Grappling with Fisher's Legacy in Social Science Hypothesis Testing: Some Comments on Denis (S'attaquer Au Legs de Fisher's Dans L'essai d'Hypothése de la Social Science: Une Partie Présente Ses Observations Sur Denis)

Jeff Gill, University of California, Davis

The Current State of the Null Hypothesis Significance Test

The null hypothesis significance test (NHST) should not even exist, much less thrive as the dominant method for presenting statistical evidence in the social sciences. It is intellectually bankrupt and deeply flawed on logical and practical grounds. More than a few authors have convincing demonstrated this (Bakan 1960, Barnett 1973, Berger, Boukai, and Wang 1997, Berger and Selke 1987, Bernardo 1984, Carver 1978, Cohen 1994, 1992, 1977, 1962, Falk and Greenbaum 1995, Gelman, Carlin, Stern, and Rubin 2004, Gigerenzer 1993, 1987, Gigerenzer and Murray 1987, Gill 1999, Gill and Meier 2000, Greenwald 1975, Howson and Urbach 1993, Hunter 1997, Hunter and Schmidt 1990, Jeffreys 1961, Lindsay 1995, Macdonald 1997, Meehl 1990, 1978, Oakes 1986, Pollard and Richardson 1987, Rosnow and Rosenthal 1989, Rozeboom 1960, Schmidt 1996, Schmidt and Hunter 1977, Sedlmeier and Gigerenzer 1989).

And now Denis (2005) comes along and does a great service by highlighting the under-appreciated fact that Fisher is not responsible for this state of affairs. This is an important essay both from the historical perspective and also in that it further reinforces the problems with continued use of the NHST. The recalcitrance of the median-sophistication social scientist here remains puzzling given the ever-increasing bulk of criticism. Hence Denis' fine essay is most welcome in that it provides additional nuance and clarity about the distinction between Fisher's *test of significance* and the NHST.

Mature scientific disciplines are highly sensitive to changes in fundamental baseline knowledge, including the process of discovery. Recently Steven Hawking reversed course and declared, after decades of denial, that information can actually escape from a black hole. He had previously claimed that the combined principles of quantum uncertainty and Einsteinian relativity assured leakage of particles from black holes (so-called *Hawking Radiation*), but this leads to a paradox because the radiation had to be random meaning that the structure of anything that was absorbed

by a black hole is lost forever violating a central tenet of quantum theory that it must be possible to see causal events back in time. The world of cosmology and physics reacted immediately and earnestly to Hawking's new work and the subsequent consequences by starting to rethink current models and assumptions. Why is it then that the social sciences fail to pay attention important and obvious methodological problems underlying nearly all empirical work in similar fashion? In both physics and social science methodology it is an issue of key baseline knowledge. The answer unfortunately lies our scientific maturity.

I Come to Praise Fisher, Not to Bury Him

Fisher is often called the father of modern statistics and more than one observer has referred to him as the most influential figure in twentieth century statistics. Stigler (1976), who is arguably the most accomplished statistical historian on the planet, notes that it was Fisher who first introduced the term "parameter" into our modern statistical lexicon in a 1922 paper and thereafter. This is perhaps merely an overt sign of his tremendous but often overlooked influence on such details. I believe that none of these admiring statements are really exaggerations, despite my personal perspective as a strong adherent of Bayesian philosophy and methods. To some degree Fisher is misunderstood simply because his name is attached to so many important concepts and these concepts develop and change over time. We are fortunate, however, that Fisher's works are all readily accessible, at least in the physical sense.

Denis points out that the key to understanding the problem with the current paradigm is to see how it differs from Fisher's original process. Fisher (1925a, 1934, 1955) posits merely a single (null) hypothesis, H_0, and a known distribution for the test statistic θ under this assumption. If this test statistic is found to lie far away from its conditional expected value, $E(\theta|H_0)$, then H_0 is declared to be implausible (i.e. unlikely to have occurred by chance). The level of significance as a guage of this implausibility is then produced by measuring the density under H_0 starting at θ going away from the expected value.

If the Fisherian level of significance is small then the null hypothesis is rejected, otherwise there is insufficient evidence for a conclusion, *even in experimental settings with randomization.* Denis finds a superbly typical example of how many social scientists violate this procedure by

declaring confirmation for the null hypothesis with large significance levels (Snow and Compton 1996). Why is this wrong as well as non-Fisherian? By failing to find evidence to reject the null with a single model and a single dataset, we have not ruled out an infinite number of competing theories about the state of some social system. Thus, as Denis notes, Fisher would not have agreed with the "modern" practice of null hypothesis significance testing. Yet many people take Fisher's perspective and subsequently add the belief that the smaller the significance level, the greater the probability that the null hypothesis is false (Carver 1978: Cohen 1994, Meehl 1990). Fisher knew that this is simply untrue, *since interpreting the significance level as a probability statement is conditional on assuming that the null is true in the first place.* He rather clearly states: "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 itself in statistical than in other kinds of scientific reasoning." (1935a).

Although I have often personally argued that the Bayesian perspective is almost uniformly superior for social science empirical work (Gill 2004, 2002, 2001, 2000, 1999), there naturally is much to admire in Fisher's contributions (perhaps not his vitriolic contempt for twentieth century Bayesian inference though), and we sometimes lose sight of this. In fact, Savage (1976, p.445), a devout Bayesian, observes that "...largely because of the vision of statistics to which his activities gave rise, was statistical training inaugurated in a few universities." Not a minor feat for someone who was never housed in a mathematics or statistics department and supervised only one dissertation in statistics (C. R. Rao's). Fisher's difficult personal qualities, nonetheless, are legendary and usually end up being balanced against the importance of his work: "A difficult genius though, one in whom brilliance usually outdistances clarity." (Efron 1976).

While many breakthroughs are anticipated by earlier, less focused or articulate, works (c.f. "Stigler's Law of Eponymy" [Stigler 1999]), a number of huge contributions can be undisputedly attributed to Fisher. These include: design of experiments, k-statistics as estimators of cumulants, sufficiency, likelihood estimation, the correct view of consistency, Fisher information (differential), ancillarity, analysis of variance, and exponential family disstributions. And this list does not include some more mechanical achievements: the derivation of the exact distribution of the sample correlation coefficient, the correction of the degrees of freedom from cross-tabulation, the discriminant function, and probability tables that endured for decades.

Denis also points out, along the way, several of Fishers admonitions that have become standard practice in the social sciences and elsewhere. These include: full enumeration of the possible outcomes of an experiment before the experiment is conducted, randomization of experimental subjects/cases, "usual and convenient" significance level thresholds, the importance of researcher expertise and intuition, sensitivity of the experiment, and the pre-eminence of theory. Denis also nicely highlights features of Fisherian statistics that have *not* positively contributed to modern practice. Some of these are: the hypothetical infinite super-population idea, denigration of statistical power, and unclear advice on strict adherence to significance level thresholds.

Fisherian Obfuscation

While much can be learned from carefully reading Fisher's work and quotations from Fisher's work, there is a substantial danger in fully trusting excerpted passages. Fisher is famously vague and self-contradicting in many writings and it is therefore possible to portray him in many lights. There are also instances where he abandons and denigrates his own ideas making it difficult to firmly credit discovery. These latter were not always minor points either; he apparently invented confidence intervals only to dismiss them later in an editorial (Owen 1962). So I take some modest discomfort from the frequent and lengthy quotes of Fisher that Denis provides (of course the reader will see that I too am guilty of this practice right here!). To be fair, Denis also notes the ambiguity of some of Fisher's work in Section 1.5 when discussing Fisher's significance level thresholds and in the wonderful table provided.

For example, Fisher was strongly against decision-theoretic work at times: "It is important that the scientific worker introduces no cost functions for faulty decisions,..." (1956, p.102-3). Yet at other times he specifically ties together costs (in the literal sense) and design, noting at one point that statistical studies are demonstrably useful when conforming to the maxim that they "conduct experimental and observational inquiries so as to maximize the information obtained for a given expenditure." (1951 p.54).

Fisher deplored Bayesian inference in the form of inverse probability with uniform priors, even while he professed admiration for Bayes' original paper. Interestingly, he (1935b) developed *fiducial inference*, as is an attempt to apply inverse probability without the uniform prior assumption. De-

spite his many attempts to defend fiducial inference, it failed to find adherents and Efron (1998, p. 105) calls it "Fisher's biggest blunder." Lindley (1958) eventually proved that fiducial inference is consistent (note the use of one of Fisher's criteria) *only* when it is made equivalent to Bayesian inference with a uniform prior, and Jeffreys (1961) demonstrated that in some cases fiducial inference produces the Bayesian result obtained assuming improper priors.

More pointedly, Denis (Sections 1.6 and 2.3.3) notes that Fisher advocates publication of both significant and non-significant results as a means of fully describing evidence for and against a particular theory. Yet Fisher also states that the researcher "should only claim that phenomenon is experimentally demonstrated when he knows how to design an experiment so that it will rarely fail to give a significant result." (1929, p.191). This seems to imply quite the opposite: scientific findings are only supportable when repeatable positive results are possible. Thus he seeks to reduce or eliminate discussion of non-significant findings even if they are occasionally relevant. I should also note that if one accepts this latter quotation as Fisher's true intent, then it invalidates Denis' claim that strict adherence to Fishers test of significance (as opposed to NHST) would likely prevent the file drawer problem (Section 1.6).

Denis also addresses Fisher's views on Randomization. Yet we can find contradictions in this area of Fisher's work too. In Section 1.2, Denis quotes the following sentence from the eighth edition of *Design of Experiments* (it appeared earlier too):

> Apart, therefore, from the avoidable error of the experimenter himself introducing with his test treatments, or subsequently, other differences in treatment, the effects of which the experimenter is not intended to study, it may be said that the simple precaution of randomization will suffice to *guarantee* the validity of the test of significance, by which the result of the experiment is to be judged. (Fisher 1966, p.21, emphasis added).

Conversely in a widely read essay written after *Design of Experiments*, Fisher states:

> ... and whereas planned randomization (1935-1953) is widely recognized as essential in the selection and allocation of experimental material, it has *no useful part to play* in the formation of opinion, and consequently in the tests of significance designed to aid the formation of opinions in the Natural Sciences. (Fisher 1956, emphasis added).

What he means by "(1935–1953)" is the span of editions of *Design of Experiments* up to that point and thus refers *his* central contribution of randomization therein. Now we get a fundamental contradiction where Fisher first asserts a key relationship between randomization and the quality of estimation then later denies such a linkage.

Where does this general level of ambiguity leave us then? I belive that have we (close students/readers of Fisher) now become the equivalent of Talmudic scholars in that the reading and interpretation of Fisher's text is an exercise in perspective and reflection. So Fisher's complexity expresses itself in abundant ways giving statistical historians multiple interpretations and ramifications. This is not necessarily bad because it has the potential to see Fisher's contributions in a more accurate light and it certainly makes modern scholars less likely to attribute direct Fisherian approval of hybrids like the NHST. So in this new light I congratulate Rabbi Denis on a wonderful contribution to our understanding of this complex body of work.

References

Bakan, David. (1960). "The Test of Significance in Psychological Research." *Psychological Bulletin* 66, 423-437.

Barnett, Vic. (1973). *Comparative Statistical Inference.* New York: John Wiley & Sons.

Berger, James O., Boukai, B. & Wang, Y. (1997). "Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12, 133-160.

Berger, James O. & Sellke, Thomas. (1987). "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." *Journal of the American Statistical Society* 82, 112-122.

Bernardo, José M. (1984). "Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis." *Journal of the American Statistical Society* 79, 510-515.

Carver, Ronald P. (1978). "The Case Against Statistical Significance Testing." *Harvard Education Review.* 48, 378-399.

Cohen, Jacob. (1962). "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65, 145-153.

Cohen, Jacob. (1977). *Statistical Power Analysis for the Behavioral Sciences.* Second Edition.

New York: Academic Press.

Cohen, Jacob. (1992). "A Power Primer." *Psychological Bulletin* 112, 115-159.

Cohen, Jacob. (1994). "The Earth is Round ($p < .05$)." *American Psychologist* December, 12, 997-1003.

Denis, Daniel J. (2005) "The Modern Hypothesis Testing Hybrid: R. A. Fisher's Fading INfluence *Journal de la Société Française de Statistique*. Forthcoming.

Efron, Bradley. (1976). Comment on Savage's "On Reading R. A. Fisher." *Annals of Statistics* 4, 483-484.

Efron, Bradley. (1988). R. A. Fisher in the 21st Century. *Statistical Science* 13, 95-122.

Falk, R. & Greenbaum, C. W. (1995). "Significance Tests Die Hard." *Theory and Psychology* 5, 396-400.

Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700-725.

Fisher, R. A. (1929). "The Statistical Method in Psychical Research." *Proceedings of the Society for Psychical Research* 39, 185-189.

Fisher, R. A. (1934). *The Design of Experiments.* First Edition. Edinburgh: Oliver and Boyd.

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

Fisher, R. A. (1935b). The Fiducial Argument in Statistical Inference. *Annals of Eugenics* 6, 391-398.

Fisher, R. A. (1951). "Statistics." In *Scientific Thought in the Twentieth Century*, (ed. A.E. Heath), p.31-55. London: Watts.

Fisher, Sir Ronald A. (1955). "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society* B, 17, 69-78.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Gelman, Andrew, Carlin, John B., Stern, Hal S. & Rubin, Donald B. (2004). *Bayesian Data Analysis.* Second Edition. New York: Chapman & Hall.

Gigerenzer, Gerd. (1987). "Probabilistic Thinking and the Fight Against Subjectivity." In Krüger, Lorenz, Gerd Gigerenzer, and Mary Morgan, eds. *The Probabilistic Revolution.* Volume 2.

Cambridge, MA: MIT.

Gigerenzer, Gerd. (1993). "The Superego, the Ego, and the Id in Statistical Reasoning." In G. Keren, and C. Lewis, eds. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, Gerd. & Murray, D. J. (1987). *Cognition as Intuitive Statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gill, Jeff. (1999). "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52, 647-674.

Gill, Jeff. (2001). "Whose Variance is it Anyway? Interpreting Empirical Models with State-Level Data." *State Politics and Policy Quarterly* 1, 313-338.

Gill, Jeff. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach.* New York: Chapman and Hall/CRC.

Gill, Jeff. (2004) "Introducing the Special Issue of Political Analysis on Bayesian Methods: (An Introduction and Overview of Bayesian Methods)." *Political Analysis* 12, Forthcoming.

Gill, Jeff & Meier, Ken. (2000). "Public Administration Research and Practice: A Methodological Manifesto." *Journal of Public Administration Research and Theory* 10, 157-200.

Greenwald, Anthony G. (1975). "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82, 1-20.

Howson, Colin & Urbach, Peter. (1993). *Scientific Reasoning: The Bayesian Approach.* Second Edition. Chicago: Open Court.

Hunter, John E. (1997). "Needed: A Ban on the Significance Test." *Psychological Science* January, Special Section 8, 3-7.

Hunter, John E. & Schmidt, Frank L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Beverly Hills: Sage.

Jeffreys, Harold. (1961). *The Theory of Probability.* Oxford: Clarendon Press.

Lindsay, R. M. (1995). "Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy." *Accounting, Organizations and Society* 20, 35-53.

Macdonald, Ranald R. (1997). "On Statistical Testing in Psychology." *British Journal of Psychol-*

ogy 88, No. 2 (May), 333-349.

Meehl, Paul E. (1978). "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Counseling and Clinical Psychology* 46, 806-834.

Meehl, Paul E. (1990). "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." *Psychological Reports* 66, 195-244.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* New York: John Wiley & Sons.

Owen, A. R. G. (1962). "An Appreciation of the Life and Work of Sir Ronald Aylmer Fisher :F.R.S., F.S.S. Sc.D." *The Statistician* 12, 313-319.

Pollard, P. & Richardson, J. T. E. (1987). "On the Probability of Making Type One Errors." *Psychological Bulletin* 102, (July) 159-163.

Rosnow, Ralph L. & Rosenthal, Robert. (1989). "Statistical Procedures and the Justification of Knowledge in Psychological Science." *American Psychologist* 44, 1276-1284.

Rozeboom, William W. (1960). "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin.* 57, 416-428.

Leonard J. Savage. (1976). "On Reading R. A. Fisher." *Annals of Statistics* 4, 441-483.

Schmidt, Frank L. (1996). "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods* 1, 115-129.

Schmidt, Frank L. & Hunter, John E. (1977). "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62, 529-540.

Sedlmeier, Peter & Gigerenzer, Gerd. (1989). "Do Studies of Statistical Power Have an Effect on the Power of Studies." *Psychological Bulletin* 105 (March), 309-316.

Snow, T. S. & Compton, W. C. (1996). "Marital Satisfaction and Communication in Fundamentalist Protestant Marriages." *Psychological Reports* 78, 979-985.

Stigler, Stephen M. (1976). Comment on Savage's "On Reading R. A. Fisher." *Annals of Statistics* 4, 498-500.

Stigler, Stephen M. (1999). *Statistics on the Table: the History of Statistical Concepts and Methods.* Cambridge: Harvard University Press.