Trials and Tribulations of Teaching NHST in the Health Sciences

Philip M. Sedgwick
Institute for Medical and Biomedical Education,
St. George's, University of London, London SW17 0RE, UK

Abstract
Null hypothesis significance testing (NHST) with a critical level of significance of 5% (P<0.05) has become the cornerstone of research in the health sciences, underpinning decision making. However, considerable debate exists about its value with claims it is misused and misunderstood. It has been suggested it is because NHST and *P*-values are too difficult to teach, and encourage dichotomous thinking in students. Consequently, as part of statistics reform it has been proposed NHST should no longer be taught in introductory courses. However, this paper will consider if the misuse of NHST principally results from it being taught in a mechanistic way, along with claims to knowledge in teaching and erosion of good practice. Whilst hypothesis testing has shortcomings, it is advocated it is an essential component of the undergraduate curriculum. Students' understanding can be enhanced by providing philosophical perspectives to statistics, supplemented by overviews of Fisher's and Neyman-Pearson's theories. This helps the appreciation of the underlying principles of statistics based on uncertainty and probability, plus the contrast of statistical with contextual significance.

Keywords: Teaching, Null hypothesis significance testing, *P*-value, Statistical inference, Probability, Uncertainty.

1. Introduction

Traditional null hypothesis significance testing (NHST) with the null and alternative hypotheses, plus critical level of significance of 0.05 (5%) needs little introduction. It's central to any undergraduate or postgraduate curriculum in the healthcare sciences. Statistical significance as indexed by $P<0.05$ has become the cornerstone of decision making, underpinning conclusions in the healthcare literature.

Whilst the *P*-value is a valuable statistical measure it is frequently misused and misinterpreted. Concern has focussed on the concept of statistical significance based on the categorisation of $P<0.05$ versus $P\geq0.05$, with the implication that statistical significance implies contextual significance. Whilst the scientific community has discussed the misuse and misinterpretation for decades, the debate has intensified within the last ten years. In particular, the journal *Basic and Applied Social Psychology* (*BASP*) banned the use of NHST in 2015. An editorial by Trafimow and Marks claimed 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'*. The ban not only included NHST, but any traces of it including *P*-values, test statistics, and statements about significant differences or lack thereof. More recently a group of statisticians and scientists recommended the abandonment of statistical significance based on the categorisation of $P<0.05$ versus $P\geq0.05$ (Amrhein *et al.*, 2019). It was claimed that such categorisation has led to *'hyped claims and the dismissal of possibly crucial effects'*. In 2016, the American Statistical Association (ASA) published a statement on *P*-Values providing guidance on context, process, and purpose (Wasserstein and Lazar, 2016). This provided a balanced approach to the discussion. It focussed on encouraging the informed use of statistical inference in science, rather than banning the practice simply because it is frequently misused and misinterpreted. Subsequently in 2019, Wasserstein *et al*. edited a special issue in *The American Statistician* titled *"Statistical Inference in the 21st Century: A World Beyond p < 0.05"*. The overarching aim was to provide guidance as to how to undertake statistical inference sensibly. It was felt that the ASA statement in 2016 concentrated too much on what not to do rather than what to do, not least because there is widespread agreement about the don'ts.

The recent discussion about the misuse and misinterpretation of NHST and the *P*-value are not new. It would appear that ever since statistical significance has been suggested based on P<0.05, there has been debate as to whether the execution of a statistical hypothesis test of significance was the ultimate objective for researchers, who paid too little attention to the sample estimates of the population parameters they are investigating. This was advocated by Fisher himself (1935), and later Yates (1951). However, it is not clear why the misuse and misinterpretation of statistical inference continues to persist? In a personal communication at an ASA forum in 2014, George Cobb no doubt highlighted why little progress has been made with regards statistical inference based on the categorisation of P<0.05 versus P≥0.05; "*We teach it because it's what we do; we do it because it's what we teach.*" Therefore, it is suggested that if change is going to happen then the best place to start is in the classroom. In this paper, I give an outline of my approaches to teaching of statistical inference incorporating NHST and the *P*-value as delivered on undergraduate and postgraduate courses. My teaching is so-called "service teaching", delivered to non-experts in the healthcare sciences. The aim is to help students appreciate good practice in their future careers when interpreting, and undertaking statistical inference. I also describe some of my experiences with teaching this topic, and possible barriers to change.

2. Teaching NHST

Hak (2014) presented arguments for not teaching NHST in (introductory) undergraduate courses. It was claimed NHST was too complex to teach, difficult for students to understand, and encouraged dichotomous thinking. Moreover, the dichotomous thinking inherent to NHST based on the categorisation of P<0.05 versus P≥0.05 was a cognitive obstacle for interpretation. Post and van Duijn (2014) also discussed how researchers and students alike find the interpretation of statistical inference based on NHST difficult. In a survey of methodology instructors, approximately 80% were found to have problems with the interpretation of *P*-values (Haller and Kraus, 2002).

Personal experience would suggest that the key challenge to teaching NHST and the *P*-value is that such concepts are based on abstract ideas. Statistical inference is based on probability, which in itself is an unintuitive and difficult concept for many (Spiegelhalter, 2010). In particular, we derive a single test statistic based on the data collected in a study using hypothesis testing, and ascertain the evidence it provides to support the statistical null hypothesis using probability based on a *P*-value as referenced against a sampling distribution. This is unintuitive since probability is based on a large number of events, and yet we have results from a single study. Moreover, there are no interpretations of the *P*-value that are simple and intuitive.

In recent years there has been much discussion about the teaching of statistics to non-specialists and the many challenges it presents. Undergraduate students studying statistics as part of their healthcare degree program often do not have an interest in the subject, and typically did not come to university to study it. Furthermore, students have differing abilities when studying mathematical subjects, and for some studying statistics or quantitative methods can generate anxiety. To help address these challenges there has been a shift in teaching to statistical thinking (and critical appraisal) rather than calculations, formulae and theory. Moreover, the shift has been to help students see the application of statistics in the real-world rather than learn theoretical contexts that do not have an obvious application. With a shift towards a curriculum based on statistical thinking, it may not be surprising that the philosophical concepts that underlie statistical theory and statistical inference are omitted for simplicity. This may have contributed to NHST and statistical significance being typically taught and seen in a mechanistic way – almost as a recipe. In particular, the only point of interest in this process has become the presence of statistical significance (P<0.05), or lack of it. It may be such that statistical significance is subsequently taught or viewed as contextual significance, not least to give simplicity to teaching and help give a real-world application.

I believe that the biggest challenge for students in understanding NHST and the P-value is the lack of appreciation of the underlying principles of statistics. Indeed, very few statistics texts even consider this. Therefore, it is essential to pay consideration to the underlying principles based on uncertainty and probability. This ultimately means more, not less teaching of the theoretical concepts underpinning statistical testing.

2.1 Teaching Uncertainty

Within my teaching the concept of uncertainty is delivered using the example of a randomised controlled trial published in the *BMJ* (Vinding *et al*., 2018). The study was a double-blind, placebo-controlled superiority trial, the aim of which was to investigate the effects of fish oil supplementation in pregnancy on infant body growth and composition. The participants were 736 expectant mothers and their offspring, recruited between 22, and 26 weeks gestation following presentation. The primary outcomes included the body mass in children at age 6 years. The purpose of using this example was to give teaching a real-world application.

Statistics is based on the theoretical concept of repeated sampling, with the same sample size, and under the same conditions from an infinite population. The sample estimate of the population parameter is the difference between the intervention and placebo groups in mean body mass of the children at age six years. By taking different samples from the population we achieve different sample estimates, each of which can be represented by an arrow on a target board. The bullseye of the target board represents the population parameter. Some arrows will be closer to the bullseye than others, a reflection of sampling at random from the population. We have just one study and the aim is for that arrow to be as close to the bullseye as possible, thereby reduce sampling error. However, we do not know where the arrow for our single study falls upon the target board. This idea illustrates there is nothing special or magical about the sample estimate from a single study, whilst it serves to develop awareness of uncertainty within statistical inference.

2.2 Sampling Distributions

Having introduced the concept of infinite sampling, students can begin to appreciate the concept of test statistic distributions. Consider testing the difference between the intervention and placebo groups in the primary outcome using the independent samples t-test. We derive a test statistic for each of the theoretically infinite number of samples. A histogram of these statistics would represent the t-distribution, similar in shape to a Normal distribution. Students are reminded that such distributions are theoretical, generated by the computer, and referred to as sampling distributions. The distribution is symmetrical about zero because the sample mean for the intervention group could be smaller, or greater in magnitude than the placebo group. It is essential to keep an open mind as to the effects of the intervention – it may be shown to be inferior to placebo. The larger the difference between treatment groups in the sample mean of the primary outcome, the larger the magnitude of the absolute value of the test statistic. Consequently, the test statistic will be further away from the point of equipoise between treatments (i.e. a difference of zero between treatment group sample means), providing less evidence to support the null hypothesis. We derive the *P*-value by referencing the single test statistic obtained for our study against the sampling distribution. We take the absolute value of the test statistic, and the *P*-value is the total area bounded by the curve of the sampling distribution to the right of the test statistic on the positive-side on the X-axis, and to the left on the negative-side. Statistical significance is achieved if the single test statistic for our study is within the tails of the distribution – that is, those 5% of samples with the largest differences.

2.3 So What Does *P* = 0.049 mean?

By teaching the underlying principles of statistics based on uncertainty and probability, it helps students appreciate what a *P*-value is. There is nothing magical about statistical significance – it is simply achieved if the single study is one of the extreme 5% of the theoretical samples achieved under repeated sampling. Hence statistical significance is a mathematical concept. For postgraduate students who have experienced statistics teaching in their undergraduate degrees, this is often something of a revelation. Typically, their teaching on NHST and *P*-values has been based on a recipe or black box approach with no instruction as to how the *P*-value is derived. Furthermore, if statistical significance is achieved (P<0.05), it somehow (magically) implies importance.

Within my teaching students are encouraged to consider the contextual significance of study results, regardless of whether statistical significance (*P*<0.05) is achieved. In particular, they should consider the data collected and the sample estimates of the population parameters being investigated for potential clinical significance. However, that is difficult for students with very limited expertise, if any, in the area of speciality that the teaching is based upon in order to give it a real-world application. No doubt that has contributed to the difficulties in teaching when encouraging students to consider the contextual importance of results when statistical significance is achieved (*P*<0.05).

2.4 History of NHST

Kennedy-Shaffer (2019) proposed that in order to appreciate the challenges facing statistical inference based on statistical significance today, and to overcome these challenges in the future, it is imperative we consider the history of the discipline. Within my teaching students are given a short historical perspective of the development of NHST plus statistical significance. Traditional NHST is a single process based on two distinct theories as proposed by Fisher (1925), and Neyman-Pearson (1933). Fisher and Neyman-Pearson were fiercely opposed in their schools of thought, and no doubt the combination of their theories represents a misunderstanding. Fisher proposed the null hypothesis and *P*-value. It was proposed that the *P*-value was the strength of evidence supporting the null hypothesis. Whilst Fisher advocated *P*=0.05 (5%) for statistical significance, it was not an absolute cut-off; in particular, interpretation was meant to be subjective and for the researcher to decide. Fisher's intention was for statistical significance to be used as a tool to indicate if the results warranted further investigation. Neyman and Pearson (1933) introduced hypothesis testing, advocating it was not possible to have a null hypothesis without an alternative one. Furthermore, the probabilities of making incorrect decisions – namely type I and II errors, were set in advance. They did not necessarily advocate a cut-off of 5% (0.05) for Type I errors (and therefore) statistical significance. Whilst the theories of Fisher and Neyman-Pearson are distinct, they have similarities which are often misunderstood and confused; it is this that has no doubt led to the combination of their theories. In particular, some of the confusion occurs because the critical region of the Neyman-Pearson theory can be defined in terms of Fisher's *P*-value. It is not clear how or why these two distinct theories were combined.

2.5 Developing a Culture of Uncertainty

Throughout all of my teaching, and in particular when teaching statistical inference incorporating NHST and the *P*-value, I remind students that the underlying principles of statistics are based on uncertainty and probability.

Data is evidence, and particular the resulting *P*-value is the evidence in support of the null hypothesis. When *P*-values are small, and in particular if we have statistical significance (P<0.05), then we have little evidence to support null hypothesis and should consider rejecting the null in favour of the alternative hypothesis. When teaching, there is avoidance of indicating that statistical hypotheses are true or false. This is because statistical hypotheses can never be proven or disproven. Statistical inference is based on infinite sampling from a population, and each sample will give a different sample estimate for the population parameter. We have no reason to believe that any sample is better or worse than any other sample in terms of the accuracy when estimating the population parameter. In particular, '*Absence of evidence is not evidence of absence'* (Bland and Altman, 1995; Sedgwick, 2014). Equally, "*Evidence of a difference is not evidence of no difference*" (Sedgwick 2021). Furthermore, I have also always avoided indicating that the inference from statistical hypothesis testing is positive (P<0.05), or negative (P≥0.05). The words positive and negative have connotations in themselves, in particular success or failure We should never consider scientific results as such, and be careful of encouraging students to think so through the use of our language, intentional or otherwise.

Students and researchers typically find it challenging that there is a difference between statistical inference and the "truth". It is not obvious if healthcare can live with the uncertainty in clinical effectiveness as presented by probability, not least because it is at odds to what healthcare practitioners want and are trained to provide. There is the underlying desire to establish if a treatment or therapeutic regimen is superior to standard care or placebo, and the dichotomous thinking offered by NHST is a convenient one. For most clinicians, '*a statistically significant P-value is the end of the search for truth'* (Banerjee, Jadhav & Bhawalkar, 2009).

3. Barriers to Change

Of particular concern are two barriers to change, which prevent the scientific community from misunderstanding and misinterpreting NHST and P-values in the application of statistical inference (P<0.05). These are *claims to knowledge*, and *erosion of good practice*.

3.1 Claims to Knowledge

Schwab-McCoy (2018) described how the teaching of statistics in undergraduate courses at many institutions is not delivered by statisticians with a formal training in the discipline. The teaching might be provided by a variety of departments including biology, sociology, or psychology, and is often delivered by non-statisticians. Compared to a formally-trained statistician, it is most likely that such individuals will not have developed the same understanding of statistics and the principles that underlie the speciality. Such claims to knowledge can lead to a wide variation in content, and pedagogy. Given the ongoing debates surrounding the misuse and misinterpretation of statistical inference, plus so-called "*P*-hacking", it is necessary to develop our understanding of the state of the teaching of statistics. By doing so, Schwab-McCoy (2018) suggested we may be able to identify areas for improvement, plus potential collaboration and partnerships with non-statisticians. Nonetheless, it is accepted that even trained statisticians, depending on their teaching, may still lack understanding of statistics and the principles that underlie the speciality. Therefore, they too would benefit from such partnerships.

3.2 Erosion of Good Practice

Erosion of good practice represents research practice in students that differs to the good practice delivered in their teaching. I have found this to be a particular challenge when teaching postgraduate students. Following my teaching they are subsequently supervised by a research supervisor who are typically not a statistician. Sometimes the supervisors have little training in statistics, and their prime interest is the execution of a statistical hypothesis test of significance. The ultimate objective is the categorisation of the *P*-value (P<0.05 versus P≥0.05), with little or no attention paid to the sample estimates of the population parameters they are investigating. Statistically significant results are coveted. Such behaviour may be driven by the notion that statistically significant results are more likely to be published, and are the only thing that journals are interested in. Nonetheless, it leads to publication and a curriculum vitae that increases in length.

Without doubt, many of the challenges presented by claims to knowledge and erosion of good practice have arisen through the concept proposed by Cobb (2014) "*We teach it because it's what we do; we do it because it's what we teach.*" We are in a vicious cycle whereby statistical inference is taught as a recipe and P<0.05 infers contextual importance, not least because we have always done that. This premise is passed down through cohorts of students. These students subsequently become teachers and researchers, who pass the same idea down to future cohorts of students.

4. Remarks

By teaching the underlying principles of statistics based on uncertainty and probability, it is my belief that it is has facilitated a deeper appreciation of the role of hypothesis testing in statistical inference. Nonetheless, such claims are difficult to verify. The assessment of such principles is difficult at the undergraduate level, not least since assessment is typically limited to short answer questions which do not permit sufficient investigation of such ideas. Whilst such ideas may be difficult for some students to appreciate, they are also essential to the understanding of confidence intervals which are central to any undergraduate or postgraduate curriculum.

Acknowledgements

I would like to thank all my undergraduate and postgraduate students in successive cohorts. Through their constructive interaction they have helped me develop my teaching materials, and approaches to learning.

References

Altman, D.G. & Bland, J.M. (1995). Absence of evidence is not evidence of absence. *BMJ*, 311, 485.

Amrhein, V., Greenland S. & McShane, B. (2019). Retire statistical significance. *Nature*, 567, 305.

Banerjee, A., Jadhav, S. L. & Bhawalkar J. S. (2009). Probability, clinical decision making and hypothesis testing. *Industrial Psychiatry Journal*, 18(1), 64–69.

Hak, T. (2014). After statistics reform: Should we still teach significance testing? In Makar, K., de Sousa, B. & Gould, R. (eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014)*. Flagstaff, Arizona, USA.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. 1st edn. Edinburgh, UK: Oliver and Boyd.

Fisher, R.A. (1935). *The Design of Experiments*. 1st edn. Edinburgh, UK: Oliver and Boyd.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, 7(1). http://www.mpr-online.de

Kennedy-Shaffer, L. Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize *p*-Values and Significance Testing. In Wasserstein, R.L., Schirm A.L. & Lazar N.A. (eds) (2019). Statistical Inference in the 21st Century: A World Beyond p < 0.05. *American Statistician*, 73(sup1), 82-90.

Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society London A,* 231, 289–337.

Post, W.J. & van Duijn, M.A.J. (2014). Teaching hypothesis testing: A necessary challenge. In Makar, K., de Sousa, B. & Gould, R. (eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014)*. Flagstaff, Arizona, USA.

Schwab-McCoy, A. (2018). Statistics Education Across the University: A Systematic Review. *Joint Statistical Meetings (JSM), Annual Meeting of the American Statistical Association*, Vancouver, July 28th-August 2nd.

Sedgwick, P. (2014). Understanding why "absence of evidence is not evidence of absence". *BMJ*, 349, g4751.

Sedgwick, P. (2021). Trials and Tribulations of Teaching NHST in the Health Sciences. *Joint Statistical Meetings (JSM), Annual Meeting of the American Statistical Association*, Virtual Conference, August 8th-12th.

Spiegelhalter, D. (2021). Why Do People Find Probability Unintuitive and Difficult? https://nrich.maths.org/7326 (Accessed: 6 October 2021).

Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology,* 37(1), 1–2.

Vinding, R.K., Stokholm, J., Sevelsted, A., Sejersen, T., Chawes, B.L, Bønnelykke, K., Thorsen, J., Howe, L.D., Krakauer, M. & Bisgaard, H. (2018). Effect of fish oil supplementation in pregnancy on bone, lean, and fat mass at six years: randomised clinical trial. *BMJ*, 362, k3312.

Wasserstein, R.L. & Lazar, N.A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2), 129-133.

Wasserstein, R.L., Schirm A.L. & Lazar N.A. (eds) (2019). *Statistical Inference in the 21st Century: A World Beyond p < 0.05. The American Statistician*, 73(sup1).

Yates, F. (1951). The influence of “Statistical methods for research workers” on the development of the science of statistics. *Journal of the American Statistical Association*, 46, 19-34.