these new "data" are then added to the equivalent new "data" of every other subject, producing newer "data." Such a procedure seldom fails to produce normally distributed "data," suitable for NHSTP. That the occurrences of each specific action (response) of each subject might show orderliness – "lawfulness" – not suitable for NHSTP methodology is ignored, even though this is easily demonstrated by research in both "psychophysics" and "learning."

The wedding of NHSTP with cognitive science, with the blessing of "theory-construction," has been successful: count the number, since 1955, of papers given at meetings and published in refereed journals, then duly summarized in "secondary sources." Count the number of kinds of memory discovered by "operationalizing."

NHSTP has enabled research to be carried out easily; computer programs can both produce and analyze data, all but untouched by human hands – or thought. Doing such research is easier than observing, counting, and classifying. That most findings are trivial, that the theories are all but irreconcilable, that answers to most questions lie buried under ten to the *n*th bytes of "information" is becoming evident. A cognitive scientist now wonders publicly whether they've been "spinning (our) wheels" for the past thirty years or so.

Behavioral science needs data on the individual behaviors of individual organisms, each finding "verified" – replicated – by data taken from a number of other individuals, one by one. The visual methods introduced by Tufte, non-parametric "quick and dirty," and descriptive statistics (excluding means and standard deviations), suffice in testing generalizations, confirming or disconfirming them.

Four reasons why the science of psychology is still in trouble

Kim J. Vicente

Cognitive Engineering Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario M5S 3G8, Canada.
benfica@mie.utoronto.ca www.ie.utoronto.ca/IE/HF/kim/home.html

Abstract: Chow's monograph exhibits four prototypical symptoms of psychology's enduring scientific crisis: (a) it equates empirical science with statistical analysis; (b) it settles for qualitative rather than quantitative theories; (c) it ignores the role of ecological validity in the generalizability of theories; and (d) it puts rigid adherence to arbitrary but documentable rules over critical thinking about the meaning of results.

Chow's exceptionally well-written monograph shows why the science of psychology is in trouble, an opinion that has been consistently expressed by prominent psychologists over a disturbingly long period (e.g., Allport 1975; de Groot 1990; Gibson 1967/1982; Hammond et al. 1986; Loftus 1996; Meehl 1967; 1978; Neisser 1976; Newell 1973). Specifically, Chow's monograph exhibits four typical symptoms of psychology's enduring crisis.

1. Empirical science = Statistical analysis. Chow equates empirical science with statistical analysis of data from highly controlled experiments. This view is partially conveyed in the very first sentence of the book: "To conduct empirical research is to engage in an exercise which requires conceptual, theoretical and *statistical* skills" (p. ix, emphasis added; see also p. 168). This view is re-emphasized in one of the last sentences of the book: "At a minimum an empirical research is good if it has statistical conclusion validity and inductive conclusion validity" (p. 187). Both of these opinions would come as a surprise to Nobel Laureate Konrad Lorenz (1973), who not only did not exhibit statistical skill in his research, but never even published a paper with a graph in it! Instead, Lorenz devoted his life to describing and identifying phenomena as they occurred in nature. His rationale was that naturalistic observation is a legitimate form of empirical science which should precede formalization, quantification, and controlled experimentation. Before one can meaningfully formalize, quantify, or experiment, one should identify a natural phenomenon that is worthwhile investigating in more detail, and categorize the dimensions of that phenomenon to know what should be manipulated experimentally. Lorenz's rich, multi-faceted view of empirical science contrasts with the impoverished, unidimensional view that Chow refers to as "*the* scientific procedure" (p. 42, emphasis added).

Perhaps because of a preoccupation with statistical analysis, Chow undervalues the role of theory and intuition in science. For example, he states that psychology is an "empirical discipline" (p. 164), that theories are speculative (pp. 46, 61), and that the term "theory" has grandiose connotations (p. 46). Holton (1988), a well-known historian of science, has observed and criticized this narrow view of science: "The younger sciences . . . are now (erroneously, in my opinion) trying to emulate the older physical sciences by restricting their area of investigation, even if artificially, to . . . phenomenic (empirical) and analytical statements" (p. 3). But as Holton's rich case studies illustrate, science is much more than this, encompassing naturalistic observation, qualitative description and categorization, inductive leaps of faith, and axioms that can never be empirically tested. Again, Chow's view of science is comparatively narrow, a fact that has important implications (see below).

2. Settling for low-hanging fruit. Chow also claims that "the exact magnitude of the effect plays no role in the rationale of theory corroboration" (p. 96). This is certainly true, if the intent is to capture most existing theorizing in psychology. However, it overlooks the fact that there are a continuum of theories, allowing us to make predictions about the impact of the independent variables on dependent variables with increasing specificity. Areas on this continuum include: categorical, ordinal, interval, and point predictions. As the preceding quote makes clear, Chow believes that these last two categories play no role in theory corroboration. But surely a theory that makes more specific predictions (e.g., Simonton 1997) is a more mature theory, and one that we should be seeking (Meehl 1978)? By overlooking this point, Chow is settling for "low-hanging fruit" rather than striving to develop more sophisticated and powerful theories.

3. A science of the laboratory. On the one hand, Chow claims that psychologists take the external validity of experiments very seriously (p. 92). On the other hand, he follows Ebbinghaus's (1885/1964) legacy, stating that ecological validity is detrimental to the validity of theory-corroboration experiments (pp. 102, 171). It is difficult to reconcile these two statements. The fact is that external validity has *not* been taken very seriously by many psychologists because of an effort to keep the experimental setting "pure" and cleansed of the rich details that characterize ecologically valid settings (e.g., Banaji & Crowder 1991). As a result, many psychological results do not generalize beyond the experimental laboratory. As Neisser (1976) puts it, "the artificial situation created for an experiment may differ from the everyday world in crucial ways. When this is so, the results may be irrelevant to the phenomena that one would really like to explain" (p. 33). And as Gibson (1967/1982) observed, "when a science does not usefully apply to practical problems there is something seriously wrong with the theory of the science" (p. 18).

4. Cargo-cult science. Finally, Chow also holds "objectivity," "rigor," and "integrity" as the holy grails of scientific criteria. Perhaps the most extreme example of this attitude is the claim that treating $p = 0.048$ and $p = 0.052$ differently is "doing the right thing" (p. 97). The fact that there is no ontological basis for doing so (Rosnow & Rosenthal 1989) is given secondary importance (see also n. 3 on p. 118). Chow prefers "rigid adherence" (p. 97) to arbitrary but documentable rules over thinking critically about the meaning of results if the latter involves criteria that are not completely objective. In doing so, Chow (p. 168) overlooks the fact that science inevitably involves making decisions that involve intuition, aesthetics, and subjective preferences (Holton 1988).

The cause of this attitude may lie in the "physics-envy" that has plagued psychology since its inception. The result is cargo-cult

science – research that seems to follow the norms of scientific investigation, but that nevertheless misses something essential (Feynman 1985). The shackles imposed by such a restricted view of rigorous science may explain, for example, why the first 100 years of memory research largely served to confirm what the average middle-class third-grader already knows about human memory (Kreutzer et al. 1975).

Conclusion. If we follow Chow's logical arguments, we will continue to have a science of psychology that holds a narrow view of science, that only seeks weak qualitative theories, that has little to say about activities outside of the laboratory, and that strives so hard to look like "real science" that it puts itself into an intellectual straitjacket. After over a 100 years of experience, we should know better than to repeat the errors of our predecessors.

ACKNOWLEDGMENT

The writing of this commentary was sponsored by research grants from the Natural Sciences and Engineering Research Council of Canada.

Statistics without probability: Significance testing as typicality and exchangeability in data analysis

John R. Vokey

Department of Psychology and Neuroscience, University of Lethbridge, Lethbridge, Alberta, Canada T1K 3M4. **vokey@uleth.ca**
www.uleth.ca/~vokey

Abstract: Statistical significance is almost universally equated with the attribution to some population of nonchance influences as the source of structure in the data. But statistical significance can be divorced from both parameter estimation and probability as, instead, a statement about the atypicality or lack of exchangeability over some distinction of the data relative to some set. From this perspective, the criticisms of significance tests evaporate.

Chow (1996) equates statistical significance with the rejection of "chance influences" as an explanation for patterning or structure in the data, such a rejection then serving a very limited but important role as inductive evidence (in the form of corroborating an experimental implication) in a hierarchical, logical argument in support of a to-be-corroborated theory. He argues cogently that by correctly recognising the different levels of this logical argument, and the position of statistical significance within it, many of the criticisms of significance tests and equally the proposed remedies offered can be seen to be either irrelevant or misplaced. This laudable, point-by-point deconstruction and refutation of critics' arguments is in line with, but exceeds in depth those of other recent defenses of null hypothesis testing (Frick 1996; Greenwald et al. 1996; Hagen 1997; Macdonald 1997), as well as earlier attempts of his own (e.g., Chow 1988). As far as these arguments go, I am in thorough agreement. But they don't go far enough.

With very few exceptions (e.g., May et al. 1990), significance tests are routinely presented in textbooks as probability-based, binary statements about population parameters. Null hypotheses are stated in terms of population parameters (e.g., $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$), and the principal result of a significance test is the conditional probability of the data, given the null hypothesis, $p(\text{data}|\text{null})$ (which includes random sampling from a specified population distribution, typically normal). Critics and defenders alike, Chow included, appear to accept this representation as canonical, generally in the example form of a *t*-test for independent samples. It would appear that this conflation of significance testing with random sampling from populations and, hence, parameter estimation and probabilistic inference is responsible for much of the debate about significance testing.

If significance testing is accepted as synonymous with this one representation, then the criticisms of significance testing outlined by Chow, such as that the null can never be true, that what is really desired is the inverse probability of the null given the data rather than the data given the null, that confidence intervals around estimates of population parameters are preferable to a simple binary decision about them, and that the Bayesian or subjective probabilistic approach is preferable to the frequentist can appear reasonable, even conclusive. But the logic of significance testing, especially for theory-corroborative research, does not *require* parameter estimation and random sampling, as attested to, inter alia, by randomisation testing and other nonparametric (e.g., rank based) permutation tests (e.g., Edgington 1966; 1995; Fisher 1935; Hunter & May 1993; Kempthorne 1955; Pitman 1937a; 1937b; 1937c), and by the fact that for theory-corroborative research – the focus of Chow's exposition – random-sampling and inferences about extant populations are not necessary, may often be undesirable (e.g., Eysenck 1975; Mook 1983), and are frequently unobtainable (or at least intentionally unobtained) in behavioural research (e.g., Hahn & Meeker 1993). These approaches are not merely approximations to parametric tests, even though they are often presented as such; they represent a fundamentally different conceptualisation of statistical inference (e.g., Camilli 1990). Clearly, if there is no random sampling then there can be no estimates of and inferences about population parameters; the null hypotheses in these cases refer to *effects* in particular samples with the statistical validity provided by an assumption (or act) of random assignment rather than random sampling.

The idea can be taken further to eliminate also the dependency on probabilistic (random) considerations, eliminating the probability-based criticisms of statistical significance in the process. Rouanet et al. (1986) discuss significance testing as the assessment of what they refer to as the *typicality* of the data relative to some specified *set* with regard to some aspect or measurement. Although the usual probability calculus is used, no probabilistic considerations are involved. Instead, the 0–1 range of probabilities is used as a scale of typicality, and the resulting "*p*-values" are not taken as probabilities, but as the proportion of the different groups of observations from the specified set that are at least as extreme as the obtained one.

For example, with the canonical *t*-test in mind, consider two groups of 4 observations each. The result that the four highest scores all fall in one group is "significant" in that of the set of all possible ways of dividing 8 scores into two groups of 4, less than 5% (1 out of 70) would be this extreme. That is, the result is *atypical* of the specified set, and remains so regardless of the basis of forming the groups in the first place (i.e., whether or not any random process was involved). If, for example, the 8 scores were the heights of 4 adult males and 4 adult females, the result of the four tallest falling in the male group is still atypical of the permutations of the set, however highly probable or expected the result is. The same would be true if the result were coded as, say, a mean difference, a *t*-statistic, or an *F*-ratio; and of the 70 permutations of the scores, less than 5% of the statistics of the set were as extreme as the observed one.

The reference set need not be restricted to values directly observed. The atypicality or "significance" of a result or collection of scores can be computed relative to any set, including infinite sets. In these cases, the distribution of the set statistics would be obtained from the more traditional sampling distributions (e.g., normal, χ-square, etc.) for the determination of the *p*-values. With the appropriate assumptions (e.g., random-sampling or randomisation), any of these *p*-values could be converted to probabilities, but it is not obvious what would be gained by such a move, unless the assumptions were true.

Similar arguments have been advanced by Draper et al. (1993), for whom significance testing is seen as an assessment of exchangeability (i.e., can the scores be seen to be exchangeable or equivalent over some distinction, e.g., sex, with respect to some aspect, e.g., height, relative to some set, e.g., the set observed, or some larger set?). From these perspectives, statistical significance is not about the rejection of "chance influences," but rather simply a statement about the presence or absence of structure (lack of