FAMILIAR OLD WINE: GREAT NEW BOTTLE

American Journal of Psychology, 1997, 146-153

Geoffrey R. Loftus and Jennifer E. McLean
University of Washington

Send Correspondence to:

Jennifer E. McLean
Department of Psychology
University of Washington
Seattle, WA 98195

email: mclean@u.washington.edu

Statistics is a subject that is often difficult and frightening for many students of the social sciences. The awareness of this phenomenon will inspire a concerned and conscientious statistics teacher to categorize common mistakes and misunderstandings and to figure out those presentation techniques that best facilitate the students' learning. Over many years, spent both teaching and observing the ongoing scientific practice, such an individual will accumulate a great deal of experience and will have collected together bits of wit, wisdom, and advice in short, a useful bag of tricks. In *Statistics as Principled Argument*, one learns what's inside Abelson's bag and can profit from his experience.

Abelson's tips and tricks are unified under his theme, articulated at the outset (p. xiii) that "the purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric." He assumes the reader to have some exposure to and practice with elementary statistics. He intends this book to help refine and enliven the understanding of statistics and to move the reader towards his image of the ideal statistician who, he asserts, has the talent for persuasive argument of a good and honest lawyer, the appealing narrative style of a good storyteller, and the logic, openness, perceptiveness, and meticulousness of a good detective. Considering statistics in this three-pronged fashion makes it more fun. Abelson seeks to revive statistical rhetoric from the lifeless, dogmatic, and ritualistic practice that it has become by suggesting a focus on the spirit, rather than just the letter, of the law.

While this appealing book has many virtures, we find ourselves in a quandary in reviewing it. A major focus in this book is on how to best make use of null hypothesis testing to understand a set of data—a technique that is relied on by many, but, for reasons that we will describe in a later section, one that we believe is fundamentally flawed. Our problem is that, while we do not want to ignore the book's obvious virtues, we also do not want to conclude by tacitly endorsing null hypothesis testing as the primary means of understanding data. Our solution is to write a trifurcated review that includes the following:

1. Given that one buys into statistical hypothesis testing as a fundamental and useful data-interpretation technique, the book is wonderful. It is difficult to imagine a better vehicle, either for rookie students or for seasoned pros to gain a variety of valuable insights into the subtle workings and common pitfalls of hypothesis testing.

2. Even if one were not interested in hypothesis testing per se, the book would be valuable for its insightful expositions of design-related issues such as randomization and counterbalancing.

3. We believe that hypothesis testing is approximately as useful as newspaper horoscopes as a technique for interpreting a set of data. In the final section of this review, we will briefly articulate the reasons for this belief.

On the Book's Many Virtues

Statistics as Principled Argument is well-written and enjoyable to read. It is witty without being unclear. It can be read alone or as a companion to a standard statistics text, then kept as a useful reference.

Throughout the book, Abelson makes excellent citations of other works. He presents the issues and gives sufficient information and perspective to make them interesting and informative, but also points the interested reader to seminal papers on the subject. In this way, the reader may benefit from his experience which has made him a knowledgeable guide through the immensity of literature that is available.

As we suggested earlier, we found the book worth reading for two categories of reasons. First, although we are dubious about the framework of null hypothesis testing, Abelson provides an intuitive understanding of it along with useful advice for making the best of it. Second, Abelson offers a good deal of wisdom on a collection of topics that are independent of framework.

Making the Best of Hypothesis Testing

Given that hypothesis testing is pervasive in the social sciences, *Statistics as Principled Argument* is useful to read for the purpose of developing a better understanding of the procedure. If one wants to use hypothesis testing, one is then better positioned to use it correctly. Conversely, if one wants to criticize hypothesis testing, it behooves one to first understand it thoroughly. Some of the high points of the book's overview of hypothesis testing are the following.

On the logic of hypothesis testing

Abelson discusses the "language and limitations of null hypothesis tests." Here he acknowledges that the there are many critics of this form of data analysis, not the least of whom are the students themselves, who (quite understandably) find its procedures to be counterintuitive and confusing. He attempts to bestow some clarity on the practice by explaining that the procedure arose from a rhetorical consideration, that of the investigator preparing in advance a counter-counter argument to an imagined critic's counter to a claim that an experimental factor has some effect. (That such an explanation indeed provides some clarity is more a testament to the inherently convoluted logic of hypothesis testing than it is to Abelson's inherent skill as an explanation provider).

Continuous versus dichotomous decision making

Abelson warns the reader not to treat null hypothesis testing as a two-valued decision process, using semantically strong terms, such as "accept" or "reject" the null hypothesis. Rather, he encourages the use of statistical tests as aids to wise judgment. For example, a p-value of .06, while not significant according to the p=.05 convention, still suggests a trend in the data that warrants attention and further investigation. He echos Tukey's (1991) suggestion to use in these circumstances "shades of wording to indicate different degrees of uncertainty" (p. 74), but his best advice is to do more research. This point constitutes a nice summary of the argument that hypothesis testing overlays the illusion of precision and objectivity on the intrinsically imprecise and subjective practice of data interpretation (cf. Loftus, 1991).

The replication fallacy

In his *Styles of Rhetoric* chapter, Abelson alerts the reader of the tendency to maintain an overconfidence in the repeatability of statistically significant results. An analysis of this *replication fallacy* by Greenwald, Gonzalez, Harris, and Guthrie (1993) shows that there is only an 80% chance of a replication yielding $p < .05$, when the original two-group study had a $p < .005$. Indeed, Abelson consistently emphasizes the importance of replication, arguing that a single null hypothesis test is never enough to establish a scientific fact, but that solid research conclusions instead arise from a cumulative process of replication.

Replication and generalizability

The replication theme continues in the *Generality of Effects* chapter, wherein Abelson discusses what types of replication studies are optimally useful and efficient. For example, he notes, a series of interrelated studies may be more useful than a set of exact replications. He addresses the issue of how much change of context is required for the replicated results to be generalizable, and to what extent. He discusses the potential constraints on generality such as the possibility that universals are rare and that there are limits to how well a laboratory situation can be compared to real life or does perhaps life imitate science imitating life.

Demystifying fixed versus random effects

In the same chapter, Abelson brings up the issue of treating contexts as a fixed versus random effect. Here he clearly explains that the assumptions of the usual two-way analysis of variance that uses the MS(within groups) as the denominator of the F statistic are that the levels of the factors are fixed and hence statistical inferences are limited to that particular set of levels. If the goal is to generalize findings to all levels of a context factor, a random effects model is required. In this case the levels must be randomly selected from the set of all levels in the population of levels to which the findings are to be generalized. Then the denominator of the F statistic is the MS(interaction). This will typically produce a smaller, and hence, perhaps nonsignificant F value for the main effect. Thus there are disadvantages to both approaches, but Abelson gives a list of suggestions for handling the dilemma, such as using many levels of the context factor which would increase the sensitivity of the test or even an exhaustive set which would be possible in the case of gender, for example. He quips, "One does not deserve a general result just by wishing it. ...There is no free hunch." (page 142).

Testing for specific patterns

In the *Articulation of Results* sixth chapter, Abelson discusses the F-testing of contrasts of means instead of carrying out only the profoundly uninformative omnibus null hypothesis test in a one-way analysis of variance. Testing contrasts allows for greater sensitivity to an expected pattern in the data and better articulation of the results. A significant omnibus F-test only indicates that not all means are the same; it does not provide any further articulation. He suggests that the effectiveness of using contrasts is undercut if there is no mention of the residual variation from the trend specified by the contrast.

In the *Magnitude of Effects* chapter, Abelson makes brief mention of confidence limits and acknowledges that they give the same information as a p-value, and more, because they give an estimate on the *size* of the difference between means. Others have made the same point as a foundation of a logical consequence: If one reports confidence intervals, there is little, if any, reason to carry out hypothesis testing (Grant, 1962; Loftus, 1991; Loftus & Masson, 1994).

Wisdom about framework-independent topics

Abelson discusses a variety of topics that are not specific to null hypothesis testing, but rather, are important no matter what framework is being employed. The wisdom and advice he offers on these topics alone may make the book worth reading, even for those opposed to hypothesis testing.

So what is "chance" anyway?

In the chapter, entitled *Elementary Arguments and the Role of Chance*, Abelson introduces the topics of random generation and random sampling, pointing out that the concept of chance is a slippery one, that is often misunderstood. He launches a useful and memorable (if a bit cutesy) metaphor of a "committee of leprechauns responsible for producing data on demand" (p. 18), and returns to this metaphor several times throughout the book. His explanation of chance is witty and accurate but not too technical, thereby providing the reader with an intuitive foundation while reserving the mathematical details for some other venue. Abelson states that the ideal is to find the most parsimonious explanation of a set of data. Further, he states that the goal of the statistician should be to determine if there is a need for a systematic factor or if the most parsimonious account would be one of chance alone. To test the adequacy of a pure-chance account, the statistician should look for statistical regularities in the data, determining whether those regularities conform to the results of a random process, or whether a systematic factor is needed in addition to chance.

While quite useful, this chapter could have been improved by referring to the oft-made criticism of hypothesis testing that rarely, if ever, can the results of some experiment be *completely* due to chance (i.e., rarely can a null hypothesis be literally true). There are always systematic, nonrandom factors at work in experiments. The major questions should be: What are these factors, and how big are they?

Red flag detectors

In Chapter 5, entitled *On Suspecting Fishiness*, Abelson gives useful and clear advice on how to snoop through data and statistical summaries to check the validity of the results. He demystifies this process by giving specific examples in actual data sets of red flags such as strange-looking distributions that include outliers, dips, gaps, cliffs, or peaks, or the occurrence of impossible or unusual scores or test statistics that may indicate mistakes of various kinds, and he explains how to investigate each potential problem. For example, an F-value that is extraordinarily small, indicating the means being compared are almost identical, could happen as a fluke or possibly as a result of a design flaw such as "the inadvertent balancing of a factor across experimental conditions" (p. 95).

Research is more than statistics

Abelson's final two chapters are called *Interestingness of Argument* and *Credibility of Argument*. Here he broadens the focus and highlights the narrative aspects of statistical argument. "Interestingness" is a key quality that has a huge impact on the direction of research: that is, the studies that receive the most attention and hence stimulate further research will typically be the most interesting ones. Yet, how is interestingness defined? Abelson suggests that "a statistical story is scientifically interesting when it has the potential to change what scientists believe about important causal relationships" (p.158). This includes surprisingness and counterintuitiveness as aspects of interestingness. Abelson goes on to say that importance, which depends on the number of connections to other issues, is also a key aspect of interestingness.

We believe that these are points whose importance cannot be overstated. Many sciences, most notably the social sciences have, we feel, become corrupted by the overbearing influence of statistical techniques: data analysis has somehow become, in large part, a search for significance rather than searches for patterns and tests of theories. As David Freedman once noted in connection with the science of economics, the "off-the-shelf techniques" that are the staple of current statistical methodology produce "off-the-shelf conclusions."

On the Role of Hypothesis Testing

To summarize, Abelson's book has two classes of virtues. First, it provides an excellent intuitive overview of both the central machinery and the subtle intricacies of the pervasive practice of hypothesis testing. Second, it provides a plethora of "platform-independent" wisdom and advice. As we have indicated, however, the book's central philosophy still remains squarely within the hypothesis-testing tradition. We have already asserted our general discomfort with a book that embraces this practice and, in this closing section, we briefly elaborate on the reasons for this discomfort.

The point has been made numerous times over the years that the practice of hypothesis testing is a barren and misleading technique for trying to extract information from a set of data[1]. A nonexhaustive list of some of its major deficits are these:

1. Hypothesis testing reduces the potentially rich information in a data set to an impoverished list of binary (reject/don't reject) decisions.

2. Because an exact null hypothesis (e.g., $\mu_1 = \mu_2 = ...= \mu_J$) can almost never be true, the question of whether or not it should be rejected is generally irrelevant.

3. Hypothesis testing affords the illusion of assigning a more-or-less precise probability (e.g., less than 0.05) to the ostensibly interesting proposition that the null hypothesis is true given the obtained data. In fact, of course, such a probability is assigned to the opposite, and considerably less interesting proposition: that the obtained data would be obtained given that the null hypothesis is true.

4. The slavish attention paid to the probability of a Type-I error diverts attention and interest from the issues of Type II errors and statistical power.

5. There is an almost irresistible urge to classify non-significant effects as non-existent effects which infuses considerable confusion into the scientific enterprise. This problem is, of course, especially acute under low-power conditions.

6. An emphasis on hypothesis testing entails a corresponding de-emphasis on trying to understand the quantitative functions that relate variables.

[1] See, for example: Bakan, 1966; Carver, 1978; Chow, 1988; Cohen, 1990; 1994; Gigerenzer, Swijtink, Porter, Daston, Beatty, & Kruger, 1989; Grant, 1962; Greenwald, Gonzalez, Harris, & Guthrie, 1995; Loftus, 1991; 1993a, b; 1995; Loftus & Masson, 1994; Lykken, 1968; Meehl, 1990; Rosenthal & Rubin, 1985; Rozeboom, 1960; Schmidt, 1994. Abelson himself admits early in the book that "Truth to tell, however, the information yielded from null hypothesis tests is ordinarily quite modest, because all one carries away is a possibly misleading accept-reject decision. Furthermore, the categorical mode of thinking encouraged by significance tests can lead to misinterpretations of comparative results.

In summary, it has been the contention of many observers that the overwhelming emphasis on hypothesis testing has severely impeded scientific progress in a variety of scientific endeavors—and, accordingly, that nothing short of a revolution will be required to escape from the methodological cul-de-sac into which the practice of hypothesis testing has led us. Such a revolution could be realized in various forms: different emphases in textbooks and classrooms (e.g., Lockhart, 1997), different editorial policies (e.g., Loftus, 1993b), and in a steady drumbeat of articles and talks designed to underscore the serious shortcomings of hypothesis testing. Such a revolution could be impeded, however, by a beautifully designed book that, while certainly criticizing many aspects of hypothesis testing, also, when all is said and done, positions itself squarely within the hypothesis testing framework.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Carver, R.P. (1978). The case against statistical significance testing. Harvard *Educational Review, 48*, 378-399.

Chow, S.I. (1988). Significance test or effect size? *Psychological Bulletin, 103*, 105-110.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994).The Earth Is Round (p < .05), *American Psychologist.*

Gigerenzer, G., Swijtink, Z, Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989).*The Empire of chance: How probability changed science and everyday life*. Cambridge England: Cambridge University Press.

Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1995) Effect Sizes and p-values: What should be reported and what should be replicated? *Psychophysiology* (in press).

Lockhart, R.S. (1997) *An introduction to statistical data analysis in the behavioral sciences.* New York: Freeman.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*, 102-105.

Loftus, G.R. (1993). A picture is worth a thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers, 25*, 250-256 (a).

Loftus, G.R. (1993). Editorial Comment. *Memory & Cognition, 21*, 1-3 (b).

Loftus, G.R. (1995). Data analysis as insight. *Behavior Research Methods, Instrumentation and Computers, 27*, 57-59.

Loftus, G.R. and Masson, M.E.J. (1994) Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review, 1*, 476-490.

Lykken, D.T. (1968). Statistical significance in psychological research. *Pschological Bulletin, 70*, 131-139.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, Monograph Supplement 1-V66.*

Rosenthal, R. & Rubin, D.B. (1985). Statistical Analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97*, 527-529.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Schmidt, Frank (1994). Data analysis methods and cumulative knowledge in Psychology: Implications for the training of researchers. APA (Division 5) Presidential Address.

Tukey J.W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116