The Changing History of Robustness

Stephen M. STIGLER

This essay, a reflection upon the changing views of robust statistics from the euphoria of the 1960s to the present day, was given as a keynote address at the International Conference on Robust Statistics (ICORS) in Prague on June 28, 2010.

KEY WORDS: Estimation; History of statistics.

That there has been a change in the area statisticians refer to as Robustness cannot be doubted. I could as one extreme example point to a recent book by my University of Chicago colleague, econometrician Lars Peter Hansen and his co-author Thomas Sargent. Its title is *Robustness* and it has a striking cover (Figure 1).

You may recognize the cover as from an early 17th century painting by Georges de la Tour, "Le Tricheur" or "The Cheat." If so, that recognition is itself an example of a robust inference, since the painting has been altered by the cover designer, including rendering it in an almost cubist style. I will return to this book and its cover at the end of this essay, but for now I only point to what to me is surprising evidence of change: This long and well-documented work contains not a single reference to work on robustness that I knew as a young researcher, nor to later developments in that same line. Furthermore, no reader of this book could claim that such a reference was required. Clearly something has changed.

I will begin with a summary of the subject of Robust Statistics as it was seen by me in 1972, sharpened of course by hindsight. The statistical uses of the word "Robust," as all statisticians who have looked into this agree, date from an article "Non-Normality and Tests on Variance" published by George Box in 1953 in *Biometrika*. There, Box wrote of the "remarkable property of 'robustness' to non-normality which [tests for comparing means] possess," a property that he found was not shared by tests comparing variances. He directed his fire particularly toward Bartlett's test, which some had suggested as a preliminary step, to check the assumption of equal variances before performing an ANOVA test of means. He summarized the results this way:

> "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!"

Stephen M. Stigler is Ernest DeWitt Burton Distinguished Service Professor, Department of Statistics, University of Chicago, 5734 University Ave., Chicago, IL 60637 (E-mail: *stigler@uchicago.edu*).

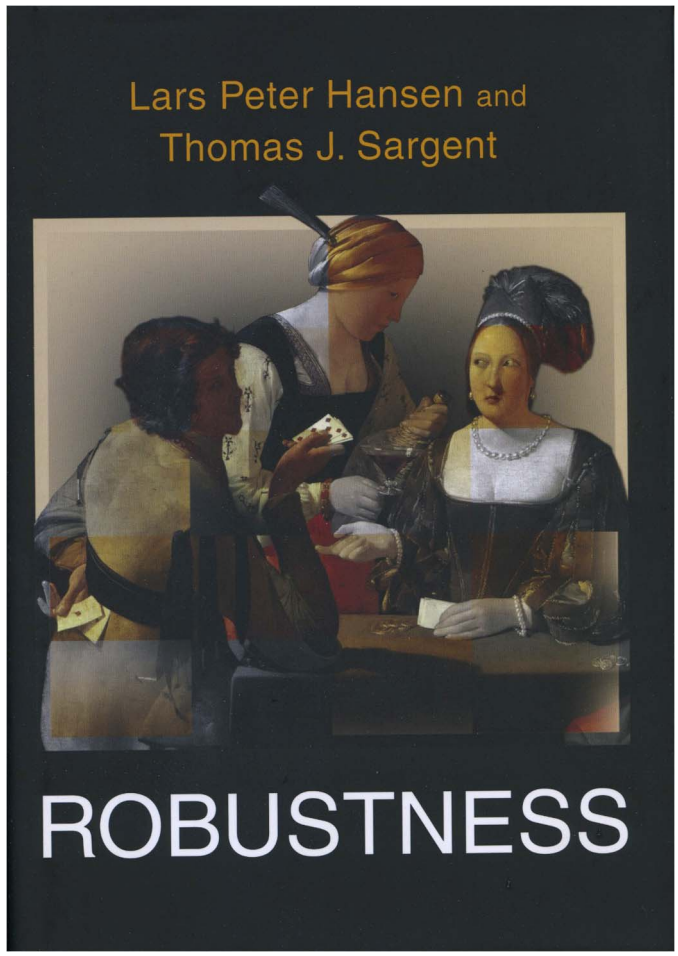

Figure 1. The front cover of the book by Hansen and Sargent (2008). The online version of this figure is in color.

After this dissection, Bartlett's test, much like a frog in a high school biology laboratory, was never the same again.

But if Box gave us the term, most people agree that robustness as a subdiscipline in statistics was primarily the creation of John W. Tukey and Peter J. Huber. For many years it was—and in review papers still is—almost a requirement to cite Tukey's (1960) contribution to the Hotelling Festschrift volume. This article is best known for a single example, where Tukey called attention to the fact that in estimating the scale parameter of a normal distribution, the sample standard deviation ceases to be more efficient than the mean deviation if you contaminate the distribution with as little as 8-tenths of a percent from a normal component with three times the standard deviation. This took most statisticians of that era as a surprise—Tukey stated that of

 DOI: 10.1198/tast.2010.10159

all those he quizzed on the question, only R. A. Fisher foresaw that such a small contamination could have such a large effect.

Tukey's article was much more than that single example—it was in fact a mathematical and philosophical summary of researches Tukey had begun nearly 20 years earlier, and it brought a full set of the tools of classical statistical theory to bear on an examination of a philosophical topic: how might we address the problem of planning for inference when the assumed model could neither be fully trusted nor adequately checked? Tukey's examination of the location parameter was only illustrative of what could be done to study trade-offs, but it was definitive on one point: He announced that in such studies it was sufficient to consider contaminated distributions as alternatives. Tukey wrote,

> "All available evidence and discussion leads the writer to believe that, although contaminated distributions are a thin representation of non-normality, they are an exploration of the most important area of non-normality. Consequently he [Tukey] judges the step from normality alone to contaminated distributions to be a large part of the step from normality alone to all smooth, nearly normal distributions." [p. 471]

The gauntlet had been thrown down, and it was not to be long before it was picked up by Peter Huber while on a research fellowship at Berkeley for the 1962–1963 academic year. Huber's article, "Robust Estimation of a Location Parameter," was published in the *Annals of Mathematical Statistics* in 1964. It is now such a familiar part of our intellectual landscape that it is hard to realize what an astonishing and totally unexpected accomplishment it was at that time. Before Huber, robustness, if the term was known at all, signified a discussion of poorly articulated trade-offs that practical statisticians had to countenance, but theoreticians could ignore. With Huber in 1964 it became a different object altogether. To general amazement, Peter Huber had shown that there was a *Best* robust answer to a widely accepted formulation of the location parameter problem!

If the newspapers of the day had covered such things, a headline might have read, "The army of optimality has landed on the beach of applied statistics, and without visible resistance erected a large base camp." Now to be sure, at Berkeley in those years that was a busy beach—the forces of nonparametrics had been there and were already moving to higher ground. And I must confess that although I was a student there from 1963–1967, only a slight clamor was audible in our classes, as we gave most of our attention to that earlier nonparametric landing. And in any case we had insufficient historical grounding to appreciate the promise Huber's one article offered. From my present vantage point, I would liken it to the initial publication of the Neyman–Pearson Lemma in potential impact: a bold approach to one problem was successful to an altogether unexpected degree, and all around lay other problems where similar tools could, one hoped, be deployed.

The next major steps were not too long in coming. In summer 1968 Frank Hampel completed his dissertation, with its formalization of the concept of robustness as approached by Huber, via a functional derivative of an estimator, and its informalization of that abstraction via the newly christened "Influence Curve," itself a cousin to Tukey's "pseudo-values." Another Berkeley dissertation by Louis Jaeckel in 1969 showed the activity was becoming more general in less fundamental directions as well: Jaeckel explored the potential use of adaptive trimmed means.

Over the next few years Huber published a series of articles and reviews, in the Netherlands, in Montreal, and then in 1972 as his Wald Lectures in the *Annals of Mathematical Statistics*. In these he expanded upon his earlier results, and by 1972 the world of robust estimation was considerably larger, with optimality results now for three classes of estimates—R estimates and L estimates joining Huber's (1964) M estimates. Huber's (1972) *Annals of Mathematical Statistics* article was quite important to me for a curious reason—it began with a nice set of historical observations, and my following up on and greatly expanding upon those led to my first formal article on the history of statistics, appearing in *JASA* in December of 1973.

Any history is a product of its time; it must necessarily take the present view of the subject and look back, as if to ask, how did we get here? My 1973 account was just such a history, and it took the 1972 world of robustness as a given. Huber had brought attention to Simon Newcomb and his use in the 1880s of scale mixtures of normal distributions as ways of representing heteroscedasticity; I enlarged and extended that to other works. I noted that Newcomb had used an early version of Tukey's sensitivity function, itself a forerunner of Hampel's influence curve. I reviewed a series of early works to cope with outliers, and I trumpeted my discovery of Percy Daniell's 1920 presentation of optimal and efficient weighting functions for linear functions of order statistics, and of Laplace's 1818 asymptotic theory for least deviation estimators. I found M estimates in 1844 and trimmed means (called "discard averages") in 1920.

None of this, I hasten to say, is recounted to undercut the striking originality of Tukey and Huber and Hampel—to the contrary. I mean it in the spirit of Alfred North Whitehead's famous statement that, "Everything of importance has been said before by somebody who did not discover it"; that is, to provide historical context, where one might now see that, for example, it was not the M estimates that were new in 1964, it was what Huber proved about them that was revolutionary.

Daniel Bernoulli provided a striking example of this phenomenon, one I came upon in the late 1970s in archives in Basel. In an unpublished 1769 manuscript Bernoulli proposed an iteratively reweighted mean algorithm with a semi-circle weight function, and he easily illustrated its use for a case with five observations. Does he belong in the pantheon of early robust statisticians? It is not so clear. He subsequently revised the manuscript for publication, and in the revision, the semi-circle became a probability density and the estimation then was done by maximum likelihood, not a notably robust method for that density. And he was so wedded to the new approach that he persisted even though he then could only handle a case with three observations (Stigler 1997). For Bernoulli, the Maximum Likelihood estimate had supplanted the M-estimate (which was only, in Huber's phrase, "maximum likelihood-like"). Similarly, Simon Newcomb's method of choice in applications was least squares, and if Percy Daniel is remembered in statistics at all today it is for the Daniel window for estimating the spectrum of a Gaussian time series.

I return to the statistical world of 1972. In a more perfect world, the next steps after 1972 would be clear. In the manner of one of Thomas Kuhn's Scientific Revolutions, the next decades would be a parade of expansions of the theory, followed by widespread adoption in statistical practice. It would be just as in the 1930s to the 1950s, when we saw the fruits of the Neyman–Pearson Lemma and Fisher's likelihood methodology with their extension to new examples, and more complex areas such as multivariate analysis. There were new wrinkles introduced to cover newly discovered difficulties, but all leading to a general consensus for testing and estimation, at least within the Berkeley School, which in 1960 covered North America and much of continental Europe. Now, by 1972 a number of the early workers in robust statistics expected that from the 1970s to 2000 we would see the same development with robust methods—extensions to linear models, time series, and multivariate models, and widespread adoption, to the point where every statistical package would take the robust method as the default and any author of an applied article who did *not* use the robust alternative would be asked by the referee for an explanation. This was, and I will call it, a Grand Plan. But that plainly is not what has occurred. What happened? Why did the statistical world not follow this script, this Grand Plan whose broad outlines seemed in 1972 so obvious and desirable to all statisticians with an interest in such matters?

Before offering a view on this, I want to issue a strong disclaimer. First, there was of course a series of really wonderful advances over those decades in exactly such a direction. Koenker and Bassett (1978) devised what I still believe is the best approach to robust methods in higher dimensional linear model problems, with their regression quantiles and subsequent extensions. In addition, Hampel brought his thesis forward, in part with co-authors Ronchetti, Rousseeuw, and Stahel, to make advances on several fronts and to show what an influence function approach emphasizing breakdown points could lead to (Hampel et al. 1986). Robust approaches to covariance matrices, to time series, to just about any problem yet encountered, with the possible exception of contingency tables, have been devised. All of this is true; much of the work is wonderful. But of the Grand Plan, only the first element has begun to be realized.

With hindsight, the first signs of trouble with the Grand Plan were evident even in 1972. In Fall 1970, a small group gathered at Princeton University's Statistics Department for a year-long seminar in robust statistics. At Tukey's invitation, Peter J. Huber and Peter J. Bickel came to Princeton as year-long visitors, and they were joined by recent Ph.D.s Frank Hampel and David F. Andrews from Toronto. Andrews's Christmas suggestion that they perform a large-scale simulation study led to the group's most tangible product, a book published in 1972 and universally known as the Princeton Robustness Study (Andrews et al. 1972). The published study presented a summary of a huge investigation in the form of tables, charts, and computer code. From the full set of 10,465 estimates of a location parameter they had considered, they reported in detail on the accuracy of 68 estimates that had received extensive study, focusing upon small samples and an inventively wide selection of 32 distributions, nearly all of which were symmetric scale mixtures of normal distributions.

There are different ways of describing the Princeton Study. At one extreme it could be seen as a path-breaking exercise in the exploration of the function space product of all conceivable estimates crossed with all conceivable distributions including all interactions, via ingenious advances in simulation technique and graphical exploratory data analysis. At another extreme it could be seen as a fruitless exercise in self-indulgent ad hockery, beating a small and uninterestingly limited problem to death by computer overkill. I see elements of truth in both views. The practical lesson that jumps out to me is that this function space is much larger and more complicated than the intrepid explorers anticipated, leaving them to forage with little to guide them, and even then they did not travel far from samples of sizes 10 and 20 or from unimodal symmetric distributions, and this only for a location parameter. In a 1997 Tukey Festschrift, Frank Hampel offered a more generous assessment of the practical lessons learned but not included in the published report. Still, no believer in the Grand Plan would have found much aid or comfort in the 1972 study as published.

An almost immediate effect of the 1972 Princeton Study was a deluge of simulation studies submitted to statistics journals. I was a *JASA* Associate Editor at the time, and it seemed like every statistician with a random number generator had a newly issued license to publish. The floodgates were open, but the quality of the overflow was generally low and sometimes appalling. Even its severest critics granted that the Princeton Study was the work of artists, but Gresham's law (that 'bad money drives out good') holds in science as well as in monetary economics. As a reaction to this threat I wrote an article in 1975 called "Do Robust Estimates Work With *Real* Data?" It was published with discussion in the *The Annals of Statistics* in November 1977.

I located a collection of datasets from early science on measures of the speed of light, on the mean density of the Earth, and on the parallax of the sun—all cases where we knew the true value now, at least to a precision that was exact compared to the uncertainty at the time the measures were made. The conclusion I drew from my analysis was that some slight trimming did improve over the sample mean (the 10% trimmed mean was the smallest trimming proportion I considered). But the sample mean itself did nearly as well, and better than many of the estimates considered in the Princeton Study and found best there.

Some of the discussants, including Peter Huber, found these results unsurprising. Others were more critical, including Frank Hampel in correspondence and a later book, complaining that in every situation there was nontrivial systematic bias, and that bias dominated the comparison, making most estimates similar in performance. My reply was that, yes, it did indeed, although you should only expect it to dilute the differences, not to reverse any order of merit. But that bias, which averaged only about a half standard deviation of a single measurement, made an important point: it represented a component of error that was ignored in all theoretical studies, a component that the methods could not and did not deal with. The robust estimates could only help with variations from assumptions that the scientists had foreseen.

By the turn of the century, no one was publicly optimistic about the future prospects for the Grand Plan. Where once there

Figure 2. Georges de la Tour's circa 1635 painting "Le Tricheur" (or "The Cheat"). The painting currently exists in two versions. This version (also known as "The Ace of Diamonds") is in the Louvre. The online version of this figure is in color.

was a hope for a Robust New World on a scale that would displace most of classical statistics, now there was no audible voicing of such a hope. The euphoria of the first decade proved to be unsustainable. Reasons for this change were not hard to find. Indeed, some were noted in the discussion of Bickel (1976), and many of them were summarized in a new chapter Peter Huber himself added to the 1996 second edition of his 1977 SIAM monograph. As models grew more complicated, so too did the question of just what "robust" meant. The potential model failures in a multivariate time series model are huge, with no consensus upon where to start. Even in such cases, important progress was made, but with lower expectations of completeness. The crisp formulation and brilliant solutions for the location parameter problem were not to be repeated. The Neyman–Pearson Lemma extends effortlessly to Banach and Hilbert spaces; Huber's (1964) result was not available for the more complex worlds that modern statisticians live in.

I suspect all of you are familiar with these issues, and I know some of them that have been at least partially overcome—this discussion has after all been going on for at least thirty years. There is a further issue that I suspect is perhaps as important, and I have not seen it brought up in this connection, as part of this discussion. It is the nature of the established methodology, the normal theory of linear models and multivariate analysis. Ever since the statistical world fully grasped the nature of what Fisher created in the 1920s with the analysis of multiple regression models and the analysis of variance and covariance—ever since about 1950—we have seen what that analysis can do and seen the *magic* of the results it permits. The perfection of that distribution theory, the ease of assessing additional variables and partitioning sums of squares as related to potential causes—no other set of methods can touch it in these regards. Least squares is still and will remain King for these reasons—these magical properties—even if for no other reason. In the United States many consumers are entranced by the magic of the new iPhone, even though they can only use it with the AT&T system, a system noted for spotty coverage—even no receivable signal at all under some conditions. But the magic available when it *does* work overwhelms the very real shortcomings. Just so, least squares will remain the tool of choice unless someone concocts a robust methodology that can perform the same magic, a step that would require the suspension of the laws of mathematics.

Where does this leave robustness? In very good condition, thank you. I hasten to reiterate and extend my earlier disclaimer—there was and is no shortage of important and exciting research on robustness. And a lot that was learned for simpler problems contributed to the much greater understanding that has been achieved in the more complex problems. Robust thinking is everywhere; it is just a different sort of thinking than might have been expected 40 years ago. Modern researchers are seeking deeper understanding from the geometry of statistical models, looking at missing data problems, giving serious attention to the diagnostic use of robustness, and addressing robustness as a point of view rather than the answer to all ills.

Let me end by returning to my colleague's book, *Robustness* (Hansen and Sargent 2008). It is a tour de force by two top econometricians, who summarize a large body of work, including their own, to develop the use of a robust control theory for analyzing economic decision making, a theory that acknowledges the real potential for model misspecification. The picture on the cover represents an evil agent, one that can justify a use of minimax criteria in certain realms of game theory, even when in most ordinary statistical problems we have discarded it as too pessimistic. But the designer of their cover has done more than cast De La Tour's scene in mottled shades: it has

been seriously trimmed to fit, and trimming as we know is not always warranted. In fact, a major character has gone missing, one whose presence changes the dynamics—the wealthy young man can now be seen as the victim of the other three who conspire against him! (Figure 2). Sometimes trimming in the name of Robustness can eliminate within-sample variation that is crucial for the accurate assessment of evidence.

My colleague's book promises a macroeconomics for the modern day. And if it cannot answer all questions, it presents a most appealing start. The 21st-century study of robustness may lack the unrealistic expectations that we had in 1970, but the brightness of its promise is no less, and the task, while difficult, has now a generally recognized importance that justifies the effort.

[Received August 2010. Revised October 2010.]

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton: Princeton University Press. [279]

Bickel, P. J. (1976), "Another Look at Robustness: A Review of Reviews and Some New Developments," *Scandinavian Journal of Statistics*, 3, 145–168. [280]

Box, G. E. P. (1953), "Non-Normality and Tests on Variance," *Biometrika*, 40, 318–335. [277]

Hampel, F. R. (1968), "Contributions to the Theory of Robust Estimation," Ph.D. dissertation, University of California Berkeley. [278]

——— (1997), "Some Additional Notes on the 'Princeton Robustness Year'," in *The Practice of Data Analysis: Essays in Honor of John W. Tukey*, eds. D. R. Brillinger and L. T. Fernholz, Princeton: Princeton University Press, pp. 133–153. [279]

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based Upon Influence*, New York: Wiley. [279]

Hansen, L. P., and Sargent, T. (2008), *Robustness*, Princeton: Princeton University Press. [277,280]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101. [278,280]

——— (1972), "Robust Statistics: A Review," *Annals of Mathematical Statistics*, 43, 1041–1067. [278]

——— (1996), *Robust Statistical Procedures* (2nd ed.), Philadelphia: SIAM. [280]

Jaeckel, L. A. (1969), "Robust Estimates of Location," Ph.D. dissertation, University of California Berkeley. [278]

Koenker, R., and Bassett, G., Jr. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [279]

Stigler, S. M. (1973), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920," *Journal of the American Statistical Association*, 68, 872–879. Reprinted in *Studies in the History of Statistics and Probability*, Vol. 2, eds. M. G. Kendall and R. L. Plackett, London: Griffin, pp. 410–417, 1977. [278]

——— (1977), "Do Robust Estimators Work With *Real* Data?" *The Annals of Statistics*, 5, 1055–1098. [279]

——— (1997), "Daniel Bernoulli, Leonhard Euler, and Maximum Likelihood," in *Festschrift for Lucien LeCam*, eds. D. Pollard, E. Torgersen, and G. Yang, New York: Springer-Verlag, pp. 345–367. Revised in *Statistics on the Table*, Cambridge MA: Harvard University Press, 1999. [278]

Tukey, J. W. (1960), "A Survey of Sampling From Contaminated Distributions," in *Contributions to Probability and Statistics*, eds. I. Olkin et al., Stanford: Stanford University Press. [277]