the number of people who need to be given a new treatment in order to get an additional positive outcome (Cook and Sackett, 1995). The injunction to focus on such measures of clinical "significance" is an important one, although we fear that because clinical and statistical significance have so often been discussed together, the true nature of the injunction is sometimes lost. It would be better if authors adopted different language when talking about effects to clearly distinguish the issues. Since the issue is about how important an observed effect might be this seems the obvious choice of language.

1.7. Conclusions

We encourage all authors and journals to adopt the policies recommended by (Hayat et al., 2019), based on the guidance of theAmerican Statistical Association. However these do stop short of encouraging the more radical changes in reporting that are widespread and longstanding. While we have offered some recent examples gleaned from the International Journal of Nursing Studies we could have found them anywhere, possibly even in our own work. The problems we have discussed are not new nor are they newly recognised, but they are remarkably persistent. Many of us will need to work hard to wean ourselves off deeply ingrained habits in talking about the results of our analyses. The issue runs deeper than simply banning significance tests and p-values and requires a fundamental change in thinking about how to interpret and talk about the results of statistical analyses.

2. A position statement for the International Journal of Nursing Studies

Based on these considerations the International Journal of Nursing Studies is not proposing to 'ban' significance testing entirely, although we are happy for authors to report point estimates and confidence intervals without giving any p-values or reference to statistical significance. To avoid confusion, authors should never use the term 'significant' without the qualifier 'statistical' and we ask that they avoid the appearance of a link between clinical and statistical significance by preferring the term 'importance' (e.g. clinical importance) when considering the importance of an effect or association.

In considering effects, we request that authors avoid limiting the discussion of effect sizes to reference to standard effect sizes and instead consider the implications of the observed effect. This may, in turn, influence how authors chose to report their outcomes and wherever possible outcomes should be reported in units whose meaning and importance is clear – original units, absolute (rather than relative) risk and numbers needed to treat. Finally, in drawing conclusions authors must consider the issue of precision, reflected in the confidence interval.

In summary, we ask authors of papers submitted to the International Journal of Nursing Studies in future to follow the guidance below.

Do:

- When reporting a p-value always report the corresponding measure of effect or association.
- When reporting measures of effect or association, report the corresponding confidence interval.
- When reporting a p value, state its exact value to an appropriate degree of precision (typically 3 decimal points).
- Authors must give confidence intervals where relevant but they need not report p values or 'statistical significance'.
- Focus on both the clinical importance and uncertainty of results.
- Prefer 'clinically important' to 'clinically significant'.
- When drawing conclusions, consider the width of confidence intervals associated with the main measures of effect/association.
- Report whether the magnitude of the effect/association is close to the null, or whether the estimated magnitude is clinically important, and how precisely it is estimated.
- If imprecisely estimated, discuss how much confidence there is in the finding, based on the data and other studies.

Don't:

- To avoid ambiguity, do not use the term 'significant' in isolation to refer to results or conclusions (so results are 'statistically significant' not just 'significant')
- Do not use p-values to conclude that there is no effect/association

- Do not use p-values as the basis for decisions about the meaningfulness/importance of an effect.

Provenance

This paper was solicited by the journal and was subject to internal editorial review. Not externally peer reviewed.

References

Alderson, P., 2004. Absence of evidence is not evidence of absence. BMJ 328 (7438), 476–477.

Anon, 2019. It's time to talk about ditching statistical significance. Nature 567 (2983), 567.

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., Longobardi, C., 2016. Misconceptions of the p-value among chilean and italian academic psychologists. Front. Psychol. 7, 1247.

Cook, R.J., Sackett, D.L., 1995. The number needed to treat: a clinically useful measure of treatment effect. BMJ 310 (6977), 452–454.

Coulson, M., Healey, M., Fidler, F., Cumming, G., 2010. Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. Front. Quant. Psychol. Meas. 1.

Cummings, G., 2018. Banning p values? The journal 'Political Analysis' does it. Secondary Banning p values? The journal 'Political Analysis' does it, https://thenewstatistics.com/itns/2018/02/03/banning-p-values-the-journal-political-analysis-does-it/. (Accesssed 30 May 2019).

Demidenko, E., 2016. The p-value you can't buy. Am. Stat. 70 (1), 33–38.

Gill, J., 2018. Comments from the new editor. Political Anal. 26 (1), 1–2.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31 (4), 337–350.

Hayat, M.J., Staggs, V.S., Schwartz, T.A., Higgins, M., Azuero, A., Budhathoki, C., Chandrasekhar, R., Cook, P., Cramer, E., Dietrich, M.S., Garnier-Villarreal, M., Hanlon, A., He, J., Hu, J., Kim, M., Mueller, M., Nolan, J.R., Perkhounkova, Y., Rothers, J., Schluck, G., Su, X., Templin, T.N., Weaver, M.T., Yang, Q., Ye, S., 2019. Moving nursing beyond $p < .05$. Int. J. Nurs. Stud..

Hughes, L.C., Ward, S., Grindel, C.G., Coleman, E.A., Berry, D.L., Hinds, P.S., Oleske, D. M., Murphy, C.M., Frank-Stromborg, M., 2001. Relationships between certification and job perceptions of oncology nurses. Oncol. Nurs. Forum .

Leineweber, C., Marklund, S., Aronsson, G., Gustafsson, K., 2019. Work-related psychosocial risk factors and risk of disability pension among employees in health and personal care: a prospective cohort study. Int. J. Nurs. Stud. 93, 12–20.

Leung, W.-C., 2001. Balancing statistical and clinical significance in evaluating treatment effects. Postgrad. Med. J. 77 (905), 201–204.

Oakes, M., 1986. Statistical Inference: A Commentary for the Social and Behavioural Sciences. John Wiley & Sons, Chichester.

Stang, A., Poole, C., Kuss, O., 2010. The ongoing tyranny of statistical significance testing in biomedical research. Eur. J. Epidemiol. 25 (4), 225–230.

Sullivan, G.M., Feinn, R., 2012. Using effect size-or why the P value is not enough. J. Grad. Med. Educ. 4 (3), 279–282.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am. Stat. 70 (2), 129–133.

Whitehead, L., Ghosh, M., Walker, D.K., Bloxsome, D., Vafeas, C., Wilkinson, A., 2019. The relationship between specialty nurse certification and patient, nurse and organizational outcomes: a systematic review. Int. J. Nurs. Stud. 93, 1–11.

Woolston, C., 2015. Psychology journal bans P values. Nature 519 (19), 9.

Yun, D., Choi, J., 2019. Person-centered rehabilitation care and outcomes: a systematic literature review. Int. J. Nurs. Stud. 93, 74–83.

Ziliak, S.T., McCloskey, D.N., 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. The University of Michigan Press, Ann Arbor, MI.