Some problems with Chow's problems with power

Deborah G. Mayo
Department of Philosophy, Virginia Polytechnic Institute, Blacksburg, VA 24061. **mayod@vt.edu.**

Abstract: Chow correctly pinpoints several confusions in the criticisms of statistical hypothesis testing but his book is considerably weakened by its own confusions about concepts of testing (perhaps owing to an often very confusing literature). My focus is on his critique of power analysis (Ch. 6). Having denied that NHSTP considers alternative statistical hypotheses, and having been misled by a quotation from Cohen, Chow finds power analysis conceptually suspect.

Standard statistical testing (null-hypothesis significance tests and Neyman–Pearson methods), while widely used in diverse sciences, have been the subject of considerable criticism and controversy among philosophers (especially Bayesians) and others.

There is no doubt that these methods are in need of defense from someone who can clarify the complex issues and disentangle the disagreements and confusions involved, especially in the psychological literature.

Chow's book (1996) raises a number of important and correct points against critics: many of the criticisms, Chow rightly notes, are based on confusing statistical inference with substantive inductive and scientific inference, on poorly designed or misinterpreted tests, and on a misplaced desire for a probability that these methods are not designed to supply: a posterior probability of a hypothesis. Chow is at his best when emphasizing what critics tend to overlook: that statistical tests concern hypotheses about a sampling distribution (e.g., of a test statistic), that such hypotheses must be distinguished from what he calls experimental and research hypotheses, and that the result of statistical testing must be distinguished from corroborating a scientific hypothesis, although progress toward theory corroboration may be afforded by combining sufficiently numerous and probative statistical tests. Correct too is Chow's distinction between the context of theory corroboration in science and the Bayesian context. These ideas warrant further attention. But first Chow should rethink the version of standard statistical testing theory worthy of being defended.

Null-hypothesis statistical-testing procedure, NHSTP. NHSTP is the hybrid of Fisherian and Neyman-Pearson (NP) tests that Chow imagines practitioners use, and it is the one he is defending. Although the essential contribution of NP theory was the introduction of alternatives to the null hypothesis and the corresponding power function – Chow discards this from NHSTP (e.g., "Fisher was correct not to consider Type II error because it plays no role in NHSTP," p. 43). What Chow keeps from NP theory is the conception of a test as a decision procedure to reject or accept a null hypothesis (of chance) according to whether data reach a preset critical value of a test statistic. To some, it might seem as if Chow's NHSTP ejects the best parts of each approach.[1] Given this view of tests, it is not surprising that Chow finds the notion of *power* problematic.

Chow's critique of power analysis. Chow faults the field of power analysis for two reasons: (1) "The a priori probability of obtaining statistical significance is said [by power analysts] to be given by the power of the test" (p. 131) but this is false; and (2) NHSTP is restricted to only the null hypothesis H_0, but power analysis depends upon alternative statistical hypotheses. Chow's charges against power analysts accordingly boil down to arguing first, that a power calculation does not give an unconditional probability (of a statistically significant result) and second, that the calculation of power is impossible for an account that excludes alternative statistical hypothesis. Both charges are correct, yet they

are not damaging to a correct use and interpretation of power in standard Neyman-Pearson testing. I will take these up in turn:

1. Chow is misled throughout by an unfortunate quote from Cohen (1987, p. 1) that "The power of a statistical test is the probability that it will *yield* statistically significant results" (Chow's emphasis; quote 6-2 on p. 120). Thus, Chow charges that in power analysis, "the power of the statistical test is treated as the probability of H_1 being true (by virtue of the fact that it represents, to power analysts, the probability of obtaining statistical significance)" (p. 124). Chow seems also to be confusing (or alleging that the power analyst confuses) the probability a test correctly rejects H_0 and accepts H_1, with the probability that H_1 is true. Anyone who treats power as either the probability of a significant result or the probability of H_1 is justly castigated by Chow – but does anyone commit such egregious errors?

2. According to Chow (p. 132), the probability of a Type II error should be defined as (i) p(Accept Chance∪not-H_0) while in power analysis it is defined as (ii) p(Accept Chance∪H_1). But (i) is not defined in NP theory unless not-H_0 is a point hypothesis, and since Chow's NHSTP excludes such alternatives it is not surprising Chow concludes that "it is impossible to represent statistical power graphically in the sense envisaged in power analysis without misrepresenting NHSTP" (p. 137). But if so, then it is NHSTP that forces a nonstandard interpretation of the probability of a Type II error. For Chow, (i) refers to the probability that the test Accepts H_0 when some (substantive) nonchance factor is really responsible – a calculation which he admits a statistical method cannot supply. We are not told why such a nonstandard notion should be preferred to the standard statistical one, nor why we should oust alternative statistical hypotheses from our methodology of testing. Moreover, since power analysts are working within NP testing theory where it is entirely appropriate to consider the power of a test to reject H_0 for various different point alternatives – that is, power curves – Chow's criticism misses its target.

By restricting himself to the single hypothesis of the Fisherian test, Chow's defense of NHSTP is forced to accept an overly limited role for statistical analysis: "NHSTP answers the question as to whether or not there is an effect. However, it is not informative about the magnitude of the effect." (p. 7). In fact, considering a test's ability to detect alternatives can provide information about the magnitude of the effect that is or is not indicated by a statistical result. For example, if a test had a high [low] power to detect an effect of a magnitude specified in H_1 then failure to reject the null hypothesis (of 0 effect) would be a good [poor] indication that the magnitude of the effect was less than H_1 asserts. Thus, power considerations offer a good way to scrutinize the meaning of statistical results,[2] and Chow has given us no reason to abandon them.

Chow overlooks the fact that, although his one-sided tests may be articulated with reference to the null hypothesis alone, their justification as good or best tests had to be derived by considering alternative statistical hypotheses (e.g., as in deriving uniformly most powerful tests). Chow's NHSTP tests are cut off from their logical foundation in NP theory.

An error regarding the goal of NP statistics. On pp. 21 (Table 2.3), 23, 42, and elsewhere, Chow asserts – quite erroneously – that the probability of interest in NP statistics is "the inverse probability, p(H∪D)" (p. 21) and declares that "the Neyman–Pearson preference for the inverse probability" is not consistent with the mathematical foundation of NHSTP (p. 42). Indeed, such an NP preference would also be inconsistent with NP theory (which was designed for cases where no such inverse probability

was even meaningful)! Chow's mistake, if uncorrected, will only supply further grist for the Bayesian mills which regularly accuse NP theory of unsoundness (alleging it to be interested in posterior probabilities while supplying only error probabilities).

NOTES

1. My own preference would be to reverse what gets ejected: to retain the NP use of alternative hypotheses while replacing the decisiontheoretic interpretation of tests with an inferential one. Power calculations are needed to specify tests, but to infer what is and is not indicated by a specific result may be achieved by calculating error probabilities using that result (rather than a preset cut-off). Suppose, for example, that the *p*-value observed is not small and so H_0 is "accepted." To interpret this one might calculate, not the usual power of the test against an alternative H_1, but

rather, the probability of a result more significant statistically than the one obtained given H_1. How this relates to using tests inferentially is discussed in Mayo 1996.

2. An analysis sensitive to the specific value of the result is also possible (see N. 1).