

Sherlock Holmes and the Death of the Null Hypothesis

Paul Wilson

Calibre Global, Australia

*Corresponding Author: bodhitree@internode.on.net

Abstract In the eighty years since R.A. Fisher's original work, null hypothesis significance testing has become the ubiquitous research methodology in fields as diverse as biology, agronomy, social science, psychology, epidemiology and most forms of medical research. Throughout this period, the method has been heavily criticised by statisticians who point to a range of problems. The most severe of these is the Bayesian fallacy of wrongly accepting an alternative hypothesis and the problem becomes most obvious in the screening test anomaly whereby a medical screening test can deliver more than 50% of false positives. This paper examines the inverse logical fallacy, that of failing to reject a null hypothesis and thereby accepting an invalid conclusion. A definition of the fundamental concept of experimental completeness leads to a *reductio ad absurdum* proof. The paper finishes with a live example of the dangerous consequences of accepting an invalid null hypothesis out of a null hypothesis significance test.

Keywords Null Hypothesis, Evidence, Methodology, Rigour, Rigor, Bayes

Introduction

In the eighty years or so since the development of statistical testing the technique of null hypothesis significance testing has come under a great deal of criticism and yet, despite its almost universal condemnation, it remains widely taught as the paradigm of choice for research in disciplines as diverse as biology, agronomy, social science, psychology, epidemiology and many areas of medicine. The criticism goes back as far as R.A. Fisher himself, the original creator of the method, and from that time the number of papers condemning it has increased exponentially. The statistical fallacy has been described many times in the literature, but what appears to have been largely overlooked, in all the discussion is that, based on Fisher's original formulation of the null hypothesis and the subsequent modification by Neyman & Pearson, there is a second way of looking at the fallacy and the very definition of the null hypothesis contains the seeds of its own destruction. So inherently damaging is the problem of definition that this

paper could have been written without the use of a single reference and it would still result in a valid conclusion. In deference to the conventions of academic publishing, I furnish it with a modest number of references, although some of those may raise a few eyebrows. The final consequence of this paper is that a proportion (and I am not prepared to guess at what percentage that might be) of null hypothesis tests ever undertaken must be considered invalid and the policy implications in all of the disciplines that continue to use it can be far-reaching.

The Null Hypotheses under Examination

In this paper, I make the distinction between two different definitions, or formulations of the null hypothesis, and these are:

Type 1: It has not been possible to establish a relationship and the experiment is inconclusive

Type 2: A relationship does not exist

In an admittedly brief review of the literature pertaining to the debate over the continued use of null hypothesis significance testing I was unable to find a single usage of the type 1 definition, except in reference to R. A. Fisher's original usage (Fisher, 1935) and the examination of a logical fallacy when we reject the null hypothesis. In every paper I read, there appeared to be an implied and sometimes stated assumption that the null hypothesis was, in the words of Anderson, Burnham and Thompson, (2000) "representing no difference between population parameters of interest." This is clearly definition type 2.

While many papers have expressed doubts about what the null hypothesis means in terms of both statistical relevance and of practical implications, I was not able to find any author who explicitly identified the two competing definitions or examined the inverse of the statistical fallacy, what happens when we fail to reject the null hypothesis, as I have done here. With over 400 papers critical of null hypothesis significance testing to choose from, I found the task of reading all of them too daunting, so it is possible that the distinction has been identified previously. Kreuger certainly hints at an alternative definition with, "the probability of a point hypothesis is indeterminate - - - and thus adds nothing to what is already known." (Kreuger,

2001).

This comment from Kreuger (2001) is consistent with the following similar comment from R. A. Fisher (1935), "For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate in statistical than in other kinds of scientific reasoning." Fisher, in his typical style does not explain the logical fallacy, but others (Cohen, 1994) have done so. Nevertheless he sits amongst several people who have pointed to the logical absurdity contained within it.

One of the more commonplace criticisms of null hypothesis significance testing is not that the null hypothesis is invalid but rather that most researchers would have had some prior knowledge or a suspicion that there is a relationship between the parameters of interest even before the experiment was devised and that, as a consequence, the null hypothesis is already thought to be untrue.

One of the most vocal opponents of null hypothesis significance testing was Jacob Cohen [b. 1923 –d. 1998] (1994) who, for 33 years until he retired in 1993, worked in quantitative psychology at New York University. Cohen's papers are frequently cited in this debate. Cohen explains the Bayesian statistical fallacy very clearly and to explain it here I have modified the scenario. The p-value is the probability of the observed pattern in the data occurring by chance given the null hypothesis. That does not mean that the probability of the null hypothesis being true given the data pattern is the same value. It isn't. If a person buys a lottery ticket, the chance of them winning the lottery might be 1 in 10 million against ($p=0.0000001$). Thus, a novice might argue that if a person wins the lottery, the chance that they actually bought a ticket is 1 in 10 million against. Clearly this is absurd as you cannot win the lottery if you do not buy a ticket.

In his "Earth is round" paper, Cohen (1994) explains very clearly the low number problem and uses the mathematics to illustrate the screening test anomaly whereby a screening test such as the schizophrenia screening test or the prostate specific antigen screening test (PSA test) can give more than 50% of false positives. The same is true of many psychology screening tests as indicated by Professor Dorothy Bishop at Oxford University (Bishop, 2010) and is most likely true of many cancer screening tests as mentioned by Professor Bishop. Cohen's explanation of the statistical fallacy shows the dangers of rejecting the null hypothesis in favour of the alternative hypothesis. Where this paper differs is that I show the dangers of the inverse; of not rejecting the null hypothesis and instead, accepting an invalid null hypothesis.

Aside from Cohen, two others of the most informative papers I found were those of Anderson, Burnham and Thompson from Colorado State University and the US Forest Service (Anderson, Burnham & Thompson, 2000) and also the one by Joachim Krueger of Brown University (Kreuger, 2001). Anderson and Thompson compiled a bibliography of over 400 publications critical of null hypothesis significance testing (Thompson, 2001).

A Thought Experiment – The Case of the Lost Car Keys

This is a thought experiment along similar lines to the famous Schrodinger's Cat experiment. Its purpose is to reveal in an informal way the absurd conclusion that would be reached by using null hypothesis type 2 as opposed to null hypothesis type 1. Imagine that you have lost your car keys. You know you had them in your hand when you walked in through the front door. Your husband remembers seeing them in your hand. No-one has left the house since. You both search the house thoroughly but you cannot find the keys.

Let us now introduce two null hypotheses and an alternative hypothesis (3) as follows:

1. The search of the house was inconclusive (null hypothesis type 1)
2. The search was conclusive and, since the keys were not found, the alternative hypothesis is rejected (null hypothesis type 2)
3. The search found the keys inside the house as expected and the null hypothesis is rejected (alternative hypothesis 3)

If we use the first type of null hypothesis (1), given we have failed to find the keys, we end up with a nil result, we have learnt nothing, although we may perhaps have learnt that we have not searched thoroughly enough so our search was incomplete.

If we use the second definition of the null hypothesis it is the opposite of the alternative hypothesis (3) and it would have to be supported by the unjustified assumption that the search was complete. Therefore if we accept the type 2 null hypothesis the keys no longer exist: they have been vapourised by an alien or teleported into a parallel universe. Clearly such a result is absurd so we are only able to use the null hypothesis with definition type 1.

The Concept of Completeness

As a child, I remember my father occasionally coming out with profound statements and one that had a singular impact on me was the quotation from Conan Doyle (1890) speaking through the persona of Sherlock Holmes, "But I have always told you (Watson) that if you eliminate the possible, whatever remains, no matter how improbable, must be the truth." This statement contains within it a fundamental rule in investigative methodology, the concept of completeness. What Holmes is saying is that in cases where there must be a cause it is necessary to investigate and eliminate every possible cause or factor. When you have investigated and eliminated the most likely, the less likely and the least likely must also be investigated.

Many years later, I came upon the same concept developed most thoroughly by Robert Pirsig in his popular philosophy book, *Zen and the Art of Motorcycle Maintenance* (Pirsig, 1974), which I have since used as a

basic reference text in my lectures on problem solving techniques. Pirsig developed the concept of completeness in the pseudo-context of troubleshooting motorcycles, but the context equally applies to forensic pathology, to criminal detection, to the investigation of air crashes or road crashes and in my own specialisation of optimising industrial plant and machinery.

What Pirsig and Holmes have said is that, in any form of investigation, an essential step is the formulation of a list of possible causes, or in the case of criminal detection a list of possible suspects. In order to thoroughly investigate every possible cause or contributing factor the initial list must include every possible one. In other words, the list must be complete.

In experimental methodology in the health sciences, including sociology and psychology attempts are made to design complete experiments by including a consideration of confounders, the use of blind experiments, the use of controls and even the removal of the placebo effect. Nevertheless as I shall explain by example later, health science investigations are rarely, if ever, complete.

Reductio Ad Absurdum

Following on from the concept of completeness, a simple proof of absurdity can be developed by using the *reductio ad absurdum* proof borrowed from mathematics. The proof follows precisely the philosophy of Karl Popper (1932). Popper says that, in general, scientific proof is impossible except in one circumstance: refutation by logical deduction. If we have a statement in some area of knowledge (such as an *always* or *never* statement), or an hypothesis, we can logically deduce a consequence. If we can show that such a consequence is absurd or impossible then it is sufficient to refute the original statement or hypothesis.

Leading up to an investigation (a test, experiment or study) we can have two possible null hypotheses as already discussed:

1. The investigation was inconclusive and no new knowledge has been gained
2. The investigation showed that the alternative hypothesis is not demonstrated and so a relationship is assumed to be absent

Taking null hypothesis type 1 the logical conclusion is that the investigation was inconclusive because the investigation (experiment or study design) was incomplete. This is a perfectly acceptable conclusion and it validates the use of null hypothesis type 1.

Taking null hypothesis type 2, the logical conclusion is that the investigation was complete. This, in turn, leads to the deduction that we have not only included every possible factor in the investigation but that we can know every possible factor. As Popper (1932) and Hume (1739) before him have indicated, we can not know everything and so completeness is impossible. Thus the deduced conclusion is

absurd and this refutes the validity of using null hypothesis type 2.

In practice we can say that the use of a type 2 null hypothesis is invalid unless we can also demonstrate that the study or experiment is complete.

Potential Sources of Experimental Incompleteness

I offer this list as part of my own toolbox in engineering investigations. It is by no means a complete list of itself but it may serve to give some indication of the reasons why studies can go wrong.

- Poor choice of measuring instrument or one that gives erroneous results
- Lack of resolution in, or failure to calibrate the measuring instrument
- Measuring an inappropriate parameter (very commonplace)
- Invalid starting assumptions
- Mistaken assumptions about the sources of noise or randomness
- Poor experiment design or investigating the wrong question
- Use of unsuitable statistical tools, particularly the choice of probability density function for any of the hypotheses
- Misunderstanding Bayesian statistics and the problem of inverse conditional probability
- Misguided selection of samples or failure to cover a large enough range of sample variability
- Failure to account for confounding factors or failure to understand underlying influences
- Failure to notice anomalies or non-linear relationships in the data
- Failure to recognise the significance of outliers, or the inappropriate removal of outliers

Example of the Dangerous Consequences of a Type 2 Null Hypothesis

It is clear that there is one significant outcome of using a type 2 null hypothesis rather than a type 1. At the conclusion of an experiment that does not support the alternative hypothesis we might conclude that there is nothing left to investigate which would be true if the experiment was complete. In reality the response ought to be to ask the question, "Did we do this correctly? Was there something we missed or was our methodology in some way incorrect?" We might, thus, abandon a potentially fruitful or important line of enquiry instead of revisiting the research and identifying the mistakes we made. Even worse, we might persuade others not to further investigate the same line of enquiry.

This is precisely what has happened in the following example.

The question of whether psychological stress causes cancer has been a source of great debate in medical circles for decades. Recent studies have shown that psychological stress is third in a list of heart attack risk factors behind cholesterol and closely behind smoking (Yusuf, et.al., 2004). Chronic Psychological stress is known to be a cause of the shortening of telomere age (Epel, et.al., 2004). Psychological stress is still regarded as causal for stomach ulcers (Levenstein, 1998) and for irritable bowel syndrome (Mayer, Naliboff, Lin Chang, & Coutinho, 2001) (Bennett, Tennant, Piessea, Badcock, & Kellowa, 1998). Given that there are some common factors such as cortisol and impairment of immune response, (Reiche, Morimoto, Nunes, 2005) it would be surprising if stress was not a factor in the development of some forms of cancer.

There have been several, but not many pre-diagnosis studies of stress and cancer. All the results have been ambiguous and contradictory (Butow, et.al., 2000) and all have tended to use the null hypothesis significance methodology. One such study was a stress / cancer study based on the Nurses' Health Study in the United States. In the published results, Schernhammer et al state (2004), "Findings from this study indicate that job stress is not related to any increase in breast cancer risk". This is a classic case of the use of null hypothesis type 2 and, as a consequence, the acceptance of a completely invalid result.

Not only was this an incomplete study, it is not difficult to find several sources of trouble. As a measuring instrument, the team used "Job strain measured by Karasek and Theorell's job content questionnaire in four categories (low strain, active, passive, and high strain)". The Karasek and Theorell model is based on three factors: job demand, personal control (empowerment) and support. Such a measuring device has just 81 possible combinations of answers of which most replies would be focused on about six. The resolution of such a device is truly awful and the questions do not even address the primary indicators of chronic psychological stress as the Cohen PSS10 stress test does (Cohen, Karmarck, & Mermelstein, 1983). Neither was the stress measuring device calibrated against a different metric such as the measurement of cortisol as others have done (Epel, et.al., 2004). The measurements used a self-reporting questionnaire, a potential source of gross error. If experts cannot agree on a definition of chronic psychological stress, what hope is there for a lay person in a self-reporting questionnaire? (Gottman, 1999) In addition, although the study also examined shiftwork, the connection between working shifts and chronic stress was not considered even though shift-work is known to increase the risk of cancer. The most kindly thing we can say about this study is that it was inconclusive; we learnt nothing.

From a range of studies similar to the Schernhammer one, and, since completeness is exceedingly difficult to achieve in studies of this type, it is reasonable to assume they were just as flawed, Garssen published the results of a meta-study in

2004 (Garssen, 2004) in which he said, "...there is not any psychological factor for which an influence on cancer development has been convincingly demonstrated in a series of studies." While Garssen is careful not to say that a relationship does not exist the whole tone of the paper implies it and that is how it has been interpreted by others. This paper has been quite widely cited in the literature.

At the end of 2006, the investigative panel into a cluster of breast cancer cases at the ABC studios in Toowong in Brisbane reported its findings (ABC, 2006). There was no doubt (1 million to 1 against) that this was no chance event. The panel was exceedingly thorough. They investigated every possible thing they could, except job-related psychological stress. Nothing was ever found and the cause remains a mystery. The argument against investigating stress was simply to cite Garssen's paper. As a result, the panel wasted a unique opportunity to learn something of value about this puzzling question and because it was the only factor not investigated, job stress just became the prime suspect.

Until recently, the Cancer Council Australia adopted a similar line to the ABC panel and to Garssen. On its iheard.com (Cancer Council Australia, pre-1913) website it stated, "There is no evidence that our state of mind could affect our risk of cancer". As of May 2013 and following representations from myself, although I have no idea as to whether the two events were related, the section about state of mind and cancer was removed and a new section about stress and cancer was included which is a brief but accurate representation of the current state of research and knowledge on the topic (Cancer Council Australia, 2014.)

Unfortunately, the "there is no evidence" mindset is all too prominent within medical research sufficient to prejudice the allocation of funding (Wilson, 2013). In the case of investigations of psychological stress as a causal agent of cancer, good scientific investigations have been hampered by access to funding primarily because of the general (incorrect) belief that the lack of any demonstrated correlation indicates that there is no correlation, a belief that this paper seeks to invalidate. Absence of evidence is not evidence of absence. The "there is no evidence" mindset derives directly from the acceptance of type 2 null hypotheses instead of the logically correct approach of "we have never been able to design a complete investigation so we just do not know".

Thus it can be seen how the misuse of the null hypothesis leads to a profound and dangerous negative impact on research funding into a potentially deadly causal link.

Conclusion

This paper now begs the question: can we believe any of the results derived from the use of null hypothesis significance testing including all screening tests, most studies in psychology and social science, field biology and agronomy, much of our epidemiology and a substantial part of the medical evidence base?

Acknowledgement

I wish to acknowledge the support of my wife Julie Wilson-Hirst, on the faculty of the Masters of Mental Health program at the University of Queensland. She first drew my attention to the inappropriate use of statistics to justify the use of cognitive behavior therapy for treating depression and to replace more effective methods. That led to the writing of this paper.

Author's Notes

Paul is the technology manager of the Industrial Technology Group within Calibre Operations, a Perth based engineering company. Paul holds a Bachelor's degree in electrical engineering, a full thesis Master's degree in system simulation and a doctorate in data analysis and artificial intelligence. He began his engineering career as a student apprentice with the United Kingdom Atomic Energy Authority in 1962.

Paul's professional specialisations include engineering management, controllability studies of process designs, loop tuning, plant optimisation, dynamic simulation and system modelling. His current professional interests lie in techniques for optimising process plant and in forensic investigative trouble-shooting, especially into the causes of poor performance.

Following 18 years as an academic at QUT, Paul maintains an academic interest in the philosophy of science and research methodology which he also applies to his forensic engineering work.

REFERENCES

- Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* 64(4):912-923. 2000.
- Australian Broadcasting Corporation. (2006). *Breast Cancer at the ABC Toowoong Queensland: Third Progress Report from the Independent Review and Scientific Investigation Panel – 21st December 2006*
- Bennett, E.J., Tennant, C.C., Piessea, C., Badcock, C-A., Kellowa, J.E. (1998). Level of chronic life stress predicts clinical outcome in irritable bowel syndrome. *Gut* 1998;43:256-261
- Bishop, D.V.M. (2010). The difference between $p < .05$ and a screening test. Retrieved from <http://deevybee.blogspot.com.au/2010/07/difference-between-p-05-and-screening.html>
- Butow, PN, Hiller, JE, Price, MA, Thackway, SV, Krickler, A, Tennant, CC. (2000). Epidemiological Evidence for a Relationship between Life Events, Coping Style and Personality Factors in the Development of Breast Cancer. *J Psychosom Res* 2000; 49: 169-181.

Cancer Council Australia (pre May 2013). Retrieved from

<http://www.iheard.com.au>

Cancer Council Australia (2014). Retrieved from <http://www.iheard.com.au>

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

Cohen, S., Karmarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behaviour*, 24, 385-396.

Doyle, Sir Arthur Conan (1890). *The sign of the four (The sign of four)*. First published 1890.

Epel, E.S., Blackburn, E.H., Jue Lin, Dhabhar, F.S., Adler, N.E., Morrow, J.D., Cawthon, R.M. (2004). Accelerated telomere shortening in response to life stress. *PNAS* Dec 7, 2004, Vol 101, No 49, 17312-17315.

Fisher, R. A. (1935). *Statistical tests*. Nature 136, 474.

Garsen, B. (2004). Psychological Factors and Cancer Development: Evidence after 30 years of research. *Clin. Psychol Rev* 2004; 24; 315-338.

Gottman, J.M. (1999). *The Seven Principles for Making Marriage Work*. Three Rivers Press, New York.

Hume, D. (1739). *A treatise of human nature*. Glasgow, Scotland: William Collins, 1978 ed.

Kreuger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*. January 2001, pp 16-26.

Levenstein, S. (1998). Stress and peptic ulcer: life beyond helicobacter. *BMJ* 1998; 316:538

Mayer, E.A., Naliboff, B.D., Lin Chang, Coutinho, S.V. (2001). Stress and irritable bowel syndrome. *American Journal of Physiology - Gastrointestinal and Liver Physiology*: 1 April 2001 Vol. 280 no. G519-G524

Pirsig, R.M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values*. Bantam Books.

Popper, K.R. (1932). *The Two fundamental problems of the theory of knowledge*. Published in paperback, Routledge, 2008.

Reiche, EM, Morimoto, HK, Nunes, SM. (2005). Stress and Depression-induced Immune Dysfunction: Implications for the Development and Progression of Cancer. *Int Rev Psychiatry* 2005, (6):515-27.

Schernhammer, ES, Hankinson, SE, Rosner, B, Kroenke, CH, Willett, WC, Colditz, GA, Kawachi, I. (2004). Job Stress and Breast Cancer Risk. *Am J Epidemiol*, 2004 160(11):1079-1086

Thompson, W.L. (2001). 402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies. Retrieved from <http://warnercnr.colostate.edu/~anderson/thompson1.html>

Wilson, P.A. (2013). Personal communications with the Cancer Council Australia.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., Liu Lisheng, on behalf of the INTERHEART Study Investigators. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004; 364: 937-52