Statistical Science
1997, Vol. 12, No. 1, 48–52

Testing Statistical Hypotheses: The Story of a Book

E. L. Lehmann

Abstract. This is an account of the life of the author's book *Testing Statistical Hypotheses*, its genesis, philosophy, reception and publishing history. There is also some discussion of the position of hypothesis testing and the Neyman–Pearson theory in the wider context of statistical methodology and theory.

Key words and phrases: Testing statistical hypotheses, Neyman–Pearson theory, textbook publishing.

1. INTRODUCTION

As author of several textbooks, I am often asked about various aspects of book writing—from the severely practical to the quite personal. These questions suggest a certain curiosity about the writing and publishing of such texts; the following is the account of the life history of one of them.

The book I shall write about, *Testing Statistical Hypotheses* (*TSH*), was my first and has now been in print for almost 40 years. Before describing its history, let me get one question out of the way which is frequently hinted at although less often asked outright: "What was the financial reward, how much money did you make from the book?" I am not being coy in saying that I do not really know since I have not kept adequate records. However, I have a rough idea in terms of a scale suggested by the number theorist Edmund Landau, who stated that from the payments for one of his books he built his house in Göttingen. I can say categorically that from the earnings of *TSH* I could not have built a house, not even a modest one, much less Landau's mansion in Göttingen. On the other hand, the proceeds would have enabled me to buy a car—in fact, a somewhat fancier one than the compact we drive. These upper and lower bounds are a bit crudė, but then mathematical bounds often are.

TSH gives an account of the small-sample theory and methodology of hypothesis testing. The idea of a small-sample approach based on an assumed model was introduced by Student in 1908 (Student, 1908) and greatly extended by R. A. Fisher in the 1920s. Fisher's 1925 book, *Statistical Methods for Research Workers*, brought the new methodology to the attention of users of statistics. A theoretical framework for testing was formulated by Neyman and Pearson (1933), who showed that the Student–Fisher tests had some optimal properties within the normal models on which they were based.

When in 1942 I became a student of Neyman, I learned this theory (with the basic ingredients of level, power, similarity, unbiasedness and optimality) at the source, and wrote my thesis (1946) on its application to a special case suggested to me by P. L. Hsu.

With this background I was very receptive to a second group of ideas, which had been strongly influenced by Neyman's approach to statistics. These were the concepts of minimaxity and admissibility of Wald's decision theory and the related invariance considerations which I learned from Charles Stein when in 1947 he became a colleague, friend and collaborator.

2. THE BLYTH NOTES

These two sets of ideas easily combined into an integrated whole which formed the basis of the graduate course on hypothesis testing which I gave in 1948. Among the students attending it was Colin Blyth, to whom the testing book to a large extent owes its existence. In a letter to me of April 1981 he describes how this came about:

> I attended your course 260A in Testing in the Fall 1948 and wrote up careful notes. (After each lecture I went home and wrote up the rough notes I'd taken

E. L. Lehmann is a senior scholar, Educational Testing Service, 15-T Rosedale Road, Princeton, New Jersey 08541, and Professor Emeritus, Department of Statistics, University of California, Berkeley, California 94720.

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_82,

> in class, putting them into readable form, and settling to my own satisfaction any details omitted or points I hadn't been clear about in class.) Sometime in the Spring of 1949 you saw these notes (I don't remember how this happened, and can think of no reason that it might have) and suggested mimeographing them for other students. I preferred to attend your course again in the summer of 1949 and put them in a form more suitable for distribution and this we did. I wrote them up, you read them carefully and made numerous suggestions, and after a final revision they went to the typist.

The notes were mimeographed and sold at cost, first by the Berkeley Statistical Laboratory and later by the University Bookstore. As our students graduated and took up positions at other universities, they recommended them or used them in their courses. As a result, orders started arriving from other colleges. In the absence of any other systematic treatment of this still fairly novel material, the notes gradually became something of an underground text. They formed a slight paper-covered package of 163 pages, which provided a skeletal account of the theory and contained only the simplest applications.

3. THE CONCEPT OF THE BOOK

An increasing demand, and references to the material in books by colleagues, gradually led to the idea of expanding the notes into a book. In the intervening years I had occasionally been teaching a graduate course along the lines of the notes, and came across questions concerning the theory which I tried to answer in my research. In particular, this included the joint work with Henry Scheffé on the completeness of sufficient statistics. In the course of this work I also became familiar with the theory of exponential families, and this provided an important unifying idea.

I mentioned earlier that the notes combined threads from two different lines of work: the Neyman–Pearson treatment of hypothesis testing and Wald's general decision theory. These provided the underlying structure. However, I did not think of the book as an account of the theory for its own sake, but rather as a means of justifying the many standard tests whose optimum properties the theory had established. Presenting these tests and their properties was my central concern. The theory makes it possible to do so in a systematic fashion. Without such a unifying structure they would form an unmotivated, disorganized collection of ad hoc procedures (see Note 1), dependent on the ability of their originators to intuit which particular test statistics would prove effective.

NOTE 1. This statement is perhaps too strong. The likelihood ratio principle, for example, does provide a systematic approach which frequently leads to the standard tests. However, it can be justified only asymptotically; there is no reason to expect the resulting tests to be satisfactory for small samples.

A question that caused me some difficulty concerned the mathematical level at which to present the material. The natural mathematical framework is provided by measure theory; however, such an advanced level threatened to erect a formidable barrier for many potential readers. This would be particularly regrettable since the measure theoretic considerations contribute little to the understanding of the statistical ideas and results. On the other hand, a clean, rigorous treatment is not possible without them.

In the end I decided to include a brief introduction to measure theory, giving the principal definitions and results, and then using them in the rest of the book where needed. Russian colleagues later told me of a surprising effect of this inclusion when a translation appeared in their country. According to these reports, it helped to legitimize statistics in the eyes of some Russian mathematicians. If it is based on measure theory, they felt, there may be something to it!

4. FROM NOTES TO BOOK

The next problem was to obtain a publisher. I am sometimes asked about this aspect, but at the time the answer was easy. The statistical textbook literature was dominated by the "Wiley Publications in Statistics," a series that contained among others such distinguished volumes as Feller's *Probability Theory*, Cochran's *Sampling Techniques*, Doob's *Stochastic Processes*, Wald's *Decision Theory*, Rao's *Advanced Statistical Methods*, Blackwell and Girshick's *Theory of Games and Statistical Decisions*, Savage's *Foundations of Statistics*, Scheffé's *Analysis of Variance* and Anderson's *Multivariate Analysis*. Furthermore, a Wiley representative regularly visited the Department and had expressed an interest in the book. Thus, it appeared in 1959 under the Wiley imprint.

The notes had been issued in 1949, the book came out 10 years later. What took it so long? I can think

of various excuses. The book was about three times as long as the notes; I was also doing other things: teaching, research, a three-year stint as Editor of the *Annals* and so on. However, the main reason can be summed up in one word: rewriting! There is no page, section or chapter that I wrote, which looking at it again a few months later I could not improve, and a few months after that improve again. This process might have gone on forever had Neyman not told me one day that I had dawdled long enough; get off the dime! So I finally took the plunge, but with great trepidation. Publication is so irreversible.

5. ERRORS

After publishing a book one waits with some anxiety for the reviews. But at least in my case another ordeal arrives much sooner: the discovery of errors, big and little, subtle and obvious—how could one have been so foolish and so careless! Over the years I have learned how to interpret letters starting with, "I have difficulty following your proof..." or "I was unable to solve problem...". These are polite ways of telling me: "You botched it again!" Only days after publication I found that one of the figures was seriously in error. And of all the figures in the book this was the one the publisher had chosen to put on the dustjacket. Most embarrassing!

The biggest source of errors was the more than 200 problems. The difficulty often resulted from some fine points or special cases that I had overlooked and that naturally caused readers much trouble when they struggled with them. Letters asking for clarification were not only painful reminders of my ineptitude but they could also take quite a bit of time and effort to answer at a time when I was no longer working in this area. I was saved from this bondage to my past errors when a heroic group of 15 Dutch statisticians decided to work through the whole collection systematically and in 1984 with Wiley's permission published the solutions as a 310-page book (Kallenberg et al., 1984). One member of the group told me later that this was the most painful job he had ever undertaken. However, from then on I was able to answer queries about the problems with a simple reference.

6. THE BOOK AS TEXT

Shortly after the book appeared, I was scheduled to teach the graduate course for which it was intended as a text. But, it was not clear to me how best to utilize it. There seemed little point in rephrasing in lectures what I had labored hard to express as clearly as possible in print, and so I announced that I would proceed differently than in the past. I would outline and highlight a section in class; the students were then to read it; and I would answer questions at the following class meeting. If that left any time I would use it to talk about auxiliary material that was of interest but not covered in the book. After a week of this program, three students appeared in my office. They told me that they were a delegation from the class which had met to discuss the situation. They felt unanimously that my new approach was unacceptable since the text was much too terse for independent reading, and that the whole class would drop the course unless I returned to the usual method of lecturing.

While the experiment was thus a failure, the availability of the book did provide a new flexibility to the course. When I taught it to an audience with diverse background, it was possible to omit some material (e.g., the measure theoretic aspects or some particular applications) and ask interested students to work this through on their own.

7. REVIEWS

It typically takes a year or so after publication before the first reviews begin to appear. Since serious critical reviews require a careful reading of the book for which reviewers often have neither time nor inclination (remember, these reviews are unpaid!), many reviews of technical books consist of a sketch of the contents and the level of difficulty, with comments about omissions and relevance for different potential groups of readers. In the present case there was one exception: a serious, highly critical review by John Pratt (Pratt, 1961).

After some friendly praise for the writing and mathematical exposition, John expressed his strong reservations concerning the Neyman–Pearson theory which constituted the subject matter of the book. His criticisms had various strands, but in the end many of them had their origin in his Bayesian orientation with its emphasis on coherence and logical necessity. In contrast, the approach of *TSH* is pragmatic. It looks at a variety of properties (unbiasedness, invariance, minimaxity and Bayes averages) which are not always compatible. Each of them is taken as far as it will carry and is illustrated on situations to which it is applicable. This is not the place to revisit this old dispute except to repeat my long-held belief that neither approach is right or wrong, that each has its place and the choice must depend on the situation. However, although I did not agree with his position, I have always been grateful to John for giving the book such serious attention.

Another discussion of the book appeared in 1970 as a section of a paper by Jack Wolfowitz on the future of mathematical statistics (Wolfowitz, 1970). His main reason for including this critique was to warn potential readers against the book's pernicious influence. He characterized it as being both bad statistics (it treats the wrong problem) and bad mathematics (it lacks deep and difficult theorems), and he concluded his review with the thought that it would be "a disaster for statistics if this book should determine the direction for research for any appreciable period of time."

8. SECOND EDITION

In fact, Wolfowitz need not have worried: by 1970 the heyday of small-sample theory was past. It had become clear by then that concepts such as unbiasedness and invariance which are so central to the book, while important and applicable to a large class of basic problems, are nevertheless severely limited. They provide results in exponential and transformation families but do not extend much further. (An exception is the minimax approach, which applies in great generality. It is computationally difficult but is finding renewed interest; see, e.g., Brown, 1994, and Donoho and Johnstone, 1995.) Thus, someone faced with a new and complex problem is not likely to resort to small-sample theory but will use maximum likelihood or one of its variants, or perhaps a Bayes procedure with a noninformative prior.

Despite the waning interest in small-sample theory as a research area, *TSH* (which had been translated into Russian, Polish and Japanese) continued to be used as a text and in 1983 was joined by a parallel volume on point estimation. However, it was now 25 years old and showing its age. It required a general updating; in addition, certain issues that had not been addressed in the original version deserved attention. One of these was conditional inference, the importance of which had been pointed out in Pratt's review. Another was multiple comparisons and other simultaneous inference procedures, a methodology dealing more realistically with problems presented traditionally as tests of hypotheses. This is often an oversimplification, which had justifiably been criticized among others by Wolfowitz.

As a result, in 1986 I brought out a second edition. In this new incarnation, the book had grown from 400 to 600 pages, with a new chapter on conditional inference and an exposition of Wijsman's beautiful theory of optimal simultaneous confidence procedures. Some time later, on a visit to China, I brought a copy of the new edition for Zheng Zhongguo, who had made many of the arrangements for our trip. He told me that he would accept it only if in exchange he could give me his copy of the pirated (English) version of the first edition. When I agreed, he signed it in a reversal of the traditional inscription, "To the author of the book with best wishes, Z.G."

9. AN ODYSSEY

For many years, both the testing and estimation books had fared well with Wiley, and relations between publisher and author had been cordial. I was therefore shocked when in 1991 I received a notification from an unknown Wiley official informing me that they were letting one of the books go out of print (the title was garbled, so that I was not sure which one). The tone was peremptory and did not invite discussion. So I accepted the verdict, requesting only that the copyright of both books should be returned to me.

A natural new home for them at this point seemed to be the Wadsworth Series in Probability and Statistics, of which my friend John Kimmel was the Sponsoring Editor and which had previously published a Festschrift for my 65th birthday. John agreed, and both books were reissued by Wadsworth with new covers but otherwise unchanged.

What I did not know was that Wadsworth was one of a number of publishing companies owned by the Thomson Corporation. Soon after the reissue it was decided (for some internal reasons) to transfer the series that Kimmel had built up over the years to a different Thomson Company, Van Nostrand Reinhold (VNR). There it lingered without an editor or much activity for nine months, and was then moved to still another Thomson Company, Chapman and Hall, New York, where John in the meantime had become Statistics Editor. In 1994 history repeated itself. The statistics program for Chapman and Hall, New York, was terminated and the list of books sent to Chapman and Hall, London. And John Kimmel also moved again, from Chapman and Hall to Springer-Verlag.

While these transfers were going on, I had begun work on a second edition of *Point Estimation*, jointly with George Casella. We both believed that the books were not likely to do well under the new arrangement and, with some effort, persuaded Chapman and Hall to relinquish the two copyrights. They are now slated to go to Springer, and will thus over a period of five years have had six different publishers! It is discouraging to see books treated as purely commercial property, and to realize how little power authors have to control the fate of their work.

10. FUTURE PROSPECTS

At the time of Blyth's notes, nearly 50 years ago, the theory of hypothesis testing was at the center of theoretical research. In the intervening years the field of statistics has expanded enormously and the once "hot" Neyman–Pearson theory has become classical. It is still struggling with important and interesting issues, for example, the combinatorial problems arising in the exact treatment of randomization tests (see, e.g., Diaconis and Holmes, 1994) and the still unresolved question of when and how far to condition. However, in the wider context of statistics as a whole these problems no longer hold center stage.

Interest in the theory of testing is threatened also from another direction: attacks on the practice of hypothesis testing itself. These attacks are partially a reaction to a period in which testing was greatly overused and misused (with some editors urging significance at conventional levels as a precondition for publication). Estimation by confidence intervals is an alternative preferred by many of the opponents of testing. (For a balanced discussion of this issue see, e.g., Bailar and Mosteller, 1992.) In the 19th century, particularly in the work of Laplace, fixed level tests, p-values and confidence intervals were used very flexibly, often in conjunction. This seems sensible and is advocated, for example, in Levin (1993).

The theory described in *TSH* should not be viewed as contradicting or inhibiting such an approach. In fact, *TSH* includes a treatment of confidence estimation and discusses the relation of the two theories. Most of the material could have been written in terms of confidence sets. The distinction between the two approaches is largely one of reporting and interpretation. Much of the theory can also be presented in terms of p-values (see Schweder, 1988).

Hypothesis testing has been around for a long time and appears to provide a solution to a type of question that seems natural in many inquiries. Thus I believe that, singly or as a component of more comprehensive strategies, testing will continue to remain an important part of statistical methodology. If this is the case, the associated theory, which together with its application is the subject of *TSH*, will continue to be of interest.

REFERENCES

BAILAR, J. C. and MOSTELLER, F., eds. (1992). *Medical Uses of Statistics*, 2nd ed. NEJM Books, Boston.

BROWN, L. D. (1994). Minimaxity, more or less. In *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.) 1–18. Springer, New York.

DIACONIS, P. and HOLMES, S. (1994). Gray codes for randomization procedures. *Statistics and Computing* **4** 287–302.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

KALLENBERG, W. C. M. et al. (1984). *Testing Statistical Hypotheses: Worked Solutions*. Centrum voor Wiskunde en Informatica, Amsterdam.

LEVIN, J. R. (1993). Statistical significance testing from three perspectives. *J. Exper. Educ.* **61** 378–382.

NEYMAN, J. and PEARSON, E. S. (1993). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A* 289–337.

PRATT, J. W. (1961). Book review. *J. Amer. Statist. Assoc.* **56** 163–167.

SCHWEDER, T. (1988). A significance version of the basic Neyman–Pearson theory for scientific hypothesis testing. *Scand. J. Statist.* **15** 225–235.

STUDENT (W. S. GOSSET) (1908). On the probable error of a mean. *Biometrika* **6** 1–25.

THOMPSON, B., ed. (1993). Statistical significance testing in practice. *J. Exper. Educ.* **61** 285–396.

WOLFOWITZ, J. (1970). Reflections on the future of mathematical statistics. In *Essays in Probability and Statistics* (R. C. Bose, ed.) 443–453. Univ. California Press, Berkeley.