PERSISTENT MISCONCEPTIONS CONCERNING NULL HYPOTHESIS SIGNIFICANCE TESTING

Bootheway, G. B. P.
St. Bonaventure University

ABTRACT

The Null Hypothesis Significance Test is a popular testing procedure used mainly in the social sciences and occasionally in the natural sciences. The procedure has been criticized on several grounds. Firstly, opponents have attacked the procedure itself, claiming that it either does not tell the researcher what he or she wants to know; and/or that the procedure is not logically valid. Secondly, opponents argue that the concept of statistical significance is flawed on the ground that is it either too arbitrary or, that statistical significance cannot confer substantive significance. This paper argues that critics of Null Hypothesis Significance Test either lack a proper understanding of how and why the test works, or they apply the test in an incomplete manner.

INTRODUCTION

The controversy surrounding the Null Hypothesis Significance Test (hereafter NHST) is the malevolent comet of statistics, not only does it return with predictable periodicity, but its intensity has now increased to such an extent that criticisms that were hitherto distinct have now become indistinguishable. The purpose of this paper is not resolution but clarification. Whilst it is true that NHST and statistical significance are logically intertwined, this does not necessitate that they become irrevocably entangled. For example Ziliak and McCloskey (1996, 2008) question the validity of NHST because of issues they have with the concept of statistical significance, however these issues seem to evaporate when they pass favourable judgment on the procedure of Estimation, which is nothing more than an algebraic re-statement of a Fisherian (as opposed to a Neyman-Pearson) NHST. The argument seems alarmingly circular.

Others (Kirk, 1996; Nickerson, 2000) object to the dichotomous nature of the NHST, claiming that critical values are simply too arbitrary. While other still (Cohen, 1994; Kline 2004) complain that NHST does not tell the researcher what they know, i.e., the probability that the null hypothesis is actually true. Cohen amongst others (including Ziliak and McCloskey) further argues that rejection of the null hypothesis is inevitable as sample size increases. And, for good measure, Cohen finally declares the NHST ”a misapplication of deductive syllogistic reasoning.”

Although NHST and Statistical significance are coupled, the former is a procedure the latter an evaluation. By examining each in isolation (as much as one may) it is hoped that some of the aforementioned objections made be defused on the grounds that: a) The procedure is misunderstood, b) The procedure is incompletely applied and c) Conclusions are misinterpreted. The paper will proceed as follows. Section I outlines the mechanics of a NHST. Section II considers the issue of Statistical significance. Section III provides a numerical example. Section IV examines the objections to NHST as a procedure. Section V examines objections to NHST via issues surrounding statistical significance. Section VI concludes.

I. T**HE NULL HYPOTHESISGNIFICANCE TEST**

Historically, there are two versions of the NHST, the version formulated by Fisher (1925) and the version formulated by Neyman and Pearson (1933). Philosophically they are different; indeed the very public rows that took place between Fisher and Pearson have become part of the lore of Statistics. However, for the purposes of this paper no distinction is made unless where noted. For purposes of exposition the Neyman-Pearson version is used since it is the more commonly accepted.

The NHST is a procedure designed to test a claim made about a population parameter. The null hypothesis (denoted H_0) represents the *status quo*, this status quo is not the truth about a parameter, but instead a presumption about the present state of the parameter. For example, in a criminal proceeding, a defendant is not innocent till proven guilty, but *presumed* innocent till proven guilty. A claim is made that contradicts the null hypothesis named the alternative, or research hypothesis (denoted either H_1 or H_a). H_0 and H_1 must be both mutually exclusive and collectively exhaustive. Evidence is collected, summarized and evaluated; if the evidence against the null hypothesis is both sufficient and deemed not to have occurred by chance, the null hypothesis is rejected. Otherwise, the null hypothesis cannot be rejected and the status quo prevails. *Ei incumbit probatio qui dicit, non qui negat* (*Digesta seu Pandectae: 22.3,2 circa AD 530-521:* proof lies on him who asserts, not on him who denies).

The salient features of the NHST are therefore as follows:

1) The test pertains only to the presumptive null hypothesis.
2) The test does not pertain to the alternative hypothesis or the data (evidence).
3) The null hypothesis represents a status quo, not truth or fact.
4) Non-rejection of the null hypothesis does not affirm a truth or a fact.
5) Rejection of the null hypothesis does not suggest that the null hypothesis is wrong or false.

Or, in the words of R.A. Fisher himself (1935): [Coining the term Null Hypothesis]

> "In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

When first introduced to Statistics, student are often befuddled by such murkiness, However, unlike mathematicians who are often afforded the luxury of dealing with identities – statements that are true always and everywhere, statisticians toil in a universe of propositions where argument may not be fully specified or dependent on other variables. The existence of such uncertainty does not relieve the statistician of the burden of philosophical underpinning and logical reasoning; it does however demand that conclusions drawn from statistical procedures be presented conservatively and with appropriate caution.

As shall be made clear in later sections, misunderstanding the procedure of NHST is often a source of the complaints.

II. STATISTICAL SIGNIFICANCE

Historically, the notion of significance can be traced as far back as Cicero (*De Divinatione, 44BC)* or, more recently to Arbuthnott (1710-12) and Laplace (1778) when investigating variation in birth gender. In the nineteenth century both Edgeworth (1885) and Venn (1887) alluded to it; but the term itself was first coined by Fisher (1925). Statistical Significance refers to the weight of evidence required to reject a null hypothesis? However, evidence obtained by statistical methods suffers from the disadvantage that it might have been caused by chance alone. Thus, results are said to be statistically significant if there is only a small probability that they could have occurred by accident. In modern statistics, three levels of significance are deemed conventional: 90%, 95% and 99%, otherwise stated, results could have occurred by chance: 10%, 5% or 1% respectively. These levels are, of course, arbitrary, and therefore controversial in and of themselves. Indeed Fisher himself was somewhat Quixotic in his choice of levels.

Fisher (1926, p. 504)

> “... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."...”

Fisher (1946, p.80)

> “In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: *Either* the hypothesis is untrue, *or* the value of χ^2 has attained by chance an exceptionally high value. The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.”

Fisher (1956, p.41-42)

> “The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.”

III. A NUMERICAL EXAMPLE OF A NHST

Suppose it is believed that the average length of a particular species of snake is less-than-or equal to ten-feet. This is our null hypothesis:

H_0: $\mu_0 \leq 10'$where μ is population men

A herpetology might claim otherwise, suggesting that the average length of said species is actually greater than ten-feet. This is our alternative or research hypothesis.

H_a: $\mu > 10'$

Evidence is collected in the form of a sample drawn from the population and summarized in the following test statistic:

$$Z = \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \qquad (1)$$

Where, $\bar{X}$ = The mean length of the sample
μ_0 = The presumptive population mean (10')
S_x = The sample standard deviation length
n = The sample size

Suppose the following results were obtained: $\bar{X}$ = 10.5' S_x = 1.25' n = 64

Then $Z = \frac{10.5-10}{1{,}25/\sqrt{64}} = 3.2$

Since the probability of obtaining a test statistic of Z = 3.2 is 0.0007 (about 1-in-1,428), the null hypothesis can be rejected at all three levels of significance 90% (one-in-ten), 95% (one-in-twenty) and 99% (one-in-one hundred)

IV. OBJECTIONS TO NHST AS A PROCEDURE

Objections to NHSTs as a procedure fall into two categories. The first claims that NHST does not tell a researcher what he or she wants to know, which according to Cohen (1994, p.997) is, "Given these data, what is the probability that the null hypothesis is true?" Given that the first word of the quote is "given" (no pun intended), it should come as no surprise that this objection is the standard bearer of those who adhere to Bayesian inference. The second objection, again due to Cohen (1994, p.997-998) purports to show through a Bayesian example, that a NHST may lead to contradictory null hypotheses. Finally, the third objection attacks the logic of the NHST by claiming that it is a misapplication of deductive syllogistic reasoning to assume something to be true (the null hypothesis) in order to demonstrate that it is false – or vice versa (Roseboom, 1960; Carver, 1978; Shaver, 1993).

The first objection has the unfortunate sound of having come from a researcher who has had one too many hypotheses rejected, and now seeks a testing procedure more amenable to his theories. The second objection is even more unfortunate since the example provided by Cohen, is not only controversial from the Bayesian point of view, but it reveals a misunderstanding of the NHST procedure.

Cohen asks his reader consider a scenario where 98% of a population is free a particular disease whilst 2% are not. Using the standard notation of probability:

$P(\overline{D}) = 0.98 \qquad P(D) = 0.02$

Cohen therefore declares his null and alternative hypotheses to be:

$H_a \qquad P(\overline{D}) = 0.98$

$H_a \qquad P(D) = 0.02$

Cohen next provides the reader with some likelihoods.

The probability that a diseased individual will test positive for the disease is 95% and the probability that a disease-free individual will test positive for the disease (a false positive) is 3%. Again, using the probability notation we write these:

$P(+|D) = 0.95 \quad P(+|\overline{D}) = 0.03$

We may now use Bayes Theorem to answer the inverse probability question: what is the probability that a an individual is disease-free, given that he or she has tested positive.

$$P(\overline{D}| +) = \frac{P(+|\overline{D}).P(\overline{D})}{P(+|\overline{D}).P(\overline{D})+P(+|D).P(D)} = \frac{(0.03)(0.98)}{(0.03)(0.98)+(0.95)(0.02)} = 0.6074$$

Cohen therefore claims he has produced two contradictory null hypotheses, the first being the original prior: $P(\overline{D}) = 0.98$, and the second being the posterior probability $P(\overline{D}| +) = 0.6074$.

The arithmetic is correct, indeed with slightly different numbers this is a standard example of Bayes Theorem. However, Cohen has erred in defining his priors as relative frequencies, as such frequencies would imply that the priors are, as Bayesians would put it, *informed* rather than *uninformed*. Thomas Bayes himself argued for the necessity of uninformed priors, and it was perhaps the only issue surrounding Bayesian analysis that Ronald Fisher and Harold Jeffries (1939, 1948, 1961) ever agreed upon. Whilst it true that more recent writers such as Jaynes (2003) have argued for the permissibility of informed priors, the stipulation has been that they be uniformly distributed (which is barely different from Jeffries' choice of 50% in an either/or situation). Thus, before even moving on to how this may impact NHST, Cohen has defined a example that is controversial to own his own Bayesian brethren.

However, let us suppose, *arguendo* that none of the previous paragraph mattered. Cohen's case against NHST fails because the relative frequencies he used to establish his initial prior (0.98), and which he then uses to establish a second, seemingly contradictory prior (0.6074), are, by their very definition, qualities of a sample – not a population. The whole point of hypothesis testing is to learn something about a parameter not a statistic. The null hypothesis must refer to a population parameter, because only a population parameter can provide us with a knowable sampling distribution against which we can compare the value of our test statistic. Cohen's null hypothesis is a statement about a sample, so it cannot tell us anything about the sampling distribution of a test statistic.

The only possible way that Cohen can rescue the situation is to argue that his original prior of

0.98 was a relative frequency drawn from the entire population, which begs two questions: firstly, if the population is known, why are we bothering with a hypothesis test in the first place? Secondly, how can the second "prior" (0.6074) be contradictory to the first when it is a statement about a completely different population? The first prior refers to a population made up of diseased and non-diseased individuals. The second prior refers only to the smaller population of those who have tested positive.

The third objection is perhaps the most surprising, since it objects to the logic of the well established mathematical device of *proof by contradiction* of which the NHST is but one of many examples, the most famous of which being Euclid's proof of that there are an infinite number of primes. Euclid starts with the false assumption that, in fact only a finite number of primes exist, and then proceeds to generate a contradiction. The logic of the NHST is no different.

It is possible, although none of the cited proponents of the "NHST is illogical" argument actually say this in black and white; that they believe that no probabilistic argument can be accorded logical validity. This may indeed be so, but an argument can still reasonable without being formally valid. For example: *If you are bitten by a black mamba (Dendroaspis Polylepis) and do not receive anti-venom, you will probably die of cardiac arrest. You were bitten by a black mamba and did not receive anti-venom. You will probably die of cardiac arres*t. The probabilistic argument is not logically valid since the one can accept the premises, but still reject the conclusion. But it does not detract from the fact that getting the anti-venom as soon as possible might be a very good idea.

V. ISSUES SURROUNDING STATISTICAL SIGNIFICANCE.

Critics of the notion of statistical significance generally fall into three camps. The first (Kirk, 1996; Nickerson, 2000) complain that the Reject/Do not reject dichotomy is too arbitrary. The second argue that statistical significance is too often misinterpreted as substantive significance (Ziliak and McCloskey, 1996, 2008). While the third points out that rejection of a null hypothesis is almost guaranteed if a large enough sample is chosen (Kadane, 2011). Each of these points will be examined in turn, but it will be argued that collectively, all three objections seem to arise because NHST are applied incompletely or without proper experimental design.

In some respects it is easy to be sympathetic to the first objection since, as we saw in Section II, Fisher himself dithered on the probabilities required for rejection. Furthermore, the current research practice of reporting all three conventional levels of confidence (0.1, 0.05 and 0.01) might provide the illusion of rigor, but it may also result in hypotheses being consigned to some *oubliette* should they pass one or two but not all three of the statistical hurdles. Indeed, over the course of his life, Fisher became more disenchanted with such hurdle rates.

Fisher (1951).

> "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."

While Fisher may have become more dissatisfied with statistical significance in the sense of a one-off test, it remained and indeed still remains, the *only-game-in-town*. It should be remembered that rejection of a hypothesis (at say the 0.05 level) merely states that the result is unlikely to have occurred by chance, it does not say that the result definitively did not occur by chance. In that sense, the term "reject" may have been a poor choice, but it one that we may have to live with. Notice also that Fisher, first and foremost a scientist, shows a marked preference for results that can, if possible, be replicated. In that sense, those who object that the arbitrariness of NHST are being too capricious in their interpretation of rejection versus non rejection. Hypotheses should be retested and replication. Rejection is one instance may be indicative, but it need not be conclusive. Perhaps those who adhere to this objection are themselves being too arbitrary.

Statistics is a practical discipline, and practical disciplines require practical solutions. Often the arbitrary simply cannot be avoided. A motorist may, after having been handed a speeding ticket, ponder why the posted limit that he or she exceeded was set at, say 30 mph. It may seem unreasonable or even stupid, but it is simply the cost of doing business. However, as will be made clear further on, the arbitrary cut-offs are in fact not so arbitrary if experiments are properly designed and result are properly interpreted.

Criticisms two and three, the claim that statistical significance does not imply substantive significance and that rejection of a null is guaranteed with a large enough sample have become intertwined in the literature (Ziliak and McCloskey, 2008), and need to be separated. Let us first consider the issue of sample size. Certainly, a cursory examination of equation (1) in section III, reveals that the test statistic “Z” will indeed , ceteris paribus, increase as the sample size increases and will therefore inevitably drift into the rejection region. This however is only the case in the rarified instance where the researcher has the luxury of knowing the population standard deviation “σ”. In almost all empirical investigations the sample standard deviation has to be used in equation (1) because σ is unknown. Thus, the variance of differences implied by H_0 and H_1 actually decreases as the sample size increases if the differences are small, causing the researcher not to reject a null which should not be rejected.

Suppose instead that the population standard deviation is known. in this case the standard error will decline as the sample size increases potentially creating a Type I error (rejecting the null when it should be not be rejected. This is the problem of statistical versus substantive significance. The problem is certainly not new, but it has received increased scrutiny since Ziliak and McCloskey (1996, 2008) who argued that many papers may have obtained statistically significant results, but those results are economically (or substantively) insignificant.

Before proceeding to the quantitative, it may be worthwhile to consider the qualitative aspects of this claim. Statistical tests are neutral in the sense that data are collected and evaluated formulaically. The results do not evaluate themselves that is the job of the researcher. For example in a regression model, a particular coefficient may statistically significant but its contribution to R^2 is minimal. In another instance, it may be found that a statistically significant correlation exists between two variables that defies logical interpretation (e.g., price of onions in Buenos Aires and number of squirrels on a university campus). Is it possible that the requirement of justification is so ingrained and obvious to researchers that pedantic elaboration is a waste of valuable journal space. In their 1996 paper, Ziliak and McCloskey *keel-haul* a series of papers that appeared in the American Economic Review (AER). It could be argued that some interesting methodological issues were raised, but certainly nothing to suggest that misrepresentation or fraud was evident. Rigor is a good thing, but excessive rigors verges on the pedantic – readers of academic journals are (or should be) perfectly capable of reading such things are regression output without the clutter of paragraphs of explanation.

Quantitatively, light can be cast on the question of substantive significance by investigating the *statistical power of a test.* Specifically the ability of a test to distinguish small discrepancies (noise) from the null from large substantive discrepancies. It is regrettable that few introductory statistical text emphasize power, choosing instead to mention it, *en passant*, as merely the probability of rejection the null hypothesis – when indeed it should be rejected. This deficiency has contributed greatly to the misunderstanding of the NHST, as evidenced by Ziliak and McCloskey (2008, p.133) who write:

> "a good and sensible rejection of the null is, among other things, a rejection with high power".

To which they then add (Ziliak and McCloskey, 2008, p152)

> "refutations of the null are easy to achieve if power is low or the sample is large enough"

Unfortunately, as Spanos (2008, p.160) points out, Ziliak and McCloskey "have it backwards". Rejection with high power is the main source of the statistical/substantive problem. At high power with a large sample, the numerous small discrepancies from the null swamp the larger substantives ones. As such, rejection of the null ends up providing less evidence of the substantial rather than more. By contrast, rejection of the null with low power allows the substantive discrepancies to be caught in the net, while the smaller discrepancies simply pass through the mesh.

VI . CONCLUSION

The NHST, either the Fisher or Neyman Person version, is essentially a proof by contraction – a times tested technique dating by to at least the mathematicians of ancient Greece. The procedure is simple to execute and easy to interpret, as such it has found favour with many disciplines that have a need for statistical testing. Many of these disciplines are in the social sciences.

The procedure comes under periodic fire from two camps. The first find fault with the procedure itself, the second find fault with the concept of statistical significance. It has been argued that those who find fault with the procedure do so because they either lack the mathematical sophistication to understand the procedure's validity, or they lack the ability to discern what the procedure is able to, or not able to deliver. In both these respect, the NHST has become a victim of its own success – it is so easy to use that it ends up being used by the very people who should not be using it.

The second camp object to the concept of statistical significance. It has been argued here that such objections stem from an incomplete use of the NHST. More specifically that the design of experiments that rely on the NHST should consider the "power" of the statistical test, not as an optional extra, but instead as a vital component.

REFERENCES

Carver, R. P. (1978). *The Case against Statistical Testing.* Harvard Education Review, 48.

Cohen, J. (1994). *The earth is round (p < .05).* American Psychologist, 49, *997-1003.*

Edgeworth, Francis Y. (1885). Methods of Statistics. *Jubilee Volume of the Statistical Society*, June 22-24. Royal Statistical Society of Britain: 181-217

Fisher, R. A. (1925). *Statistical Methods for Research Workers* 8th ed. New York: G. E. Stechart and Co. 8th edition.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1955). “Statistical Methods and Scientific Induction.” Journal of the Royal Statistical Society, Series B (Methodological), Vol. 17, No. 1..

Fisher, R. A. (1956). *Statistical Methods and Scientific Infere*nce. New York: Hafner. Second edition.

Jaynes, E. T. *The Logic of Science*. Cambridge University, Press, Cambridge, UK.

Jeffreys, H. (1939). *Theory of Probability*, 1st ed. The Clarendon Press, Oxford.

Jeffreys, H. (1948). *Theory of Probability*, 2nd ed. The Clarendon Press, Oxford.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. *Oxford Classic Texts in the Physical Sciences*. Oxford Univ. Press, Oxford.

Kadane, J. B. (2011). *Principles of Uncertainty*. Chapman and Hall/CRC Press,

Kirk, R. E. (1996). *Practical significance: A concept whose time has come*. Educational and Psychological Measurement, 56, 746-759.

Kline, Rex B. , (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC, US: American Psychological Association

Ziliak, Stephen T., and Deirdre N. McCloskey (2008). *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. Ann Arbor (MI): The University of Michigan Press.

McCloskey, Deirdre, and Stephen Ziliak. (1996). *The Standard Error of Regressions*. Journal of Economic Literature 34(March): 97-114.

Neyman, J. and Pearson, E. S. (1933). *On the Problem of the Most Efficient Tests of Statistical Hypotheses*’, Philosophical Transactions of the Royal Society of London.

Nickerson, R. S (2000). *Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy*. Psychological Methods, 5, 241-301

Rozeboom , W. W. (1960). *The fallacy of the null-hypothesis significance test*. Psychological Bulletin, 5 7.

Shaver, J. P. (1993) *What Statistical Testing is, and what it is not.* Journal of Experimental Education, 61.

Spanos, A. (2008). *Review of S. T. Ziliak and D. N. McCloskey’s The Cult of Statistical Significance.* Erasmus Journal for Philosophy and Economics, 1 (1): 154-164. http://ejpe.org/pdf/1-1-br-2.pdf