Received: 28 February 2022 | Revised: 31 March 2022 | Accepted: 3 April 2022

DOI: 10.1111/aogs.14366

COMMENTARY

Current controversies: Null hypothesis significance testing

Philip M. Sedgwick[1] | Anne Hammer[2,3] | Ulrik Schiøler Kesmodel[4,5] | Lars Henning Pedersen[3,6,7]

[1]Institute for Medical and Biomedical Education, St George's, University of London, London, UK

[2]Department of Obstetrics and Gynecology, Gødstrup Hospital, Gødstrup, Denmark

[3]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[4]Department of Obstetrics and Gynecology, Aalborg University Hospital, Aalborg, Denmark

[5]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

[6]Department of Biomedicine, Aarhus University, Aarhus, Denmark

[7]Department of Obstetrics and Gynecology, Aarhus University Hospital, Aarhus, Denmark

Correspondence
Philip M. Sedgwick, Institute for Medical and Biomedical Education, St George's, University of London, London SW17 0RE, UK.
Email: p.sedgwick@sgul.ac.uk

Abstract

Traditional null hypothesis significance testing (NHST) incorporating the critical level of significance of 0.05 has become the cornerstone of decision-making in health care, and nowhere less so than in obstetric and gynecological research. However, such practice is controversial. In particular, it was never intended for clinical significance to be inferred from statistical significance. The inference of clinical importance based on statistical significance ($p < 0.05$), and lack of clinical significance otherwise ($p \geq 0.05$) represents misunderstanding of the original purpose of NHST. Furthermore, the limitations of NHST—sensitivity to sample size, plus type I and II errors—are frequently ignored. Therefore, decision-making based on NHST has the potential for recurrent false claims about the effectiveness of interventions or importance of exposure to risk factors, or dismissal of important ones. This commentary presents the history behind NHST along with the limitations that modern-day NHST presents, and suggests that a statistics reform regarding NHST be considered.

KEYWORDS
null hypothesis significance testing, statistical significance, $p < 0.05$, clinical significance

1 | INTRODUCTION

Traditional null hypothesis significance testing (NHST), incorporating the null hypothesis and two-sided alternative, *p* value, plus critical level of significance of 0.05 needs little introduction. Statistical significance ($p < 0.05$) underpins decision-making in medicine, and none less so than in obstetric and gynecological research. The presence of statistical significance has become the gold standard for establishing whether exposure to a risk factor is important, or determining if a newly developed medical strategy, drug, device, surgical approach, or alternative way of using a known treatment is effective.[1] Furthermore, lack of statistical significance ($p \geq 0.05$) is inferred as evidence that exposure to a factor is not important, or an intervention is not effective. However, such practice has invoked much discussion because it represents misuse of NHST and misunderstanding of the *p* value.[2] In particular, it was never intended that clinical significance be inferred based on statistical significance.[3]

2 | HISTORY OF NULL HYPOTHESIS SIGNIFICANCE TESTING

To understand the challenges that inferences based on NHST pose and avoid them in the future, it is important to consider the history of NHST.[4] Traditional NHST is a single procedure constructed

Abbreviations: ASA, American Statistical Association; NHST, null hypothesis significance testing.

from two theories as suggested by Fisher in 1925,[5] plus Neyman and Pearson in 1933.[6] Although the concept of the *p* value is credited to Pearson in 1900,[7] it was Fisher who formalized it. Fisher suggested the statistical null hypothesis and *p* value; the *p* value was the strength of evidence provided by the data supporting the null hypothesis. In particular, the *p* value was the probability of obtaining a result at least as extreme as that observed given the null hypothesis—that is, if the position of equipoise (no difference between groups in outcome) existed in the population. Although Fisher suggested a value of *p* less than 0.05 for statistical significance, it was not advocated as an absolute cut-off. The intention was for statistical significance to be used as a tool to indicate if the results warranted further investigation. Interpretation and subsequent inferences should be subjective and for the researcher to decide. It was never intended that clinical significance be inferred based on a value of *p* less than 0.05. Neyman and Pearson subsequently proposed statistical hypothesis testing, suggesting that it was not possible to have a null hypothesis without an alternative one. Furthermore, they suggested the maximum probabilities of making incorrect decisions—that is, type I and II errors—should be set in advance. A type I error occurs if based on the sample the statistical null hypothesis of equipoise is rejected in favor of the alternative, when in the population the position of equipoise holds. A type II error occurs if based on the sample the null hypothesis of equipoise is not rejected in favor of the alternative, when in the population the position of equipoise does not hold and there is a difference between groups in outcome. Type I and II errors are discussed in further detail elsewhere,[8] but the implications of such errors are described below.

The theories of Fisher and Neyman–Pearson have been combined to give traditional NHST. It is not entirely clear why this happened. Fisher and Neyman–Pearson were strongly opposed in their schools of thought, and traditional NHST represents a misunderstanding of their theories. Although their theories are different, there are similarities that may have led to their combination. Although Neyman and Pearson never advocated a probability of 0.05 for type I errors, the probability of one occurring can be defined in terms of statistical significance ($p < 0.05$) as suggested by Fisher. Rejection of the null hypothesis could be a type I error, and therefore a critical level of significance of 0.05 represents the maximum probability of a type I error given the statistical model. It has been suggested that the commercialization of textbooks, and the desire of publishers to promote cookbook recipe approaches rather than encouraging scrutiny of the data also played an important role in the combination of the two theories.[9]

3 | IMPLICATIONS OF TYPE I AND II ERRORS

Type I and II errors are major limitations to traditional NHST. Both types of error are conceptual, and it will not be known if they have occurred. Such errors are a necessary evil of traditional NHST. This is because rejection of the null hypothesis in favor of the alternative could be a type I error, and failure to reject the null hypothesis in favor of the alternative could be a type II error. With a critical level of significance of 0.05, the maximum probability of a type I error for a single statistical hypothesis test is 0.05. However, the probability of a type I error increases rapidly when multiple hypothesis tests are performed.[10] For example, when 20 hypothesis tests are performed the probability of a type I error is at least 0.64. In such situations, Bonferroni's correction factor is a simple approach to minimizing the probability of type I errors occurring.[10] Type II errors are generally attributed to small sample sizes, and their probability of occurring is difficult to predict.

Key message

Traditional null hypothesis significance testing incorporating the critical level of significance of 0.05 is used in clinical decision-making. However, it was never intended for clinical significance to be inferred from statistical significance.

Type I errors will lead to false claims, for example, about the effectiveness of an intervention, whereas type II errors lead to the dismissal of potentially important interventions. The concept of type II errors led to the phrase "*absence of evidence is not evidence of absence*".[11] That is, just because NHST fails to find statistical significance, it does not mean that a clinically important difference does not exist in the population.[12]

The potential for type I and II errors is considered to be high, and contributed to the idea that most research findings based on NHST are false.[13] Although it is possible to limit the probability of type I and II errors occurring, researchers rarely do. Coupled with the misunderstanding that statistical significance implies contextual significance, failure to acknowledge such errors impacts the validity of inferences in research.

4 | NHST AND SENSITIVITY TO SAMPLE SIZE

Studies with large sample sizes are important because as sample size approaches the population size, the sample estimates have increased accuracy when estimating the population parameters. However, NHST is sensitive to sample size, and large sample sizes are more likely to lead to statistical significance. As sample size increases, it results in increasingly smaller differences between groups being observed as statistically significant. Therefore, NHST guarantees that any difference, no matter how small or irrelevant, will be statistically significant if the sample size is large enough.[14] Conversely, statistical significance is less likely when sample sizes are small. Sample size considerations are important when planning a study in order to ensure meaningful effects are observed as statistically significant if they exist in the population.

5 | THE CALL FOR A STATISTICS REFORM

The controversy surrounding NHST and misconception that contextual significance can be inferred from statistical significance has been discussed for decades. In 1951, Yates commented on the misuse of NHST and suggested that the ultimate objective for researchers had become establishing if statistical significance existed.[15] This detracted from the primary objective of research—that is, interpreting the results and considering their potential contextual significance. It has been proposed that NHST be banned because of the high probability of type I errors and the implications for clinical practice.[16] However, others have been more restrained and suggested a greater understanding of NHST and the *p* value is needed, while using the process more cautiously.[17]

The debate and argument for a statistics reform has intensified within the last 10 years. In 2015, the editors of the journal *Basic and Applied Social Psychology* banned NHST and statistical significance based on the dichotomy of $p<0.05$ vs $p\geq0.05$.[18] They declared NHST was "invalid" and "We believe that the $p<0.05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research". In addition to NHST and statistical significance, the ban included any *p* values, test statistics, and statements about significant differences or lack thereof.

In 2016, the American Statistical Association (ASA) published a statement regarding NHST and statistical significance providing guidance on the context, process, and purpose of *p* values.[19] The statement focused on the informed use of statistical significance and inference in research, rather than banning NHST because it is frequently misused. In 2019, *The American Statistician* published a special issue titled *Statistical Inference in the 21st Century: A World Beyond $p<0.05$*.[20] The aim was to provide further guidance to the ASA statement on the use of NHST and statistical inference. This represented the views of the editors of the special issue and was not a statement on behalf of the ASA. The main message was "…'statistically significant'—don't say it and don't use it." Moreover, statistical significance was never meant to imply contextual importance. The aim was to stop studies being published with "$p = 0.049$" and "$p = 0.051$" that were directly opposing in their inferences.

Shortly after the special issue in *The American Statistician*,[20] a prominent article was published in *Nature* echoing similar views.[21] It advocated abandoning statistical significance based on the categorization of $p<0.05$ vs $p\geq0.05$, because it had led to "…hyped claims and the dismissal of possibly crucial effects". In particular, emphasis should be placed on the contextual importance of the study results. It was not recommended that *p* values be banned, but rather used alongside an emphasis upon sample estimates and the accuracy in them. Nonetheless, it has been suggested that to do so would ultimately promote bias.[22] Inferences based on subjectivity are prone to conflicts of interest. For example, researchers will be biased toward inferences that support their beliefs.

6 | RESPONSE TO THE CALL FOR A STATISTICS REFORM

Despite the calls for a statistics reform regarding the use of NHST and statistical significance in decision-making having been ongoing for decades, the practice has persisted. A major challenge is that there are no interpretations of these concepts that are simple and intuitive.[23] George Cobb commented at an ASA forum (2014) "We teach it because it's what we do; we do it because it's what we teach",[24] highlighting the circularity to the challenges.

Journal editors have become increasingly concerned about the use of NHST. Some journals have updated their statistical reporting guidelines in response to the ASA statement in 2016,[19] including *The New England Journal of Medicine*.[25] Although it is difficult to encapsulate the general approach adopted, it would appear that editors are cautious in their response to the call for a statistics reform. The process of NHST and statistical significance generally remains, but authors are increasingly encouraged to present sample estimates plus confidence intervals so as to describe their accuracy. Second, there is a greater emphasis on minimizing the potential for type I and II errors and therefore controlling the limitations of NHST. However, there is limited guidance on discussing the contextual importance of study results over and above the presence of statistical significance.

7 | SUMMARY

The use of NHST and its role in decision-making in healthcare research is controversial, not least because it is typically misunderstood and misused. The idea that an intervention is effective, or exposure to a risk factor is only important if the value of *p* is less than 0.05 is a reductionist view that does not always reflect clinical importance. There have been frequent calls for a statistics reform regarding the use of NHST in decision-making, including no longer using the concept of statistical significance. Nonetheless, the binary approach to decision-making is a convenient one, and its use has remained ubiquitous. Regardless of the future for statistical significance, there are calls for greater focus on the magnitude and accuracy of the observed effects and their clinical importance. This ultimately seems sensible and accords with the original intentions of Fisher.[5] However, to do so will bring challenges not least because of the subjectivity that will exist when interpreting study results. Journals have been cautious in their approach to calls for a statistics reform, and it appears that statistical significance will continue to play a role in decision-making also in obstetrics and gynecology. This may be acceptable if we continue to educate ourselves in the role of statistics, including controlling the probability of type I and II errors plus the importance of sample size. In particular, statisticians, researchers, and clinicians all need to recognize that a statistical answer based on NHST to the question posed is not necessarily an answer to the scientific question asked. Statistical inference does not automatically reflect clinical inference.

8 | PRIOR PUBLICATIONS ON RELATED ISSUES

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133.
2. Wasserstein RL, Schirm AL, Lazar NA, eds. Statistical inference in the 21st century: a world beyond p < 0.05. *Am Stat*. 2019;73(sup1):1–401.
3. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature*. 2019;567:305–307.

CONFLICT OF INTEREST

None.

AUTHOR CONTRIBUTIONS

Philip M. Sedgwick created the original idea, and led on the writing and revision of the manuscript. Anne Hammer, Ulrik Schiøler Kesmodel, and Lars Henning Pedersen all helped develop the idea for the manuscript, and contributed to the writing and revisions of the manuscript.

ORCID

Philip M. Sedgwick https://orcid.org/0000-0001-8859-2175
Anne Hammer https://orcid.org/0000-0002-4616-9827
Ulrik Schiøler Kesmodel https://orcid.org/0000-0003-3868-106X
Lars Henning Pedersen https://orcid.org/0000-0001-6726-1991

REFERENCES

1. Nahm FS. What the *P* values really tell us. *Korean J Pain*. 2017;30:241-242.
2. García-Pérez MA. Thou shalt not bear false witness against null hypothesis significance testing. *Educ Psychol Meas*. 2017;77:631-662.
3. Ziliak ST. The validus medicus and a new gold standard. *Lancet*. 2010;376:324-325.
4. Kennedy-Shaffer L. Before *p* < 0.05 to beyond *p* < 0.05: using history to contextualize *p*-values and significance testing. *Am Stat*. 2019;73:82-90.
5. Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd; 1925.
6. Neyman J, Pearson E. IX. On the problem of the most efficient tests of statistical hypotheses. *Proc R Soc, Lond, Ser A*. 1933;231:289-337.
7. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond Edinb Dublin Philos Mag J Sci*. 1900;50:157-175.
8. Sedgwick P. Pitfalls of statistical hypothesis testing: type I and type II errors. *BMJ*. 2014;349:g4287.
9. Gigerenzer G. Statistical rituals: the replication delusion and how we got there. *Adv Meth Pract Psychol Sci* 2018;1(2):198-218.
10. Sedgwick P. Multiple hypothesis testing and Bonferroni's correction. *BMJ*. 2014;349:g6284.
11. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
12. Sedgwick P. Understanding why "absence of evidence is not evidence of absence". *BMJ*. 2014;349:g4751.
13. Ioannidis JPA. Why Most published research findings are false. *PLoS Med*. 2005;2:e124.
14. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front Hum Neurosci*. 2017;11:390.
15. Yates F. The influence of "statistical methods for research workers" on the development of the science of statistics. *J Am Stat Assoc*. 1951;46:19-34.
16. Hunter JE. Needed: a ban on the significance test. *Psychol Sci*. 1997;8:3-7.
17. Cohen HW. *P* values: use and misuse in medical literature. *Am J Hypertens*. 2011;24:18-23.
18. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psychol*. 2015;37:1-2.
19. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129-133.
20. Wasserstein RL, Schirm AL, Lazar NA, eds. Statistical inference in the 21st century: A world beyond p < 0.05. *Am Stat*. 2019;73(sup1):1-401.
21. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature*. 2019;567:305-307.
22. Ioannidis JPA. Retiring statistical significance would give bias a free pass. *Nature*. 2019;567:461.
23. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-350.
24. Cobb G. Personal communication: ASA Discussion Forum. Mount Holyoke College; 2014.
25. Harrington D, D'Agostino RB, Gatsonis C, et al. New guidelines for statistical reporting in the *journal*. *N Engl J Med*. 2019;381:285-286.

How to cite this article: Sedgwick PM, Hammer A, Kesmodel US, Pedersen LH. Current controversies: Null hypothesis significance testing. *Acta Obstet Gynecol Scand*. 2022;00:1–4. doi: 10.1111/aogs.14366