

Book Review

*Journal of Official Statistics*

04-30-98

**Chow, S.L., *Statistical Significance: Rationale, Validity and Utility*. SAGE Publications, London, 1996. ISBN 0 7619 5205 5, xi+205pp.**

Significance testing has been criticized over the years on logical and philosophical grounds, as well as blamed for the slow progress of “soft” psychology (Harlow, Mulaik and Steiger, 1997; Meehl, 1978; Morrison and Henkel, 1970). Chow’s primary thesis in this book is that the null-hypothesis significance-test procedure (NHSTP) has been criticized for the wrong reasons. Through a comprehensive treatment of criticisms of the NHSTP, Chow outlines statistical and philosophical arguments that support his contention that these criticisms are unwarranted.

The book is organized into eight chapters, with the first chapter summarizing the basic criticisms of the NHSTP. Chapter 2 describes the features of the NHSTP and its mathematical foundations, particularly its origins in the works of Fisher and of Neyman and Pearson. Subsequent chapters address the criticisms in detail, providing a rationale for why they might be invalid. The criticisms are interwoven throughout the book, rather than addressed in separate chapters, and they revolve generally around these issues: (a) problematic aspects of both the null and research hypotheses; (b) ambiguities in the meaning of statistical significance; (c) uninformative nature of the significance test in terms of effect size, likelihood of theory corroboration, and usefulness of results; (d) what significance testing cannot do; and (e) arguments that the null hypothesis is never true.

Chapters 3 and 4 are the toughest reading in the book, because the concepts are difficult to comprehend for those untrained in the philosophy of science and metatheoretical notions. Yet, the concepts are central not only to Chow’s arguments in favor of the NHSTP, but also to the reader’s grasp of how the remaining chapters support the book’s primary thesis. Chapter 3 makes a distinction between substantive and statistical hypotheses, which Chow argues is central to an understanding of the NHSTP and its criticisms. A description of the logical relations among hypotheses in four kinds of experiments (theory corroboration, utilitarian, clinical, and generality) is presented to show that the NHSTP criticisms, while unwarranted in every case, are even more inappropriate for some kinds of experiments than for others. The main point of Chapter 4 is to demonstrate that NHSTP is not an inductive procedure that can or should be used for theory corroboration. Instead, Chow maintains that the only role for the NHSTP is limited to assessing whether chance influences can be disregarded in hypothesis testing. At the same time, however, the NHSTP is an essential step in the process of securing evidential support for theory.

In Chapter 5 Chow argues that, because of the ambiguous and anomalous nature of the NHSTP, many of the complaints about it are based on non-statistical reasoning. Particular examples here include claims that the null hypothesis of no difference between two populations can never be true and that statistical significance does not provide information about the substantive value of the research. Criticisms of the NHSTP posit that statistical significance is largely (a) a matter of sample size, (b) related to effect size in anomalous ways, and (c) uninformative with regard to the

practical value of the results. Not surprisingly, given his argument in Chapter 4, Chow also rebuts the common wisdom of abandoning statistical significance in favor of reporting effect size because of its greater value in providing evidential support.

The power of a statistical test to yield statistically significant results is related mathematically to the NHSTP, and the use of power to determine sample size has become standard operating procedure. Chapter 6 challenges the validity of power as well as the notion that power should be used instead of, or in addition to, statistical significance. Chow argues that (a) the meaning of Type II error is different for power and for the NHSTP, (b) those two meanings cannot be reconciled, and (c) graphical representation of power is inconsistent with that of the NHSTP.

Chapter 7 evaluates the Bayesian approach to statistical decision-making and its implications for the NHSTP. Chow's view is that the Bayesian approach requires a different data gathering procedure--which he refers to as the "sequential-sampling procedure"--than that typically used in a theory corroboration experiment. This requirement limits the usefulness of Bayesian methods and argues for continued use of the NHSTP.

The final chapter re-summarizes the basic criticisms of the NHSTP in terms of a set of 14 questions that could be asked of data. Chow suggests that it is unreasonable to expect that the NHSTP can provide an answer to each of these questions. Instead, the major role of significance testing is to allow a decision about whether chance can be ruled-out as an explanation of the data. Arguing that the criticisms of the NHSTP are flawed because basic metatheoretical concepts are misunderstood, Chow concludes that the NHSTP has a limited--but very important--role in empirical research. These conclusions are echoed in a more recent and broader treatment of this topic (Trout, 1998), which likewise argues that most of the standard criticisms of the NHSTP are unwarranted.

Chow's work fills a void in the literature on significance testing and provides a concise, but thorough, review of the significance test controversy. Specifically, graduate students in the social and behavioral sciences generally are introduced to significance testing in their first course in statistics, but few students have the opportunities this book offers to explore the methodological rationale and the philosophical underpinnings of statistical significance testing. However, this book may require several readings to appreciate fully the complex arguments that are made.

## **References**

- Harlow, L., Mulaik, S., & Steiger, J. (Eds.) (1997) What If There were No Significance Tests? Lawrence Erlbaum: Mahwah, N.J.
- Meehl, P. E. (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.
- Morrison, D. E. & Henkel, R. E. (Eds.) (1970) The Significance Test Controversy: A Reader.

Aldine Publishing, Chicago.

Trout, J. D. (1998) Measuring the Intentional World. Oxford University Press, New York.

John Tarnai  
Social & Economic Sciences Research Center  
Washington State University  
Pullman, WA