On the Third Edition of *Testing Statistical Hypotheses*

Joseph P. Romano

Abstract In this article, the history of the third edition of *Testing Statistical Hypotheses* will be discussed.

Sometime during the fall of 1998, Erich Lehmann and Julie Shaffer visited the Statistics Department at Stanford during one of our Tuesday joint colloquia with the Berkeley Statistics Department. While having coffee prior to the seminar, Erich asked me if we could have a quick chat. During this seemingly impromptu conversation, Erich asked if I would be willing to work on the revision of his testing book. Coming from a man I so truly admire, I was genuinely taken aback by this offer. After all, Erich's testing book was rightly regarded as one of the most important books in statistics, and it has had an enormous and enduring impact since the first edition appeared in 1959. Overwhelmed by the offer, I told Erich I would soon get back to him.

My immediate thoughts were that this was an extraordinary opportunity for me to work with Erich on the book, but it also was an enormous responsibility to live up to the high standard of excellence that the first and second editions already held. There was also the matter that I was already working on a book on subsampling with Dimitris Politis and Michael Wolf, and I never imagined I would undertake another book project, at least not so soon.

Over the years at Stanford, I had taught estimation and testing from Erich's texts. But, I wondered why Erich chose to ask me to work on this project. As a student at Berkeley, I took Erich's first-year graduate sequence, Statistics 210, with Javier Rojo and Dorota Dabrowska as the teaching assistants for the course that year. I was quite proud that I got perfect scores on all six exams which I took from Erich during my first year at Berkeley. But after graduating in 1986, I had not had much contact with Erich,

J.P. Romano

Professor of Statistics, Departments of Statistics and Economics, Stanford University,
Stanford, CA 94305-4065

e-mail: romano@stanford.edu

aside from seeing him at the joint colloquia. In any case, I called Erich the next day or so, and told him that I was indeed quite excited about the project, but that I could not begin until the subsampling book was complete, which I expected to be in early 1999. Erich was fine with the delay, and we began in earnest during the spring of 1999.

Our first meeting took place at Erich's home in Berkeley. Erich made a couple of points that have stayed with me. First, one of the nice things about working on hypothesis testing is that the notion of power is so compelling as a measure of performance that one need not bring in arbitrary loss functions. Second, even though the first and second editions are written in such a way where optimal tests are derived as examples of a general approach (or theorem), in fact almost all the tests considered in the text were already proposed prior to any optimality treatment. A prime example of this is the t -test, dating back to Gossett's publication in 1908 under the name Student, which was invented without regard to such properties as uniformly most powerful among invariant, or unbiased, tests. One of Erich's main goals was to be able to find some optimality properties or objective justification for ad hoc procedures.

We initially discussed various topics that could be incorporated into the new edition, such as foundational issues and the Bayesian criticism of the Neyman-Pearson approach, new developments in multiple testing, and the emerging literature of tests claiming superiority over the likelihood ratio, as discussed in Perlman and Wu (1999). Erich even pointed out sections from the second edition which he thought could be improved. But for the most part, I felt the topic which needed the most attention was large-sample theory, and Erich agreed. Starting with large-sample theory had several advantages from my point of view. First, I was very reluctant to rewrite anything that Erich had already published, given the book's success and influence. Why mess with a good thing? Second, based on my experience with teaching hypothesis testing, I would already supplement Erich's book with more modern asymptotic approaches. (The second edition did include a little asymptotics; for example, the robustness of the t -test was discussed using the Central Limit Theorem.) The drawback to finite-sample theory is that it offers amenable solutions in only a restricted class of models. On the other hand, a large-sample approach applies much more generally, and easily offers solutions to such classic problems as the Behrens-Fisher problem and the optimality of the Chi-squared goodness of fit test, as well as many others where finite-sample considerations are less fruitful.

Of course, a reasonably general theory of asymptotic optimality requires the development of new machinery, such as contiguity, the Hellinger metric, quadratic mean differentiable families, and convergence to a normal limiting experiment. Such concepts, no doubt inspired by Wald, were developed by Hájek, Le Cam, and others. In fact, such ideas were no longer that recent, and it was important to bring in the new technology. Naturally, the development of large-sample theory would require a substantial addition to the book. We mistakenly thought the project would take about two years.

Basically, the plan was that I would start writing some sections, and Erich would read, criticize, and sometimes rewrite them. I would visit Erich every couple of weeks in Berkeley, usually on a Thursday morning. Without fail, as I approached Erich's building, I would see Erich watching from his balcony at the appointed time, so that

he could meet me downstairs and let me in. (Fortunately, I am punctual.) Then, we would work on the book until we were hungry for some lunch. Usually, Julie would join us for lunch, and we would discuss statistics, but also politics and music. Erich would often lend me some of his audio discs, so that I might better appreciate some of his favorite composers as well as fill some gaps in my musical knowledge. Despite my being a self-proclaimed amateur musician, I greatly appreciated being exposed to the Shostakovich string quartets and some of the works of Schnittke.

It was absolutely clear that Erich was actually quite concerned with pedagogical issues. He really tried to view the writing from a student's perspective, and was concerned with clarity, motivation, and having sufficiently many examples. He was also concerned that we not raise the level of the text, so that the book remains accessible to first year students. For example, is almost sure convergence a concept that is really needed, or does convergence in probability suffice for statistical applications? Do quadratic mean differentiable models greatly increase our ability to provide methods for important models? Do bootstrap and resampling methods really fit in a book concerned largely with optimality?

At some point, it was nice to realize that not only was I learning from Erich, but that I had something to offer as well. On such occasions, Erich would always ask relevant, insightful questions, and it became obvious that any new material would be treated in an honest and scholarly manner. If I could not convince Erich of the value of a section, it risked being abandoned. For the most part, our philosophical views were in agreement, and it was just a matter of reaching the point where our writing styles were in harmony.

Erich's concerns were evident in the pages of critique he would mail to me or hand me on one of my visits. (I still have a file full of them.) Some sections required several rounds of rewriting. Of course, we made some compromises, but we tried to empathize with the reader. As an example, Erich convinced me to include some material on Edgeworth expansions, because he thought a few introductory pages on the subject could be of enormous benefit in exposing students to a topic which they might not otherwise see.

Other areas of the book required major reworking. After including such tests as the Kolmogorov-Smirnov test and the Chi-squared goodness of fit test as examples in various chapters of the manuscript, Erich thought it might be a good idea to have an entire chapter on goodness of fit. This chapter was probably the most challenging (and least satisfying), because the available results are not completely satisfactory. Even so, Erich viewed goodness of fit as such a fundamentally important problem that, by laying out some key considerations and discussing some currently used tests, such a chapter could be quite valuable.

Soon, another problem emerged. The length of the manuscript was getting to be quite long. At one point, we decided there should be two volumes, with the first volume devoted to finite-sample theory and the second to large-sample theory. We even signed a new contract with Springer, and we were fortunate enough to work with a patient editor, John Kimmel. But, we then became concerned that some departments might only adopt the first volume. After some back and forth, we decided upon one volume if we could limit the number of pages to 800, which we barely succeeded in doing.

(It was particularly surprising to see how many pages were saved by moving all the references to the end of the book, rather than to the end of each chapter.)

Another change to the third edition was the addition of a separate chapter on multiple testing and simultaneous inference. In fact, the book became an impetus for me to direct some of my research effort on multiple testing. Our discussions led to a couple publications in the *Annals of Statistics* in 2005, one of which was coauthored with Julie Shaffer as well. I am extremely proud that this joint work appeared, extending the number of consecutive decades where Erich has published in the *Annals* to an incredible 7, dating back to the 1940s (when the original name of the journal was the *Annals of Mathematical Statistics*).

Toward the end of the project, Erich proposed that we dedicate the book to two of the great pioneers in our field, Lucien Le Cam and John Tukey. The reasons were clear. Tukey inspired much of the work on multiple testing, due to his 1953 mimeographed volume on the subject, which was only officially published in 1994 as Volume VIII in *The Collected Works of John W. Tukey*. Le Cam's groundbreaking work on large-sample optimality culminated in his 1986 volume entitled *Asymptotic Methods in Statistical Decision Theory*. Given my own personal memorable interactions with Tukey as an undergraduate at Princeton and Le Cam as a graduate student at Berkeley, I thought Erich's idea to dedicate the book to their memory was perfect.

By the time the third edition appeared in 2005, the original plan of a two year project had turned into six. But, I view the time as well spent. Working on this project with Erich was certainly one of the most rewarding academic experiences I have undertaken. I would like to think the end result is a success, with the whole exceeding the sum of the individual parts.

References

- [1] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- [2] Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33**, 1138-1154.
- [3] Lehmann, E., Romano, J. and Shaffer, J. (2005). On optimality of stepdown and stepup procedures. *Annals of Statistics* **33**, 1084-1108.
- [4] Perlman, M. and Wu, L. (1999). The Emperor's new tests. *Statistical Science* **14**, 355-381.
- [5] Student. (1908). The probable error of a mean. *Biometrika* **6**, 1-25.
- [6] Tukey, J. (1953). The problem of multiple comparisons. Published in *The Collected Works of John W. Tukey: Multiple Comparisons, Volume VIII. (1999)*. Edited by H. Braun, CRC Press, Boca Raton, Florida.