Chow's defense of null-hypothesis testing: Too traditional?

Robert W. Frick

Department of Psychology, State University of New York at Stony Brook, Stony Brook, NY 11790. **rfrick@sunysb.edu** **www.psy.sunysb.edu/rfrick/**

Abstract: I disagree with several of Chow's traditional descriptions and justifications of null hypothesis testing: (1) accepting the null hypothesis whenever $p > .05$; (2) random sampling from a population; (3) the frequentist interpretation of probability; (4) having the null hypothesis generate both a probability distribution and a complement of the desired conclusion; (5) assuming that researchers must fix their sample size before performing their study.

Critics of the null-hypothesis statistical-testing procedure (NHSTP) do not tend to criticize one another, despite differences in their positions. For example, NHSTP is criticized but power analyses are not, even though a power analysis assumes the existence of NHSTP. Researchers are advised to report effect size in statistical units such as Cohen's *d* (e.g., Schmidt 1996) or to report confidence intervals (e.g., Loftus & Masson 1994), but they are not told to report confidence intervals for effect size reported in statistical units. Cohen (1994) criticized the underlying logic of NHSTP but then suggested that researchers report confidence intervals because that accomplished NHSTP for all possible null hypotheses.

One might expect the defenders of NHSTP to ally, but this alliance too would be unnatural. I agree with Chow that NHSTP plays an essential and irreplaceable role in science (Frick 1996). I agree with many of his points, especially that effect size is not relevant in the theory-corroboration experiment. However, I disagree with many of the justifications Chow provides for NHSTP. In this commentary, I will focus on ways that Chow is in a sense too traditional. In assessing NHSTP, the actual practice of researchers must be distinguished from the way it is described in textbooks and the attempts to justify that practice logically. In each of the following criticisms, Chow has defended the traditional description or justification of NHSTP rather than the actual practice of researchers.

First, Chow implies that the null hypothesis is accepted whenever $p > .05$. Good researchers do sometimes argue that their evidence supports a hypothesis of no effect or no difference, but they use more evidence than just $p > .05$ (e.g., Frick 1995).

Second, Chow uses random sampling from a population to justify the construction of the requisite probability distribution. This implies that researchers should sample randomly from populations and that the business of statistical testing is making claims about populations. I disagree. To make a claim about a pattern in the data, such as that one treatment is more effective than another, the researcher must address the possibility that this observed pattern occurred just by chance. As Chow notes, statistical testing accomplishes this, with *p* being a measure of the strength of the evidence against the just-by-chance hypothesis. The outcome of statistical testing and a lack of artifacts – which I call the finding – is a conclusion about the subjects tested. No assumption of randomsampling is needed for this interpretation (Frick, in press b).

Third, Chow defends the frequentist interpretation of probability, in which probability is defined as the limiting ratio of an infinite sequence of trials. This definition confuses probabilities with the method of measuring probabilities. In other words, it is the operationalism Chow decries (p. 153). A propensity definition of probability better justifies the procedures of NHSTP (Frick, in press b).

Fourth, in the traditional justification of NHSTP, the null hypothesis plays two roles – it generates the probability distribution underlying the determination of *p*, and it is the complement of the researcher's desired conclusion. These two roles are incompatible. To generate the probability distribution, a point hypothesis, for example, $\mu_1 = \mu_2$ is needed. However, the complement of this is $\mu_1 \neq \mu_2$, which is not the claim researchers make and – as critics of NHSTP are fond of noting – not even a claim worth making. Researchers in practice make a directional claim, such as $\mu_1 < \mu_2$. To allow this claim, Chow describes the null hypothesis as being directional, for example, $\mu_1 \leq \mu_2$. However, this leads Chow to the awkward position of primarily defending the use of a one-tailed test, which researchers rarely use. This definition also does not support the definition of *p* as the probability of achieving the observed results or larger given the null hypothesis.

A solution is this: A point hypothesis is used to generate the probability distribution. Following the conventional rules of science, $p < .05$ allows rejection of this hypothesis, and it would also allow rejecting the hypotheses even more discrepant from the observed data. Therefore, a directional conclusion can be made. This is exactly the process Chow describes (and Fisher before him), but it cannot be described with a single null hypothesis serving two roles.

Fifth, Chow equates NHSTP with the fixed-sample stopping rule, in which the number of subjects is determined in advance. Do researchers actually use the fixed-sample stopping rule? Do researchers never (a) give up part way through a study because the results were discouraging, (b) test less than the planned number of subjects because *p* was already less than .001, or (c) test more subjects than planned when *p* was slightly greater than .05? These actions seem rational to me, but they violate the fixed-sample stopping rule. Fortunately, the alternatives to the fixed-sample stopping rule – sequential stopping rules in which the number of subjects is not fixed in advance – are compatible with NHSTP. Because of their increased efficiency and practicality, sequential stopping rules should usually be preferred to the fixed-sample stopping rule (Frick, in press a).

We need statistical thinking, not statistical rituals

Gerd Gigerenzer

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany. **gigerenzer@mpib-berlin.mpg.de**

Abstract: What Chow calls NHSTP is an inconsistent hybrid of Fisherian and Neyman-Pearsonian ideas. In psychology it has been practiced like ritualistic handwashing and sustained by wishful thinking about its utility. Chow argues that NHSTP is an important tool for ruling out chance as an explanation for data. I disagree. This ritual discourages theory development by providing researchers with no incentive to specify hypotheses.

Future historians of psychology will be puzzled by an odd ritual, camouflaged as the *sine qua non* of scientific method, that first appeared in the 1950s and was practiced in the field for the rest of the twentieth century. In psychology and education textbooks of this period they will find this ritual variously referred to as "statistical inference," null hypothesis testing, significance testing, and most recently, NHSTP. These historians will be surprised to learn that the ritual was quickly institutionalized, although (1) the eminent psychologists of the time – including Sir Frederick Bartlett, R. Duncan Luce, Herbert Simon, B. F. Skinner, and S. S. Stevens – explicitly wrote against its use (Gigerenzer & Murray 1987); (2) the statisticians Sir Ronald Fisher, Jerzy Neyman, and Egon S. Pearson would all have rejected NHSTP as an inconsistent mishmash of their ideas (Gigerenzer et al. 1989, Chs. 3 and 6); (3) hardly any eminent statistician of the time endorsed it; and (4) although it was presented to psychologists as the scientific method, it never caught on in the natural sciences.

Chow (1996) responds to a paper (Gigerenzer 1993) in which I used a Freudian analogy to capture how the conflicts between Neyman and Pearson's doctrine (the superego), Fisher's null hypothesis testing (the ego), and the Bayesians's approach (the id) have been projected into the psyches of textbook writers and researchers in psychology. The results are wishful thinking, sup-

pression of conflicts, and a statistical practice – null hypothesis testing – that resembles ritualistic handwashing. For instance, many textbook authors and the majority of experimenters do not understand what its final product – a *p*-value – actually means (see Acree 1978; Gigerenzer 1993; Oakes 1986; Sedlmeier & Gigerenzer 1989). Chow acknowledges this, but argues that if we can strip NHSTP (his term for an inconsistent hybrid of Fisherian and Neyman-Pearsonian ideas) of the mental confusion associated with it, something of limited but important use is left. According to Chow, NHSTP's usefulness is "restricted to deciding whether or not research data can be explained in terms of chance influences" (p. 188). This sounds like a reasonable and modest proposal, and Chow succeeds in pointing out many sources of confusion about significance testing. I do not, however, believe that even in this purified form NHSTP has much value for psychological research. Rather, this ritual undermines progress in our field by giving researchers no incentive to specify their hypotheses and by replacing statistical thinking with a mindless statistical procedure.

Is testing unspecified hypothesis against "chance" a good research strategy? No. The single most important problem with null hypothesis testing is that it provides researchers with no incentive to develop precise hypotheses. To perform a significance test, one need not specify the predictions of either one's own research hypothesis or those of alternative hypotheses. All one has to do is test an unspecified hypothesis (H_1) against "chance" (H_0). In my experience, the routine of testing against chance using NHSTP promotes imprecise hypotheses.

To be sure, there are cases where testing against chance makes sense, such as in parapsychology.[1] But read John Arbuthnot's proof of God against chance in 1710 – the earliest null hypothesis test of which I know – and you see the flaws in this program (Gigerenzer & Murray 1987, pp. 4–5). In a science striving for precise process models, one needs methods that test the predictions of one model against those of alternative models, not a ritual that tests an unspecified hypothesis against chance.

Recall that statistical thinking involves making an informed choice among the various techniques available. Avoiding statistical thinking in the name of "objectivity," as Chow's implicitly advocates, has produced blind spots in research (Gigerenzer 1987). There is a toolbox of statistical methods for testing which of several predictions, if any, comes closest to the data. For certain problems least squares are useful, for others maximum likelihood, Neyman-Pearson analysis, Wald's sequential analysis, or Bayesian models. But even simple descriptive statistics can be better than null-hypothesis testing at discriminating between hypotheses. For instance, Anderson and Cuneo (1978) proposed two hypotheses about the processes underlying children's estimates of the area of rectangles ("adding" versus "multiplying" height and width). Following the null hypothesis-testing ritual, they identified one with chance ("adding") and did not specify the predictions of the other. Because the anova test was not significant, they took this as evidence for the "adding" process. However, had the authors specified the precise predictions of *both* hypotheses, they would have seen that the data pattern was in fact close to that predicted by the "multiplying" process and not by the null hypothesis (see Gigerenzer & Murray 1987, p. 100; Gigerenzer & Richter 1990). This example illustrates one blind spot that results from using NHSTP, which requires that the prediction of only one hypothesis be specified. Hypothesis testing should be symmetric, not asymmetric.

NHSTP allows researchers to get away with imprecise hypotheses and predictions. Testing an unspecified hypothesis against chance may be all we can do in situations where we know very little. But when used as a general ritual, this method ironically ensures that we continue to know very little.

Compulsory rules. Chow proclaims that null hypothesis tests should be interpreted mechanically using the conventional 5% level of significance. This is what Fisher suggested in his 1935 book, a practice that was subsequently codified by many textbook writers into a religious doctrine of "objectivity." Later, this practice was rejected by both Fisher and Neyman and Pearson, as well as practically every other eminent statistician (Gigerenzer et al. 1989). The reason Fisher adopted a conventional level of significance of 5% (or 1%) in the first place seems to have been that he had no table, for other significance levels, partly because his professional enemy, Karl Pearson, refused to let him reprint the tables Pearson had. In the 1950s, Fisher rejected the idea of a conventional significance level: "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; rather he gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1956, p. 42). He then recommended reporting the exact level of significance instead (e.g., $p = .03$, but not $p < .05$).

In my opinion, statistical thinking is an art, not a mechanical procedure. Chow's view reminds me of a mechanical maxim regarding the critical ratio, the predecessor of the significance level: "A critical ratio of three, or no Ph.D."

What we need to teach our students is neither NHSTP nor any other statistical ritual. We need to teach them statistical *thinking:* how to generate bold hypotheses, derive precise alternative predictions, set up experiments to minimize real error (rather than just to measure and insert error into the F-ratio), analyze data for each individual separately if possible rather than automatically aggregating, and perform sound descriptive statistics and exploratory data analysis. And we need to teach them that there are several important statistical schools and tools, rather than pretending that *statistics is statistics is statistics is statistics.*[2] We should give students examples of situations where each tool works and where each does not work. Students should learn why Neyman believed that null hypothesis testing can be "worse than useless" in a mathematical sense (e.g., when the power is less than alpha), and why Fisher thought that Neyman's concept of Type II error reflects a "mental confusion" between technology (such as in Stalin's 5-year plans) and science (Fisher disdained the Russian-born Neyman; see Gigerenzer 1993). We can make statistics fun and interesting by scrapping the thoughtless ritual advocated by Chow and instead teaching students about the real statisticians and controversies behind the diverse array of statistical tools we have. Choosing among these tools requires statistical thinking, not rituals.

NOTES

1. Null-hypothesis testing (*t*-test and anova) was first applied in parapsychology and education, from which it spread to basic research. Danziger (1990) offers an interesting argument for why this happened in the United States and not in Germany.

2. Chow acknowledges that there exist different logics of statistical inference. But at the same time he falls into the it's-all-the-same illusion when he asserts: "To K. Pearson, R. Fisher, J. Neyman and E. S. Pearson, NHSTP was what the empirical research method was all about." (p. xi). This statement is incorrect. Neyman and Pearson spent their careers arguing against Fisher's null hypotheses testing and developing their own alternative, which rests on two precise hypotheses (rather than one null hypothesis) and the concept of Type-II error (which Chow declares not germane to NHSTP). Furthermore, for Fisher (1955; 1956), null-hypothesis testing was only one of several useful statistical methods, such as maximum likelihood and fiducial probability (Gigerenzer et al. 1989, Ch. 3; Hacking 1965).

Stranded statistical paradigms: The last crusade

Judith Glück and Oliver Vitouch

Institute of Psychology, University of Vienna, A-1010 Vienna, Austria.
judith.glueck@univie.ac.at; oliver.vitouch@univie.ac.at

Abstract: Chow tries to show that for the case of hard-core experimentation, the criticisms of NHST are not valid. Even if one is willing to adopt his epistemological ideology, several shortcomings of NHST remain. We argue for a flexible and thoughtful application of statistical tools (including significance tests) instead of a ritualized statistical catechism that relies on the magic of α.