Erich Lehmann's Work on Asymptotics

David Draper

Erich Lehmann's contributions to mathematical statistics have been both broad and deep, and his influence on applied statistics has been more widespread than perhaps even he knew. As an example, I used `Google scholar` to look at the citation patterns of Chernoff and Lehmann (1954), one of the articles I discuss below. As of this writing (November 2011), this paper has been cited

- quite steadily from 1954 to the present, in fact a total of 327 times, by workers in archaeology, astronomy, biology, biostatistics, business, computer engineering, computer science, ecology, econometrics, education, electrical engineering, environmental sciences, finance, genetics, gerontology, machine learning, mathematical statistics, mechanical engineering, oncology, operations research, physics, probability theory, psychology, reliability theory, remote sensing, sports sciences, traffic safety, and urban planning;
- by researchers writing in English, French, German, Italian, Korean, Romanian, Russian, and Spanish;
- by investigators working in Australia, Belgium, Canada, China, Denmark, France, Germany, Greece, India, Italy, Japan, Korea, the Netherlands, New Zealand, Romania, Russia, Spain, Taiwan, the U.K., and the U.S.; and
- by both frequentists and Bayesians.

A small (non-random) sample of the titles of the works that cite this paper, chosen to illustrate the breadth of application areas, includes the following (alphabetically by the first letter of the title):

- "Antecedents and implications of search engine use as pre-purchase information tools";

D. Draper
Department of Applied Mathematics and Statistics, Baskin School of Engineering,
University of California, 1156 High Street, Santa Cruz CA 95064 USA
e-mail: draper@ams.ucsc.edu

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_46,

- "A downscaling approach for air quality at a mid-latitude site using circulation patterns and surface meteorology";
- "Bayesian item fit analysis for unidimensional item response theory models";
- "Communication of emotion in mediated and technology-mediated contexts: face-to-face, telephone, and instant messaging";
- "Gender equity and `foxsports.com`: a coverage analysis of the 2007 NCAA Division I basketball tournament";
- "Guiding architectural SRAM models";
- "Is the universe expanding?";
- "Labor market segmentation: the case of Ukraine and Russia";
- "Mechanical state estimation for overhead-transmission power lines with level spans";
- "Nonlinear multisystem physiological dysregulation associated with frailty in older women";
- "Opinion retrieval from blogs";
- "Principle of detailed balance and convergence assessment of Markov Chain Monte Carlo methods and simulated annealing" (this topic is absolutely central to contemporary Bayesian computing);
- "Probabilistic approach for durability design of reinforced concrete in marine environments";
- "Production of antibody fragments in *Arabidopsis* seeds";
- "Statistical problems in ancient numismatics";
- "Supermarket customers segments stability";
- "Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development";
- "Universal Donsker classes and metric entropy";
- "Vergleich verschiedener Postremissionsstrategien bei dur akuten myeloischen Leukämie mit normalem Karyotyp"; and
- "Whom we laugh with affects how we laugh."

In what follows I examine a portion of Lehmann's contributions to asymptotic methods in statistics, by revisiting and commenting upon six of his papers in this field.

Chernoff and Lehmann (1954) *The use of maximum likelihood estimates in* χ^2 *tests for goodness of fit.* The problem considered by the authors of this paper is the use of the χ^2 machinery to test for goodness of fit to a specified distribution such as the Poisson or normal, a method that at the time they wrote the paper had been in use for more than 50 years (Pearson (1900)). Suppose (to illustrate the issues) that you have data values $(y_1, \ldots, y_n)$ on the real line that you think are like a random sample from a normal distribution with unknown mean μ and standard deviation σ, and you want to make a repeated-sampling calculation that will assess the plausibility of this model. If you knew μ and σ, Pearson would tell you to (a) partition the real line into k intervals (although he was not so clear on suggesting what k should be and what cut-points to use); (b) calculate

$$D = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \tag{1}$$

where n_j is the number of data values in interval j and p_j is the probability that the $N(\mu, \sigma^2)$ distribution assigns to that same interval; and (c) refer D to its asymptotic χ^2 distribution with $(k-1)$ degrees of freedom to judge whether the assumed model is plausible. If you don't know μ and σ, it's natural (in this frequentist setting) to estimate them from the sample $(y_1, \ldots, y_n)$, using a method such as maximum likelihood, and then pretend that

$$\hat{D} = \sum_{j=1}^{k} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j} \tag{2}$$

also has a large-sample χ^2 distribution, this time with $(k-s-1)$ degrees of freedom, where $\hat{p}_j$ is the MLE of p_j and where (in this case) $s = 2$ (the number of parameters estimated). Chernoff and Lehmann note that "this is in fact the procedure recommended in many textbooks," and it is the purpose of their paper to show that this approach may in practice be flawed, because "the test statistic $[\hat{D}]$ ... is stochastically larger than would be expected under the χ^2 theory." They work out the actual asymptotic distribution of $\hat{D}$ and provide some numerical examples of how far wrong one can go using the χ^2_{k-s-1} distribution; for instance, with the normal setup above and the $k = 4$ intervals obtained by cutting the real line at $\{-1, 0, +1\}$ they obtain a lower bound of 0.12 for the actual significance level of a nominal 0.05 test, showing that "in the normal case the use of maximum likelihood estimates in χ^2 may lead to [a serious] underestimate of the probability of type I error." The reverberations of this elegant paper, only eight journal pages long, are still being felt today.

Fix, Hodges and Lehmann (1959): *The restricted χ^2 test.* This is another paper about χ^2, but in this case the authors are interested in examining variants on the usual χ^2 method that have greater power against a class of specified alternatives than the unrestricted χ^2 test (while of course at the same time sacrificing power against other alternatives). Since the basic idea of hypothesis tests is to construct a measure of the distance between {how the data came out} and {how the data should have come out if the null hypothesis were true}, and since the usual χ^2 measure is clearly constructed to be a distance between observed and expected counts, it's natural for Fix, Hodges and Lehmann (FHL) to adopt a geometric perspective in examining their restricted χ^2 tests, and they do so in a way that again brings to mind the word "elegant." Letting $X = (X_1, \ldots, X_k)$ be multinomial with parameters n and $p = (p_1, \ldots, p_k)$, FHL note that the usual χ^2 measure when testing the simple hypothesis that $p = \pi$ can be written as

$$n \sum_{i=1}^{k} \frac{(R_i - \pi_i)^2}{\pi_i}, \tag{3}$$

where $R = (R_1, \ldots, R_k) = \frac{X}{n}$. This is a special case of the general expression

$$D_c(a, b) = n \sum_{i=1}^{k} \frac{(a_i - b_i)^2}{c_i} \tag{4}$$

for vectors a, b and c of values between 0 and 1 and summing to 1; in particular (3) is $D_\pi(R, \pi)$. When the null hypothesis is composite, say $p \in T$ where T is a t-dimensional surface on the hyperplane $\sum_{i=1}^{k} p_i = 1$, it's again natural to use the data X to find the best estimate π^* of the location of p if the null were true, and to base the test on $D_{\pi^*}(R, \pi^*)$. FHL note that Pearson (1900) got the asymptotic distribution of this test statistic wrong and that the error was not corrected until Fisher did so about 25 years later (in fact this was one of the strongest battles in Fisher's relentless campaign to show that he was the rising young star in British statistics and Pearson was yesterday's man).

Fix, Hodges, and Lehmann (1959) assume that the alternative hypotheses can be represented by a surface S of dimension s that contains the null surface T. Neyman (1949) had shown earlier that if P is a good estimate of p under restriction T and Q is a good estimate of p under restriction S (where "good" includes possibilities such as maximum likelihood and minimum χ^2), then $U(R) = D_R(R, P) - D_R(R, Q)$ is an asymptotically valid statistic for testing $H_0 : p \in T$ against $H_A : p \in S$ with limiting null distribution χ^2_{s-t}. FHL investigate the behavior of this test against a sequence of alternative values of p approaching T at a $\sqrt{n}$ rate and prove a theorem, identifying the non-central χ^2 distribution obeyed by $U(R)$ asymptotically, that permits an investigation of the power of the test. In numerical work that was not easy for the late 1950s, they provide new tables of the non-central χ^2 distribution and a plot of the power of the test against the degrees of freedom $f = (s - t)$ for a variety of values of the non-centrality parameter λ. They conclude the paper by illustrating the use of Neyman's test in a setting in which the null hypothesis is that events of interest occur uniformly in time against the alternative that their occurrence is cyclical, noting that at significance level 0.01 for a particular alternative considered the restricted test has power 0.90 when the usual χ^2 test would only have power 0.62.

There are many points of interest in this paper even with 40 years of hindsight: it features a non-Bayesian use of prior information (in the precise formulation of the alternative hypothesis); the proofs, which involve a nice degree of geometric insight, are discussed in a clear, intuitive way rather than enclosing them in a straitjacket of formalism; FHL display an appealing command of the history of statistics in their narrative (for example, they note at one point in a proof that the idea they're about to mention dates back to Gauss); in an interesting demonstration of flexibility on the setting of type *I* and type *II* error targets, the numerical tables give entries for significance levels from 0.001 all the way up to 0.5; and again the question of what to choose for k comes up (and this time FHL have useful advice to offer the reader).

Lehmann (1963a) and **Lehmann (1964):** *Asymptotically nonparametric inference: an alternative approach to linear models* and *Asymptotically nonparametric inference in some linear models with one observation per cell.* I've known these papers rather well for a long time; they were the basis for part of my dissertation work with Lehmann in the late 1970s and early 1980s (Draper (1988)). The context of these articles is as follows.

It's hard to argue with the viewpoint that parametric frequentist inferential methods were greatly strengthened in the 1920s and 1930s by the work of Fisher on the one

hand (e.g., Fisher (1922)) and Neyman and Pearson on the other hand (e.g., Neyman and Pearson (1933)). Given a particular parametric model

$$(y_i|\theta, f) \overset{\text{IID}}{\sim} f(\cdot|\theta), \tag{5}$$

where θ is a vector of real numbers and the functional form of f is assumed known, these investigators sought methods of estimation and hypothesis testing that were in some sense "best" in repeated sampling from the presumed model (5). As people started to use these methods, a new class of questions arose: what should you do if you don't know the "right" parametric model? One idea would naturally be to continue to use one of the methods that Fisher and Neyman-Pearson had developed for standard choices of f, and then to investigate how badly these methods perform when f was not the actual data-generating mechanism; this approach (studying the *robustness* of the standard methods) was pursued vigorously by many people, including Pearson (1931) himself. A second idea would be to look for new methods that worked rather well "in the neighborhood of f" and never did much worse than the "best" methods (if somehow you knew what f really was). This gave rise to *nonparametric* (or *distribution-free*) techniques, such as the Wilcoxon (1945) rank tests in the one- and two-sample problems. These tests were shown by Pitman (1949) and others to have an *efficiency* (the asymptotic ratio of sample sizes needed to achieve the same power against the same alternative at the same significance level) (a) of approximately 96% when f is normal and (b) never lower than about 86% no matter what f is, but they had the limitation that, as tests, their final products (P-values) live on the probability scale, rather than on the scale of the data (where most good scientific inference actually resides (or at least *should* reside) in practice). Lehmann wondered: how can point and interval estimates (on the data scale) be developed that in some sense arise naturally from the nonparametric tests?

It's a common idea in (frequentist) statistics that if you have a good method for estimating a parameter then you can derive a good test from that estimate; it's equally true (but less often used) that if you have a good test you can work backwards from it (*invert* it) to get a good (point or interval) estimate. In the early 1960s, during a highly productive period, Lehmann — sometimes working alone (e.g., Lehmann (1963b)), sometimes with Joe Hodges (e.g., Hodges and Lehmann (1963)) — used this idea, of inverting the Wilcoxon procedures, to produce rank-based point and interval estimates in a large variety of interesting inferential problems. The two papers I discuss here are examples of this: Lehmann (1964) and Lehmann (1963c) cover analysis of variance models with one and several observations per cell, respectively.

Following Draper (1988), the model that Lehmann (1963c) considers for ANOVA with several observations per cell can be written

$$Y_{ij} = \mu_i + e_{ij}\ , \left\{ \begin{array}{l} i = 1, \ldots I \\ j = 1, \ldots, n_i \end{array} \quad \sum_{i=1}^{I} n_i = N \right\}, \tag{6}$$

in which the e_{ij} are IID continuous real-valued random variables with density g satisfying

$$\theta = \int_{-\infty}^{\infty} g^2(y)\, dy < \infty \tag{7}$$

and $\sigma^2 = V(e_{ij}) < \infty$; here μ_i is a measure of center for the ith of I total cells, Y_{ij} is the jth of the n_i observations in cell i and N is the total number of observations. This looks like a one-way ANOVA model, but more complicated layouts can be handled just by numbering the cells from 1 to I.

Lehmann noted that most inference in ANOVA is based on contrasts among the cell centers, and any contrast

$$\phi = \sum_{i=1}^{I} c_i\, \mu_i, \quad \sum_{i=1}^{I} c_i = 0, \tag{8}$$

can be expressed in terms of the cell centers:

$$\sum_{i=1}^{I} c_i\, \mu_i = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} b_{ij} (\mu_i - \mu_j) \tag{9}$$

(note that the b_{ij} are not unique). To estimate $(\mu_i - \mu_j)$, Hodges and Lehmann (1963) showed that inverting the Wilcoxon rank-sum test yields the robust estimate

$$T_{ij} = \text{med}\{Y_{ik} - Y_{jl} : k = 1, \ldots, n_i; l = 1, \ldots, n_j\}, \tag{10}$$

which inherits the efficiency properties of the Wilcoxon test noted above. This turns out to be unsatisfactory as a basis for inference in ANOVA, however, because the T_{ij} don't satisfy the linearity constraints obeyed by the quantities they're trying to estimate:

$$(\mu_i - \mu_j) + (\mu_j - \mu_k) = (\mu_i - \mu_k), \quad \text{but} \quad T_{ij} + T_{jk} \neq T_{ik}, \tag{11}$$

since the operations of subtraction and taking a median don't commute. Lehmann's suggested fix for this problem, improved slightly a few years later by Spjøtvoll (1968), was to express $(\mu_i - \mu_j)$ as $[(\mu_i - \bar{\mu}) - (\mu_j - \bar{\mu})]$, where $\bar{\mu} = \frac{1}{N} \sum_{i=1}^{I} n_i \mu_i$ is the grand mean, and estimate $(\mu_i - \bar{\mu})$ by $\bar{T}_i = \frac{1}{N} \sum_{k=1}^{I} n_k T_{ik}$; this linearizes the estimates by making all comparisons relative to $\bar{\mu}$. With this approach, $(\mu_i - \mu_j)$ can now be estimated by $W_{ij} = (\bar{T}_i - \bar{T}_j)$, leading to the contrast estimate

$$\hat{\phi} = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} b_{ij} W_{ij}; \tag{12}$$

the above linearization ensures that even though the b_{ij} are not unique, all choices of b_{ij} lead to the same value of $\hat{\phi}$.

It's now straightforward to "robustify" virtually any standard normal-theory-based ANOVA procedure, simply by (a) replacing the usual quantity $(\bar{Y}_i - \bar{Y})$ — in which $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{I} n_i \bar{Y}_i$ — by its rank-based analogue $\bar{T}_i$, and (b) replacing the normal-theory estimated error variance $\hat{\sigma}^2$ by its analogue $\hat{\sigma}_R^2$. Lehmann showed that the asymptotic analogue of σ^2 in this procedure is

$$\sigma_R^2 = \frac{1}{12\,\theta^2} = \frac{1}{12(\int g^2)^2}. \tag{13}$$

In my dissertation work with Lehmann I developed two methods for estimating $\int g^2$ — one (suggested by Lehmann (1963b)) derived from the lengths of confidence intervals based on the same differences $(Y_{ik} - Y_{jl})$ whose median is T_{ij} (see equation (10)), and another involving density estimation — and conducted a large simulation study to see how this machinery works in small samples. The conclusions were that interval estimates from this approach (a) have about the same coverage as the standard methods with normal and non-normal data, (b) are about 2% wider when the data really are normal, but (c) can be noticeably narrower with samples from non-normal distributions (e.g., normal-theory intervals are about 40% wider than the rank-based intervals with data from t distributions with small degrees of freedom).

In parallel with this paper, Lehmann (1964) is principally interested in ANOVA models with one observation per cell:

$$X_{i\alpha} = \nu + \xi_i + \mu_\alpha + U_{i\alpha}, \quad \sum_{i=1}^{c} \xi_i = \sum_{\alpha=1}^{N} \mu_\alpha = 0; \tag{14}$$

here the factor of interest, represented by ξ, is at c levels, μ represents a blocking factor at N levels, and the $U_{i\alpha}$ are IID with CDF F. Lehmann again focuses on contrasts $\theta = \sum_{i=1}^{c} c_i \xi_i$ (with $\sum_{i=1}^{c} c_i = 0$); he defines $V_\alpha = \sum_{i=1}^{c} c_i X_{i\alpha}$ and notes that these random variables are also IID, with CDF (say) G_c. If F (and therefore G_c) are symmetric, then a natural place to start in estimating θ from a rank-based point of view is with the *Walsh averages* $\text{med}\left[\frac{V_\alpha + V_\beta}{2}\right]$, where the median is taken over all $\alpha \le \beta$. As in the paper discussed above, these quantities are linearly incompatible, so Lehmann again focuses instead on $(\xi_i - \xi_j)$ and estimates these differences with

$$Y_{ij} = \text{med}\left[\frac{(X_{i\alpha} - X_{j\alpha}) - (X_{i\beta} - X_{j\beta})}{2}\right]. \tag{15}$$

Even these quantities are not compatible, so he linearizes them with

$$Z_{ij} = (Y_{i\cdot} - Y_{j\cdot}), \tag{16}$$

where the dot signifies averaging with respect to the indicated subscript, and bases his inferences about contrasts on the Z_{ij}. The asymptotic distribution of the Z_{ij} again

involves $\int g^2$, where this time g is the density of $(X_{i\alpha} - X_{j\beta})$, and efficiency gains relative to the standard normal-theory methods are similar to those found in Lehmann (1963a). Taken together, these two papers — and the other related articles published around the same time — provide the foundation for *R-estimation*, an approach to robust inference that's competitive with the other two leading methods developed subsequently (*L-estimation* (e.g., Bickel (1973)), based on linear combinations of order statistics, and *M-estimation* (e.g., Huber (1973)), which generalizes maximum likelihood in a robust fashion).

Hodges and Lehmann (1970): *Deficiency*. In this paper Hodges and Lehmann (HL) are interested in comparing the number of observations required by two different inferential methods to achieve the same accuracy, as a measure of how much one method is better than the other. As they note, if methods A and B require n and $k_n > n$ observations to arrive at the same level of performance, one natural way to summarize this is through the ratio $\frac{k_n}{n}$, whose limit e as $n \to \infty$ will generally be stable (this is the basis of the familiar concept of *asymptotic relative efficiency* (ARE)). If $e > 1$, its value can often give reliable guidance as to how much method A will be better than B with finite (and even rather small) n; but what if (as frequently occurs) $e = 1$? It may still be that A is better than B in small samples, and a natural way to quantify this is through the difference $(k_n - n)$, which HL call the *deficiency* of B relative to A; if $d = \lim_{n\to\infty}(k_n - n)$ exists, they call this the *asymptotic deficiency*.

The paper is devoted to an exploration of this idea in a series of examples, using expected squared error of point estimators as the performance measure. They show, for instance, that if you and I are both estimating the population variance σ^2 based on a sample $(X_1, \ldots, X_n)$, and I pretend that I know the population mean ξ and use $M_n = \frac{1}{n}\sum_{i=1}^n (X_i - \xi)^2$ while you make no such assumption and use $M_n' = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ (where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$), both of us will be using estimators that are unbiased (in repeated sampling) with the same large-sample variance $\frac{\gamma\sigma^4}{n}$, where $(\gamma + 1) = \frac{\mu_4}{\sigma^4}$ is the standardized fourth central moment of the population distribution F. Thus in this case $e = 1$, but they show that the asymptotic deficiency of M_n' relative to M_n is $\frac{2}{\gamma}$. With normal data $\gamma = 2$, so (as they put it) in that case "it costs only one observation to protect against an erroneous value of ξ," whereas the deficiency can be arbitrarily large for non-normal populations since γ can be as close to 0 as you might want to make it.

Other examples analyzed include comparisons of (a) biased and unbiased estimates when drawing inferences about the population variance, (b) medians versus quasi-medians in estimating the center of symmetry of a symmetric distribution, (c) several methods for creating confidence sets for normal means, (d) the one-sample t test versus the analogous z procedure obtained by pretending the population variance is known, and (e) Bayesian versus unbiased estimation of a normal mean. I'm a bit uncomfortable with their only other example not on this list, however: with X as the (binomial) number of successes in n IID Bernoulli trials with unknown success probability p, they compare the usual unbiased estimator $M_n = \frac{X}{n}$ with the minimax estimator

$$M_n' = \frac{\sqrt{n}}{1 + \sqrt{n}} M_n + \frac{1}{2(1 + \sqrt{n})}; \tag{17}$$

Table 1 *Expected squared errors for the unbiased (V_n) and minimax (V_n') estimators of a Bernoulli/binomial p, for various values of n and p.*

	$p = 0.25$		$p = 0.4999$		$p = 0.5$	
n	V_n	V_n'	V_n	V_n'	V_n	V_n'
1	$1.88 \cdot 10^{-1}$	$6.25 \cdot 10^{-2}$	$2.50 \cdot 10^{-1}$	$6.25 \cdot 10^{-2}$	$2.50 \cdot 10^{-1}$	$6.25 \cdot 10^{-2}$
10	$1.88 \cdot 10^{-2}$	$1.44 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$	$1.44 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$	$1.44 \cdot 10^{-2}$
100	$1.88 \cdot 10^{-3}$	$2.07 \cdot 10^{-3}$	$2.50 \cdot 10^{-3}$	$2.07 \cdot 10^{-3}$	$2.50 \cdot 10^{-3}$	$2.07 \cdot 10^{-3}$
1,000	$1.88 \cdot 10^{-4}$	$2.35 \cdot 10^{-4}$	$2.50 \cdot 10^{-4}$	$2.35 \cdot 10^{-4}$	$2.50 \cdot 10^{-4}$	$2.35 \cdot 10^{-4}$
10,000	$1.88 \cdot 10^{-5}$	$2.45 \cdot 10^{-5}$	$2.50 \cdot 10^{-5}$	$2.45 \cdot 10^{-5}$	$2.50 \cdot 10^{-5}$	$2.45 \cdot 10^{-5}$

these estimators have expected squared errors

$$V_n = \frac{p(1-p)}{n} \quad \text{and} \quad V_n' = \frac{1}{4(1+\sqrt{n})^2}, \tag{18}$$

respectively. You can see that if $p \neq \frac{1}{2}$ the asymptotic relative efficiency of M_n' to M_n is $\frac{p(1-p)}{4} < 1$, so that for large n the unbiased estimator is better (with expected squared error as the figure of merit); but with $p = \frac{1}{2}$ the ARE is 1 and $V_n' < V_n$, and HL show that in this case the asymptotic deficiency of M_n relative to M_n' is infinite. This seems misleading to me from a practical point of view, as illustrated in Table 1. HL note that "the deficiency computation thus shows that the [unbiased] estimator requires a much larger sample size than the minimax [estimator] if they are to have the same expected squared error at $p = \frac{1}{2}$, in spite of the fact that the corresponding asymptotic efficiency is 1." While this may strictly speaking be true, to me a better summary of what's going on (with reference to Table 1) would be just to say that (a) for p near 0.5, M_n' is better for all realistic sample sizes n, but the amount that it's better goes to 0, and (b) for p not near 0.5, M_n' is better for small n and then from a certain point on (as n grows), M_n is better.

Lehmann and Loh (1990): *Pointwise versus uniform robustness of some large-sample tests and confidence intervals*. This paper is about the sensitivity of hypothesis tests to departures from the distributional assumptions made to derive them. Lehmann and Loh (LL) note, for instance, that the one-sample t test of the null hypothesis that the population mean is 0 against positive alternatives is designed to have exact type-I error rate α under normality, and it's natural to wonder how it behaves when the data are not normal. If the data $(X_1, \ldots, X_n)$ are IID from CDF F, then as long as F has finite variance the first-order asymptotic answer is pleasant: with $\alpha_n(F)$ as the probability of rejecting $H_0 : \mu(F) = 0$ (where $\mu(F)$ is the mean of F), it's easy to show that $\alpha_n(F) \to \alpha$ as $n \to \infty$, which people typically describe by saying that the the level of the t test is *robust* against non-normality. But **Q:** how big does n need to be for $\alpha_n(F)$ to be close to α?

Looking at the problem a bit more generally, LL consider testing

$$H_0' : \text{the mean } \mu_F \text{ of the unknown } F \text{ is } 0 \tag{19}$$

against the alternatives (indexed by F) that μ_F is positive. When testing H_0' the level of the t test with n observations is $\alpha_n = \sup_{\mathcal{F}} \alpha_n(F)$, with $\mathcal{F}$ as the set of all CDFs with mean 0 and finite variance. It would be nice if $\lim_{n\to\infty} \alpha_n = \alpha$; if this were true we could say that the t test is *uniformly robust* over $\mathcal{F}$. But it's been known for a long time (Bahadur and Savage (1956)) that $\sup_{\mathcal{F}} \alpha_n(F) = 1$ for all $\alpha > 0$, so it's necessary to seek the answer to question **Q** above a bit differently: the Bahadur-Savage result says that, given a small $\epsilon > 0$, you cannot find a sample size n_0 such that

$$|\alpha_n(F) - \alpha| < \epsilon \text{ for all } n > n_0 \tag{20}$$

when F is allowed to vary over $\mathcal{F}$, but can you find a *subset* $\mathcal{F}_0$ of $\mathcal{F}$ (that's big enough to be relevant to actual statistical practice) for which (20) is true? LL devote much of this paper to showing, unhappily, that across natural and obvious choices of $\mathcal{F}_0$ — for example, all distributions in a fixed Kolmogorov-Smirnov neighborhood of normal densities with mean 0, and all continuous distributions with support contained in a bounded interval — the t test fails to achieve uniform robustness; in fact, the only class $\mathcal{F}_0$ that LL can find that works is the set of absolutely continuous $F \in \mathcal{F}$ with the rth standardized absolute moment about the mean uniformly bounded for some $r > 2$. They also obtain a negative result for inference about the success parameter p in Bernoulli/binomial sampling. Overall the paper serves to show how hard it is to achieve uniform robustness as a performance criterion for repeated-sampling hypothesis tests.

An overview. Together (as of this writing) these papers have been cited more than 560 times, by researchers working in a wide variety of fields; the resulting citation rate per paper (about 93) is above average for an author who is himself highly-cited (the citation software `Publish or Perish` finds that Lehmann has been cited almost 22,000 times over a 66-year period, with an average number of citations per paper of about 62). All of these papers are well-written, which (besides high impact) is another general feature of Lehmann's work. Two threads run through these articles, and serve to draw my comments to a close.

- They're all written from the repeated-sampling (frequentist) point of view, although Bayesian methods sometimes get a mention. A Bayesian would approach the problems addressed in these articles differently — for example, (1) today I would solve the linear-model problems in Lehmann (1963b), Lehmann (1964) not by seeking robust estimates but by direct Bayesian parametric or nonparametric modeling of the (non-normal) error distributions, which would yield inferential procedures with repeated-sampling robustness properties similar to those exhibited by Lehmann's methods, and (2) in my view many uses of hypothesis-testing in practice should really be reformulated as Bayesian decision-theory problems with utility structures (a) that are sensitive to the real-world context of the problem and (b) that will often be found to differ from the loss function inherent in the usual type *I* and *II* error story — but everyone would agree that the problems posed here (Does this model fit the data? Does this procedure still work well, even though some of the assumptions on which it's based are violated? How much better is method A than B in extracting signal from the data?) are important.

- They all feature an interest in how often statistical procedures get the right answer. This is a fundamental scientific question to which, in my view, all statisticians — whether they work in the Bayesian or frequentist paradigms — need to pay constant attention (I learned the importance of this topic both from Lehmann and from his mentor Jerzy Neyman). I began my research career in the frequentist paradigm (which was the only story readily available at Berkeley, at least when I was there in the late 1970s and early 1980s), and came to appreciate the value of Bayesian methodology and thinking later (in the mid 1980s). The standard position taken by many statisticians in the 20th century was that you had to (a) choose between the Bayesian and frequentist paradigms and then (b) defend your chosen approach against attacks from people who had chosen the other paradigm, but this is a flawed formulation of the problem: when you look closely you find that each paradigm has both strengths and weaknesses, so — in my view — in the 21st century it's the job of all working statisticians to try to construct a fusion of the two paradigms that emphasizes the strengths and de-emphasizes the weaknesses. My own personal fusion, combining elements of my Berkeley training and post-Berkeley study, is (i) to reason in a Bayesian way when formulating my inferences, predictions and decisions (because the Bayesian approach seems to me to be the most successful method so far invented for capturing and quantifying all relevant sources of uncertainty, whether arising from an inherently repeatable process or not) and (ii) to reason in a frequentist way when evaluating the procedures in (i) (because we all need to pay attention to how often we get the right answer, and this is an inherently frequentist question). I'm not sure what Lehmann's position was on this issue toward the end of his long career, but I *am* sure that, if he were still here to give us his thoughts on the subject, they would be well worth listening to.

References

[1] Bahadur, R. R. and Savage, L. J. (1956). The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *The Annals of Mathematical Statistics*, **27**, 1115–1122.

[2] Bickel, P. J. (1973). On Some Analogues to Linear Combinations of Order Statistics in the Linear Model. *Ann. Statist.* **1**, 597–616.

[3] Chernoff, H., and Lehmann E. L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statist.*, **2**, 5579–586.

[4] Draper, D. (1988). Rank-Based Robust Analysis of Linear Models. I. Exposition and Review. *Statistical Science*, **3**, 239–257.

[5] Fisher, R. A. (1922) On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. R. Soc. Lond. A*, 309–368

[6] Fix, E., Hodges, J. L., and Lehmann, E. L. (1959). The restricted chi-square test, Probability and Statistics. *The Harald Cramer Volume*, 92–107.

[7] Hodges, J. L., and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.*, **34** (1963), 593–611.

[8] Hodges, J. L., and Lehmann, E. L. (1970). Deficiency. *The Annals of Mathematical Statistics*, **41**, 783–801.

[9] Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799–821.

[10] Lehmann, E. L. (1963a). Asymptotically nonparametric inference: an alternative approach to linear models. *Ann. Math. Stat.*, **34**, 1494–1506.
[11] Lehmann, E. L. (1963b). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Stat.*, **34**, 1507–1512.
[12] Lehmann, E. L. (1963c). Robust estimation in analysis of variance. *Ann. Math. Stat.*, **34** 957–966.
[13] Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Stat.*, **35**, 726–734.
[14] Lehmann, E. L. and Loh, W-Y. (1990). Pointwise vs. uniform robustness of some large-sample tests and confidence intervals. *Scand. J. Statist.*, **17**, 177–187.
[15] Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. R. Soc. Lond. A*, 289–337.
[16] Neyman, J., (1949). Contribution to the theory of χ^2 test. In: *J. Neyman, ed., Proceeding of the First Berkeley Symposium on Mathematical Statistics and Probability*, 239-273.
[17] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine, 5th Series*, **50**, 157–172.
[18] Pearson, E. S. (1931). The Analysis of Variance in Cases of Non-normal Variation. *Biometrika*, 114–133.
[19] Pitman, E. J. G. (1949). Lecture notes on nonparametric statistical inference. Unpublished.
[20] Spjøtvoll, A. (1968). A Note on Robust Estimation in Analysis of Variance. *Ann. Math. Statist.*, **39**, 1486–1492.
[21] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83.