CONTEMPORARY PSYCHOLOGY, 1991, VOL. 36, NO. 2 (PP. 102-105)
NOTE: This reprint created by the G.Loftus. Page numbers do not correspond to those in the original article.

On the Tyranny of Hypothesis Testing in the Social Sciences

Gerd Gigerenzer, Zeno Swijink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Krüger, **The Empire of Chance: How Probability Changed Science and Everyday Life.** Cambridge England: Cambridge University Press, 1989. 340 pp. IBSN 0-521-33115-3. $44.50

Review by:
Geoffrey R. Loftus

Gerd Gigerenzer, professor of psychology at the University of Salzburg (Austria), is coeditor, with L. Krüger, L.J. Daston, and M. Heidelberger, of The Probabilistic Revolution, Vol. 1: Ideas in History. •*Zeno Swijtink is assistant professor f history and philosophy of science at Indiana University Bloomington.* •*Theodore Porter is associate professor of history at the University of Virginia (Charlottesville) and author of* The Rise of Statistical Thinking. •*Lorraine Daston is professor of the history of science at the University of Göttingen (Germany) and author of* Classical Probability in the Enlightenment. •*John Beatty is associate professor of the history of science and technology at the University of Minnesota (Minneapolis).* •*Lorenz Krüger is professor of philosophy at the University of Göttingen (Germany).* •*Geoffrey R. Loftus, professor of psychology at the University of Washington (Seattle), is coauthor, with E.F. Loftus of* Essence of Statistics (2nd ed.).

The Empire of Chance is about the history and current use of probability theory and statistics. The book provides a broad treatment of these topics; one could, accordingly, read or review it from quite a variety of different perspectives. Because this review is for psychologists, I will organize it around the book's insights into a question that I believe is at the heart of much malaise in psychological research: how has the virtually barren technique of hypothesis testing come to assume such importance in the process by which we arrive at our conclusions from our data?

In what follows, I first describe why this question is timely and important. I then provide a brief synopsis of the book. And finally, I detail the book's answers to the question.

The Ascent of Hypothesis Testing

Since the 1940s, the practice of hypothesis testing has been seeping into all nooks and crannies of social-science methodology. Today, hypothesis testing constitutes the major foundation of data analysis in experimental psychology: it is used to justify conclusions from data in over 90% of articles in major psychology journals. Gigerenzer *et al.* underscore hypothesis testing's importance again and again - perhaps most dramatically when they paraphrase the editorial edicts issued by Arthur Melton (1962) upon assuming editorship of the august *Journal of Experimental Psychology.*

"Melton's message was, in short, that manuscripts that did not reject the null hypothesis were almost never published, and that results significant only at the 0.05 level were barely acceptable, whereas those significant at the 0.01 level deserved a place in the journal. Psychology students could no longer avoid statistics, and the experimenter who hoped to publish could no longer avoid a test of significance." (p. 206)

The Molding of Young Minds

We psychologists all learned about hypothesis testing during our undergraduate days. Many of us remember thinking at the time that it seemed kind of backward and perverse. You establish a null hypothesis, usually something like "the population means are the same in all the experimental conditions," or "the population correlation is zero." Some statistic like an *F* or a *t* or a *z* then provides evidence for or against the null hypothesis's plausibility. Depending on some arbitrary value of the statistic's magnitude, you either reject or fail to reject the null hypothesis.

No one ever seemed to know exactly what hypothesis testing could tell you that was at all interesting or important. Many budding psychologists (perhaps in wishful desperation) came to believe in a variety of murky and generally incorrect implications of the process (for example, that the p value's magnitude tells us something about an effect's magnitude or an effect's replicability or the probability that the null or alternative hypothesis is true or false). Somewhere along the line, however, we all internalized one lesson that is entirely correct: the more you reject the null hypothesis, the more likely it is that you'll get tenure.

Deficits of Hypothesis Testing

Despite the stranglehold that hypothesis testing has on experimental psychology, I find it difficult to imagine a less insightful means of transiting from data to conclusions. A list (by no means exhaustive) of its more glaring deficits are as follows.

What's the point of rejecting a hypothesis you know is false to begin with? One is hard-pressed to think of a situation in which a null hypothesis might plausibly be true. Consider a typical experiment in which, say, one is examining the difference between two clinical treatments. One group of people is given Treatment A and the other is given Treatment B. The null hypothesis is that the population mean outcome measures are exactly the same for the two treatments.

No one would seriously consider this hypothesis to be literally true. So the results of a hypothesis test can only tell you is whether you have sufficient experimental power to detect the presence of whatever treatment effect must inevitably exist. Yet the conclusions from the experiment (and its suitability for publication) rest entirely on whether one is able to reject the null hypothesis. It's bizarre. Gigerenzer *et al.* provide an apt summary from Nunnally (1960): "if rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data." (p. 210)

In fairness I should issue two caveats. First, some writers, (e.g., Hays, 1973) suggest the use of a "null range" rather than an exact null hypothesis. However a null range (1) is rarely suggested (Hays himself tucks it away in an end-of-the-book Bayesian-methods section), (2) is even more rarely (if ever) used in practice, and (3) is more a kind of contorted interval-estimation technique than a bona-fide hypothesis-testing technique.

The second caveat is that, while difficult, it is not *impossible* to concoct experimental situations in which a null hypothesis may be plausible and rejection of the null hypothesis interesting (attempts to demonstrate the existence of parapsychological phenomena provide examples). However, the vast majority of social-science experiments are of the Treatment A/Treatment B variety rather than of the plausible-null-hypothesis variety.

Where are the error bars in social-science journals? The emphasis on hypothesis testing produces a concomitant *de emphasis* on an alternative technique for coping with statistical error that is simple, direct, intuitive, and has wide acceptance in the natural sciences: the use of confidence intervals. Whereas hypothesis testing emphasizes a very narrow question (do the population means fail to conform to specific pattern?) the use of confidence intervals emphasizes a much broader question (what *are* the population means?). Knowing what the means are, of course, *implies* knowing whether they fail to conform to a specific pattern, although the reverse is not true. In this sense, use of confidence intervals subsumes the process of hypothesis testing.

The complications of experimental power. The emphasis on rejecting the null hypothesis also produces a de emphasis on experimental power (e.g., Guilford's widely read, *Fundamental Statistics in Psychology and Education* declared power, as late as 1956, to be "too complicated to discuss.") Things have improved in the intervening 35 years, but only marginally so. When power *is* discussed, it is typically discussed within the context of hypothesis testing. The focus, accordingly, is on the probability of a Type-II error which is fairly meaningless, given (1) the almost universal lack of quantitative alternative hypotheses and (2) the implausibility of the null hypothesis to begin with. (All readers whose students can't seem to understand power when it's discussed in this way, please raise your hands. Ha! No wonder Guilford gave up). A more enlightening exposition of power would relate it to confidence intervals: the more power you have, the smaller are your confidence intervals, i.e., the better your knowledge of where population means are.

Biases against the null hypothesis. There is a profound asymmetry between conclusions issuing from rejecting the null hypothesis on the one hand, and failing to reject it on the other (see Greenwald, 1975 for a lengthy discussion of this bias). Somewhat ironically, the main reason for this asymmetry is that, as noted, accepting the null hypothesis is almost always a guaranteed error. However, if there is sufficient experimental power, then failing to reject a null hypothesis can be interesting anyway, as it implies that some specific model can account for the data reasonably well. Although there is a small tradition within psychology of conclusion-by-root-mean-square-error, this tradition is overwhelmed by the alternative of conclusion-by-p-value.

Off-the-shelf assumptions. Finally the process of hypothesis testing forces (or at least nudges) theoretical psychology into a fairy-tale land wherein the assumptions and traditions of hypothesis testing (usually parametric hypothesis testing) are all correct. Variances are unaffected by treatments, variables are distributed normally, monotonic relationships are linear relationships, and all results are reduced to binary statements about the presence or absence of main and interaction effects within a linear model.

It isn't actually the validity of these assumptions that I worry about. It's well known that most relevant sampling distributions are robust against most assumption violations (and, of course, there are always non-parametric tests). Rather, the problem is that our early (and continuing) imprinting on these assumptions and traditions engenders strong biases against formulating theories incorporating other, perhaps more interesting and realistic assumptions. Thus, psychological theory becomes generic analysis-of-variance theory and the potential for insight is lost. To paraphrase Freedman, Rothenberg, and Sutch (1983), who leveled analogous complaints about standard regression analysis in econometric theory, it seems that off-the-shelf assumptions produce off-the-shelf conclusions.

A Synopsis of the Book

The point of this lengthy prologue has been to argue that, given its rather severe shortcomings, the question of "why hypothesis testing" is a baffling one. What light do Gigerenzer *et al.* have to shed on this question? Let me start the answer by providing a brief sketch of the book itself.

Historical Origins

The Empire of Chance accomplishes two goals that, roughly, constitute the first and second halves of the book. The first goal is to provide a history of the twin disciplines of probability and statistical theory. The saga begins at the beginning with Pascal in the 17th century. Then, driven by the triple influences of economics (on what basis should we set insurance rates?), science (how are we supposed to deal with variability in our data?), and philosophy (what is *meant* by chance, anyway?), it wends its way to modern times when, in the words of the penultimate chapter title, "Numbers rule the world."

The authors emphasize controversy. Right from the start, they point out, different thinkers and practitioners of probability and statistics had rather different ways of viewing both fundamental concepts and their applications to practical issues. These differences never seem to show up in the social sciences either in textbooks or in lectures. I had never, for example, quite realized how much Ronald Fisher and Egon Pearson disdained each other's viewpoints. The descriptions of their differences are instructive for the practitioner because they underscore two very different ways of applying probability theory to statistical issues that have somehow become merged and misbegotten.

(The descriptions are also entertaining as gossip).

Current Practice

The book's second goal is to illustrate applications of probability and statistics in four fields: biology, physics, psychology, and "the real world" (e.g., sports). In each of these instances two distinct themes emerge. The first carries over the book's historical motif, dealing with the question: how did the current usage of probability and statistics emerge over time? The second theme involves the interplay of data-analysis techniques on the one hand and theoretical development on the other. Within psychology, for example, a prominent view arising from the statistical tradition, that of "mind as statistician," permeates subfields as diverse as signal-detection theory and causal reasoning.

On Hypothesis Testing

The Empire of Chance is a wide-ranging book. Its insight about the ascent of hypothesis testing constitutes but one of several themes of interest to scientists in general and psychologists in particular. But to me it is the most interesting theme and its inclusion is the most important reason that psychologists should read the book. As described by Gigerenzer *et al.*, there are two major reasons why hypothesis testing has come to enjoy its present stately position.

Miscast Objectivity

First, for mid-20th-century experimental psychologists (status hungry, perhaps, amidst their natural-science colleagues) hypothesis testing provided the illusion of endowing psychological data - which are intrinsically complicated, messy, multidimensional, and subjective - with a seductive simplicity and objectivity. Banished was the slovenly panoply of "eyeballing curves, personal judgment, description without inference, or bargaining with the reader" to be replaced by one clean, simple, refreshing rule: if p < .05 (or so) an effect is real; otherwise, it's not. In the domain of pure research, this switch eased the decision process for journal editors, while in the domain of practical applications (particularly educational and military applications) it provided researchers with a new concept, easily explainable to policy makers, that could be used to justify (or denounce) the implementation of novel techniques.

Statistical Newspeak

The second reason was that legions of "Statistics for the Social Sciences" textbook writers simply ignored history. They left enormous gaps in their teachings (such as Fisher's development of interval-estimation techniques or his long-forgotten distinction between a significant result and the demonstration of a natural phenomenon) and they bestowed upon their subject matter a consensus of its founders that was simply never there to begin with. Gigerenzer *et al.* illustrate this latter process as it applied to the controversy between Fisher (who emphasized the binary test of a single null hypothesis) and Pearson and Neyman (who viewed a statistical test as a means of choosing among a slate of candidate hypotheses). The authors note that,

> ...almost no [social science statistics] text presented Neyman and Pearson's theory as an alternative to Fisher's, still less as a competing theory. The great mass of texts tried to fuse the controversial ideas into some *hybrid* statistical theory...Of course this meant doing the impossible. But...statisticians were eager to sell, and psychologists were eager to buy *the* method of inductive inferences. The statistical texts now taught hybrid statistics, of which neither Fisher nor, to be sure, Neyman and Pearson would have approved. The type-II errors became added to null hypothesis testing (although it could not be determined in this context), Neyman and Pearson's interpretation of the level of significance as the proportion of type-I errors in the long run became mishmashed with Fisher's and so on. Whatever the textbooks taught, it was *not* indicated that some of the ideas stemmed from Fisher, others from Neyman and Pearson. The hybrid statistics was presented anonymously, as if it were the only truth, as if there existed *only one type of statistics*. There was no mention of the existence of a deep controversy, much less of the controversial issues, nor of the existence of alternative statistical theories...(p. 208)

Students and practitioners of statistics were accordingly left with the impression that "statistics is statistics" and all that is really necessary is to learn the rules - which leaves very little incentive or historical precedent for innovation and creativity. This remarkable state of affairs is analogous to engineers teaching (and believing) that light consists only of waves, while ignoring its particle characteristics - and losing in the process, of course, any motivation to pursue the most interesting puzzles and paradoxes in the field.

Conclusions

The negative tone that I've adopted in this review shouldn't be construed to reflect an opposition to the teaching and use of quantitative methods. Social science research is, perforce, cursed with more than its share of statistical error and inaccessible measures, which have to be dealt with somehow. A thorough knowledge of (at the very least) probability theory, measurement, and descriptive statistics is a critical component of any social scientist's methodological arsenal.

The Empire of Chance provides an historical context that constitutes a refreshing counterpoint to the tedious catechism delivered by the vast majority of social-science statistics textbooks that have appeared over the past fifty years. It is not a textbook itself, but it includes most of the background information that any social scientist should have in order to realize that probability theory and statistics is far more than just a static collection of dry equations and inviolate rules.

References

Freedman, D.A., Rothenberg, T. and Sutch, R. (1983). On Energy Policy Models. *Journal of Business and Economic Statistics, 1*, 24-36.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 83*, 1-20.

Guilford, J.P. (1942). *Fundamental Statistics in Psychology and Education*. First Edition, 1942; third edition, 1956. New York: McGraw-Hill.

Hays, W. (1973). *Statistics for the Social Sciences (second edition)*. New York: Holt.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*, 641-650.

A Hypothesis Test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. Whenever we want to make claims about the distribution of data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests. There are two possible outcomes: if the result confirms the hypothesis, then youâ€™ve made a measurement.Â By now we understand that the entire hypothesis testing works on based on the sample that is at hand. We may come to a different conclusion if the sample is changed. There are two types of errors that relate to incorrect conclusions about the null hypothesis. 7. (a). Type-I Error