THE DEVELOPMENT OF THE IDEA OF THE NULL HYPOTHESIS IN RESEARCH AND TEACHING

John Truran, University of Adelaide, Australia

This paper examines the development of Fisher's concept of the null hypothesis and ways in which it has been misunderstood by statistical workers and teachers. It provides examples of how the idea has been incompletely presented in textbooks, and argues that the omissions have led to an emphasis on algorithmic competence at the expense of discursive analysis.

This paper is a personal odyssey as well as an historical analysis, so requires a brief *raison d'être* . I have been teaching statistics for some thirty years, but without formal training in the subject. Most of my knowledge has come from books, not always particularly clear ones. When I arrived as a fresher at St Mark's College, University of Adelaide, it happened that Professor Sir Ronald Fisher arrived at the same time to work in his retirement. We never talked, but I heard him lecture and he made statistics sound so simple. Only at his funeral in 1962 did I find out some of his achievements. It was much later still that I linked the rigid formalities of classroom teaching of the "null hypothesis" to Fisher's succinct exposition of statistical inference. So in this paper, I trace the transmission of this concept to new students and users of advanced stochastic thinking.

This restricted topic provides a good basis for seeing how statistical ideas are passed on to future practitioners. It also allows us to use the approach of Hefendehl-Hebeker (1991) which uses the historical development of a subject as a basis for clarifying pedagogical difficulties. I do not discuss here the relative merits of the different formal theories of inference—I am a teacher and historian, not a statistician—but I try to see how Fisher's has been variously reported in many texts, including some consideration of how it has been set into the broader picture. I use this historical cameo to question whether the standard checks and balances of academia ensure that students receive accurate summaries of standard ideas. I use history to enhance pedagogy.

Pedagogy needs to be enhanced. Shaughnessy (1983) has summarised research into inferential strategies. Many use processes quite different from those employed in other forms of human decision-making, which are often based on a trial-and-error approach. He adds:

> Very few statisticians or teachers of probability and statistics are even aware of the literature of judgment and decision making, or of the implications of this research for teaching. It is not enough to look at the student from only one side of the coin, and decide, "They are just statistically ignorant. We only have to teach them things." Our students also have intuitive heuristics

> and schemas for dealing with inferences, and many of these are psychologically based. (p. 343)

Not only do some current pedagogies disregard such intuitions, but they may also be quite restrictive. Williams (1997, p. 591) has found that some tertiary students see drawing a statistical conclusion as more important than the meaning of that conclusion within a given context and has claimed that “typical textbook questions limit students’ statistical experience”. Vallecillos (1996, p. 248) has found that tertiary students find the critical distinction between sample and population difficult. Among practising psychologists, John (1992) has claimed that many practising psychologists not only use statistics merely in an algorithmic way, but frequently have a poor grasp of what may reasonably be inferred from a set of data. He attributes some of these limitations to the way in which statistics are used to preserve power and confidence within the psychological community and to reduce cognitive discomfort. Falk and Greenbaum (1995) present similar arguments.

How were Fisher’s ideas received by researchers and teachers? Very early he decided that practitioners were more likely to use them than teachers, ascribing this to the theoretical nature of traditional statistics teaching (Bennett, 1990, p. 330; Fisher to Maclean, 3 March 1930). For Fisher understanding was more likely to be gained from “the actual body of data to be examined” than from mathematical background (Bennett, 1990, p. 332; Fisher to Thom, 8 October 1941). But both groups developed significant misconceptions.

The term “null hypothesis” was coined (Simpson and Weiner, 1989) in *The Design of Experiments* as part of Fisher’s (1935) well-known discussion of testing the ability to taste tea. He wrote:

> The two cases of results which are distinguished by our test of significance are … those which show a significant discrepancy from a certain hypothesis; namely, in this case, the hypothesis that the judgements given are in no way influenced by the order in which the ingredients have been added; and … results which show no significant discrepancy from this hypothesis. This hypothesis … is again characteristic of all experimentation. … 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.

Confusion arose early. The issue of whether the null hypothesis may be accepted is well known. Fisher was quite clear on this point, as shown above, and also said:

> 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 itself in statistical than in other kinds of scientific reasoning. (Fisher, 1935)

Much later he said (Fisher, 1955, pp. 73–75): “The fashion of speaking of a null hypothesis as ‘accepted when false’ … shows real ignorance of the research workers’ attitude, by suggesting that in such a case he has come to an irreversible decision.… At most a null hypothesis may be “confirmed or strengthened.… What we look forward to in science is further data, probably of a somewhat different kind, which may … form an enlarged basis for induction.” The key words here are “in science”, which for Fisher is quite different from industry, where acceptance procedures must be used because irreversible decisions have to be made, and where the term “error” is quite appropriate.

But there was also confusion about whether the null hypothesis:

(a) was an hypothesis that the difference between two groups was null; or

(b) came from the term ‘nullify’, because a null hypothesis is any hypothesis set up for the purpose of being nullified. (Bennett, 1990, p. 322)

Fisher concurred with the latter.

> I chose the term ‘null hypothesis’ without particular regard for its etymological justification by analogy with a usage, formerly and perhaps still current among physicists, of speaking of a null experiment, or a null method of measurement, to refer to a case in which a proposed value is inserted experimentally in the apparatus and the value is corrected, adjusted, and finally verified, when the correct value has been found; because the set-up is such, as in the Wheatstone Bridge, that a very sensitive galvanometer shows no deflection when exactly the right value has been inserted. … One might put it by saying that if the null hypothesis is exactly true no amount of experimentation will easily give a significant discrepancy, or, that the discrepancy is null apart from the errors of random sampling (Bennett, 1990, pp. 321–322; Fisher to McGuigan, 8 April 1958).

These ideas have been poorly preserved. In the Oxford English Dictionary Simpson and Weiner (1989) see the null hypothesis as “a hypothesis that is the subject of a significance test, esp. the hypothesis that there is no difference between the specified populations (any apparent difference being due to sampling or experimental error).” Frick (1995), in a paper specifically questioning the claim that the null hypothesis cannot be accepted, states that the term “has come to refer to the ‘null’ effect of the experimental manipulation”. Even in a specialist dictionary (Kendall and Buckland, 1982, p. 139), we find:

> Null Hypothesis. In general this term relates to a particular hypothesis under test, as distinct from the alternative hypotheses which are under consideration. It is therefore the hypothesis which determines the probability

> of the Type I error. In some contexts, however, the term is restricted to a hypothesis under test of ‘no difference’.

None of these examples discusses the importance of context, none emphasised the distinction between proof and confirmation, and, more importantly, none picks up Fisher’s point that “null” refers to no difference between the *sample data and the result predicted by the null hypothesis,* and not necessarily to no difference between populations.

The situation is no better in textbooks. There is only space here to cite individual examples. The ones chosen reflect my personal odyssey and have an Anglo-Australian bias. However, most have been popular among significant groups of teachers and so are sufficient evidence for demonstrating the *existence* of misconceptions among many textbook authors. This existence is corroborated by other workers. Frick (1995, p. 132) found four different approaches to the null hypothesis in 15 textbooks. At a deeper level, Falk and Greenbaum (1995, pp. 83–84) provide several examples of inconsistent text-book presentation. Ortiz de Haro (1996) found that many Spanish textbooks provide only enough practical examples of elementary stochastic ideas to define a concept, and then give mainly algorithmic exercises on the calculus of probabilities. In the absence of a rich approach it is easy for misconceptions to remain.

Consider first two popular school texts from the 1950s and 1960s. In one: “[t]he null hypothesis is the *assumption* which is made when applying a significance test” (Loveday, 1961, p. 134). In the other: “it is necessary to formulate some hypothesis about the behaviour of the coin and then carry out an experiment to discover whether in fact the hypothesis can account for the results of the experiment. This hypothesis is to be tested on the assumption that it is true and is called the *null hypothesis”* (Brookes and Dick, 1951/1966, p. 145). As part of a χ^2 example Loveday (1961, p. 44) states: “Are the differences between the observed and expected frequencies great enough to force us to reject the null hypothesis as false, or are they small enough to allow us to accept it as true?” This correctly emphasises the importance of small discrepancies but it remains deterministic in its approach, especially in its use of words like “force” and “accept”. Brookes and Dick (1951/1966, p. 148) are more precise, but do not emphasise the importance of preferring “do not reject” to “accept”. Neither text satisfactorily explains why a null hypothesis is necessary, where it is appropriate, or why it is called what it is.

A similar under-emphasis is found in a generally thoughtful text from the time of the “new mathematics” when some authors were trying to clarify ideas in terms of deep underlying structures. Durran (1970, p 80) merely describes the null hypothesis as a “sort of Aunt Sally” which enables us to “compare the real results with what *would* have happened if the connexions were absent”.

Examination of some modern texts makes it clear that there not been no consistent trend towards more precise presentation. In Jaeger (1990, pp. 164–167) we read

> You've begun with an initial statement of belief.…A statement of belief such as yours is known in statistical language as a *Null Hypothesis.* … In hypothesis testing, belief in the value of the null hypothesis continues, unless evidence collected from a sample is sufficient to make continued belief unreasonable. … [D]epending on the value of the sample statistic, decide that the null hypothesis is true, or reject the null hypothesis in favor of the alternative hypothesis. … If the alternative hypothesis is true, but the decision maker decides to stick with the null hypothesis, a *Type II error* results.

Initially, Jaeger emphasises the provisional nature of the null hypothesis, then moves to a suggestion that it may be true. This rapid transition from researcher to decision maker is likely to pass unnoticed by students. Selvenathan et al (1994) is designed for economics students. It seeks to teach students "how to recognise which statistical technique to use" (p. v) It explains (pp. 263–271) that a null hypothesis cannot be accepted, but also state that in some circumstances "enough statistical evidence [might exist] to allow us to conclude that the alternative hypothesis is true" (p. 271). Its algorithmic approach, focussed on a precise decision rule, can lead, in my experience, to students' reporting "we must retain H_0 until we gather further evidence and can draw a conclusion". This bizarre response, probably based on material provided in lectures, seems to assume that an alternative hypothesis will eventually be accepted: it is only a matter of having enough evidence.

Two more discursive modern texts are more comprehensive. In Quadling (1987, pp. 107–109) we find a juridical analogy which emphasises that defined evidence is used to evaluate an hypothesis which "remains tenable until it has been shown to be unacceptable" and draws an analogy between retaining a null hypothesis and bringing down a Scottish verdict of "not proven". Moore and McCabe (1989, pp. 461–464) emphasise the provisional nature of any stochastic judgement and the distinction between science and industry, but present a rather confused view of the origin of the word "null".

> Confidence intervals … are appropriate when our goal is to estimate a population parameter. The second common type of inference is directed at a quite different goal: to assess the evidence provided by the data in favor of some statement. … The statement being tested in a test of significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of "no effect" or "no difference".

Clearly some of these texts make good accurate points. What I have not found is any text which covers all of the points clearly and comprehensively.

CONCLUSION

This paper has discussed the understanding and transmission of Fisher's idea of the null hypothesis. It has not discussed alternative approaches, nor recent criticisms by Falk and Greenbaum (1995) and Frick (1995) of the logic of the argument, as opposed to its philosophy. The striking features of the cases examined here have been how the meaning of "null" has become changed, and how neither teachers nor researchers have adequately addressed the complexity or context of Fisher's ideas, preferring to concentrate on the deterministic aspects of what is really a very non-deterministic form of thinking.

Hawkins (1997) claims that most school statistics teachers have little formal background in statistical education and few tertiary teachers in educational training. Now that statistics is commonly seen as part of a general R–12 education, I would suggest that it is now the case that most school teachers have no formal training in statistics. But this paper has demonstrated the existence of some other, less well acknowledged, issues.

Hotelling (1951) claimed that Fisher's 1930s ideas had become well integrated into textbooks by 1950. This paper has argued that they are still not well integrated, and that, despite a more thoughtful approach in a few recent texts, some seem to have been lost. We need to investigate how it is that standard academic procedure have not ensured a more precise transmission of Fisher's ideas. I am quite sure that my teaching would have been much better had they not been lost, though it remains to be verified whether this would have been true for most teachers. But this historical cameo has been able to demonstrate the existence in texts of sufficient confusion of concepts and approaches to give such an assertion *prima facie* validity. It has shown how an historical analysis of ideas and pedagogy can provide helpful insights into current pedagogical practice.

REFERENCES

Bennett, J.H. (Ed.) (1990). *Statistical Inference and Analysis—selected correspondence of R.A. Fisher* Oxford: Clarendon Press.

Brookes, B.C. and Dick, W.F.L. (1951/1966). *Introduction to Statistical Method* London: Heineman.

Durran, J.H. (1970). *Statistics and Probability* Cambridge: Cambridge University Press

Falk, R. and Greenbaum, C. W. (1995). Significance Tests Die Hard. The Amazing Persistence of a Probabilistic Misconception. *Theory and Psychology*, 5, 75–98.

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

—– (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society, B17*, 69–78.

Frick, R. W. (1995). Accepting the Null Hypothesis. *Memory and Cognition, 23(1),* 132–138.

Hawkins, A. (1997). Discussion: *Forward* to Basics! A Personal View of Developments in Statistical Education. *International Statistical Review, 65(3),* 280–287.

Hefendehl-Hebeker, L. (1991). Negative Numbers: Obstacles in Their Evolution from Intuitive to Intellectual Constructs. *For the Learning of Mathematics 11(1),* 26–32.

Hotelling, H. (1951). The Impact of R.A. Fisher on Statistics. *American Statistical Association Journal, 46 (253),* 35–46.

Jaeger, R. M. (1990). *Statistics: A Spectator Sport* Newbury Park CA:

John, I. D. (1992). Statistics as Rhetoric in Psychology. *Australian Psychologist 27(3),* 144–149.

Kendall, M. G.; Buckland, W. R. (1982). *A Dictionary of Statistical Terms. 4th Edition. Revised and Enlarged.* London: Longman.

Loveday, R. (1961) *A Second Course in Statistics* Cambridge: Cambridge University Press.

Moore, D. S. and McCabe (1989). *Introduction to the Practice of Statistics* New York: W.H. Freeman and Company.

Ortiz de Haro, J. J. (1996). *Significados de los Conceptos Probabilísticos Elementales en los Textos de Bachillerato* Granada, Spain: University of Granada.

Quadling, D. (1987). *Statistics and Probability* Cambridge: Cambridge University Press.

Selvenathan, A., Selvenathan, S., Keller, G., Warrack, B., and Bartel, H. (1994). *Australian Business Statistics* Melbourne, Australia: Nelson

Shaughnessy, J. M. (1983). The Psychology of Inference and the Teaching of Probability and Statistics: Two Sides of the Same Coin? In R.W. Scholz (Ed.) *Decision Making under Uncertainty* (pp 325-350), North-Holland: Elsevier

Simpson, J.A. and Weiner, E.S.C. (1989). *The Oxford English Dictionary 2nd Edition* Oxford: Oxford University Press.

Vallecillos, J. A. (1996). *Inferencia Estadística y Enseñanza: Un análisis didáctico del contraste de hipótesis estadísticas* Granada, Spain: Comares

Williams, A. (1997). Students Understanding of Hypothesis Testing: the Case of the Significance Concept. In F. and K Carr (Eds.), *People in Mathematics Education.* Proceedings of the Twentieth Annual Conference of the Mathematics Education Research Group of Australasia Incorporated (pp522-529) held at Rotorua, New Zealand, 7–11 July 1997.