JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ERNEST KWAN

MICHAEL FRIENDLY

Discussion and comments : strong versus weak significance tests and the role of meta-analytic procedures

Journal de la société française de statistique, tome 145, nº 4 (2004), p. 47-53

<http://www.numdam.org/item?id=JSFS_2004__145_4_47_0>

DISCUSSION AND COMMENTS

Strong versus Weak Significance Tests and the Role of Meta-Analytic Procedures

Ernest KWAN & Michael FRIENDLY *

ABSTRACT

As Paul Meehl once pointed out, there is both a strong form and a weak form of significance tests in the appraisal of scientific theories. Null hypothesis significance testing in psychology is the weak form. The strong form resembles more of Fisher's original inception of hypothesis testing, and it is a much more appropriate method of theory appraisal. We review the distinction between weak and strong tests. While it may be difficult to formulate strong tests in psychological research, we suggest that the adoption of meta-analytical procedures may be a promising first step in the right direction.

RÉSUMÉ

Ainsi que Paul Meehl l'a remarqué, il y a deux formes de test de signification en matière d'évaluation des théories scientifiques, une forte et une faible. Le test d'hypothèse nulle en psychologie est la forme faible. La forme forte ressemble davantage à la conception originelle de Fisher et convient bien mieux à l'évaluation d'une théorie. Nous réexaminons la distinction entre formes faible et forte des tests. Dans la recherche en psychologie il peut être difficile de formuler un test fort, mais nous suggérons que l'adoption de procédures « méta-analytiques » peut être un premier pas dans la bonne direction.

1. Introduction

We are grateful to Dan Denis (2004) for raising important issues regarding the origins of null hypothesis significance testing (NHST) in Fisher's writings, and its transformation over time [1]. We believe his analysis of the components of NHST (summarized in his Table 1) and the changes with Neyman-Pearsón, Bayesian and current practice (at least in the social sciences) does much to clarify the historical development of these ideas and logical problems

* Send correspondence concerning this article to Ernest Kwan, Department of Psychology, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada; email: ernest@yorku.ca

[1] Of course, the first use of a NHST-like procedure may be attributed to John Arbuthnot (1667-1735), who essentially used a sign test to infer that the guiding hand of a divine being could be discerned in the nearly constant ratio of male to female births in London for 1629-1710 (e.g., Gigerenzer & Murray, 1987, p. 4-5).

encountered today in both pedagogy and application of NHST (*e.g.*, Harlow, Mulaik & Steiger, 1997). This brief commentary aims to provide another way out of this present quagmire. For concreteness, our focus is on statistical inference in psychology, although these comments would apply equally to other applied areas (*e.g.*, educational research; see Carver, 1978, 1993).

In psychology there has been much debate over the appropriateness of the approach to testing quantitative hypotheses (see Nickerson, 2000, for a recent summary), and this exchange has been referred to as the NHST controversy. Since Fisher (1925, 1935) was the first to introduce significance tests to psychology (*e.g.*, Gigerenzer, 1993), it is natural to lay the blame for the present discord and inadequacies of NHST solely on his head. Yet as Denis (2004) has argued, we believe this verdict is somewhat unfair. Our goal is to provide a further mitigation of Fisher's role behind the vexing nature of NHST.

The crucial point we put forth is Paul Meehl's distinction between two formulations of significance testing (*e.g.*, Meehl, 1967). Whereas one formulation is problematic and almost void of scientific value, the other, in contrast, can be a valuable tool for theory appraisal. We suggest it is thus not so much Fisher's fault as it is how Fisher has been applied. Reviewing this distinction will effectively illustrate why NHST as typically employed in psychology is of questionable merit. This distinction will also illustrate how recent proposals of reform (Wilkinson & the Task Force on Statistical Inference, 1999) may lead to more meaningful tests of theories.

2. Fisher's null hypothesis test

Formally speaking, Fisher has been credited with *significance* testing, while the work of Neyman and Pearson came to be known as *hypothesis* testing (*e.g.*, Huberty, 1993). That these different sources of contribution be recognized has important implications (*e.g.*, Gigerenzer, 1993; Kwan, 2004), and we shall point out the distinction shortly. For our discussion, however, we will not rely on "significance" and "hypothesis" to demarcate Fisher from Neyman-Pearson. Where it is relevant, we will clarify to whose procedure we are referring.

First, Fisher's null hypothesis test (*e.g.*, Gigerenzer & Murray, 1987, p. 8-12) is conducted by proposing the hypothesis H_0: $\omega = k$, where ω is a parameter of interest, k is a numeric value, and H_0 is the null hypothesis. On the basis of the proposal, one derives a sampling distribution of ω's estimator. By noting how far the sample estimate of ω falls away from the center of this sampling distribution (as reflected by the p-value), one obtains information to evaluate the veracity of H_0. If the p-value is low (estimate is far off in the tail), one may regard the sample data as evidence against H_0.

From Fisher's formulation, we like to point out two important implications. As the sample size (N) used to estimate ω increases, the sampling distribution decreases in variability and a given degree of discrepancy from k will appear less and less likely. All else being equal, by using a larger N to conduct the test, stronger evidence against H_0 will be obtained. Secondly, as ω is based on a continuous variable, it is theoretically impossible that ω is equal to k (*e.g.*, Jones & Tukey, 2000; Nester, 1996). Thus H_0 taken literally is false. Upon finer and finer precision in the estimation of ω, it will be shown that $\omega \neq k$. These are inherent properties of Fisher's null hypothesis test; whether the test leads to meaningful or not so meaningful results depends pivotally on how it is used.

3. Strong versus weak significance tests

As Paul Meehl pointed out (1967) and subsequently elaborated upon (1978, 1990, 1997), consider what happens if a theory, T, has been used to predict H_0: $\omega = k$. Of course one already knows this prediction is literally false, and so should the competent researcher from whom the claim had originated. Underlying the researcher's hypothesis then is the awareness that while T has predicted $\omega = k$, T's veracity is not dependent on this exact equivalence. Rather, one considers the closeness to which ω is to k. If ω is in reality very close to k, then T is of high verisimilitude.

Suppose one carries out a series of tests to evaluate T. If the experimentation of such tests improves (greater N, better measurement), then one is subjecting H_0 to increasing risk of being rejected. Because H_0 is a derivation of T, one is correspondingly subjecting the veracity of T to greater and greater challenge. With infinite power and zero measurement error, H_0 will of course be rejected. But short of this asymptotic state of affairs, should H_0 remain unrefuted despite efforts to make the test easier to refute H_0, a great deal of impressive corroboration would have accumulated for T. After all, this series of outcomes implies that ω is indeed very close to k. Meehl (1990) has labeled this form of significance testing the "strong" test.

As may be apparent to anyone familiar with statistical inference in psychology, the strong test resembles little what is commonly done. Rather, psychologists carry out "weak" significance tests, and this is the NHST that has provoked the uproar of critics. To illustrate NHST, we adapt Meehl's diagram (1990) in Figure 1. Based on T, one derives (arrow 1) the hypothesis H_1 that a difference or relationship exists (e.g., H_1: $\omega \neq 0$, where ω is the parameter of difference or association). Then one forms the logical complement of H_1 (arrow 2) that claims the absence of this difference or relationship (H_0: $\omega = 0$). H_0 derives (arrow 3) the sampling distribution, and as before, pending where the sample estimate of ω falls, one may reject H_0 (arrow 4). A rejected H_0 is proof of H_1 (arrow 5), and proof of H_1 is used as proof of T (arrow 6).

Imagine what happens if one uses a series of NHST to appraise T. As in the strong test, should experimentation improve, H_0 is subjected to easier and easier refutation. But because the confirmation of T arises from rejecting H_0, it also means that T will become easier and easier to acquire corroboration! Not only does the weak test lead to a counter-scientific scenario where theory confirmation solely depends on increasing N or using finer instruments, the

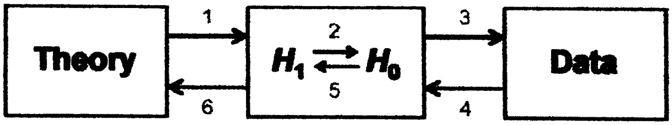

FIG 1. — Sequence of thinking in NHST. Solid arrows denote derivations: (1) From substantive theory to H_1; (2) from H_1 to H_0; (3) from H_0 to expectations of the observed data. Broken arrows denote inference: (4) from observed data to decision on H_0; (5) from decision on H_0 to decision on H_1; (6) from decision on H_1 to decision on the theory. Adapted from Meehl (1990).

entire enterprise is downright deplorable as no aspect of T is being tested at all!

It is true that by rejecting H_0, the test does seem to speak for T because T asserted the opposite of H_0. But one needs to be reminded again of the falsity of H_0: $\omega = 0$. In the domain of psychological phenomena, it is further empirically false that populations are absolutely equal on some characteristic, or that variables are completely unrelated (*e.g.*, Cohen, 1994; Meehl, 1997). Thus by only predicting a difference or an association, T has essentially done nothing of substance; confirmation of such a prediction is likewise an empty feat.

There is one small redemption for weak tests. If T asserts a directional hypothesis, *e.g.*, H_1: $\omega > k$ (thus, H_0: $\omega \leqslant k$), the procedure becomes a one-tailed test. The H_0 sampling distribution is then based on the closest value of ω to the range in H_1 without being in that range, *i.e.*, $\omega = k$. One rejects H_0 if the estimate of ω is much bigger than k in this sampling distribution. Unlike the previous non-directional test, H_1 may or may not be true. It is thus more of an accomplishment for T if the test does reject H_0.

Suppose, however, that T is completely void of merit. Then T's derivation of a directional H_1 is like a random coin toss: T has a 50% chance of claiming the correct direction. Consequently as experimentation improves, the chance of confirming baseless theories approaches 50%. Relying on one-tailed weak tests for theory appraisal thus still leaves much to be desired.

4. Fisher and the misuse of significance tests

As we have shown, Fisher's inception of the null hypothesis test is much more similar to the strong test than it is to the weak test. Thus to the extent that applications of the weak test (*i.e.*, NHST) have been criticized for being flawed and fruitless, it seems hardly fair to associate Fisher with these inadequacies.

For example, all too often, the failure of a weak test to reject H_0 has misled many to believe that there is no effect present. This belief of course overlooks the influence of power, or the tenability of a zero effect H_0 in the first place. To make matters worse is that a conclusion of "no effect" generally does not promote the calculation and reporting of effect size estimates. Then as far as accumulating evidence for a research program, resources have been wasted and progress has been impeded. Please see, *e.g.*, Schmidt's detailed discussion (1996) of such problems.

Perhaps the most justified criticism of Fisher is in suggesting the "$p < 0.05$" convention (Fisher, 1935) that has since been ingrained into the mind of psychologists (*e.g.*, Rosenthal & Gaito, 1963). To many, 0.05 is a definitive cut-off and pending on which side one's p-values fall, it could mean *proof, respectability, publication*, or despairingly, the *lack of.* Sure enough, given the mighty role p-values play, data analysis in psychology has evolved into p-value tabulations that are primarily concerned with whether or not statistical significance has been reached. Please see, *e.g.*, discussions by Gonzalez (1994), and Hallahan and Rosenthal (2000).

It is important to point out that a major contribution to such misuses is the failure to distinguish between Fisher's null hypothesis test from the method of Neyman and Pearson (*e.g.*, Gigerenzer, 1993). NHST is in fact a horrible conflation of these two incompatible schools of inference (*e.g.*, Denis, 2004; Gigerenzer, 1993; Huberty, 1993). The ideas of competing hypotheses (H_0 vs. H_1), decision errors, power, and critical values came from Neyman-Pearson; their formulation is meant for reaching a decision to guide a course of action, not the evaluation of scientific theories (*e.g.*, Gigerenzer & Murray, 1987, p. 12-17). Thus by using NHST to appraise theories, it is of no wonder why psychology has cultivated such counter-scientific customs and beliefs (*e.g.*, Rozeboom, 1960).

5. Conclusion: Recommendations of reform

In 1996, as a response to the increasing debates over NHST, the Board of Scientific Affairs of the American Psychological Association assembled the Task Force on Statistical Inference (TFSI) to investigate and address the controversy (Wilkinson & the TFSI, 1999). Members of the TFSI included both prominent psychologists and statisticians, and in 1999 the TFSI reported its recommendations (Wilkinson & the TFSI, 1999). While the scope of this report went far beyond the original mandate, the central advice pertaining to NHST is that one should not rely on accept-reject decisions alone. Instead, one ought to make more use of graphics and adopt a "meta-analytic" perspective towards data analyses (Wilkinson & the TFSI, 1999; also see elaborations by Cumming and Finch, 2001, Panicker, 2000, and Thompson, 2002).

Meta-analysis is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). To take a meta-analytic perspective in research is to not draw definitive conclusions based on isolated investigations alone. The emphasis instead is on replication and the comparison and accumulation of many studies. One regards any single study as just a modest contribution to a body of evidence. Through confidence intervals and effect size estimates, one integrates results across studies to obtain a more comprehensive and accurate description of the effect under study.

To conduct research meta-analytically is thus a drastic improvement over the unscientific decision-making nature of NHST. But furthermore, meta-analytic procedures can be of value by facilitating the estimation of parameters and effect sizes. While in many areas of psychology it may not be easy for theories to derive precise numeric predictions (*e.g.*, Meehl, 1978, 1997), devoting more effort to estimation and description is a start. As psychology obtains better quantification over the phenomena of its research, the day may come when psychological theories will be able to assert a lot more than the mere existence of differences or associations alone.

References

CARVER R. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.

CARVER R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287-292.

COHEN J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, 49, 997-1003.

CUMMING G., & FINCH S. (2001). A primer on the understanding, use, and calculations of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532-574.

DENIS D. J. (2004). The modern hypothesis testing hybrid: R. A. Fisher's fading influence. *Journal de la Société Française de Statistique*, 145, 4, 5-26.

FISHER R. A. (1925). *Statistical methods for research workers.* London: Oliver & Boyd.

FISHER R. A. (1935). *The design of experiments.* New York: Hafner.

GLASS G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, 5(10), 3-8.

GIGERENZER G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

GIGERENZER G., & MURRAY D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

GONZALEZ R. (1994). The statistics ritual in psychological research. *Psychological Science*, 5, 321, 325-328.

HALLAHAN M., & ROSENTHAL R. (2000). Interpreting and reporting results. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 125-149). San Diego, CA: Academic Press.

HARLOW L. L., MULAIK S. A. & STEIGER J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, New York: Lawrence Erlbaum Associates.

HUBERTY C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, 61, 317-333.

JONES L. V., & TUKEY J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5, 411-414.

KWAN E. (2004). The null hypothesis significance testing controversy and the teaching of introductory statistics. Submitted to *Teaching of Psychology*; under revision.

MEEHL P. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science*, 34, 103-115.

MEEHL P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.

MEEHL P. (1990). Appraising and amending theories: The strategy of Lakatosian defense and the two principles that warrant it. *Psychological Inquiry*, 1, 108-141.

MEEHL P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-425). Mahwah, New York: Lawrence Erlbaum Associates.

NESTER M. R. (1996). An applied statistician's creed. *Applied Statistics*, 45, 401-410.

NICKERSON R. S. (2000). Null hypothesis significance testing: A review of an old· and continuing controversy. *Psychological Methods*, 50, 241-301.

PANICKER S. (2000). Narrow and shallow. *American Psychologist*, 55, 965-966.

ROSENTHAL R., & GAITO J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55, 33-38.

ROZEBOOM W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416-428.

SCHMIDT F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods*, 1, 115-129.

THOMPSON B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31, 24-31.

WILKINSON L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594-604.