The Annals of Statistics
1977, Vol. 5, No. 6, 1055–1098

DO ROBUST ESTIMATORS WORK WITH *REAL* DATA?[1]

BY STEPHEN M. STIGLER

University of Wisconsin at Madison

Most studies of robust estimators of location parameters have relied upon mathematical theory, computer simulated data, or a combination of these. This paper presents a comparison of the performances of eleven estimators using real data sets. Twenty sets of data from 1761 determinations of the parallax of the sun, from 1798 measurements of the mean density of the earth, and from circa 1880 measurements of the speed of light, are employed in the study, with the current values of these physical constants being compared with the estimators' realized values. We find that light trimming provides some improvement over the sample mean, but that the sample mean itself compares favorably with many recent proposals. The bias and nonnormality of the data sets is considered, and the data sets are presented and discussed in an appendix.

1. Introduction. Phoney real data (or, if you prefer, genuine phoney data) has played an important role in the development of statistical theory and practice for some time. Well before the first time an electronic computer spewed forth a thousand pseudo-random numbers in a simulation study, tables of "random" numbers compiled by L. H. C. Tippett and by H. Wold had been used by mathematical statisticians for the study of sampling distributions in situations too complex for analytical treatment [26, 27]. As the costs of computation have declined in recent years, the use of the computer for the production of pseudo-random numbers has increased, to the point where the volume of synthetic data produced in universities and research laboratories may even surpass the Census Bureau's production of the real thing.

The advantages of the use of simulation, particularly in robustness studies, have been clear for over 40 years, ever since E. S. Pearson pioneered its use in a series of papers in *Biometrika* about 1930. By suitable transformation of pseudo-random numbers, an investigator can mimic a sample from any mathematically definable probability distribution; he is not constrained by problems of analytic tractability. He therefore has great flexibility in the specification of distributions (and can choose some which may serve as suggestive of distributions of "real" data), and he has the advantage of knowing exactly what mechanism produced his "data" and hence can easily evaluate the performance of statistical procedures for these "data." If an estimator is designed to estimate the mean of a population, and its value is calculated for simulated samples known to have come from

Received November 1975; revised December 1976.

[1] This research was supported by the National Science Foundation under Grant No. SOC 75-02922.

AMS 1970 *subject classifications*. Primary 62G35; Secondary 62-02.

Key words and phrases. *M*-estimators, trimmed means, simulation, Monte Carlo, median, bias, adaptive estimators, skewness, kurtosis.

a Student's t distribution with 5 df, symmetric about zero, then the performance of the estimator can be judged on the basis of how distant the numbers it gives are from zero.

The principal disadvantage of such simulation is that no matter how clever the investigator is in his choice of specifications for sampling distributions, there is no guarantee that the pseudo-samples he generates are actually representative of real data. Indeed, most simulation studies of the robustness of statistical procedures have concentrated on a rather narrow range of alternatives to normality: independent, identically distributed samples from long-tailed symmetric continuous distributions. But why should real data not be expected to be correlated, biased, asymmetric, heterogeneous and exhibiting some discreteness (or granularity)? The formidable difficulties to be encountered by any attempt to deal adequately with these possibilities in a study using simulated data suggest that an alternative be tried: why not evaluate the performance of statistical procedures with *real* data? It is the aim of this paper to do just this for a variety of robust estimates of location.

Real data have appeared before in papers on robustness, but only for purposes of illustration. For example, Hampel [10] evaluates a variety of robust estimators for the Cushny and Peebles 1904 data [25] on the effect of soporific drugs, but without specific information on the actual effects of the drugs the comparison is indeterminate: we do not know whether excluding (or discounting) the outlier helps or harms the accuracy of the estimate. In the present study we have taken an approach using historical data sets which largely overcomes this difficulty, and permits the use of real data for an objective comparison of robust estimators.

2. The present study. As already mentioned, real data can exhibit many characteristics not allowed for in most simulation-based robustness studies. An unforeseen bias may be present. Despite attempts at independent replications in collecting the data, a serial correlation or a time trend may exist. Tied values are likely, a denial of continuity. The distribution may be asymmetric, and this asymmetry may be related to the bias in a perverse manner which suggests that elimination of outliers is *not* beneficial.

In order to attempt a study in which these various factors were allowed to enter in a mix appropriate to the real world, data sets were sought from a variety of sources which could satisfy the following three requirements: (1) The integrity of the data set must be above suspicion: the values must be reported as they occurred, and recorded in their entirety—not after a prior screening to eliminate values not in accord with the experimenter's prejudices. That is, the actual data set as it was encountered by the experimenter must be available. (2) The data must be measurements of a well-defined physical quantity which, while possibly not well determined at the time the measurements were made, can be assumed to be known today to a degree of accuracy that corresponds to certain knowledge in comparison with the accuracy of the older measurements.

Furthermore, the definition of the quantity today must be the same as the definition of the quantity at the time the measurements were made, otherwise the notion of bias becomes ambiguous. (3) The experiment must have been conducted as an honest attempt to learn about nature from a standpoint of relative ignorance about the quantity measured. If not, the experimental procedure might itself be affected by preconceptions or biases not present in anticipated applications.

A collection of data sets satisfying these conditions would permit a comparative study of robust estimators that would have the advantages of a simulation study without the drawbacks. The data would *be* real data from real experiments, and not just illustrative of some investigator's often very narrow view of what real data should be like. But, as in a simulation study, the most important aspect of the mechanism which generates the data, the "parameter of interest," can now be assumed known, and estimators can be evaluated by their actual, not hypothetical, performance. In addition, by an examination of the data sets some indication of the characteristics of real data (bias, nonnormality, etc.) can be obtained.

Unfortunately, and this is the major drawback of the present approach, satisfactory data sets are rare. Several potentially useful sources have been ruled out by the requirements. Contemporary laboratory experiments could hardly be characterized as having been carried out in ignorance of the quantity measured, at least in cases where the true value of the quantity is known with relative certainty in comparison with the accuracy of the measurements. Another source, subsampling from large data sets to estimate the mean of the data set, is really a variation on ordinary simulation, one whose freedom from bias is artificial even if the values sampled are genuine "real data." Rather, our aims seem to be well served by the use of historical data from the early years of quantitative experimentation, and completely documented data sets of the type sought are not common.

The present study is based on data from 18th century attempts to determine the distance from the earth to the sun and the density of the earth, and 19th century attempts to determine the speed of light. It would have been desirable to include a much broader spectrum of real data, and attempts were made to do this. Early experiments to determine the atomic weights of elements were investigated, but it was found that then (as now) chemists were highly selective in their decision as to which measurements to report. Other possibilities which have not yielded useful data include experiments to determine the speed of sound, and early work of Maxwell and Boltzmann on the kinetic theory of gases. I hope that others will find acceptable sources of data in other fields, but the present collection of 20 data sets of size approximately 20 each, and 4 larger sets, seems sufficient to advance some preliminary and tentative conclusions which are somewhat at variance with those reached in simulation studies, including the Princeton study [2].

The data sets we employ in this study are all taken from famous experiments, and all seem to satisfy the criteria presented above. In addition, they present us with an interesting chance to assess the role of statistics in science: all of the experiments considered were designed to answer important questions; would the scientists involved have fared better with the sole additional advantage of a statistical technique from one or two centuries later? Of the twenty basic sets, eight (Table 4) come from James Short's analysis of observations made in 1761 of the transit of Venus, an event which furnished the first reliable estimate of the mean distance from earth to sun. Nine (in Tables 5, 6, 7) come from the first experiments which succeeded in determining the speed of light with precision, experiments performed in 1879–1882 by A. Michelson and by S. Newcomb. And three (Table 8) are based on Cavendish's 1798 investigation of the mean density of the earth. These sets and the manner in which they were handled are discussed in detail in an appendix.

In all cases, it was felt that the data sets could be trusted to be a complete record of the investigator's measurements. The investigators would frequently discard or discount outlying values, but these were duly reported. In all cases, a serious attempt was made to deal with the data sets as the original investigator would have, given the modern estimation procedures. The order of the values was not disturbed, and, for example, where Michelson had converted recorded times to velocities before averaging and Newcomb had averaged times before converting to velocities, we also have dealt with Michelson's data as velocities and Newcomb's as times. In the case of the speed of light data, the original investigators' own correction factors have been employed to convert the present value for the speed of light in vacuum (299,792.5 km/sec) to "true values" for the speed of light in air. The only "tampering" done has been to divide four large data sets into smaller sets of approximately 20 measurements each, to permit a more direct comparison with the results of the large simulation study [2] which focused on sample size 20. The original data sets, however, have been included as Data Sets 21–24 and analyzed separately as "large samples." The division into smaller sets was accomplished at natural breaking points or at intervals of twenty measurements if no natural break existed.

3. The estimators considered in the study. The choice of which estimators to include in a study of this type is a difficult one. It would seem desirable to include all candidates that have proved themselves worthy of consideration in other contests, together with such new proposals as are thought likely challengers to the older winners of the "Most Robust Estimator" title. Yet if no more exacting standard is imposed, there is the risk that the field of entrants will grow beyond reasonable bounds.

For two reasons, it was decided to keep the number of entrants in this study as small as possible, consistent with the desire to provide a fair representation of the most important types of estimators. The first reason for this limitation

was economy. Without a large team of workers and an even larger government grant, it would be impossible to rival the complete coverage the 1971 Princeton study gave to the list of estimators then available. Rather, it was hoped that an informative study could be carried out on a shoestring budget, a study that at least in principle could be completed by one individual on a hand calculator. (Actually, to simplify calculations and to take advantage of the programs in [2], a computer was used for most of the study. The total cost for machine computation was under $ 10.)

A second reason for limiting the size of the study was to attempt to reduce the size of the selection effect. Any such study, whether it used simulated data or real data, must base its comparisons upon a finite number of situations, and if excessive flexibility is allowed in the class of estimators there is a real (and perhaps unquantifiable) danger that an estimator selected as "best" for the given situations would perform miserably in others. For example, most of the sample size 20 comparisons in the Princeton study were based on less than 20 population distributions (all symmetric and heavy tailed), and involved consideration of 65 basic estimators (which provided information for up to 10, 465 estimators, see [2, page 28]). The limitation to 65 estimators was based on the specification of a large number of parameters, scale factors, weights and "constants" whose values were sometimes chosen after preliminary trials with some of the same population distributions (or even the same samples) used in the study. The extraordinarily large number of estimators considered would seem to at least raise the possibility that some of that study's conclusions might have been distorted by a selection effect, and that the excellent performances of the best estimators in that study would not be duplicated in another study with different (but qualitatively similar) distributions.

As the principal conclusions of this present study are based on only 20 data sets, and these data sets are not all independent (in fact three of them overlap considerably), our study is particularly vulnerable to charges that good performance is due to a selection effect rather than true merit. While this possibility cannot be denied, the following points should be considered: 1. Attention was limited to only 11 estimators, selected before the study was performed. 2. The estimators did not require the specification of parameters based on the same data sets used in the study. 3. No study can be entirely free of this effect.

Eleven estimators were selected for inclusion in the study. Ten were selected as among the best of the types of estimators being most prominently considered by current researchers (M-estimators, linear functions of order statistics, adaptive estimators), and one (the "outmean") was selected as an estimator that would perform poorly for long-tailed distributions, with respect to which the performance of the others could be gauged. The eleven estimators included in the study were (letting $X_1 \leqq \cdots \leqq X_n$ denote the ordered measurements):

1., 2. The *mean* $\bar{X}$ and the *median* $\tilde{X}$, two ancient and popular favorites without which no study would be complete.

3., 4., 5. *The* 10%, 15%, *and* 25% *trimmed means.* The $100\alpha\%$ trimmed mean was taken as defined by

$$\bar{X}_\alpha = \frac{(p\{X_{[\alpha n+1]} + X_{n-[\alpha n]}\} + \sum_{i=[\alpha n+2]}^{n-[\alpha n+1]} X_i)}{n(1-2\alpha)},$$

as in [2, page 7], where $p = 1 + [\alpha n] - \alpha n$.

6., 7., 8. *Huber's* P15, *Andrews'* AMT, *and Tukey's Biweight.* These are all versions of what have come to be known as M-estimators, and are found as a solution (T) to the equation

$$\sum_{i=1}^{n} \phi\left(\frac{X_i - T}{s}\right) = 0, \tag{1}$$

where s is an estimate of spread (here a multiple of the median absolute residual about the median, or about an earlier value of the estimator if this equation is solved by iteration), and ϕ is a function to be specified.

Huber P15 (see [2, page 13]) is a "one-step" M-estimator, where $\phi(u) = \min(k, \max(-k, u))$, $k = 1.5$, s is the median absolute residual about the median and only one step in the iteration is performed.

Andrews' AMT (*or* SINE) (see [1, page 524], or [2, page 15]) is an M-estimator where

$$\begin{aligned} \phi(u) &= \sin(u/2.1) && |u| < 2.1\pi \\ &= 0 && \text{otherwise,} \end{aligned}$$

and s is the median absolute deviation about a previous value of T (starting at the median), revised every third iteration.

Tukey's Biweight (see [4, page 15]) is an M-estimator where $\phi(u) = uw(u)$,

$$\begin{aligned} w(u) &= (1 - u^2)^2 && |u| \leqq 1 \\ &= 0 && |u| > 1. \end{aligned}$$

It was calculated by the iteration

$$T_{i+1} = \frac{\sum_j w\left(\frac{X_j - T_i}{cS_i}\right) X_j}{\sum_j w\left(\frac{X_j - T_i}{cS_i}\right)}$$

performed six times starting at the median, with $c = 6.0$, and $S_i = \text{median}\{|X_j - T_i|\}$.

The latter two of these estimators were suggested for inclusion in the present study by D. F. Andrews and J. W. Tukey, before they were informed of the nature of the study. Huber P15 was selected for inclusion on the basis of its generally good performance in [2].

9. *Edgeworth* is a weighted average of the lower quartile, the median, and the upper quartile, with weights in proportions 5 : 6 : 5 (see [24], where the weights are mistakenly described as being in proportions 5 : 7 : 5). This estimator

was proposed by Edgeworth [9] in 1893, and is similar to estimators studied more recently by Mosteller, Tukey, and Gastwirth (see [2]). The version used here took the quartiles as defined in [2, page 18], i.e., the "hinges," and was calculated as a modification of the trimean [2, page 8].

10. *Outmean*, or $\bar{X}^c_{.25}$, is essentially the average of those measurements discarded in the computation of $\bar{X}_{.25}$. It may be taken as defined by

$$\bar{X}^c_{.25} = 2\bar{X} - \bar{X}_{.25} .$$

11. *Hogg's* T_1 is an adaptive estimator proposed in [11] and suggested for inclusion in the present study by R. V. Hogg before he was informed of the nature of the study. It is defined by

$$\begin{aligned} T_1 &= \bar{X}^c_{.25} && \text{if } Q < 2.0 \\ &= \bar{X} && \text{if } 2.0 \leqq Q \leqq 2.6 \\ &= \bar{X}_{3/16} && \text{if } 2.6 < Q \leqq 3.2 \\ &= \bar{X}_{3/8} && \text{if } 3.2 < Q \end{aligned}$$

where Q is a measure of the "weight in the tail" of the sample given by

$$Q = \frac{U(.05) - L(.05)}{U(.5) - L(.5)} ,$$

where $L(\alpha)$ and $U(\alpha)$ are averages of the lower and upper $100\alpha\%$ of the X_i's, respectively.

Where possible, the estimates were calculated by use of the programs given in an appendix to [2] or minor modifications of these programs.

4. Comparisons. The values of each of the eleven estimators were calculated for each of the 24 data sets. The choice of a method for comparison of the relative performance of the estimators is akin to the specification of a loss function in a multiple decision problem. The choice is made difficult by the unimportance of analytic tractability (we are not automatically led to squared error) and by the diversity of the problems considered: is an error of a half-second of a degree in a determination of the sun's parallax (which might make a 5,000,000 mile difference in the estimated distance from earth to sun) more or less serious than an error of 200 km/sec in an estimate of the speed of light?

We have followed most other studies (but not [5]) in treating the problem as a contest or tournament between rival estimators, and except for changes of scale, treating all situations identically. To this end, to allow each data set an equal opportunity to influence the outcome, it is desirable that whatever measure of performance is adopted should be invariant with respect to changes in the origin and scale of the unit of measurement. Thus the conclusions of the study will not be affected by Short's use of seconds of a degree (rather than decimal fractions of a degree) as a unit of angular measurement, Cavendish's use of water as a standard for the measurement of density, or the possibility that Michelson might have coded his data by subtracting 299,000 km/sec (as we have done).

While a number of different indices of performance have been considered, only two have been selected for presentation. One of the two indices was selected as indicative of the average (across data sets) magnitude of the error of estimation for the estimator, the other as reflective of the rank of the estimator's performance among the eleven considered. Comparisons were made separately for small samples (data sets 1–20, ranging in size from 17 to 29 with all but three being between 17 and 23) and large samples (data sets 21–24, ranging in size from 53 to 100).

The index of relative error. This index was designed to measure the absolute magnitude of an estimator's error relative to the sizes of the errors achieved by other estimators for the same data set. For each data set (say data set j) the quantity

$$s_j = \tfrac{1}{11} \sum_{i=1}^{11} |\hat{\theta}_{ij} - \theta_j| \tag{4.1}$$

was computed, where θ_j denotes the "true value" for the jth data set, and $\hat{\theta}_{1j}, \cdots, \hat{\theta}_{11,j}$ are the values of the eleven estimators for the jth data set. Thus s_j is the average absolute error realized for the jth data set. The performance of estimator i for data set j was then measured by its relative error,

$$e_{ij} = |\hat{\theta}_{ij} - \theta_j|/s_j \,. \tag{4.2}$$

The computed values of e_{ij} are given in Table 9. A value of e_{ij} less than (greater than) one means that for data set j, estimator i made a smaller (larger) error than the average error for the eleven estimators. The "index of relative error" RE(i) for estimator i was then computed by averaging across data sets:

$$\text{RE(i)} = n^{-1} \sum_{j=1}^{n} e_{ij} \,. \tag{4.3}$$

TABLE 1

The indices of relative error and of relative rank RE(i) (4.3) *and* RR(i) (4.5), *computed separately for small and large data sets (together with measures of spread* SE(i) (4.4) *and* SR(i) (4.6) *between data sets, in parentheses). Small values of* RE *and* RR *connote good performance*

	Relative Error (RE)		Relative Rank (RR)	
	Small Samples	Large Samples	Small Samples	Large Samples
Mean	.931 (.20)	.924 (.19)	4.9 (3.2)	6.0 (4.6)
Median	1.149 (.28)	1.152 (.18)	7.1 (4.1)	8.1 (3.8)
Edgeworth	1.018 (.08)	.945 (.07)	6.4 (3.2)	3.9 (1.5)
Outmean	1.038 (.58)	.774 (.50)	5.1 (4.8)	6.0 (5.8)
10% Trim	.916 (.20)	.944 (.06)	4.6 (2.2)	4.5 (2.4)
15% Trim	.983 (.10)	.991 (.04)	6.0 (1.7)	5.5 (0.6)
25% Trim	1.039 (.08)	1.073 (.12)	6.8 (3.0)	6.1 (3.5)
Huber P15	.922 (.20)	.985 (.05)	5.3 (2.8)	5.5 (1.7)
Andrews AMT	.966 (.14)	1.032 (.13)	6.2 (2.5)	6.0 (3.4)
Tukey Biweight	1.023 (.13)	1.097 (.17)	6.6 (3.1)	7.0 (3.9)
Hogg T1	1.014 (.07)	1.084 (.13)	6.8 (2.5)	7.4 (2.5)

Table 1 gives the values of RE(i) separately for small and large data sets; the s_j are given in Table 9.

(In data sets 3 and 5 the mean and outmean fell on the opposite side of the "true value" from the other estimates, otherwise all estimates fell on the same side of the "true value." Thus the actual errors $\hat{\theta}_{ij} - \theta_j$, though not given here, can be easily recovered from Table 9.)

The numbers in parentheses are descriptive measures of the spread of the e_{ij}'s:

$$\text{SE(i)} = \{(n-1)^{-1} \textstyle\sum_{j=1}^{n} (e_{ij} - \text{RE(i)})^2\}^{\frac{1}{2}} . \tag{4.4}$$

A large value of SE(i) (such as that for the outmean) represents considerable variation in the estimator's performance from data set to data set, a small value reflects consistent performance.

The index of relative rank. The second index selected ignored the actual errors of estimation and considered only their ranks for each data set. For each data set j the rank r_{ij} of estimator i was found ($r_{ij} = 1$ for the estimator with smallest error $|\hat{\theta}_{ij} - \theta_j|$ for data set j, $r_{ij} = 11$ for the estimator with largest error $|\hat{\theta}_{ij} - \theta_j|$ for data set j). Tied estimators were given the average of the ranks tied for. The index was then computed by averaging across data sets:

$$\text{RR(i)} = n^{-1} \textstyle\sum_{j=1}^{n} r_{ij} . \tag{4.5}$$

These are given in Table 1. Again, the numbers in parentheses are descriptive measures of the spread of the r_{ij}'s across data sets:

$$\text{SR(i)} = \{(n-1)^{-1} \textstyle\sum_{j=1}^{n} (r_{ij} - \text{RR(i)})^2\}^{\frac{1}{2}} . \tag{4.6}$$

TABLE 2

Qualitative groupings of eleven estimators based upon the indices of relative error (4.3) *and relative rank* (4.5)

	Small Samples		Large Samples	
	Relative Error	Relative Rank	Relative Error	Relative Rank
Best	10% Trim Huber P15 Mean	10% Trim Mean	Outmean Mean 10% Trim Edgeworth	Edgeworth 10% Trim
Good	Andrews AMT 15% Trim	Outmean Huber P15	Huber P15 15% Trim	15% Trim Huber P15
Average	Hogg T1 Edgeworth Tukey Biweight Outmean 25% Trim	15% Trim Andrews AMT Edgeworth Tukey Biweight 25% Trim Hogg T1	Andrews AMT 25% Trim Hogg T1 Tukey Biweight	Mean Outmean Andrews AMT 25% Trim
Poor	Median	Median	Median	Tukey Biweight Hogg T1 Median

The results of these calculations lend themselves to a rough but revealing grouping, which is presented in Table 2. This grouping suggests three possibly surprising conclusions. First, regardless of the index considered or the size of the samples, the simple 10% trimmed mean—the average of the 80% of the sample remaining after trimming 10% on either end—emerges as one of the best of the estimators considered, with the sample mean itself a close competitor. The second surprise is the generally weak performance of another old favorite, the sample median, an estimator often considered only slightly inefficient but highly robust and usually selected as a starting point for the calculation of the more complicated iteratively determined estimators. The third surprise is the generally mediocre performance of the estimators selected from among the best modern proposals for dealing with real data in a robust but efficient manner, a suggestion that the alternatives to an independent identically normally distributed sample that have been considered by modern workers (e.g. [2]) are too restricted or too exaggerated to reflect accurately "real" data. We shall comment further on this in Section 6.

These three generalizations—that a slightly trimmed mean is best, that the median is inefficient, and that modern estimators are not worth the time necessary to compute them—are open to contest even within the restricted confines of the present study. Two modern estimators (Huber P15 and Andrews AMT) do show some promise as estimators of location for real data. Huber P15 does particularly well: for normal distributions its performance is asymptotically equivalent to that of a 6.7% trimmed mean, and indeed its performance index behaves similarly to that of the 10% trimmed mean. Also, Edgeworth's simple estimator does quite well for the 4 large data sets, although its performance is only average for the smaller sets. The outmean, which usually comes at one extreme or another of the set of values of the estimates for a given data set, is highly variable in its performance (and even best in one category) and could not be considered as a reliable estimator. It is the only estimator whose index values are significantly affected (for the worse) if a measure such as $\sum_j e_{ij}^2$ or $\sum_j r_{ij}^2$ is adopted which emphasizes large errors as disasters. The best that one can say for the poor median is that there *are* three data sets (15, 16, and 19) for which it is best.

Readers interested in trying other estimators on these data sets can do so using the values of s_j in Table 9. The differences between the calculated estimates and the true values, divided by the given s_j, will give a quick evaluation of the relative errors; the fact that the new estimator did not contribute to the s_j should make little difference. An index of relative error for the Hodges–Lehmann estimator $\text{med}_{i \leq j} (X_i + X_j)/2$ [see 2, page 25] was calculated in this manner and found to be .95 for the small data sets, only slightly larger than that of the sample mean.

It is interesting to evaluate the original scientists' determinations by this method. A comparison for the small data sets was not attempted, since the

scientists did not break some of the larger sets up as we have done, and no "experimenter's mean" is available in many cases. For the large data sets 21–24, however, the scientists did make determinations, involving informal outlier rejection, trimming and subjectively weighted averages. Using the values of s_j from Table 9, I find an index of relative error of .99 for these ad hoc procedures, about the same as the 15% trimmed mean. The largest error arose from Short's estimate of 8.55 for data set 22, where the "true value" is 8.798. Had Short used a mean (8.63) or 10% trimmed mean (8.57), he could have reduced his error from 2,500,000 miles to about 1,700,000 miles or 2,300,000 miles, looking at the problem as one of determining the distance from earth to sun.

5. The effect of bias. Readers of an early draft of this study have expressed reservations about a number of points, but comments have most frequently been focused on the possibility of bias in the choice of data sets, and the existence of systematic bias in the measurements themselves. On the one hand, it has been remarked that the data considered are high quality physical science data, collected by famous scientists, and likely to admit fewer abnormalities and irregularities than do data sets encountered in routine experimentation or in the social sciences. On the other hand, many of the experiments show evidence of systematic biases which may lead to difficulties in interpreting the results. To a degree these comments are inconsistent (how can data be of abnormally high quality *and* atypically biased?), but both points speak to important issues.

That the data sets are "high quality" cannot be denied. As we remarked in Section 2, the objectives of the study seem to require well-documented data collected by excellent scientists, but it does not follow *necessarily* that the data given here are more "gentle"—more normal or regular—than "average quality" data collected by "the average scientist." In fact, there are some reasons for suspecting the contrary to be true. In the cases of Cavendish, Michelson and Newcomb, the scientist was dealing with an experimental apparatus that was novel in its construction, and not familiar in its idiosyncrasies. A contemporary scientist, however, engaged in routine analyses and interested in a robust estimate, has likely been carefully instructed in the use of an experimental apparatus whose characteristics are well known. It is true that a routine analysis will be less carefully attended to than a pioneering one, but how these factors influence results on balance is a question we cannot answer with the evidence at hand. In the next section we shall consider the normality of our data sets in more detail, but the question as to whether or not these data are representative of data sets where one might actually wish to use a robust estimator, must be deferred to other investigations.

The other issue that has been consistently raised is the existence of systematic bias in the measurements. It is well known that measurements of physical constants are susceptible to systematic as well as random error; see Youden [29] for one discussion of this. These data sets also show evidence of this bias.

TABLE 3

The values of the coefficients of skewness $(\beta_1)^{\frac{1}{2}}$ *and kurtosis* (β_2), *of Hogg's Q, and of a coefficient of bias* $((\bar{X} - \theta)/s)$ *for* 24 *data sets. The corresponding population values of* $(\beta_1)^{\frac{1}{2}}$, β_2, *and Q for a normal population are* 0, 3, *and* 2.58

Data Set	Size n	$(\beta_1)^{\frac{1}{2}}$	β_2	Q	Bias	Data Set	Size n	$(\beta_1)^{\frac{1}{2}}$	β_2	Q	Bias
1	18	0.53	3.20	2.90	−0.23	13	20	0.34	1.97	1.96	1.99
2	17	−1.22	6.38	4.07	−0.48	14	20	−1.28	4.87	3.57	1.40
3	18	1.38	4.84	3.06	0.05	15	20	−0.04	1.90	1.98	1.45
4	21	0.82	2.99	2.62	−0.29	16	20	0.64	2.97	2.66	1.79
5	21	1.17	3.61	2.70	0.12	17	23	0.37	4.03	3.02	0.43
6	21	0.84	3.25	2.75	−0.37	18	23	0.14	2.31	2.31	−0.18
7	21	0.32	3.67	3.22	−1.72	19	29	−0.44	3.10	2.57	−0.31
8	21	0.39	4.06	3.49	−0.84	20	29	0.03	2.44	2.36	−0.17
9	20	−2.82	11.00	4.69	−0.64	21	53	−0.39	6.40	3.52	−0.24
10	20	0.15	2.08	2.10	−0.87	22	63	0.66	3.22	2.78	−0.21
11	26	0.11	4.05	3.20	−1.13	23	66	−4.49	29.40	4.34	−0.63
12	20	−0.89	3.15	2.66	1.66	24	100	−0.02	3.26	2.66	1.49

Table 3 presents calculated values of a coefficient of bias, the difference between the sample mean and the "true value" divided by the sample standard deviation. We see that the bias is most pronounced in Michelson's speed of light experiments (sets 12–16, 24), and that in 50% of the data sets $\bar{X}$ differs from θ by less than half an estimated standard deviation of a single measurement. (If $n = 20$ and the data are distributed $N(\theta, \sigma^2)$ then $P(|\bar{X} - \theta|/s \geqq .5) = P(|t(19 \text{ df})| \geqq 5^{\frac{1}{2}}) \cong .04$). Thus the existence of biases in these sets averaging about half the standard deviation of a single measurement is plausible. That these systematic biases lead to misleading confidence statements has long been known; indeed Newcomb's own analysis of set 23 included a three-fold increase in his estimate of the probable error $(= .64\sigma)$ to allow for possible "constant errors" [16, page 201–2].

The questions we must face are, what is the effect of this bias on our comparisons, and can or should an attempt be made to eliminate this bias? The exact effect of the bias is hard to quantify, but at least qualitatively one can see that the predominate effect is to dilute the comparisons. The error of estimation which enters into the index RE(i) has two components, random and systematic error. If the populations sampled were all symmetric, then since all estimators considered are translation invariant, symmetric functions of the data, the systematic component for each data set would be the same for all estimators. If squared error (rather than the more robust absolute error) had been employed, one would then find (via the familiar "mean-square-error = variance + bias-squared") that the bias had added equally to the expected values of both numerator and denominator of e_{ij}, and severe bias would render all estimators equally inaccurate, biasing e_{ij} toward 1.0. The same should be at least approximately true for the measure actually used, and indeed, in Table 9 we see that the more

biased sets tend to produce relative errors near 1.0 for all estimates. The index RE(i) was in fact recomputed for the small samples without the more biased speed-of-light sets 12–16, and the results were as predicted: the differences already noted were actually increased, with no change in order other than the interchange of the positions of Edgeworth and Tukey Biweight. A similar effect on the index RR(i) would be expected.

If the populations sampled are not symmetric, the situation is not so simple. In this case the estimators may have different expected values, and thus different biases. If the ratio of the largest bias to the smallest bias is small, the effect would be qualitatively the same as the symmetric case, but in some situations this may not be true. It is even possible that a systematic error would tend to be related to the asymmetry in such a manner as to penalize overly harsh trimming.

One example of how the heavy tail could be in the "right" direction and counterbalance systematic error, is the observation of transits of Venus. In this case, a major source of error was the "black drop" effect. As Venus passed into the disc of the sun an optical illusion not anticipated by the observers was seen: Venus seemed to pull a "black drop" of space with it into the sun's interior, and by the time Venus actually was seen to break contact with space and be isolated in the sun's interior, it had in fact passed some distance beyond the real contact point, where the two discs were tangent. The magnitude of this effect differed at different observation posts, and observers dealt with it individually as best they could, often guessing at the point of tangency. These guesses could easily produce a "heavy tail" that would counterbalance the black drop bias, and in some data sets this may have occurred. Seven of Data Sets 1–8 are positively skewed; six are negatively biased.

This example points up one of the hazards of attempting to sharpen the comparisons of estimators by reducing systematic error through the introduction of current corrections for factors overlooked (or dealt with incorrectly) by the early scientists. If the data sets are as samples from symmetric populations, and a careful reading of the experimental procedure by a present day expert indicates a systematic error that can be eliminated, then the differences between estimators could in principle be magnified without unfairly changing the ranking of performance. The actual ranking may change due to the small number of data sets employed, but the expected ranking would not be affected. But with even slight asymmetry this would not be true, and the introduction of corrections not known to the original investigator might have a major effect on both actual and expected performance; different present day experts might make different, equally defensible, corrections and arrive at totally different results.

Another hazard in attempting to eliminate systematic errors through hindsight is that we may be misled as to the actual difference in performance between estimators. One of the lessons of this study is that even the greatest scientists, exercising every ounce of their ingenuity, are unable to eliminate all bias. As

we have seen, a systematic error of about half the standard deviation of a single measurement remains. An attempt to eliminate this bias, even if successful, may give a quite misleading view of the estimators' relative efficiencies for real data sets.

For these reasons, and because we lacked an unambiguous scheme for proceeding otherwise, no attempt has been made to impose anachronistic scientific expertise on these early data sets. Rather, the point of view adopted has been one of treating the data as the original scientist might have, given the robust estimator (but no additional twentieth century knowledge). The estimators may have been designed to determine efficiently the center of a symmetric distribution, but they will be employed in less than ideal situations. Even when these estimators are used for purposes other than the determination of objectively defined parameters, say for comparative studies, asymmetries and unknown (and unknowable) biases will be present. There can be no guarantee that the situations studied here are representative of current applications, but that is not adequate reason for basing our assessments on unrealistically idealistic assumptions.

6. Are real data normal? An often quoted remark which Poincaré [19, page 171] attributed to "Lippmann" (probably the French physicist Gabriel Lippmann) can be roughly translated as follows: "Everyone believes in the normal law, the experimenters because they imagine it a mathematical theorem, and the mathematicians because they think it an experimental fact." The past few decades of statistical research have seen a near reversal of these sentiments. Some experimenters believe that the assumption of normality should be dispensed with because of the development of techniques (including nonparametric techniques) which do not require it, while mathematical statisticians develop "robust" techniques which would be appropriate for the heavier tailed distributions that they believe (perhaps on skimpy or nonexistent evidence) to occur commonly in practice.

That some real data sets with symmetric heavy tails do exist, cannot be denied. One excellent example is a set of 684 residuals based on observations of transits of Mercury which Simon Newcomb presented in 1882 [15, see also 24]. But, while many isolated examples of both heavy and light tails could be cited, the frequency with which heavy tails occur, the actual heaviness of the tails of real data, and the seriousness of their impact upon the performance of statistical procedures, do not seem to have received adequate systematic study. No such systematic study will be attempted here, but it is nonetheless interesting to consider the question of just how much the uncensored Data Sets 1–24 deviate from normality, according to some common measures.

Table 3 presents the values of the sample coefficients of skewness and kurtosis, defined by $(\beta_1)^{\frac{1}{2}} = \mu_3/\mu_2^{\frac{3}{2}}$ and $\beta_2 = \mu_4/\mu_2^2$, where $\mu_k = n^{-1}\sum (X_i - \bar{X})^k$ is the sample kth moment. The values of Hogg's measure of tail weight Q (see Section 3 and [11] for its definition) are also given; the value of the correspondingly

defined population value $g(.05)/g(.5)$, with $g(\alpha) = f(F^{-1}(\alpha))/\alpha$, is 2.58 for normal distributions, with larger values representing heavier tails.

While the values given in Table 3 do show some evidence of heavy tails or skewness in some of the data sets, they can hardly be taken as indicative of the disasterously heavy (even Cauchy-like) tails envisaged by the more pessimistic of the modern mathematical statisticians. (See, for example, the emphasis upon the Cauchy and similar distributions in [2].) Among the small sample data sets, the most extreme values are those of set 9($(\beta_1)^{\frac{1}{2}} = -2.82$, $\beta_2 = 11.00$) which are due in part to a single low value (-44 in Table 5), a value that Newcomb incidentally gave no weight to in his own analysis. Without this value, $(\beta_1)^{\frac{1}{2}} = 1.44$, $\beta_2 = 6.40$. The most extreme values among large sets occur in set 23, where they can be seen to be due to the combination of Data Sets 9–11, with means 22, 29, 28 and standard deviations 18, 5, 5. A plausible explanation for this shift in mean and decrease in variance after set 9 is that Newcomb was becoming more familiar with his apparatus. It could be argued then that this time trend in variance which accounts for the high kurtosis in the combined sample is not appropriately modelled as a random sample from a heavy-tailed distribution.

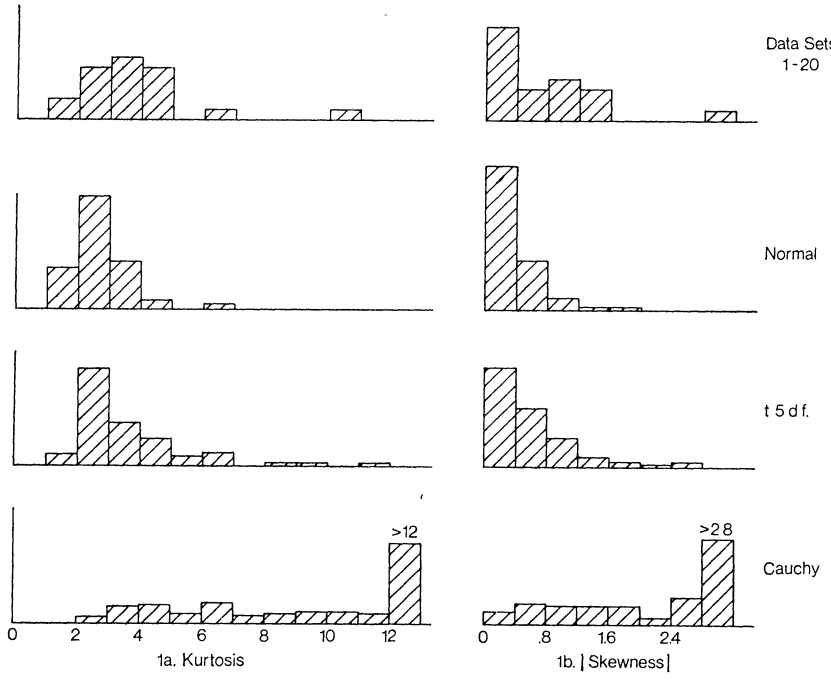

FIG. 1. Empirical frequency distributions of the sample coefficient of kurtosis and of the absolute value of the sample coefficient of skewness from Data Sets 1–20, and from 100 pseudorandom samples of size 20 from the normal distribution, Student's t distribution (5 df), and the Cauchy distribution.

It would of course not be accurate to treat Data Sets 1–20 as a "random sample" of small sets of real data, but the behavior of the sample statistics $(\beta_1)^{\frac{1}{2}}$ and β_2 in random samples of size 20 provides a useful reference with which to compare the values of Table 3. Figure 1 presents histograms of the values of $|(\beta_1)^{\frac{1}{2}}|$ and β_2 from sets 1–20, and values obtained by generating 100 independent pseudo-random samples of size 20 from each of the following distributions: (1) Normal (0, 1), (2) Student's t with 5 df, (3) Cauchy (t with 1 df).

Table 3 and Figure 1 suggest that the data sets considered tend to have slightly heavier tails than the normal, but that a view of the world through Cauchy-colored glasses may be overly-pessimistic. Probability plots (not presented here) show no remarkable abnormalities. Outliers are present in small quantities, but a small amount of trimming (no more than 10%) may be the best way of dealing with them.

7. Conclusions. While the present study may be the first evaluation of modern robust estimators to rely on real data, the small number of data sets employed and the narrow range of fields they are selected from requires that any conclusions be only tentative. Nevertheless, these conclusions are in sufficient accord with common sense and statistical tradition, that we feel they can be embraced with at least a moderate degree of confidence.

We have found that real data do exhibit behavior somewhat different from that of the simulated data used in most robustness studies, and that this affects the consequent recommendations for choice of an estimator and assessments of the relative performances of estimators. The data sets examined do exhibit a slight tendency toward more extreme values than one would expect from normal samples, but a very small amount of trimming seems to be the best way to deal with this. The 10% trimmed mean (the smallest nonzero trimming percentage included in the study) emerges as the recommended estimator; the mean itself does rather well. The more drastic modern remedies for feared gross errors recommended in [2] lead here to an unnecessary loss of efficiency.

In a sense, this study is a vindication of a vague procedure recommended by Legendre in his original publication of the method of least squares in 1805, where he recommended the application of least squares after rejecting those observations whose errors "are found to be such that one judges them too large to be admissible." (See [24].) Edgeworth reached a similar conclusion in 1887, but put it more colorfully: "The Method of Least Squares is seen to be our best course when we have thrown overboard a certain portion of our data—a sort of sacrifice which has often to be made by those who sail upon the stormy seas of Probability." [8, page 269].

Acknowledgments. I am grateful to many individuals for their comments, criticisms and suggestions at several stages of this investigation. I would particularly like to thank David Andrews, George E. P. Box, Churchill Eisenhart, Frank Hampel, Robert Hogg, Peter Huber, Erich Lehmann, Robert R. Newton,

John Pratt, George J. Stigler and two referees. I would also like to thank Lien-Ju Chao for her assistance with the computation.

APPENDIX

The data sets employed in this investigation were drawn from sources which are not accessible to many readers, and sometimes appeared in formats that are not easily adaptable to this type of analysis. In order to provide full documentation for the paper, and for those readers who wish to perform alternative analyses (or test alternative estimators), we present both the data sets and brief descriptions of the experiments which produced them. The data are presented in the original time sequence, and where large data sets have been broken up or rearranged, this is explained in the descriptions. Thus a reader wishing to view Short's data as a two-way table, treat Michelson's data as a time series, or analyze Cavendish's in terms of reciprocals (as determinations of G), may do so.

A. *Short's determinations of the parallax of the sun.* By the early years of the 18th century, astronomers had fairly well determined the relative dimensions of the solar system, the relative distances between planetary orbits and between the planets and the sun. However, they lacked precise information on the *absolute* dimensions of the solar system, and were eager to determine even one such distance in miles, in particular the mean distance from the earth to the sun, from which all others could be found. Actually, the quantity that 18th century astronomers chose to pursue was the parallax of the sun, the angle subtended by the earth's radius, as if viewed and measured from the surface of the sun. From this angle and available knowledge of the physical dimensions of the earth, the mean distance from earth to sun (or astronomical unit) could be easily determined.

The astronomer Edmund Halley (1656–1742) is generally credited with having been the first to suggest that the parallax of the sun could be determined by observing a "transit of Venus," the apparent passage of the planet Venus across the face of the sun, as viewed from earth. If observers were dispatched to all corners of the globe from which this transit would be visible, and they carefully recorded their positions and the elapsed time of the transit (on the order of 5.5 hours), then each pair of observers would furnish one determination of the parallax of the sun.

Unfortunately for the implementation of Halley's plan, transits of Venus are quite rare, owing to the 3°36′ inclination between the orbits of Venus and Earth. The first recorded transit of Venus occurred on 1639 and was observed only in England; the next transits were due in 1761 and 1769, with later transits due in 1874, 1882, 2004 and 2012. By 1761, interest in the forthcoming transit was high, and observations were made at the Cape of Good Hope, and in Calcutta, Rome, Stockholm and most other European observatories. A good account of the transits of 1761 and 1769 can be found in Woolf [28]; and excellent discussion of the data generated by these transits is given by Newcomb [17].

The data we analyzed are from a contemporary analysis of the 1761 transit by James Short. Short [23] presented several different calculations of the parallax of the sun based on different sets of pairs of observations, data presented in Table 4.

The numbers are based on pairs of observations of the transit, and thus are usually not independent. Data sets 1, 2, 3, 7, 8 are each based on comparisons of observations at a single observatory with a long list of others. Data sets 4, 5, 6 come from a pairwise comparison of 7 observing stations with 9 others. Some readers may wish to analyze this as a 9×7 two-way classification; this format can be recovered by breaking the sets after each seventh number.

Data Set 21 consists of 1, 2, 3 together, and was analyzed by Short in a "robust" manner: he took the mean of all $n = 53$ determinations (8.61), then rejected all results differing from 8.61 by more than 1.00 and obtained the mean of the remainder (8.55), then rejected all results differing from 8.61 by more than .50 and obtained the mean of the remainder (8.57); finally he took the mean of 8.61, 8.55, 8.57 to obtain the sun's parallax as 8.58. He applied a similar analysis to Data Set 22 (consisting of 4, 5, 6 together) and obtained 8.55 again and similarly analyzed Data Sets 7 and 8 separately to obtain 8.56 and 8.57. We shall adopt as our "true value" that given by Woolf [28, page 197], namely 8.798. In a private communication, Robert R. Newton informs me that recent radar determinations would lead to a value of 8.794, but this slight difference would not have a significant effect upon the comparisons.

B. *Cavendish's determinations of the mean density of the earth.* Newton's law of gravitation states that the force of the attraction (f) between two particles of matter is given by the formula $f = Gmm'/r^2$, where m and m' are their respective masses, r the distance between their centers of gravity, and G is the gravitational constant, independent of the kind of matter or intervening medium. From the late eighteenth through the nineteenth centuries, a large number of experiments were performed in order to determine G. These experiments were usually designed to determine the earth's attraction of masses, and described as experiments to determine the mean density of the earth: If the earth is supposed spherical with radius R and g is the acceleration toward the earth due to gravity, then Newton's law becomes $G\Delta = 3g/(4\pi R)$, where Δ is the mean density (g/ccm) of the earth. Since g and R could be supposed known, determination of Δ could be viewed as equivalent to determination of G. A large number of these experiments are described in [3], [20], and [21].

Of all the early experiments, that of Cavendish [7], performed in 1798 using a torsion balance devised earlier by Michell, is generally considered the best. The completeness of his description of his experiments and the excellence of his methods are often described as an ideal example of scientific experimentation. Cavendish concluded his memoir by presenting 29 determinations of the mean density of the earth; these are presented in order in Table 8. A word of explanation of

our handling of these data is in order. After the sixth of these determinations, Cavendish changed his experimental apparatus by replacing a suspension wire by one that was stiffer. In his analysis of these determinations (which amounts to little more than taking means), Cavendish considered this change as potentially important, and we have followed him in this respect: Data Set 19 consists of all 29 determinations, while Data Set 18 consists of only those last 23 made with the stiffer wire (these 23 measurements are also presented by Brownlee ([6], page 223). To further complicate matters, Cavendish erred in taking the mean of all 29 by treating the value 4.88 as if it were in fact 5.88. This was first pointed out by Baily in 1843 [3], and was overlooked by Laplace in an early statistical analysis of these data [12]. As robust methods are supposed to be able to cope with gross errors, we have also analyzed the data (as Data Set 20) with all 29 determinations and 5.88 replacing 4.88. Cavendish presented the value 5.48 as the mean of the 29 (as well as of the 23) determinations. As corrected by Baily, this figure becomes 5.448; we shall follow the most recent (1974) *Encyclopedia Britannica* and take the "true value" as 5.517 (Macropedia, "Earth, Mechanical Properties of," **6** page 37). In a private communication, R. R. Newton has suggested an alternative value of 5.513 as more appropriate to Cavendish's experiment, but this slight change would have negligible effect.

C. *Michelson's and Newcomb's measurements of the velocity of light.* Verification of the fact that light travels at a finite velocity, and is not transmitted instantaneously as early scientists (including Kepler and Descartes) had thought, is generally credited to Ole Römer, who in 1676 made comparative measurements of the times of eclipses of Jupiter's satellites from two different relative positions of Earth and Jupiter. But another two centuries passed before the experiments of Michelson and Newcomb in 1879–1882 provided what are considered the first accurate determinations of the velocity of light in vacuum.

In 1849 and 1850, the French physicists Fizeau and Foucault had separately devised methods of measuring the velocity of light. Foucault's method, as refined and improved by Newcomb and Michelson, was the source of the more accurate subsequent determinations. Foucault's method (see [18]) consists in essence of passing light from a source off a rapidly rotating mirror to a distant fixed mirror, and back to the rotating mirror. The velocity of light is then determined by measuring the distances involved, the speed of the rotating mirror and the angular displacement of the received image from its source.

We have included, as Data Sets 9–17, 23, 24, the results of three of the best of the early experiments made with Foucault's method. In 1879, A. A. Michelson proposed modifications to a plan of Simon Newcomb's and made 100 determinations of the velocity of light in air (given in Table 6, from [13]), working over a distance of 600 meters. Over the succeeding three years (1880–1882), Simon Newcomb carried out a more extensive experiment that improved on Michelson's in a number of respects. Newcomb's 1882 series of measurements is presented

TABLE 4

Short's 1763 *determinations of the parallax of the sun (in seconds of a degree), based on the* 1761 *transit of Venus. From* [23], pages 310, 316, 325. *Data Sets* 1–3 *together constitute Data Set* 21 ($n = 53$), *Data Sets* 4–6 *together constitute Data Set* 22 ($n = 63$). *The corresponding "true value" is* 8.798

Data Sets							
1 ($n = 18$)	2 ($n = 17$)	3 ($n = 18$)	4 ($n = 21$)	5 ($n = 21$)	6 ($n = 21$)	7 ($n = 21$)	8 ($n = 21$)
8.50	8.65	8.50	8.70	8.43	8.63	8.54	8.54
8.50	8.35	8.80	9.66	9.09	10.16	8.56	8.58
7.33	8.71	8.40	8.50	8.50	8.50	8.54	8.54
8.64	8.31	8.82	8.65	8.44	8.31	8.74	8.94
9.27	8.36	9.02	10.33	9.71	10.80	8.91	9.24
9.06	8.58	10.57	8.07	8.07	7.50	8.40	8.30
9.25	7.80	9.11	8.50	8.36	8.12	8.40	8.33
9.09	7.71	8.66	8.60	8.60	8.42	8.57	8.59
8.50	8.30	8.34	9.61	9.11	9.20	8.69	8.81
8.06	9.71	8.60	8.50	8.66	8.16	8.55	8.56
8.43	8.50	7.99	8.35	8.58	8.36	8.51	8.50
8.44	8.28	8.58	10.15	9.54	9.77	8.57	8.58
8.14	9.87	8.34	7.77	8.34	7.52	8.58	8.58
7.68	8.86	9.64	8.23	8.55	7.96	8.63	8.68
10.34	5.76	8.34	7.92	9.03	7.83	8.56	8.57
8.07	8.44	8.55	8.42	10.04	8.62	8.41	8.33
8.36	8.23	9.54	7.75	9.04	7.54	8.64	8.62
9.71		9.07	8.23	8.71	8.28	8.43	8.37
			8.90	10.48	9.32	8.28	8.03
			7.35	8.31	6.96	8.70	8.85
			7.68	8.67	7.47	8.60	8.74

TABLE 5

Newcomb's Third Series of Measurements of the passage time of light, made July 24, 1882 *to Sept.* 5, 1882 (*from* [16], pages 187–191). *The given values* $\times 10^{-3} + 24.8$ *are Newcomb's measurements, reading down the columns, recorded in millionths of a second. The entire table constitutes Data Set* 23 ($n = 66$). *The corresponding "true value" is* 33.02

Data Set 9 ($n = 20$)				Data Set 10 ($n = 20$)				Data Set 11 ($n = 26$)				
28	−44	29	30	24	28	37	32	36	27	26	28	29
26	27	22	23	20	25	25	36	23	31	32	24	27
33	16	24	29	36	21	28	26	27	27	32	25	28
24	40	21	31	32	28	26	30	27	26	24	32	29
34	−2	25	19	36	29	30	22	28	33	39	25	16
												23

TABLE 6

Michelson's determinations of the velocity of light in air, made June 5, 1879—*July* 2, 1879 (*from* [13], page 135–138). *The given values* +299000 *are Michelson's determinations in* km/sec. *The entire table constitutes Data Set* 24 ($n = 100$). *The corresponding "true value" is* 734.5

Data Set 12 ($n = 20$)	Data Set 13 ($n = 20$)	Data Set 14 ($n = 20$)	Data Set 15 ($n = 20$)	Data Set 16 ($n = 20$)
850	960	880	890	890
740	940	880	810	840
900	960	880	810	780
1070	940	860	820	810
930	880	720	800	760
850	800	720	770	810
950	850	620	760	790
980	880	860	740	810
980	900	970	750	820
880	840	950	760	850
1000	830	880	910	870
980	790	910	920	870
930	810	850	890	810
650	880	870	860	740
760	880	840	880	810
810	830	840	720	940
1000	800	850	840	950
1000	790	840	850	800
960	760	840	850	810
960	800	840	780	870

TABLE 7

Michelson's supplementary determinations of the velocity of light in air, made Oct. 12—*Nov.* 14, 1882 (*from* [14], page 243). *The given values* +299000 *are Michelson's determinations in* km/sec, *reading down columns. The corresponding "true value" is* 710.5

Data Set 17 ($n = 23$)				
883	711	578	696	851
816	611	796	573	809
778	599	774	748	723
796	1051	820	748	
682	781	772	797	

TABLE 8

Cavendish's 1798 *determinations of the density of the earth (relative to that of water). From* [6], page 520. *The entire table constitutes Data Set* 19 ($n = 29$). *Data Set* 20 ($n = 29$) *consists of this table with the third value* (4.88) *replaced by* 5.88. *Data Set* 18 ($n = 23$) *consists of the last* 23 *measurements* (i.e., *omitting* 5.50 *through* 5.55). *The corresponding "true value" is* 5.517

Data Set 19 ($n = 29$)					
5.50	5.55	5.57	5.34	5.42	5.30
5.61	5.36	5.53	5.79	5.47	5.75
4.88	5.29	5.62	5.10	5.63	5.68
5.07	5.58	5.29	5.27	5.34	5.85
5.26	5.65	5.44	5.39	5.46	

TABLE 9

The realized values of the relative error e_{ij} (4.2) *and the average error* s_j (4.1) *for eleven estimates and twenty-four data sets*

Data Set	Mean	Median	Edgeworth	Outmean	10% Trim	15% Trim	25% Trim	Huber P15	Andrews AMT	Tukey Biweight	Hogg T1	s_j
1	.80	1.44	1.09	.46	.97	1.02	1.14	.97	.99	1.03	1.07	.207
2	1.05	1.10	.93	1.15	.93	.95	.96	.93	.98	.99	1.02	.399
3	.29	1.77	1.08	1.67	.56	.87	1.08	.64	1.00	1.03	1.00	.095
4	.73	.94	1.16	.32	.97	1.06	1.14	1.04	1.07	1.41	1.15	.319
5	.92	1.76	.92	2.92	.19	.61	1.09	.18	.45	1.04	.92	.078
6	.77	1.07	1.18	.42	.98	1.06	1.11	.99	1.05	1.28	1.09	.455
7	.99	1.00	.98	1.00	1.01	1.00	.98	1.01	1.01	1.02	.99	.238
8	.96	.98	.95	.92	1.00	1.01	1.00	1.01	1.11	1.04	1.01	.222
9	1.34	.90	.90	1.81	.90	.89	.87	.90	.83	.77	.88	8.40
10	.98	1.10	1.03	.90	.98	.99	1.06	.97	1.00	1.00	.98	4.56
11	.97	1.12	1.01	.89	.99	.99	1.04	.99	1.00	.99	1.00	5.36
12	.94	1.10	1.02	.81	1.00	1.03	1.06	.99	1.00	1.02	1.04	186.3
13	1.02	.93	.93	1.09	.99	.98	.95	1.01	1.01	1.00	1.09	119.2
14	.92	1.00	1.03	.82	.98	1.03	1.02	1.03	1.09	1.09	1.00	120.3
15	1.01	.95	.96	1.03	1.00	1.00	.99	1.01	1.01	1.01	1.03	85.1
16	1.06	.82	1.01	1.14	1.01	1.00	.97	1.01	1.00	.98	1.00	91.9
17	.95	1.32	1.04	.76	.91	1.00	1.14	1.00	.88	.90	1.08	48.0
18	.88	1.50	1.09	.57	1.02	1.13	1.19	.86	.92	.97	.88	.038
19	1.09	.90	.95	1.24	.93	.94	.94	.98	.99	.93	1.09	.063
20	.93	1.27	1.10	.83	.99	1.08	1.04	.90	.95	.96	.93	.037
21	.82	1.34	.83	.50	.86	.96	1.14	.94	1.16	1.18	1.25	.222
22	.72	1.26	1.00	.24	.96	1.05	1.20	1.04	1.11	1.29	1.12	.236
23	1.15	1.01	.96	1.32	.95	.95	.97	.95	.85	.91	.97	5.93
24	1.01	.99	.99	1.04	1.00	1.00	.98	1.00	1.00	1.00	.99	117.1

in Table 5 (from [16]), and is based on observations of light passed over a distance of 3721 meters and back, from Fort Myer on the west bank of the Potomac to a fixed mirror at the base of the Washington monument. The final series presented (in Table 7 from [14]) consists of supplementary measurements made by Michelson after the completion of Newcomb's experiment.

In all cases every attempt has been made to handle the data in the form it was dealt with by the original investigator. Michelson preferred to reduce his measurements to velocities in air, and to then combine these velocities to obtain a final determination by an unweighted or weighted mean. Newcomb preferred to reduce his measurements to times, combine the times through a weighted mean, then incorporate the known distance with this result to produce a final determination of the velocity of light in air. In all cases, the measurements actually given are derived from sets of often widely disparate numbers of observations, a circumstance that one might expect would produce a mixture of standard deviations of the sort modern robust methods are designed to deal with.

The "true values" have been taken to be 33.02 for Newcomb's data, 734.5 for Michelson's 1879 data, and 710.5 for Michelson's 1882 data. These were arrived at by taking the "true" speed of light in vacuum to be 299, 792.5 km/sec. as presented in [22], and incorporating the investigator's own corrections to adjust this figure to a velocity in air or, in Newcomb's case, a "true" time (i.e., the measured time which, when converted to a velocity and corrected as Newcomb did in [16] to obtain a velocity for light in vacuum, would yield 299, 792.5).

REFERENCES

[1] ANDREWS, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16** 523–531.

[2] 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 Univ. Press.

[3] BAILY, F. (1843). An account of some experiments with the torsion-rod, of determining the mean density of the earth. *Mem. Royal Astronomical Soc.* **14**.

[4] BEATON, A. E. and TUKEY, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16** 147–185.

[5] BOX, G. E. P., and TIAO, G. C. (1964). A note on criterion robustness and inference robustness. *Biometrika* **51** 169–174.

[6] BROWNLEE, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd. ed. Wiley, New York.

[7] CAVENDISH, H. (1900). Experiments to determine the density of the earth, *Philosophical Transactions of the Royal Society of London for the year* 1798 (Part II) **88** 469–526. Reprinted in *The Laws of Gravitation* (A. S. Mackenzie, ed.) American, New York.

[8] EDGEWORTH, F. Y. (1887). The choice of means. *Philosophical Magazine* **24** Ser. 5, 268–271.

[9] EDGEWORTH, F. Y. (1893). Exercises in the calculation of errors. *Philosophical Magazine* **36** Ser. 5, 98–111.

[10] HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **27** 87–104.

[11] HOGG, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69** 909–922.
[12] LAPLACE, P. S. (1820). Sur la densité moyenne de la terre. *Ann. Chemie et de physique* **14** 410–416. English translation in *Philosophical Magazine*, **56** Ser. 1, 321–326.
[13] MICHELSON, A. A. (1882). Experimental determination of the velocity of light made at the United States Naval Academy, Annapolis. *Astronomical Papers* **1** 109–145, U.S. Nautical Almanac Office.
[14] MICHELSON, A. A. (1891). Supplementary measures of the velocities of white and colored light in air, water, and carbon disulphide. *Astronomical Papers* **2** 231–258, U.S. Nautical Almanac Office.
[15] NEWCOMB, S. (1882). Discussion and results of observations on transits of Mercury from 1677 to 1881. *Astronomical Papers* **1** 363–487, U.S. Nautical Almanac Office.
[16] NEWCOMB, S. (1891). Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880–1882. *Astronomical Papers* **2** 107–230, U.S. Nautical Almanac Office.
[17] NEWCOMB, S. (1891). Discussion of observations of the transits of Venus in 1761 and 1769. *Astronomical Papers* **2** 259–405, U.S. Nautical Almanac Office.
[18] NEWCOMB, S. (1911). Light, III. Velocity of light. Article in *Encyclopedia Britannica*, **16** 623–626.
[19] POINCARÉ, H. (1912). *Calcul des Probabilités*, 2nd ed. Gauthier-Villars, Paris.
[20] POYNTING, J. H. (1894). *The Mean Density of the Earth*. Griffin, London.
[21] POYNTING, J. H. (1910). Gravitation. Article in *Encyclopedia Britannica* **12** 384–389.
[22] SANDERS, J. H. (1965). *The Velocity of Light*. Pergamon, Oxford.
[23] SHORT, J. (1763). Second paper concerning the parallax of the sun determined from the observations of the late transit of Venus; in which this subject is treated of more at length, and the quantity of the parallax more fully ascertained. *Philos. Trans. Roy. Soc. London* **53** 300–345.
[24] STIGLER, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation, 1885–1920. *J. Amer. Statist. Assoc.* **68** 872–879.
[25] "Student" (W. S. Gosset) (1908). The probable error of a mean. *Biometrika* **6** 1–25. Reprinted in "*Students*" *Collected Papers*, (1958). (E. S. Pearson and J. Wishart, eds.) Cambridge Univ. Press.
[26] TIPPETT, L. H. C. (1927). *Random Sampling Numbers*. Tracts for Computers. **15** Cambridge Univ. Press.
[27] WOLD, H. (1948). *Random Normal Deviates*. Tracts for Computers. **25** Cambridge Univ. Press.
[28] WOOLF, H. (1959). *The Transits of Venus*. Princeton Univ. Press.
[29] YOUDEN, W. J. (1972). Enduring values. *Technometrics* **14** 1–11.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

DISCUSSION

D. F. ANDREWS

University of Toronto

This paper is a welcome contribution to the literature on robust estimation. It illustrates quite convincingly that there is no practical difference between the class of good robust estimators and the arithmetic mean when these are applied

to "clean" sets of data. What is impressive about the results is that, with but few exceptions, the choice of estimator does not matter for these sets of data. Most of the entries in Table 9 are embarrassingly close to 1.0!! The measures of spread in Table 1 are very large compared with the differences in the entries.

Thus satisfied that much does not matter, we are free to concentrate on what does. It is unfortunate that the analysis here obscures the few important differences that do exist. The measure of relative error used does not reflect useful information about absolute errors. Thus for example, in Table 9, the median appears to deteriorate when moving from Data Set 9 to Data Set 10 although it gives a more accurate estimate in the latter case. Furthermore, the last column of Table 9 suggests that the differences in the estimators are much more important for Data Set 9 than for Data Set 10 (where the differences just do not matter). The analysis takes no account of this.

Now Professor Stigler is an expert on matters historical. He argues convincingly for the use of "old" data on "known" constants to compare estimators. However, he also implies that "modern" data are comparable in other respects such as kurtosis. It has been my experience all too frequently to find computer summaries of my data yield reasonable medians but unprintable means, to say nothing of variances and higher moments. Often data are collected with instruments and the quantities of data collected are large. Such data typically contain gross disturbances as, for example [1], when a telescope measuring background infrared radiation while tumbling through near-space points for a moment toward Earth. Such disturbances are well described by Cauchy-like distributions (although these will never be seen through Cauchy-coloured glasses).

In simple, one-parameter models, such disturbances may be easily identified. But the purpose of much of data analysis is rarely to ascertain any absolute quantity (as was the purpose of the experiments described in the paper) but to discover relations between variables, to build and assess models. In these more complex situations, the disturbances may only be detected after first fitting the data with a procedure insensitive to such gross errors. Thus robust fitting procedures are required for more complex models. The family of M-estimates readily extends to linear and nonlinear models. These procedures operate the way Legendre rather than Edgeworth suggest. The other procedures studied here are not as useful in these situations.

Professor Stigler suggests that the use of these M-estimates results in an unnecessary loss of efficiency. I note that he does not comment on the statistical significance of this apparent loss. And Professor Stigler is a statistician.

This paper confirms that these procedures may be safely applied to clean sets of data, for in these cases even an expert cannot establish the difference. But for severely contaminated data the story is very different.

REFERENCE

[1] PIPHER, J. L., HOUCK, J. R., JONES, B. W. and HARWIT, M. (1971). Submillimetre observations of the night sky emission above 120 kilometres. *Nature* **231** 375–378.

GEORGE A. BARNARD

University of Waterloo

The main point that I take from Stigler's excellent paper is that we statisticians have been neglecting the empirical basis of our science. In the nineteenth century there was a great deal of discussion as to whether errors of observation could be taken to be normally distributed, and by the end of the century men like Edgeworth and Newcomb were able to give a rather heavily qualified yes; they both began investigating other than normal distributions. The biometric school, under the leadership of Galton and Karl Pearson, continued through the turn of the century to study the forms of distributions occurring in nature. But in the 1920's it was, perhaps, the elegance and generality of Fisher's results based on the normal distribution which led many statisticians to play down the possibility of departures from normality—when Egon Pearson raised the question of nonnormality in relation to the innovations of R. A. Fisher, the latter, with understandable but unfortunate supersensitivity, interpreted the question as an attack. At the same time, and characteristically, Fisher set himself, and encouraged others, to work, for example, on densities whose logarithm is quartic rather than quadratic; but he soon saw that the computational problems involved were beyond the capacity of the facilities then existing. Thus in the 1930's and 1940's, the prevailing view was, that provided gross departures from normality were avoided, for example, by logarithmic or square root transformations, it was lack of independence and inequality of variance that were more to be feared. Nonparametric procedures failed to win general acceptance because, among other things, while removing assumptions about distributional form, they leaned heavily on independence and identity of distribution.

The revival of Bayesian approaches enabled Box and Tiao (loc. cit. supra) to carry through a pioneering analysis of the true effect of assumptions about distributional form on a classical set of (Darwin's) data; and with the help of Fraser's (1976) necessary analysis, or my own pivotal inference (1977), it is now possible for non-Bayesians to match Box and Tiao's work in its essentials. Thus we can all deal, without undue difficulty, with nonnormality and lack of independence—so long as we know what the nonnormal and nonindependent joint distributions are. But it is this empirical information which is lacking.

I agree with Stigler's suggestion, that it is likely that the errors in the cases he has taken were more nearly normally distributed than has been supposed by most recent workers, though the lack of independence, for instance, exhibited by Michelson's data is remarkable. The fact that shrewd and skeptical nonstatisticians with practical responsibilities, such as Sir George Airy, were persuaded to accept the normality doctrine is evidence in favour of this idea. But it is easy to think of reasons why typical distributional forms of errors, or of variability should have changed between then and now.

We urgently need, therefore, to collect information on lack of independence

and on distributional form. Fortunately, this should not be difficult. Since most data processing is now done on computers, all we need do is to form the habit of computing 3rd and 4th moments, proportions of outliers, and, say, first order autocorrelations, as routine, and to file the results along with a note about the type of data involved. We could then, working together, hope soon to reach a situation where we could say: "As these data are of such and such a kind, a reasonable range of distributional forms, and dependencies, is this." We could go on, either with "The sample now to hand is robust, in the sense that the resulting range of inferential statements is very narrow" or "The sample now to hand is sensitive to the plausible range of distributional forms, so that further information on this form should, if possible, be collected."

I would therefore urge, if I may, the setting up of a Committee for the Study of Sample Configurations, to propose standard forms for the collection of information, to act as a clearinghouse for the information collected, and perhaps to suggest ways in which the resulting information can best be classified.

It seems to me that it is only in this way that the problems which Stigler raises can be properly dealt with. With the information on distributional form, etc., we would not be restricted, as Stigler is in this study, to looking at an estimate purely from the point of view of how close it comes to the true value; we could also deal with the almost equally important question of providing a reliable assessment of the error of our estimate.

We should not forget that when the "error" distribution represents real variability in a real population, we may well be interested in the middle part of the population much more than in the tails. The Halothane study is a case in point. The "heavy-tails" estimators that have been studied in recent years often will represent these middle parts better than location estimates for the whole population. It would not be the first time that a proposal intended to serve one purpose turned out to serve another.

The science of cosmology seems no nearer to definitive formulation than the science of statistics, so that E. A. Milne's suggestion of forty odd years ago, that the velocity of light is now less than it was, should not go unnoticed.

REFERENCES

AIRY, SIR G. B. (1849). Letter to Augustus de Morgan. de Morgan Collection, Univ. of London Library.

BARNARD, G. A. (1977). *Foundations of Statistical Inference.* Aarhus.

FRASER, D. A. S. (1976). Necessary analysis and adaptive inference. *J. Amer. Statist. Assoc.* **71** 99–113.

O'TOOLE, A. L. (1933). On the system of curves for which the method of moments is the best method of fitting. *Ann. Math. Statist.* **4** 1–29.

G. E. P. BOX

University of Wisconsin at Madison

In applying the criterion of usefulness to robust estimators, Steve Stigler is to be congratulated for the clarity of his thought in a confusing world.

To say that a procedure (estimator) is robust means that, not only under ideal assumptions, but when exposed to *real data*, it *usually* does *about* what is expected of it. Words italicized in the above have meaning as they relate to the real world and hence to the domain of applied mathematics.

The vital concerns of honest pure mathematics on the one hand and honest applied mathematics on the other are so different, that, lack of communication and understanding between workers in two disciplines, having such similar names, is perhaps inevitable. It is agreed that in considering the question "Given proposition A does proposition B necessarily follow?", pure mathematics need not concern itself with whether proposition A has truth or meaning in the real world or whether proposition B is of any use. When proposition A is the applied mathematician's tentative assumption, it does not need to represent truth or practical reality either. Assumptions which are grossly discordant with fact (for example, that certain particles have no mass) frequently produce useful physical laws. Furthermore, the degree of departure from the ideal is not a unique determinant of performance, for the same assumptions that lead to practically useful results for one problem (for example, standard Normal assumptions leading to analysis of variance for comparison of means) can by the same route lead to much less useful results in another (for example, Bartlett's test for equality of variances). The ultimate justification of the assumptions and methods of applied mathematics is their ability to yield results useful in practice. While one counterexample disproves a logical proposition, the existence of functions for which some useful method of, say, numerical integration or function optimization does not work is, of course, inevitable; so is the existence of distribution functions for which certain estimators do not work. The only question of interest is how often such functions occur and what are warnings of their probable occurrence.

Testing estimators with extreme distributions like the Cauchy implies an argument of the minimax type and is no more convincing there than it is in other contexts. Indeed it calls to mind the following from the mathematician C. L. Dodgson.

> "I was wondering what the mouse-trap was for," said Alice. "It isn't very likely there would be any mice on the horse's back."
>
> "Not very likely, perhaps," said the Knight; "but, if they *do* come, I don't choose to have them running all about."

I hope this paper marks a renewal of interest in how statistical procedures work in the circumstances of the real world which implies an explicit discussion of what these circumstances really are. I think this may lead to a realization that emphasis on nonnormality of the marginal distributions has been rather overdone and that considerably more attention should be paid to the assumption of independence (or distribution symmetry).

Whether we recognize it or not, as applied mathematicians we are motivated

in our researches by "prior distribution of reality" which implies how often functions of this or that type are likely to occur. If we try to operate independently of this prior (as the pure mathematician in us might like to do) we get into trouble. This is evidenced in other contexts by the properties of Stein's estimators and by "distribution free" tests.

Sometimes it helps to argue backwards. By asking what will endow with virtue the estimators chosen by the experts we might deduce their subconscious distribution of distributions. After Winsorizing out the nightmares and suitable polishing, this prior might then be used to solve the problem using Bayes theorem.

D. R. Cox

Imperial College, London

I have read Dr. Stigler's paper with much pleasure. He raises some important general issues. One is the need for more empirical study, presumably as an incidental to other work, of the shape of distributions. (In a study I made some years ago of various kinds of routine laboratory tests in textiles, distributions with negative kurtosis occurred about as often as those with positive kurtosis).

On the more general issue of the role of robust procedures, the "traditional" approach is, I suppose, a combination of preliminary inspection of the data together with study at the end of the analysis of whether there are aspects of the data and assumptions reconsideration of which might change the qualitative conclusions. This approach seems to have a great deal to commend it, especially when it allows simple methods to be used on relatively complex problems. It may often be inapplicable to the analysis of large bodies of data and here the case for automatically robust methods is much stronger; whether recent work is right to concentrate strongly on robustness to longtailed contamination rather than, for example, robustness to correlations among errors, is not so clear.

Edwin L. Crow

National Center for Atmospheric Research

Professor Stigler's study is an interesting and valuable contribution to the field of robust estimation. However, the title may suggest more generality than is stated in the body of the paper, in particular in the three requirements for inclusion of data sets and in Figure 1. Thus the study is necessarily limited to fairly precise physical measurements. Figure 1 shows "that the data sets considered tend to have slightly heavier tails than the normal" and that the skewness tends to differ only moderately from that of normal samples. One must be very careful, then, not to generalize the conclusions to all real data.

The good performance of the mean and the poor performance of the median would seem to be associated with the fact that the data sets of the study differ only moderately from normal samples.

Even though Stigler suggests that the Princeton study [his reference 2] may have included a greater variety of distributions than should be as judged by his data sets, they were still limited to symmetric unimodal distributions. Real data may exhibit great asymmetry or perhaps even multimodality. Some meteorological data, in particular precipitation data, have *J*-shaped distributions [2, 3]. These shapes can often be altered to approximate normality by transformation, such as taking a root or the logarithm, but the applier of statistical methods may get the impression that robust estimates can be used without concern for transformation. By the time an experimenter comes to formal statistical analysis he surely has a rough idea of the shape and thus what transformation, if any, to use. Alternatively, perhaps a few broad classes of distributions should be specified, such as bell-shaped, *J*-shaped, and *U*-shaped, and methods robust *within these classes* can be recommended. It may not matter a great deal whether Michelson's and Newcomb's data (end of Section 2) are analyzed as times or velocities, but one should surely consider whether the reciprocal should be taken before applying even robust estimates.

Neyman and Scott [3] have distinguished between "outlier-resistant" and "outlier-prone" families of distributions and pointed out that "elimination of a sample member [in a domain of study in which substantial samples of observations can be fitted only by some outlier-prone distribution], merely because its value deviates considerably from those of the others, cannot be justified." They state two theorems: "The family of Gamma [respectively lognormal] distributions is outlier-prone completely." Thus in footnote to Stigler's conclusions, there are cases of real data in which outliers should not be rejected either explicitly or implicitly by the choice of estimator.

Crow and Siddiqui [1] independently proposed a linear combination of the median and two other symmetric order statistics (similar to the Edgeworth estimator) to estimate the mean and determined the optimum (MVUE) for the rectangular, normal, double exponential, and Cauchy families and all combinations thereof for sample sizes 5, 9, 17 and ∞.

Although the median shows up poorly as an estimator in these data sets, it might have been considered in the measures of performance in both (4.1) and (4.3) to reduce the effect of outlying estimator measures and outlying data sets. In both cases it suffers the slight disadvantage of discreteness relative to the means used in (4.1) and (4.3). However, it is doubtful that this change would have much effect on the results of the study. More generally, for a summary measure of the sampling distribution of an estimator, what advantage other than possible ease of calculating does a (mean-)unbiased estimator have over a median-unbiased estimator?

REFERENCES

[1] CROW, E. L. and SIDDIQUI, M. M. (1967). Robust estimation of location . *J. Amer. Statist. Assoc.* **62** 353–389.

[2] National Research Council (1973). Weather and Climate Modification: Problems and Progress, 205–206, 227–258. National Academy of Sciences.

[3] NEYMAN, J. and Scott, E. L. (1971). Outlier proneness of phenomena and of related distributions. *Optimizing Methods in Statistics* (J. S. Rustagi, ed.) 413–430, Academic Press, New York.

CHURCHILL EISENHART

National Bureau of Standards

When I began reading a pre-publication copy of this paper, I looked forward with keen interest to seeing how the respective "robust estimators" compared with each other in each instance, and how they compared with the "mean" adopted by the original experimenter as his "best value"—that is, to seeing whether the original experimenters would have been led to much the same, or quite different, conclusions had their data-analysis tool kits included these "robust estimators." I was doomed to disappointment: Nowhere does the author tabulate the values of the individual "robust estimators" for the respective data sets analyzed, nor does he ever give the "mean" or "best value" adopted by the original experimenter. Instead he simply gives a table (Table 9) of "the realized values of the relative error" of each of the eleven estimators considered, for each of the twenty-four data sets considered.

As Stigler points out in the parenthetical remark following equation (4.3), "the actual errors $\hat{\theta}_{ij} - \theta_j$" of the respective estimators $\hat{\theta}_{ij}$, $(i = 1, 2, \cdots, 11)$, relative to the "true value" θ_j adopted for that set "can be easily recovered from Table 9." The "true value" θ_j in each case being given explicitly, the observed values of the estimators $\hat{\theta}_{ij}$ can be recovered also. In this manner I ascertained the values of the nine estimators other than the mean and median—which I evaluated directly—for Data Sets 23 (Table 5), 24 (Table 6) and 17 (Table 7). The results obtained for Data Set 23 are shown in the accompanying figure, together with the basic sixty-six time-of-passage values in frequency distribution form, and the value of the "mean" that Newcomb used.

It should be noted that the values of Newcomb's "mean" and all of the estimators considered, except the outmean, lie in the central narrow interval (26.2, 27.9) indicated at the left of the figure, whereas the "true value", 33.02, that Stigler adopted for these data (indicated by the arrow in the figure) is quite distant and corresponds to about the 88th percentile of the data. A similar situation prevails in the case of Data Set 24, for which Michelson's "mean," 852, and all of the estimators considered by Stigler, except the outmean, lie in a central narrow interval (849, 852.4), whereas the "true value" Stigler adopts, 734.5, lies in the lower left "tail," between the 5th and 8th percentiles of the data. Likewise, in the case of Data Set 17, all of the estimators considered, except the median and the outmean and Michelson's "mean," 771, lie in the interval (754, 765), whereas the "true value" Stigler adopts, 710.5, lies at about the 32nd percentile of the data. "True values" such as these that lie "out in the

wings" are clearly unsuitable for judging the relative merits of a group of estimators that are "arguing" over which value in a central "core" of a set of data "best" summarizes the "evidence" of the set as a whole.

The "true values" that Stigler adopted for these three data sets were all derived by "working backwards" from a 1965 "best" value for the velocity of light in a vacuum, 299, 792.5 km/sec, which agrees to the number of significant figures given with the value officially adopted in 1973 by the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions. The reason for the unsuitability of the "true values" thus derived for the objectives of the present paper is that Michelson's 1879 and 1882 determinations of the velocity of light (Data Set 24 [and Sets 12–16], and Data Set 17, respectively), and velocity of light values derived from Newcomb's 1882 time-of-passage measurements (Data Set 23 [and Sets 9–11]) are all affected by enormous positive systematic errors, as were all velocity-of-light determinations prior

Data Set 23
Newcomb's Time-of-Passage Data (τ μsec)
(Stigler's Table 5)

$10^9\,\tau$-24800 Gsec	Freq.	Cum. Freq.		
−44	×	1	Mean	: 26.212
−2	×	2	Median	: 27
16	××	4	Edgeworth	: 27.3
17				
18			[Outmean	: 25.2]
19	×	5		
20	×	6	10% Trim	: 27.4
21	××	8		
22	××	10	15% Trim	: 27.4
23	×××	13		
24	×××××	18	25% Trim	: 27.3
25	×××××	23		
⌈26	×××××	28	Huber P15	: 27.4
27	××××××	34		
⌊28	×××××××	41	Andrews AMT	: 27.9
29	×××××	46		
30	×××	49	Tukey Biweight	: 27.6
31	××	51		
32	×××××	56	Hogg T1	: 27.3
→33	××	58		
34	×	59	*NOTE*: Mean and Median determined	
35			directly; all others, by inversion of	
36	××××	63	Stigler's Table 9, Data Set 23.	
37	×	64		
38				
39	×	65		
40	×	66	Newcomb's "mean": 27.4	

to 1906; see the figure and table on page 66 of the August 1955 issue of the *Scientific American*. Furthermore, since actual causes and precise nature of these systematic errors are not known today, it is impossible to bring the Michelson and Newcomb determinations "into line" with modern values, or alternatively to "work backwards" from a currently accepted value of the velocity of light to derive trustworthy determinations of the "true values"(?) that their respective data sets were striving to indicate. This inability to "work backwards" from current "best" values to trustworthy "true values" for historic data sets will, I fear, always be with us, and will render dubious efforts to compare the *accuracies* of alternative estimators such as Stigler has attempted in this paper. (Even when an attempt is made to duplicate an experiment as exactly as possible at another time or place, it is usually found that the results of the first fail to predict accurately the results of the second; and in the case of experiments that differ in procedure, the disagreement is usually more pronounced.)

Consequently, I feel that Table 3 and Figure 1 comprise the solid message of this paper. As Stigler says in the final paragraph of Section 6: "[They] suggest that the data sets considered tend to have slightly heavier tails than the normal, but that a view of the world through Cauchy-colored glasses may be overly-pessimistic" and that "a small amount of trimming (no more than 10%) may be the best way of dealing with them." I hope this section of his paper will encourage others to prepare similar presentations of the characteristics of run-of-the-mill real-life data that they encounter in various fields of science, and explicit comparisons of the behavior of selected estimators, as I have done, to guide us all making wise choices between traditional and new-fangled estimators.

DAVID C. HOAGLIN
Harvard University

Professor Stigler has made a thought-provoking attack on the problems surrounding two questions: "How should one compare estimators in terms of robustness?" and "What is the connection between the situations used in simulation studies and situations faced in the real world?" Still, in both of these areas it seems to me that serious questions remain.

First, when a data set involves a shift, or a bias, it is not at all clear whether that bias should be charged against the performance of any estimator. To do so implies that the estimator should be able to see beyond the data to the "true" value of the physical quantity, or equivalently the estimator must introduce a bias of its own to compensate for the bias in the data. Using only data on a physical quantity which is now accurately known and whose definition is unchanged is intended to eliminate or reduce this difficulty, but allowing for novel experimental apparatus opens the question all over again. Stigler discusses bias at length in Section 5, but it would still be informative to see the results of some tentative attempts to quantify and remove the biases. (Table 3 is a beginning.)

Second, it may be desirable to base the relative-error performance measure e_{ij} on a more "robust" summary than s_j (as in (4.2) and (4.1)). As e_{ij} is defined, the constraint $\sum_{i=1}^{11} e_{ij} = 11$ is forced, and a disastrous error by one estimator on a particular data set will make the performance of the other estimators appear more nearly similar. There may be evidence of this in Data Sets 4 and 9 (see Table 9). Of course, it is possible to object that the need to robustize s_j has not been demonstrated, but the situation is not the same as in the data sets on which the study is based. The 11 estimators deliberately represent considerable variety, and the outmean (as mentioned in Section 3) is essentially a planned outlier.

Third, the data used for this study is certainly real, but one must still ask whether these data sets are in any sense broadly representative of the sort of behavior one might generally have expected at the times when the underlying experiments were carried out, let alone at the present and in the immediate future (for which the conclusions would likely have the greatest impact). The three requirements in Section 2 clearly narrow the "real world" to the physical sciences, and anyone planning to act on the conclusions must keep this restriction firmly in mind. The scarcity of sets of real data which meet the three requirements suggests that simulation studies are likely to continue as the primary method of assessing robustness. It should be possible to build on the results of the Princeton robustness study and simulate other aspects of real data besides gross errors and longer-tailed contamination. Any such simulation study would do well to emulate the careful documentation which Professor Stigler has provided for the real data and his treatment of it. Tables 4 through 8 reproduce all the data, and Table 9 makes it possible (with modest effort) to reconstruct the $\hat{\theta}_{ij}$ and try other performance measures or other analyses.

Finally, even if we set aside the question of how well these particular data sets represent "the real world," the heavy weight given to only four or five data sources is cause for concern. The 20 data sets are far from independent, with the overlap of Data Sets 18, 19 and 20 (Table 8) as the most extreme example. Even after warning the reader of this (in Section 3), Stigler analyzes the results (in Tables 1 and 9) as if they *were* independent. There may have been no feasible alternative, but it is difficult to know how the reader should adjust his interpretations of results and conclusions to allow for the lack of independence.

On balance, the emphasis on closer contact with real data is welcome, and more efforts to bridge the gaps between all kinds of real data and simulation studies are needed.

ROBERT V. HOGG

University of Iowa

Not only would I like to congratulate the author for such an interesting article, but also the editor for the decision to publish it because it is quite different from most of the articles appearing in the *Annals of Statistics*. But such a decision

seems consistent with the intent of the recent officers and council of IMS to broaden the statistical coverage of the Institute.

As Stigler so carefully points out, Andrews, Tukey, and Hogg suggested their statistics *before* they were informed of the nature of the study. It seems that each of us had on his "Cauchy-colored" glasses and tried too hard to protect against outliers. As most of the data that Stigler examined was more like that arising from distributions which are close to the normal, such protection was really not required. In only two samples, namely 9 and 23, was there an advantage in using the descending M-estimators suggested by Andrews and Tukey. Note that in those samples the measures of skewness and kurtosis were $(-2.82, 11.00)$ and $(-4.49, 29.40)$, respectively; this, of course, illustrates that these descending M-estimators can be extremely useful with data with heavy tails.

On first inspection, I thought that these results would discourage the use of adaptive estimators in practical situations due to the rather poor performance of T_1. But upon closer investigation (plotting dot diagrams, etc.), I came to realize that adaptive estimators based upon a measure of skewness (possibly along with a measure of kurtosis) were really better than ones based upon a measure of kurtosis alone. While I have a more complicated statistic that illustrates this point, for simplicity I suggest here the statistic

$$\begin{aligned} T_2 &= \bar{X}^c_{.25} && \text{if } |(\beta_1)^{\frac{1}{2}}| < 1 , \\ &= \bar{X}_{.1} && \text{if } 1 \leqq |(\beta_1)^{\frac{1}{2}}| < 2 , \\ &= \bar{X}_{.25} && \text{if } 2 \leqq |(\beta_1)^{\frac{1}{2}}| . \end{aligned}$$

Clearly, *with these data*, this is a better statistic than any of the other 11 statistics; its relative error in small samples is 0.796 as compared to the 0.916 of the 10% trimmed mean.

I recognize that I have cheated in selecting this statistic T_2 because not only do I now know the nature of the study but have actually seen the analysis. Nevertheless, the huge improvement provided by T_2 does suggest that adaptive estimators could be beneficial in the future.

In the last year, I have been extremely impressed with certain analyses of other real data using Huber's M-estimators or Andrews' descending M-estimators (or a combination of them, say, using Huber's on the first iterations and Andrews' on the last 2 or 3 iterations). These schemes, in regression situations, provide a most useful way of spotting outliers, if they exist, because outliers will have small or zero weights after several iterations. However, I also believe that if the statistics of Huber and Andrews are good, then adaptive ones would be better. For example, let the k of the Huber statistic decrease as measures of skewness and kurtosis (not necessarily $(\beta_1)^{\frac{1}{2}}$ and β_2) increase; that is, for illustration, use Huber's P20, P15 and P12, respectively, as those measures increase. Users of the M-estimators actually seem to do this in practice anyway by changing the values 1.5 and 2.1 of the respective statistics of Huber and Andrews to suit the problems at hand; I am simply suggesting that we formalize that procedure.

Finally, I have two other suggestions, neither of which will probably receive much favor because statisticians are skeptical of placing too much weight on the extreme values (although recall the outmean was extremely good in many of these samples) or of using asymmetrical estimators. The first is, in cases of *small* skewness and kurtosis, use an "out-Huber," like that associated with

$$\varphi(u) = u - d \min [1.0, \max (-1.0, u)], \qquad 0 \leqq d \leqq 1;$$

for example, if $d = 1$, then

$$\begin{aligned} \varphi(u) &= u + 1.0, \quad u < -1.0, \\ &= 0, \qquad -1.0 \leqq u \leqq 1.0, \\ &= u - 1.0, \quad 1.0 < u. \end{aligned}$$

Of course, the corresponding M-estimators place more weight on the extreme order statistics. The second is, in skewed situations, use an M-estimator associated with an asymmetric $\varphi(u)$. For a simple example, if

$$\begin{aligned} \varphi(u) &= -1, \quad u < 0, \\ &= 3, \quad 0 < u, \end{aligned}$$

then the associated M-estimator would be the 75th percentile of the sample. Seemingly, this would be highly desirable in many instances (including regression situations), and I am surprised that more statisticians have not wanted to estimate various percentile regression curves.

I have really enjoyed studying Stigler's article; it's one that challenges your imagination. And since it contained real data, I did do some data analysis (something that I never would have done with Monte Carlo values), and that effort was most instructive. I am certain others will find the article as stimulating as I did.

PETER J. HUBER

Swiss Federal Institute of Technology

This paper throws a very interesting sidelight on some issues of robust estimation. At least for me, the results were not unexpected. Though, I would like to interpret them in a slightly different way.

In my experience with real data, "good" samples from the physical sciences seem to suggest trimming rates between 1% and 10%, i.e., a trimmed mean with a suitably adjusted trimming rate in this range should be an almost efficient estimate [1, page 1057]. It is not surprising therefore that the 10% trimmed mean performed well; possibly, a 5% trimmed mean might have been even better on the average.

In my view the main purpose of robust procedures is to "prevent the worst," i.e., to prevent a catastrophe due to an occasional bad sample. Therefore, I am not entirely happy with Stigler's method of combining and interpreting the

data: this should *not* be done in a robust way by combining ranks (4.5), and even the average absolute error (4.3) may be too lenient with regard to occasional poor performances of some estimators. One should also, and perhaps primarily, put emphasis on performance in worst cases. To check this, I determined, in each column of Table 9, the largest value of e_{ij} and ranked the estimators according to these values. Since the extremal value might be a freak, the ranking was repeated also with the 2nd and 3rd largest ones. The results were as follows:

Estimators ranked according to largest absolute errors.

largest e_{ij}		2nd largest e_{ij}		3rd largest e_{ij}	
10%	1.02	10%	1.01	10%	1.01
P15	1.04	P15	1.04	P15	1.03
15%	1.13	15%	1.08	15%	1.06
AMT	1.16	AMT	1.11	Mean	1.09
Edgew.	1.18	Hogg	1.15	Edgew.	1.10
25%	1.20	Mean	1.15	AMT	1.11
Hogg	1.25	Edgew.	1.16	Hogg	1.12
Mean	1.34	25%	1.19	25%	1.14
Biweight	1.41	Biweight	1.29	Biweight	1.28
Median	1.77	Median	1.76	Median	1.50
Outmean	2.92	Outmean	1.81	Outmean	1.67

Of course, I was elated to find that the intersection of Stigler's choice of estimates and of the collection I had recommended in [1, page 1063f.], namely the 10% and 15% trimmed mean and P15, carried gold, silver and bronze in my analysis! (I swear this happened in the first attempt.)

It seems to me that Stigler's samples contain fewer gross clerical errors (copying errors, misclassifications, etc.) than average data sets, in particular fewer than those one tends to meet in the life and social sciences. The worst real data I have seen so far, a batch of some 50 ancient astronomical observations, contained about 15 (or 30%) gross errors, but, of course, this set would not meet Stigler's stringent selection criteria [2].

The Princeton study may have given an unintended and undue prominence to heavy, Cauchy-like tails and to matching estimators. It is good that this paper helps to correct this bias, but I hope the pendulum will not swing too far: heavy-tailed real data does exist also!

REFERENCES

[1] HUBER, PETER J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041-1067.

[2] HUBER, PETER J. (1974). Early cuneiform evidence for the planet Venus. Presented at the AAAS Annual Meeting, San Francisco. To be published in *Scientists Answer Velikovsky* (1977). (D. Goldsmith, ed.). Cornell Univ. Press.

JOHN W. PRATT

Harvard University

Stigler's paper [2] is certainly fascinating and provocative. I hope the reactions provoked will include serious thought and discussion about the real purposes of "robust estimation" and of studies of robustness. The emphasis in applications, I believe, will be on parameters of comparison (differences, other contrasts, regression slopes etc.), on complex situations with data scanty in relation to the complexity, and on good efficiency for a mix of anticipated and unanticipated possible error behaviors. The implications for these concerns of studies of repeated measurements of absolute physical constants and simulated simple random samples from symmetric distributions are not at all immediate, though such studies may be the best way to begin.

Two possible attitudes in these and other one-sample location problems are: (a) A specific physical quantity is to be estimated, with no substitutes allowed. (b) It doesn't matter what an estimator estimates, as long as it is a location parameter. The appropriate attitude varies with the real situation at hand, and may lie anywhere between these, or elsewhere altogether. Stigler's attitude here is (a), which is appropriate for the situations he considers but not for choosing "representative" situations to study. Yet the way he introduces requirement (2) in Section 2 implies that (a) is appropriate in almost all practical situations.

Most work in the field goes to the other extreme, at least implicitly. I would argue that the spirit in which the "robust" estimators Stigler studies are put forward and studied elsewhere is close to (b). In the absence of symmetry, they are (consistent) estimators of different parameters (and mostly complicated ones). It is rare that symmetry is guaranteed not merely under a null hypothesis (where randomization sometimes guarantees it) but also under alternatives. Yet typical robustness studies compare the estimators exclusively in terms of variability, without concern for which location parameter they estimate, and are thus so incomplete as to be almost meaningless except from a point of view near (b). This includes the high-powered blockbuster [1]. (Incidentally, its results are described on page 1 as "comprehensive, not exhaustive." On page 226, Tukey says that asymmetric situations are an important job for the future, but "we were not able to agree, either between or within individuals, as to the criteria to be used.")

Even if a point of view close to (b) is usually appropriate, is it enough to look at the variability of estimators? Unfortunately, other concerns may have comparable importance in most situations, when the problem is examined carefully. Consider even the simple example of matched pairs, treatment and control. (Similar comments would apply to more complex situations, regression and beyond, where we particularly hope for improved robustness in the future.) Typically, the treatment effect, if there is one, is different in different pairs, and the treatment-control differences are not symmetrically distributed. The sensitivity

of a "robust estimator of location" applied to these differences depends not only on its variability but also on the size of the corresponding location parameter for the particular asymmetric alternative at hand. The mean may be a more variable statistic than the median, but the population mean may also be larger than the population median. The latter aspect will even dominate some sensitivity comparisons for large samples and a fixed alternative. Neither dominates in the situation envisaged by Pitman's asymptotic relative efficiency, which is the product of two factors, the ratio of the asymptotic variances and the square of the derivative of one location parameter with respect to the other for alternatives approaching the null hypothesis.

Thus, when attitude (b) is appropriate, one can define away any bias, but a related aspect of the problem remains. The bias in Stigler's situations is the counterpart and he is right to include it. Still, the effect of bias in his situations may not be representative of the corresponding effects in more usual situations. I will elaborate briefly on this and other limitations of his study, despite his recognition of them and his Section 5.

The number of fully distinct experiments Stigler considers is small—5 by a generous count. Thus the empirical comparisons are really based on a nonsample of size 5. In at least 3, the measurements fall predominantly to one side of the true value. In Data Sets 1–8 (1761 transit of Venus), which I count as one experiment since they apparently have a large component of bias in common, the true value is exceeded by only 40 out of 158 measurements and thus appears to be near the upper quartile of the measurement distribution (between the 67th and 81st percentile with confidence 95% if the measurements are independently identically distributed). In the experiments to measure the velocity of light, which I count as three since they apparently have very different biases, the true value is exceeded by $\frac{8}{66} = 12\%$, $\frac{95}{100} = 95\%$ and $\frac{17}{23} = 74\%$ of the measurements. Even excluding the latter, and the fifth experiment (Cavendish), there are three experiments, accounting for 20 of the 24 data sets, which are so lopsided that any method of comparing the errors of estimators will inevitably reflect almost exclusively their differing biases in these particular experiments.

Stigler's use of absolute error emphasizes bias even more than squared error would. If an estimator is inconsistent, then it falls on one side of the true value with a probability which typically approaches 1 exponentially fast as $n \to \infty$. Hence, asymptotically, absolute error = absolute bias + an exponentially small contribution due to variability. By contrast, mean-square-error = bias-squared + a contribution (the variance) of order $1/n$ due to variability. The contribution due to variability approaches 0 in either case, but much faster for absolute error. In Stigler's examples, indeed, with two exceptions mentioned after (4.3), in each data set, all estimates fall on the same side of the true value, whence $e_{ij} = (\hat{\theta}_{ij} - \theta_j)/(\hat{\theta}_{\bullet j} - \theta_j)$, with algebraic error in place of absolute error.

In short, Stigler's comparisons reflect bias almost exclusively (however much or little they are diluted by it; cf. Section 5). He has a nonsample of only 5

biases. At least 3 of these are so large that it is hard to view the problem as a statistical one of robust estimation at all. Their effects may be quite different from the corresponding effects in more usual situations.

Nevertheless, Stigler has moved bravely and strongly, from simulation to reality, and beyond symmetry, which assumes away an important aspect of the problem. It is easier to find fault, both with Stigler and with simulation under symmetry, than to do better.

So much for what history can contribute to modern statistical thought. What could modern statistics have contributed to the original scientists' analyses? We now have clear concepts of sampling variance and interval estimation, accompanied by highly developed methodology. Suspect values can now be handled more systematically and with more understanding of the consequences. (Nevertheless, neither blind trimming nor sophisticated Huberizing should eliminate investigation of their background or sensitivity analysis of their effects.) The choice of sample size can now be analyzed more fully perhaps (completing a frequentist analysis Bayesianly if necessary. Under mean squared error, the value of reducing the variance is the same whether the bias is large or small, so another criterion might be preferred.) In hindsight, Short's standard error of the mean ($s/n^{\frac{1}{2}}$) is only 28% of the absolute bias, Newcomb's 20%, Michelson's 7% (in Table 6) and Cavendish's 60–123% for the three versions of the data Stigler gives (Data Sets 18–20). The smallness of these percentages does not in itself suggest bad judgment beforehand about sample size. It does point again to the central difficulty: how best to make inferences in the face of significant uncertainty about both the shape and the bias of the distribution sampled from. Estimators differ in both bias and variability. The bias of a particular estimator can be encompassed formally in inference (Bayesianly), but we are not yet ready practically to choose a best estimator or to analyze several simultaneously. Neither robustness nor Bayesianism provides a magic cure. If we could consider all estimators the same as far as bias is concerned, then minimizing variability would be the whole game after all. It is unreasonable to do so, however: Bayesianly, if the mean squared bias of three linearly dependent estimators is the same (e.g., mean, outmean, and 25% trimmed mean), then the bias must be known. Modern statistics still cannot meet all the challenges posed by these data.

REFERENCES

[1] 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 Univ. Press.

[2] STIGLER, S. M. (1977). Do robust estimators work with *real* data? *Ann. Statist.* **5** 1055–1098.

ROY E. WELSCH

Massachussetts Institute of Technology

Professor Stigler has made an important contribution to statistics by focusing attention on the need to evaluate new statistical procedures on both real *and* synthetic data. We can now hope that others will help to develop a larger collection of real data (including multivariate examples) that satisfy Stigler's conditions (page 1057).

Certainly the data in this paper give more support to the arguments that, in practice, the assumption of a Gaussian error distribution is suspect, and that we should not slavishly adhere to the conclusions derived from such an assumption.

However, having opened the door to a real world that is non-Gaussian, how are we to proceed? Stigler states his point of view on page 1070:

> "The data sets examined do exhibit a slight tendency toward more extreme values than one would expect from normal samples, but a very small amount of trimming seems to be the best way to deal with this."

The implicit, if not explicit, assumption here is that the goal of research on robust methods is to find an estimator to replace the mean. Stigler finds that his sets of real data are somewhat non-Gaussian, so he suggests replacing the mean with the 10% trimmed mean (or perhaps the Huber P15 which has a better breakdown bound [3]).

I do not feel that our goal should be to replace the mean (least-squares) with some new estimate. Rather, I feel that the theory of robust estimation is a way to provide a coherent family of logical alternatives to least-squares.

These alternatives should be used to diagnose the sensitivity (stability) of our results to moderate (extreme in some of the cases studied in [1]) departures from our assumptions (such as the Gaussian error model). It will still be up to the statistician to decide which alternatives represent good analyses of the data. (Relles and Rogers [7] have conducted an interesting experiment in this regard.) The emphasis here is on *analysis*—to replace the mean by a trimmed mean and proceed blindly as we often have with least-squares is to turn us again into data processors rather than data analysts [8].

It is revealing to take a closer look at one type of robust estimate—the M-estimates. If we start with $\bar{y}$ (often other starts are used but the ideas are similar), then an M-estimator examines $y_i - \bar{y}$, compares this to some measure of the spread of $\{y_j - \bar{y}\}_{j=1}^n$ and decides how much weight should be given to the ith data point in the final estimate (or next iteration). If $|y_i - \bar{y}|$ is very large, it may be given no weight at all.

Now in the location case with least-squares start we have

$$y_i - \bar{y} \sim \bar{y} - \bar{y}_{(i)} = \hat{\beta} - \hat{\beta}_{(i)}$$

where (i) denotes the fact that the ith data point has been removed from the

computation. While robust estimation has emphasized the residual, $y_i - \bar{y}$, I prefer to emphasize the change in the least-squares estimate of β, $\hat{\beta} - \hat{\beta}_{(i)}$, caused by deleting a data point. Deletion is clearly a type of perturbation of one of the inputs to a statistical analysis, and our hope might be that small changes in input (in this case the data) should lead to small changes in output (in this case the coefficient estimates). This viewpoint is especially useful in regression where, as a number of researchers have discovered [2, 4, 5, 6], looking at just the residuals, $y_i - \bar{y}$, is much less useful than examining various functions of $\hat{\beta} - \hat{\beta}_{(i)}$ and the projection matrix, $X(X^TX)^{-1}X^T$.

A family of robust procedures gives us a quick way to scan the output changes, $\hat{\beta} - \hat{\beta}_{(i)}$, and compute new estimates by downweighting some of the data. If the robust estimate differs markedly from least-squares (or whatever the conventional wisdom is), we should find out why—a detailed analysis of the $\hat{\beta} - \hat{\beta}_{(i)}$ may be called for and no current robust procedure can be used to automatically replace such an analysis.

Now that our eyes have been opened to robust, diagnostic, and other alternative procedures, we cannot allow them to be blinded by looking only at a 10% trimmed mean or any other single estimation procedure.

REFERENCES

[1] ANDREWS, D. F., *et al.* (1972). *Robust Estimates of Location: Survey and Advances.* Princeton Univ. Press.

[2] ANDREWS, D. F. *et al.* (1976). Outlier rejectors or *M*-estimators; a comparison for linear models. Unpublished manuscript.

[3] HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **27** 87–104.

[4] HANDSCHIN, E., SCHWEPPE, F. D., KOHLAS, J. and FIECHTER, A. (1975). Bad data analysis for power system state estimation. *IEEE Trans. Power Apparatus and Systems*, **94** 329–337.

[5] HILL, R. W. (1976). Robust regression when there are outliers in the carriers. Unpublished Ph.D. thesis, Department of Statist., Harvard Univ.

[6] HOAGLIN, D. C. and WELSCH, R. E. (1977). The hat matrix in regression and ANOVA. WP 901-77, Sloan Sch. Manag., M.I.T.

[7] RELLES, D. A. and ROGERS, W. H. (1977). Statisticians are fairly robust estimators of location. *J. Amer. Statist. Assoc.* **72** 107–111.

[8] TUKEY, J. W. (1972). Data analysis, computation, and mathematics. *Quart. Appl. Math.* **30** 51–65.

REPLY TO DISCUSSANTS

I am grateful to the discussants for their thoughtful and stimulating comments. I agree with much of what has been said, and since the main points of difference have already been addressed in the paper itself, I shall aim for brevity here.

The comments of Professors Barnard, Box, Cox, Hogg, and Welsch really require no reply, as we are in nearly total agreement! The advancement of statistical methods to the present state has depended critically upon the interaction of mathematics with real data, and we cannot but benefit if more attention

is given to the characteristics of the real world than has been done in recent decades. Barnard's suggestion of a formal study of sample configurations seems to be a good one. I wonder if a modest start might be made by asking the compiler of the computer at a large research center (such as Bell Labs, the National Bureau of Standards or the University of Wisconsin) to spy on the data sets it processes in standard statistical program packages, and to record sample sizes and measures of skewness and kurtosis. Such data might be hard to interpret, and some well-meaning civil liberties organization might protest the invasion of the kurtosis's privacy, but at least it would be a start. On the other hand, more studies in the present vein will also be beneficial, and, for example, allow us to judge if the performance of Hogg's T_2 is only an artifact of post data selection, or if his term "out-Huber" will find a home in the lexicon of statistics as other than a slogan for recalcitrant partisans of nonrobust statistical methods.

One principal focus of the remaining discussants is the subject of bias, already discussed at some length in Section 5 of the paper. Dr. Eisenhart, and Professors Hoaglin and Pratt all have cogent comments on this topic. Of the three, I feel that Pratt's views come closest to my own; his observations should be of great assistance in helping careful readers appreciate the limitations as well as the strengths of the present study. I frankly think Eisenhart may be a bit too sanguine about the absence of large bias in modern data, a feeling that I think is implicit in his dissatisfaction with my "true values." He is correct (and the reference he gives to *Scientific American* illustrates this nicely) that systematic errors have decreased since 1906. But so have measurement errors, and I am not at all convinced that relative bias is smaller now than in 1880. For example, a major attempt to measure the velocity of light by Michelson, Pease and Pearson, completed in 1933 and reported on in 1935, made nearly 2900 determinations of the velocity of light. Their systematic error was considerably less than that of early attempts (see *Scientific American*), but the mean determination was still 1.7 standard deviations of a single determination below the "true value." (These data are reproduced in my reference [22].) On the other hand, I would recommend the parenthetical remark at the end of Eisenhart's penultimate paragraph to those who, like Pratt, wonder if the biases found here could also be expected in comparative experiments.

To further understand the effect of bias on performance, and to relieve the drought of theorems in this paper, I offer the following result. Let $\hat{\theta}$ be an estimator of θ with symmetric distribution $F((x - \mu)/\sigma)$, where $\mu = \theta + B$ is the point of symmetry, B the bias, σ a scale parameter. Let $L(\hat{\theta} - \theta)$ be the loss function, and let $r(\sigma) = EL(\hat{\theta} - \theta) = EL(\sigma Z + B)$ be the risk function, viewed as a function of σ, where Z has distribution F.

THEOREM. *If L is convex, then $r(\sigma)$ is a nondecreasing function of $\sigma > 0$.*

PROOF. Consider $r(\sigma) = EL(\sigma Z + B)$ as a function of σ for all real σ. The convexity of L implies $r(\sigma)$ is convex, and the symmetry of F implies $r(\sigma)$ is an

even function ($r(-\sigma) = EL(-\sigma Z + B) = EL(\sigma(-Z) + B) = r(\sigma)$), so $r(\sigma)$ is a nondecreasing function on the positive half-line. □

The implication I would draw from this is that *if* two estimators have symmetric distributions that differ only by a scale parameter (as may be approximately true, if we look at their asymptotic distributions) *and* they have the same bias, then the bias will not reverse the order of preference. The better estimator for zero bias will be better regardless of B, although the margin of preference may change. It is easy to show that this result may fail for nonconvex L, such as $L(x) = I_{[|x|\geq 1]}$. In fact, if $L(-x) = L(x)$ is nondecreasing in $|x|$ and $f(x) = F'(x)$, integration-by-parts gives

$$r'(\sigma) = \sigma \int_0^\infty [\phi((w - B)/\sigma) + \phi((w + B)/\sigma)]\, dL(w)\,,$$

with $\phi(x) = xf(x)$, which if F is $N(0, 1)$ will only be positive for all $L(x)$ if $|B| < \sigma$. Of course, if the original measurements have asymmetrical distributions, different estimators usually have different biases and the above theorem is irrelevant, but this is a major point of my Section 5 and Pratt's comments.

The second major focus of the comments of Professors Andrews, Hoaglin, Huber and Dr. Crow is the question of whether these data adequately reflect anticipated applications, for example, in the social sciences. I have little to add to my previous comments on this, other than to reiterate that I would welcome evidence on this point. Individuals' recollections and the isolated examples they produce can be highly selective, and I doubt that disasters and unprintable means occur in potential applications for these procedures with anything approaching the frequency they do in Andrews' memory. I do welcome Crow's reminder that often a transformation is necessary before one should even contemplate using one of these procedures.

While I prefer the measures of average performance I presented to the more pessimistic measure Huber gives, I am heartened that the results are not much different, with the notable exception of the outmean. In my comparisons the unreliability of the outmean was only indicated by the large values of SE(i) and SR(i). Andrews also criticizes the measure of relative error I have used, stating that it "does not reflect useful information about absolute errors." My view is that if the analysis of estimators' performances is to be meaningful, the absolute errors are irrelevant. Even though Data Sets 9 and 10 were gathered under superficially similar circumstances, it is clear that the actual circumstances may have been quite different. What matters, I feel, is the use the estimators made of the information at hand, and that this can only be judged by measuring errors on scales relative to the individual data sets. Those who, like Andrews, might prefer an index that does not weight Data Set 9 equally with the more biased set 10, may be more comfortable with indices such as Huber's.

The open-mindedness displayed, and the variety of interpretations and proposals for new work made by my distinguished discussants are most reassuring for the future of work on robust estimation.