In summary, the book is essentially an elaboration on the point that there is really nothing to defend NHSTP against. If one agrees with this, one should not read the book.

Some statistical misconceptions in Chow's *Statistical significance*

Jacques Poitevineau[1] and Bruno Lecoutre[2]

[1]*LCPE, C.N.R.S., 92120 Montrouge, France.* **jacques.poitevineau@ens.fr.**
[2]*C.N.R.S. et Université de Rouen, 76821 Mont-Saint-Aignan Cedex, France.*
bruno.lecoutre@univ-rouen.fr http://epeire.univ-rouen.fr/labos/eris

> I know of no field where the foundations are of such practical importance as in statistics.
>
> (Lindley 1972)

Abstract: Chow's book makes a provocative contribution to the debate on the role of statistical significance, but it involves some important misconceptions in the presentation of the Fisher and Neyman/Pearson's theories. Moreover, the author's caricature-like considerations about "Bayesianism" are completely irrelevant for discarding the Bayesian statistical theory. These facts call into question the objectivity of his contribution.

Frequentist theories. In presenting current practices (NHSTP), Chow rightly refers to the "hybridism" of Fisher and Neyman/Pearson's theories; this was identified long before Gigerenzer (1993), although this particular term was not used: see, for example, Morrison and Henkel 1970, p. 7. But to say that these authors may be responsible for the hybridism because they "sometimes changed positions" and "incorporated their opponent's ideas" (p. xi) is far from the truth, since their basic theories evolved little.

Chow rightly emphasizes that α, β, and *p* are conditional probabilities in the frequentist framework. This follows from the fundamental mathematical notion of sampling distribution (conditional on the parameters). But, with regard to this distribution, the assertion that "the standard error of the mean . . . is . . . $s/\sqrt{n}$ if σ is not known" (p. 27) is nonsense, since *s* is in this case a random variable that varies from one sample to another.

Fisher. Nothing is said about Fisher's concepts of probability, which are of direct importance for the objectives Fisher assigned to statistical methods. If, in early writings, he effectively mentioned the idea of a "conventional α" (Chow's Table 2.3, p. 21, line 9), he later came to repudiate any systematic predetermined level of significance (Fisher 1956/1990, p. 45). Moreover, he argued against the interpretation of *p* as the relative frequency of error when sampling repeatedly in a same population (Fisher 1956/1990, pp. 81–82).

Last, contrary to what Chow implies, Fisher was evidently interested in p(H|D) (although this probability cannot, of course, be identified with *p*) as it emerges from his work, not only on the fiducial theory (e.g., Fisher 1935b; 1956/1990), but also on the Bayesian method (a fact often ignored) in his last years (Fisher 1962).

Neyman/Pearson. It is first rather surprising that the single reference is the (Neyman & Pearson) 1928 paper that presents only the premises of their theory. In Neyman's own words: "Our first paper on the subject was published in 1928, over twenty years ago. However, it took another five years for the basic idea of a rationale theory to become clear in our minds." (Neyman 1952, p. 58). Actually, some verbal formulations in the introduction of this paper could be misleading, and above all the concept of statistical power did not appear. Relevant references are the 1933 papers (Neyman & Pearson 1933a; 1933b).

Contrary to what is indicated in [Table 2.3, line 5], the probability of interest is not "the inverse probability, p(H|D)." Neyman and Pearson devised a precise method which is independent of the prior probabilities p(H) and hence cannot say anything about p(H|D) (Neyman & Pearson 1933b). Moreover, as a radical frequentist, Neyman later came to discard *p*(H), and consequently *p*(H|D), when it could not be assigned a frequentist meaning (Neyman 1950; 1952).

Also H_0 is not "the hypothesis of zero difference" (Table 2.3, p. 21, line 6); on the contrary, it should be the hypothesis of interest, the one for which it is more important to avoid error. This fact is clear and more than implicit, even in the 1928 paper; and it was firmly stressed by Neyman (Neyman 1950, p. 263). Consequently, H_1 is not "the substantive hypothesis itself" (Table 2.3, line 7) but the challenging one.

Moreover, "Multiple H_1's" (Table 2.3, line 8) are not necessarily assumed. The alternative hypothesis (and also H_0) may be simple as well as multiple. More important again is the fact that H_0 and H_1 are always assumed to be mutually exclusive and exhaustive (what Chow considers to be only a Fisherian feature), so that "not-H_0" is really equivalent to H_1 (or H_1's), contrary to Chow's claim (p. 132). Thus it cannot be said that "the identity of the non-null hypothesis (viz., not-H_0) is not known" (p. 186). Even in the case of a multiple hypothesis (e.g. $\mu > 0$), the possible values of the parameter under the alternative hypothesis are perfectly known (e.g., the strictly positive part of the real line). What is indeterminate is only the knowledge of the *true* value.

It is basically nonsense to say that "the meaning of 'Type II Error' is changed when statistical power is introduced" (p. 131). The concept of statistical power was introduced by Neyman and Pearson as the complement of type II error (Neyman & Pearson 1933b), and plays a basic and explicit role in their theory (a Neyman-Pearson's test consists of building a critical region that minimizes β, or equivalently that maximizes power, for a fixed α). Power, with regard to any simple hypothesis H_i, is the probability that the test statistic falls into the critical region (given that H_i is true), and is consequently equivalent to $1 - \beta_i$ (β_i denoting the probability of committing a type II error given H_i). In the case of multiple H_i's, power can be calculated for each of the simple hypotheses (which is perfectly determined, as indicated earlier), thereby leading to the well-known *power function*.

Concerning the implicit appeal to TSD by power analysts (Ch. 6), Chow seems to be unaware that it is signal detection theory that drew upon Neyman-Pearson's theory (Green & Swets 1966, p. 1), and not the reverse.

Last, Chow forgets to report the fact that the sample size *N* must be fixed *a priori* (before experiment); and, curiously enough, he does not mention the criticisms raised about the "decision characterization" of the Neyman-Pearson's theory (e.g., Rozeboom 1960).

Statistical Bayesian theory. Chapter 7 is only Chow's personal views about "Bayesianism." This has nothing to do with the rational statistical Bayesian theory, and is consequently completely misleading. Chow ignores the foundations of Bayesian statistics and appreciable theoretical developments that are ever more strongly challenging the frequentist approach in all domains (for recent accounts, see in particular Bernardo & Smith 1994; Robert 1995), including the related methodological debates peculiar to medicine (e.g., see Ashby 1993). Moreover, realistic uses of Bayesian methods for analyzing experimental data have been proposed (e.g., Spiegelhalter et al. 1994; Lecoutre et al. 1995; Rouanet 1996).

Null-hypothesis tests are not completely stupid, but Bayesian statistics are better

David Rindskopf

Educational Psychology, City University of New York Graduate Center, New York, NY 10036. **drindsko@email.gc.cuny.edu**

Abstract: Unfortunately, reading Chow's work is likely to leave the reader more confused than enlightened. My preferred solutions to the "controversy" about null- hypothesis testing are: (1) recognize that we really want to test the hypothesis that an effect is "small," not null, and (2) use Bayesian methods, which are much more in keeping with the way humans naturally think than are classical statistical methods.