The Influence Curve and Its Role in Robust Estimation

FRANK R. HAMPEL*

This paper treats essentially the first derivative of an estimator viewed as functional and the ways in which it can be used to study local robustness properties. A theory of robust estimation "near" strict parametric models is briefly sketched and applied to some classical situations. Relations between von Mises functionals, the jackknife and U-statistics are indicated. A number of classical and new estimators are discussed, including trimmed and Winsorized means, Huber-estimators, and more generally maximum likelihood and M-estimators. Finally, a table with some numerical robustness properties is given.

1. INTRODUCTION AND SUMMARY

This article gives a preliminary account of the properties and interpretation of the influence curve of an estimator and its use in the theory and practice of robust estimation. The influence curve is essentially the first derivative of an estimator, viewed as functional, at some distribution (in an infinite-dimensional space), and it is shown how it can be used not only to derive asymptotic variances, but also to study several local robustness properties which are defined and intuitively interpreted. A number of classical examples of robust and non-robust estimators are investigated.

The article is designed both for mathematical and applied statisticians. Sections 4, 6, and 7 are mainly for mathematical statisticians, Sections 3, 5, and 8 mainly for applied statisticians. Section 2 provides the basic definitions of the influence curve, first more specialized, then more abstract; Section 5 includes some basic definitions in the theory of robust estimation. Sections 2–4 deal with the influence curve *per se* (definition, interpretation with examples, place in mathematics and history of statistics). Sections 5–8 discuss the influence curve in the framework of robust estimation (basic intuitive concepts in robust estimation with examples, the statistical theory of robust estimation, a mathematical optimality result, some numerical examples and their discussion). Mathematical rigor has been emphasized less than intuitive meaning, and there are indeed still many open mathematical details (like regularity conditions) to which, it is hoped, some mathematical statisticians will address themselves. It is also hoped that applied statisticians will be able to derive, study and interpret other influence curves than those given here, in Hampel [10] and in Andrews, *et al.*, [1].

The study of influence curves serves to deepen our understanding of estimators (e.g., of the relation between trimmed means, Winsorized means and Huber-estimators). It also serves to derive new estimators with prespecified robustness properties (e.g., the three-part descending M-estimators, or the optimal robust estimators of scale, with the median deviation as limiting case). There are close relations to Tukey's jackknife, and to Hoeffding's U-statistics. The study of various norms connected with the influence curve (based mainly on Huber's work with the gross-error-model) leads to a theory of robust estimation "near" parametric models which is meant to supplement Fisher's classical theory of estimation in strict parametric models.

The specific problems of simultaneous estimation and estimation in structured designs (like robust regression) are not discussed here. Although there has been some progress recently and methods which are likely to turn out useful for practice do already exist, the precise theoretical aims do not yet seem to be clarified enough to be included here.

2. DEFINITION OF THE INFLUENCE CURVE

2.1 Special Case

Let R be the real line, let T be a real-valued functional defined on some subset of the set of all probability measures on R, and let F denote a probability measure on R for which T is defined. Denote by δ_x the probability measure determined by the point mass 1 in any given point $x \in R$. Mixtures of F and some δ_x are written as $(1 - \epsilon)F + \epsilon\delta_x$, for $0 < \epsilon < 1$. Then the influence curve $\mathrm{IC}_{T,F}(\cdot)$ of (the "estimator") T at (the "underlying probability distribution") F is defined pointwise by

$$\mathrm{IC}_{T,F}(x) = \lim_{\epsilon \downarrow 0} \{T[(1 - \epsilon)F + \epsilon\delta_x] - T(F)\}/\epsilon$$

if this limit is defined for every point $x \in R$.

* Frank R. Hampel is professor, Department of Statistics, Abt. 9, Swiss Federal Institute of Technology, Zürich, Switzerland. This article contains revised excerpts from the author's dissertation written at the University of California, Berkeley. Research for this article was partly supported by NSF Grant GP-7454 and ONR Contract No. N00014-67-0151-0017. The author gratefully acknowledges what he has learned about the theory of robust estimation from Peter J. Huber and about robust data analysis from Cuthbert Daniel and John W. Tukey. There were also stimulating discussions with P.J. Bickel, R. Gnanadesikan, E.L. Lehmann, C.L. Mallows and the late F. Stephan. For providing the opportunity to work in this field, thanks are due to R. Gnanadesikan, P.J. Huber, L. LeCam, E.L. Lehmann, V. Strassen and G.S. Watson. Suggestions for improving the article were given by C. Daniel, C. Eisenhart, P.J. Huber and C.L. Mallows, and also by an associate editor and a referee.

© Journal of the American Statistical Association
June 1974, Volume 69, Number 346
Theory and Methods Section

Examples: (i) The arithmetic mean $T = \int x dF(x)$ is defined for all probability measures with existing first moments (in particular for all empirical distribution functions and for all finite mixtures of distributions for which it is defined). Let the mean of F exist and be equal to μ. Then the influence curve of T is defined at F and is given by

$$\mathrm{IC}_{T,F}(x) = \lim_{\epsilon \downarrow 0} [(1 - \epsilon)\mu + \epsilon x - \mu]/\epsilon = x - \mu \quad (x \in R).$$

(ii) The variance $T = \int (x - \mu)^2 dF$ at an F with existing variance σ^2 and (known) mean μ has the influence curve

$$\mathrm{IC}_{T,F}(x) = \lim_{\epsilon \downarrow 0} [(1 - \epsilon)\sigma^2 + \epsilon(x - \mu)^2 - \sigma^2]/\epsilon$$
$$= (x - \mu)^2 - \sigma^2 \quad (x \in R).$$

2.2 General Case

Let Ω be a complete separable metric space, let T be a vector-valued mapping from a subset of the probability measures on Ω into the k-dimensional Euclidean space R^k, and let F lie in the domain of T. Let δ_ω denote the atomic probability measure concentrated in any given $\omega \in \Omega$. Then the vector-valued "influence curve" of T at F is defined pointwise by

$$\mathrm{IC}_{T,F}(\omega) = \lim_{\epsilon \downarrow 0} \{T[(1 - \epsilon)F + \epsilon\delta_\omega] - T(F)\}/\epsilon$$

if this limit is defined for every $\omega \in \Omega$.

3. HEURISTIC INTERPRETATION AND PROPERTIES

3.1 Elementary Examples

Let us consider a sample of real-valued observations $x_1, \cdots, x_n$ and their arithmetic mean $\bar{x}_n$ as estimator. Our question is: How does the mean change if we throw in an additional observation at some point x? Obviously, the difference between the new mean $\bar{x}_{n+1}$ and the old mean $\bar{x}_n$ is

$$\bar{x}_{n+1} - \bar{x}_n = (n\bar{x}_n + x)/(n + 1) - \bar{x}_n$$
$$= (x - \bar{x}_n)/(n + 1).$$

It is worthwhile to ponder over this simple result. The influence of a single observation with value x, e.g., a gross error, on the mean of the rest is roughly inversely proportional to the original sample size, and for any fixed sample size, it increases linearly with the difference between x and the original mean over all bounds.

Now let us consider the estimation of the variance of n observations with known expectation equal to zero. How is the estimated variance s_n^2 affected by adding a single observation in x? We have $s_{n+1}^2 = (ns_n^2 + x^2)/(n+1)$, hence

$$s_{n+1}^2 - s_n^2 = (x^2 - s_n^2)/(n + 1).$$

Again the result should be contemplated. If $|x| < s_n$, the variance is decreased, though at most by $s_n^2/(n + 1)$. For $|x| \to \infty$, on the other hand, the resulting increase in variance grows not only linearly, but even quadratically over all bounds.

3.2 The Trimmed Mean and the Median

A slightly less simple example is provided by the α-trimmed mean, one of the "classical" robust substitutes for the arithmetic mean, designed to be less dependent on a few wild stray values. The α-trimmed mean (for $0 < \alpha < \frac{1}{2}$) is obtained in the following way: order the n observations; delete the αn smallest and the αn largest observations; take the arithmetic mean of the rest. If αn is not an integer, the usual practice is to round it to a neighboring integer, e.g., to $[\alpha n]$, the largest integer $\leq \alpha n$. Now let us ask for the effect of an additional observation x on the trimmed mean.

First, let us consider the case where $0 < [\alpha n] = [\alpha(n + 1)] = h$, say; define

$$y = y(x) = x^{(h)} \text{ (the } h\text{th order statistic) if } x \leq x^{(h)},$$
$$= x \text{ if } x^{(h)} < x < x^{(n-h+1)},$$
$$= x^{(n-h+1)} \text{ if } x^{(n-h+1)} \leq x.$$

Denote the α-trimmed mean of n observations by $\bar{x}_{\alpha,n} = (\sum' x^{(i)})/g$, where $g = n - 2h$ and $\sum'$ means summation over the $x^{(i)}$ with $h + 1 \leq i \leq n - h$. Then $\bar{x}_{\alpha,n+1} = (\sum' x^{(i)} + y)/(g + 1)$ (using the notation for the original sample), and

$$\bar{x}_{\alpha,n+1} - \bar{x}_{\alpha,n} = (\sum' x^{(i)} + y - \sum' x^{(i)}$$
$$- (1/g) \sum' x^{(i)})/(g + 1)$$
$$= (y - \bar{x}_{\alpha,n})/(g + 1).$$

This result should be sketched as a function of x.[1] Compared with the corresponding graph for the arithmetic mean, it shows a higher "influence" of the additional observation in the central part of the sample, and constant (hence bounded) nonzero changes of the estimate if x falls into the left or right tail. In particular, the "influence" of an extreme outlier on the value of a trimmed mean is *not* zero, as one would naively expect (arguing that the outlier will be "thrown out"); rather it is equal to the influence of an additional x at $x^{(h)}$ resp. $x^{(n-h+1)}$, i.e., to the maximum possible influence on each side!

A similar result can be obtained for the case $0 < [\alpha n] = [\alpha(n + 1)] - 1$. To unify and simplify the formulas, we now do one standardization and two passages to the limit. First, we divide the effect of a "contaminating mass" (allowing $k \geq 1$ additional observations in x) by the "size" of the contaminating mass, namely, $k/(n + k)$ (or k/n, which eventually leads to the same limit). Then we let the sample size n tend to infinity, but still keeping the fraction of contaminating mass fixed. For example, we may double, triple, . . . all $x^{(i)}$, always also doubling, tripling, . . . the contaminating observation at x; the estimates and our standardized "finite sample influence curve" will quickly tend to certain limits. But the empirical cumulative distributions of our samples (without the contamination) may also approximate any fixed probability distribution, such as the normal distri-

[1] Cf. also the graphs [1, p. 97f]; [38, p. 4–31f].

bution, and often it makes sense to define the estimator also for such "nonsamples." E.g., the α-trimmed mean of *any* distribution F can be defined by

$$\bar{x}_\alpha(F) = \int_\alpha^{1-\alpha} F^{-1}(t)dt/(1 - 2\alpha),$$

with the natural interpretation for discontinuities of F; if, for F, we plug in an empirical distribution function F_n, the integral becomes a finite sum, which "trims" also fractions of observations if αn is not integer-valued. Once we have thus extended the range of definition of our estimator, we may compute the standardized effect of any contaminating point mass of fixed size ϵ in x on the value of the estimate "at" the distribution F, as a function of x. In our second passage to the limit, we now let ϵ tend to zero and obtain, for each x, the standardized effect of an "infinitesimal" contamination on the estimate which is nothing but the influence curve as defined previously. In our special case of the α-trimmed mean, the influence curve at an F with continuous unimodal density symmetric around zero (cf. [10]; [1, p. 34]) is

$$\begin{aligned} \mathrm{IC}_{\bar{x}_\alpha, F}(x) &= F^{-1}(\alpha)/(1 - 2\alpha) \quad \text{for } x < F^{-1}(\alpha) \\ &= x/(1 - 2\alpha) \quad \text{for } F^{-1}(\alpha) \leq x \leq F^{-1}(1 - \alpha) \\ &= F^{-1}(1 - \alpha)/(1 - 2\alpha) = -F^{-1}(\alpha)/(1 - 2\alpha) \\ &\qquad \text{for } F^{-1}(1 - \alpha) < x. \end{aligned}$$

This function may be compared with the previous finite-sample version.

If we let $\alpha \to \frac{1}{2}$, we obtain the median $F^{-1}(\frac{1}{2})$ as a limiting case of the trimmed means. Assume for simplicity that F has a continuous unimodal density f symmetric around zero. The normalized influence of a contaminating point mass in x on the median depends only on the sign of x, provided the mass is small enough, and on the distribution of F near 0, i.e., in the limit only on $f(0)$. In this way we obtain for the median IC $(x) = \text{sign } (x)/[2f(0)]$. We notice that the IC is bounded[2] and monotone. Furthermore, we notice the jump at 0 (which, e.g., causes trouble for the jackknife), but we also realize that the jump occurs only as the limit of a very steep central part (unlike most estimators which contain rejection rules, where there is a discontinuity even for finite sample size). And the increasing steepness itself is only caused by the normalization (division by the size of the contaminating mass); the actual change of the estimate goes to zero for vanishing size of the contamination.

3.3 The Winsorized Mean

We have realized above that the α-trimmed mean does not really "throw out" outliers, in the sense of ignoring them completely, but in effect "brings them in" towards the bulk of the sample. But what about the α-Winsorized mean which had been designed specifically to "bring in" outliers? It can be obtained by replacing the $[\alpha n] = h$ extreme observations on either side by $x^{(h+1)}$ resp. $x^{(n-h)}$, hence counting these values $(h + 1)$ times, and taking the arithmetic mean of this modified sample. An asymptotically equivalent definition, which holds for all distributions F (empirical or otherwise), is

$$\int_\alpha^{1-\alpha} F^{-1}(t)dt + \alpha[F^{-1}(\alpha + 0) + F^{-1}(1 - \alpha - 0)] \qquad (0 < \alpha < \tfrac{1}{2}).$$

Its IC ([10]) came as a surprise. For an F with a continuous unimodal density f symmetric around zero, it is equal to x for $|x| \leq F^{-1}(1 - \alpha)$ and to $\{F^{-1}(1 - \alpha) + \alpha/f[F^{-1}(1 - \alpha)]\} \cdot \text{sign } (x)$ for $|x| > F^{-1}(1 - \alpha)$. Again the graph should be sketched by the reader. The IC is indeed bounded, the outliers "brought in," but there is a jump at $F^{-1}(\alpha)$ and $F^{-1}(1 - \alpha)$. Furthermore both slope in the center and supremum differ from that of the α-trimmed mean. On second thought the reason becomes clearer. The "jump," for one thing, appears again only as the limit of a very steep slope, as with the median. But the main point is that the mass of the tails is put on single order statistics resp. single points in the limit, and shifting them or wiggling them causes appreciable fluctuations of the Winsorized mean which are determined solely by the density in (and near) these points. A contamination in the central part, on the other hand, has the same influence as on the arithmetic mean, while the trimmed mean spreads the influence of outliers evenly over the central part, thus giving it a higher weight.

Thus, we discover that both the trimmed mean and the Winsorized mean restrict the influence of outliers, but in different ways. While the IC of the former is always continuous, the IC of the latter is discontinuous and very sensitive to the local behavior of the true underlying distribution at two of its quantiles. If the underlying distribution is strictly normal, the trimmed mean has a smaller asymptotic variance than the Winsorized mean with the same supremum of the IC (see [10]) (it is even optimal in this respect), and the α-trimmed mean is less sensitive to outliers than the α-Winsorized mean; on the other hand the α-Winsorized mean has a smaller asymptotic variance than the α-trimmed mean, and even for finite samples it is almost optimal among linear functions of the order statistics with constant influence (zero weight function) for a fraction α in the tails (see [32]). It has long been known in life-testing[3] that a one-sided Winsorized mean is strictly optimal in the latter sense already for finite samples if the underlying distribution is exponential; only the ways of looking at this fact (with regard to "robustness," "breakdown point," etc.; see [10, p. 89]) seem to be new.

3.4 General Approach

After these examples, let us consider what we have done in general terms. First, we have replaced a given

[2] Actually, as it turns out, it is bounded by the lowest bound possible (cf. [10] and Section 7; also [1, p. 35]).

[3] See, e.g., [5, p. 41]; the formula for the variance contains an obvious misprint.

sequence of estimators (such as the usual α-trimmed means, which round α to a multiple of $1/n$ for each sample size n) by a simple functional on the space of probability distribution functions which is independent of the sample size. Some sequences of estimators, e.g., maximum likelihood estimators, are already of such a form; others, like trimmed means or the empirical variance for unknown mean, differ only slightly from such a functional, and typically these differences are due to the discreteness of the observations or to some bias correction term of the order $1/n$. A third group of sequences of estimators, which includes Bayes estimates and density estimates, are more delicate to handle. They may be "asymptotically" equivalent to some functional (e.g., Bayes estimates to maximum likelihood estimates), but for fixed small or moderate n, the discrepancy may still be so large that it may be better to use a different functional for each sample size as approximation. On the other hand, if we are given any functional, we can use it as an estimator by applying it to the empirical cumulative distribution.

Once we have described, at least approximately and for roughly the sample size considered, our given sequence of estimators by a functional on the space of probability distributions, we can try to do what every physicist, engineer or other practitioner of mathematics would regard as a standard tool: namely, try to linearize the estimator locally, i.e., approximate it by a linear functional in the neighborhood of some probability distribution. In other words, we can try to use a Taylor expansion, and for this to obtain the derivative of a functional in the infinite-dimensional space of all probability measures or even bounded signed measures. Now we remember that "the derivative" of a function on n-dimensional Euclidean space (i.e., of n variables)—if it exists at all—can be described numerically by the n partial derivatives in the directions of the coordinate axes. Analogously, "the derivative" of a functional (if it exists, e.g., in the sense of Volterra, or of Fréchet; see Section 4) can be described by the infinite set of partial derivatives in the directions of the point masses, i.e., along mixtures of the form $(1 - \epsilon)F + \epsilon\delta_x$, where δ_x is the point mass 1 in x; but this set of partial derivatives is nothing but the influence curve.

One question remains to be discussed: at which distribution(s) should the influence curve be evaluated? The simplest and most natural answer is: at the (theoretical) underlying probability distribution (e.g., at a normal distribution with specified parameters). This, indeed, gives us the greatest gain in information obtainable from a single distribution, as it describes not only many qualitative features, but even the quantitative behavior in the limit of very large n ("asymptotically"). To be more realistic, one would prefer to compute finite sample versions ($\epsilon = 1/(n + 1)$ instead of $\epsilon \downarrow 0$) at empirical distribution functions; however, there are too many of the latter, and if one takes averages (expectations) of such finite-sample influence curves, one often loses sight of some important features of the estimator. Therefore, it seems to be a reasonable strategy to obtain the exact quantitative asymptotic curve (which is simple enough) and as much qualitative and semi-quantitative information as possible about the behavior at other distributions (especially empirical ones in a larger or smaller neighborhood of the underlying probability distribution), in order to be able to make reasonable extrapolations back from the simpler but somewhat simplified (slightly unrealistic) pictures at "sample size infinity."

3.5 Some Properties of the Influence Curve

Let us now briefly state some useful formulas (cf. [10, 1]) involving the IC of an estimator T at an underlying distribution F (disregarding the necessary regularity conditions). If G is close to F, then

$$T(G) \approx T(F) + \int \mathrm{IC}_{T,F}(x)d(G - F)(x)$$
$$= T(F) + \int \mathrm{IC}_{T,F}(x)dG(x).$$

In particular, $\int \mathrm{IC}_{T,F}(x)dF(x) = 0$. And the (usual) asymptotic variance of T under F equals $\int \mathrm{IC}^2_{T,F}(x)dF(x)$. As a consequence, two estimators whose IC's at some F happen to coincide, will have the same local "asymptotic" properties under this F (as is the case for suitable pairs of trimmed means and Huber-estimators), but they may still differ widely under some other G or in more global properties like the breakdown point (cf. [10]).

We note in passing that it is sometimes possible to extend the Taylor expansion beyond the first term (even into an infinite series); another approach to determine finite (as opposed to infinitesimal) changes is to use suitable integration over influence curves. Finally, we realize that the jackknife (cf. [31]) can be viewed as a finite-sample version of an IC, computed at the empirical distribution (see also [20]).

To sum up: The influence curve can be drawn and looked at, and its various properties (qualitative shape, supremum, maximal slope, points and heights of jumps, points and intervals where it is zero, etc.; see also the later sections) together with a bit of qualitative information about type or regularity of the estimator (how the influence curve behaves in a neighborhood, and how the limit which defines it is approached) tell us a lot about the detailed behavior of the estimator and about how the separate observations contribute to the estimated value.

4. MATHEMATICAL AND HISTORICAL ASPECTS

The influence curve has already penetrated the statistical literature in disguised form to a surprising extent, perhaps most clearly as the integrand in the first term of the von Mises expansion (cf. [39]), and perhaps most frequently as the square root of the

integrand in the formulas for asymptotic variances of asymptotically normal estimators and as the normalized multiple of the score functions of maximum likelihood estimators. It occurs (as normalized multiple) in the score generating functions of rank tests and estimators derived from them as well as in the defining ψ-functions of M-estimators (in the sense of Huber [15]), and its derivative appears in the weight functions for linear functions of the order statistics. More indirectly, we can find it also in the likelihood part of the *a posteriori* distribution of Bayes estimates. In general, it seems to appear in all the "normal" cases which can be linearized locally in standard ways.

Apparently the first to stress the importance and usefulness of Volterra-type derivatives for estimators (in a different context) was von Mises (see [39] and earlier papers, or [40]). The influence curve, in its essence, is nothing but a simplified description of the first term of the von Mises expansion, stripped of all the messy regularity conditions and defined in such a simple and direct way that it can be easily calculated for most estimators.[4]

It seems that the first who tried to use derivatives of estimators and norms based on them in connection with the problem of robustness was Takeuchi [34]. In retrospect, we find an intuitive notion of an influence curve in [12] (cf. also [35]). The author's work [10] was first used by Huber [20] and his students [28, 22]; a number of applications appeared in Andrews *et al.* [1]. The "Princeton robustness seminar 1970/71" also clarified the relations between influence curves and their finite-sample versions (including the jackknife) with which Tukey had been experimenting in a parallel direction (cf. [38, Ch. 4] and [1, p. 96]). The relation between von Mises derivatives and U-statistics [14] has been investigated by C. Mallows [30]. Among other things, he showed that the finite series into which a U-statistic can be expanded, corresponds to a finite von Mises expansion without "diagonal" terms.

Finally, the close ties between influence curves and Huber's M-estimators [15] have to be reemphasized. Let us only consider the one-dimensional case. An M-estimator is essentially described by a suitable class of functions $\{\psi_\theta\}$ on (a subset of) the real line, and the estimate is then the (or a selected) solution θ of $\int \psi_\theta dF = 0$, where F denotes the empirical cumulative distribution function. Let, for some F, $\int \psi_\theta dF = 0$, let G be close to F (in the sense of the weak* topology) and let ψ_θ have a derivative ψ'_θ with respect to θ. Assuming suitable regularity conditions, we find for the estimate $\theta + \Delta$ at $G = F + H$ the approximate relations

$$0 = \int \psi_{\theta+\Delta} dG \approx \int (\psi_\theta + \Delta\psi'_\theta)(dF + dH)$$

$$\approx \int \psi_\theta dH + \Delta \int \psi'_\theta dF$$

(ignoring the higher-order term $\Delta \int \psi'_\theta dH$); hence,

$$\Delta \approx -\int \psi_\theta dH \Big/ \int \psi'_\theta dF.$$

Therefore, $-\psi_\theta / \int \psi'_\theta dF$ is the IC and derivative, in whatever sense it exists, at F, and ψ_θ is seen to be just a multiple of the influence curve ([10, p. 47f]). This opens many possibilities of defining new estimators with prescribed properties [10]; some successful first applications (notably the three-part descending M-estimators) can be found in [1].

5. SOME BASIC CONCEPTS IN ROBUST ESTIMATION

5.1 Main Concepts

There are two basic "operations" which occur in discussing robust estimation of some quantity with a single unstructured sample (Tukey and Hampel, Princeton Robustness Seminar). One is to "throw in" a small mass of arbitrary "contamination" which may be anywhere and which includes "outliers," "gross errors," "bad values" or whatever one wants to call them. It is known by now that the proportion of gross errors in data, depending on circumstances, is normally between 0.1% and 10%, with several percent being the rule rather than the exception (cf. [37, p. 14], also [4], [10] and various oral remarks). These are the "non-missing values" (Cuthbert Daniel), the ones that should not be there. The dangers even of "hidden" contamination for standard analyses have also been known for some time now (see [36]). The "inverse" operation of "throwing in" is to "take out" some observations, as it is done, e.g., in the jackknife. The second basic operation is to "wiggle around" with the observations, i.e., to change their values slightly (as happens in rounding and grouping and due to some local inaccuracies, e.g., of the measuring instrument).[5] We note that the second operation (which includes its own inverse) can be reduced to the first one: to shift an observation slightly to some neighboring point is the same as to remove it and to put it in at the new point; hence, the effect of "wiggling" is about the difference of the effects of "throwing in" at two neighboring points. Now we remember that the (normalized) effect of "throwing in" a small contamination at some point is approximately measured by the value of the influence curve in that point, and moreover we now realize that the (normalized) effect of "wiggling" somewhere is approximately measured by a normalized difference or simply the slope (first derivative) of the influence curve in that point. This explains the central role of the influence curve in the study of "local" robustness problems.

We still need some guidance as to how far "local" extends, up to what distance we may still try to linearize.

[4] For more details, cf. [10]; for details on regularity conditions, cf. [6, 26, 27].

[5] The similarity of the study of these two operations to "sensitivity tests" and "perturbation theory" in engineering and other fields was immediately noticed by some practical statisticians when the operations were explained to them. It is surprising that it took statistics so long to arrive at such simple and basic tools.

A first hint is given if we know which distance is already "far" away. Now we know precisely the borderline beyond which the estimator is totally unreliable, namely, the "*breakdown point*" of the estimator. Loosely speaking, it is the smallest percentage of free contamination which can carry the value of the estimator over all bounds.[6] The breakdown point thus measures a "global" aspect of robustness. For our local linearization, it may serve us as a "distant" orientation point: as long as the amount of contamination is "small" compared with the breakdown point (as a first wild guess: still with ratios of these two quantities around $\frac{1}{4}$, or perhaps even above $\frac{1}{2}$), the linear approximation given by the IC may prove useful (which is not to say accurate, of course).

We may try to look at the influence curve as a whole, but we may also try to summarize its most important features in a few numbers, especially in cases in which the type and qualitative behavior of the estimator is known. One such summary value, the expected square of the IC (in particular under the "ideal" distribution of the assumed parametric model), is already well known (independently of the IC) and widely used as *asymptotic variance*. It approaches its minimum as the IC approaches a certain multiple of the log likelihood derivative.

The second most important summary value of the IC, and the most important for robustness considerations, is the supremum of the absolute value of the IC, the "*gross-error-sensitivity*"

$$\gamma^* = \sup_x |\mathrm{IC}(x)|$$

(called σ^* in [10]). It measures the worst approximate influence which a fixed amount of contamination can have on the value of the estimator (hence it may be regarded as an approximate bound for the bias of the estimator). Often it is infinite, but under some natural restrictions there will be a positive lower bound, the gross-error-sensitivity of the, or a, "most robust" estimator in this respect. Typically, putting a bound on the IC is the most important step in "robustifying" an estimator (e.g., the maximum likelihood estimator), and this will often conflict with the requirement of asymptotic efficiency: the lower the bound, the larger the smallest variance that can be achieved. Thus, we get a class of "admissible robust estimators," namely, those which cannot be improved simultaneously with respect to asymptotic variance and gross-error-sensitivity. In the case of the location of the normal distribution, this class contains the trimmed means and the Huber-estimators, with the arithmetic mean and the median as limiting cases. For the general one-dimensional case, it is shown below how to find such "admissible robust estimators." A number of examples, including the binomial, Poisson, exponential and Cauchy distributions, are discussed in [10].

5.2 The Median Deviation

Example: The following example may find particular interest. The scale estimate which is the counterpart of the median as "most robust" estimator of location (both with regard to gross-error-sensitivity and to breakdown point) is *not* (a multiple of) the interquartile range, but the median of the absolute deviations from the median: $\mathrm{med}_i|X_i - \mathrm{med}_j X_j|$. In analogy to standard deviation and mean deviation, we may call it median deviation. Its IC (e.g., at the normal distribution) equals $\pm$ a constant, with jumps at those two points symmetrical to the median which include half of the total mass between them. The median deviation agrees with half the interquartile range for symmetric samples, and locally (asymptotically) at symmetric underlying distributions, but in general they are different. In particular, the breakdown point of the interquartile distance is only $\frac{1}{4}$, while the breakdown point of the median deviation is $\frac{1}{2}$, which is the largest possible value for scale estimates (to be sure: as long as false dispersions arbitrarily close to zero are considered as bad as those arbitrarily close to infinity; otherwise the bound would be one). For those who may be doubtful about the importance of the difference between the two breakdown points, the Monte Carlo results in [1], especially the comparison of the estimators $D15$ and $P15$ (see [1, e.g., pp. 253, 260]), provide ample evidence for it even though the scale estimates here are only auxiliary ones for the location parameter and are nearly identical "at" the underlying symmetrical distributions.

It may be noted that the median deviation would be the natural (nonparametric) estimator of the "probable error" of a single observation. Since the probable error of estimates was used widely in the early days of statistics, it seems rather surprising that the median deviation was apparently hardly, if ever, applied and only rarely considered theoretically. It was, in fact, mentioned briefly by Gauss [9], who noted its simplicity, but with the same breath, dismissed it because of its low asymptotic efficiency of about 40 percent for strictly normal data (cf. also [33]). However, out of seven scale estimates Gauss [9] considered, the median deviation was the only robust one; on the other hand, the "optimal" standard deviation may easily have an asymptotic efficiency as low as 40 percent with fairly "good," "approximately normal" real data (even without big blunders), e.g., if their distribution, following Jeffreys ([24, Ch. 5.7]), is represented by a t-distribution with about 5 degrees of freedom. Moreover, the median was and still is commonly used despite its inefficiency (cf. also the most interesting discussion of mean deviation *vs.* standard deviation in [36]).

The median deviation was rediscovered and derived in [10] as the M-estimate of scale with the smallest

[6] For a precise definition and examples, see [10] and, slightly improved, [11]; for some simple early examples, see [12]; and for some finite-sample cases, see [1]. A short but excellent discussion of the intuitive meaning of both influence curve and breakdown point is given in Huber [20] by drawing the analogy to the stability aspects of, say, a bridge.

possible gross-error-sensitivity at the normal (and many other) models and as the limiting most robust scale estimate in Huber's gross error model (for $c \to 0$ in [15]). It can also be derived as the maximum likelihood estimator for the scale families with distribution $F_\theta(x) = 0$ for $x \leq 0$, $F_\theta(x) = \frac{1}{2}(x/\theta)^\alpha$ for $0 \leq x \leq \theta$, $F_\theta(x) = 1 - \frac{1}{2}(x/\theta)^{-\alpha}$ for $x \geq \theta$ ($\theta > 0$) and the shape parameter $\alpha > 0$. In particular, the Cauchy distribution (which also supplies a very nice scale estimate, compare [10]) is seen to be a smoothed-out version of the distribution with density (in its standardized form) $f(x) = \frac{1}{4}$ for $|x| \leq 1$ and $f(x) = \frac{1}{4}x^{-2}$ for $|x| \geq 1$, in the same sense in which the logistic distribution is a smoothed-out version of the double-exponential. In passing, we note the different tail behavior of the maximum likelihood estimates for location and scale in the normal case and in the cases just mentioned, its relation to monotone likelihood ratios, and its consequences for robustness (cf. [29, p. 331; 10, p. 94; 3]).

The possible uses of the median deviation, as a crude but simple and safe scale estimate, include:

(i) as a rough but fast scale estimate in cases where no higher accuracy is required;
(ii) as a check for more refined computations;
(iii) as a basis for the rejection of outliers; and
(iv) as a starting point for iterative (and one-step) procedures, especially for many other robust estimators.

5.3 Further Concepts

While the gross-error-sensitivity is typically (and especially for M-estimators) a good quantitative measure of robustness, there are circumstances in which it has to be taken with a grain of salt. In particular, R-estimators (estimators derived from rank tests; cf. [1]) may have an infinite gross-error-sensitivity and still behave fairly stably (e.g., the normal scores estimator), while on the other hand L-estimators (linear combinations of order statistics) may have a bounded influence curve and still break down at once (e.g., the asymptotically optimal L-estimator for the logistic distribution). Thus, it becomes necessary to consider first a continuity (qualitative robustness) property of the estimators before going on to some sort of differentiability and even quantitative properties of the derivative. Such a general concept of *qualitative robustness* is discussed and defined in [10, 11]; it makes use of the nice statistical properties of the Prokhorov distance and essentially says that a weak*-continuous functional (or sequence of estimators) is qualitatively robust; this relation is almost an equivalence. Roughly speaking, if two empirical cumulatives are "close" to each other (in the weak sense which is used, e.g., in the central limit theorem), then the estimates based on these samples should be close to each other.

Having defined a measure for the worst possible effect of "throwing in" contamination, we may sometimes also want a measure for the worst (approximate) effect of "wiggling" the observations. Such a measure is provided by the smallest Lipschitz constant which the influence curve obeys, the *"local-shift-sensitivity"*

$$\lambda^* = \sup_{x \neq y} |\mathrm{IC}(x) - \mathrm{IC}(y)|/|x - y|.$$

It is particularly relevant when one considers the local effects of rounding or grouping and it also comes up in a comprehensive discussion of rejection of outliers. For the proper interpretation of λ^*, however, one has to keep in mind that it refers only to standardized local changes of the value of the estimator, so that even an infinite value of λ^* may refer only to a very limited actual change, and also that an infinite λ^* may or may not be the limit of finite slopes for finite sample sizes. To take an example, the median has $\lambda^* = \infty$ at the normal (and many other) distributions which is only the limiting result of a very narrow and (if standardized) very steep slope in the center of a sample; but this value of λ^* is intimately connected with the failure of the jackknife for the median. Other examples, where a jump occurs even in finite samples, are provided by most rejection rules (e.g., Tukey's skipping procedures [38, 1]).

It is often of interest to know whether an estimator rejects outliers and, if so, at what distance. There may be a region outside of which the influence curve is identically zero, and the boundary of the smallest region with this property plays a special role. The distance of its most distant point from some suitable center of the distribution (e.g., center of symmetry) may be called the *"rejection point"* ρ^*. Thus, all observations farther away than the rejection point (and possibly others, too) are rejected completely.[7]

5.4 Descending M-Estimators

Example: A class of estimates which was constructed to possess all the robustness properties mentioned—qualitative robustness, highest breakdown point possible, bounds on gross-error-sensitivity, local-shift-sensitivity and rejection point, and at the same time rather high efficiency at the normal (and other) distributions—are the three-part descending M-estimators. They are defined in the following way. Let d be the (unscaled) median deviation. Let $\psi_{abc}(x) = \psi(x)$ be defined by

$$\begin{aligned}
\psi(x) &= x && \text{for } |x| \leq a,\\
&= a \cdot \operatorname{sign}(x) && \text{for } a \leq |x| \leq b,\\
&= a \cdot [x - c \cdot \operatorname{sign}(x)]/(b - c) && \text{for } b \leq |x| \leq c,\\
&= 0 && \text{for } c \leq |x|
\end{aligned}$$

(e.g., $a = 2$, $b = 4$, $c = 8$ as multiples of the median deviation—not the standard deviation). Then the location estimate θ is the solution (closest to the median) of $\sum \psi[(x_i - \theta)/d] = 0$. It can be found iteratively in a few steps (or very well approximated by just one step) starting with the median. These estimates were proposed

[7] Note, however, that some estimators, such as the maximum likelihood estimator of location for the Cauchy distribution, give very little influence to extreme observations and behave almost like estimators with low rejection point, though their rejection point is infinite.

by the author and were entered, with various suitable choices of the three parameters a, b, c, in the Princeton Monte Carlo project where they turned out to be highly successful (see [1]).

5.5 Application of the Concepts

We note that all of the local concepts (including the IC itself) may and sometimes should be normalized or standardized in various ways, e.g., by dividing them by the asymptotic standard deviation of the estimator or by the minimum possible value of the robustness concept considered. This is especially important if several parameters or transformations of parameters are being contemplated. Undoubtedly, future practical use will show which possible variants are most reasonable in given circumstances.

Now it is easier for us to summarize the basic intuitive requirements for robust estimates. They should react little to small perturbations—corresponding to qualitative robustness—and they should be safe even in the presence of large contamination (or many gross errors)—corresponding to a high breakdown point. They should keep a bound on the maximal relative influence of any fixed amount of contamination—corresponding to a low gross-error-sensitivity. And possibly they should also react smoothly to rounding and grouping—meaning a low local-shift-sensitivity—and separate obviously extreme observations from the bulk of the data—meaning a low rejection point. Under these side conditions, they should still estimate about the right quantity—which is formalized later by requiring consistency in Fisher's original sense—and should have as small a variance as possible under the ideal parametric model—which means being highly correlated with the maximum likelihood estimator.

We note that finite rejection point and finite local-shift-sensitivity may not always be appropriate: for a first crude estimate one needs something that looks at all observations in some monotone way in order to find the bulk of the sample to start with; from there one may proceed to reject outliers. And if one wants a dichotomic separation into "completely proper" observations and "clear outliers," as opposed to "smooth rejection," one normally has to pay by a jump in the influence curve.

It is extremely instructive to study the multitude of existing estimates (not only those for the model of normality!) under the aspects just mentioned. A short list of examples for nonnormal models can be found in [10]. Any attempt to present even a scant selection of further examples would go beyond the scope of this article.

6. ROBUST ESTIMATION IN A PARAMETRIC MODEL: THEORETICAL BACKGROUND

This section sketches very briefly the outline of a theory of robust estimation as developed in [10]. As a theory, it is not to be taken literally in applications, but it is an attempt to provide the necessary general conceptual framework which has been lacking so far. It is related to (though logically independent of) Huber's two early approaches [15, 16, 17]; moreover, Huber is developing a theory of robust tests (see [16, 18, 19, 20, 21]; cf. also [22]) which opens new ways in another direction.

Let $\{F_\theta, \theta \in \Theta\}$ denote a parametric model (of probability distributions on some space) where Θ is nice, the F_θ's are nice, and the mapping $\theta \to F_\theta$ is also nice (which may mean to imply, e.g., that all F_θ's have densities f_θ with respect to some σ-finite measure which are differentiable with respect to θ). Note that strong regularity conditions and conceptual simplicity are intrinsic properties of reasonable parametric models; otherwise they would not be used. A functional (estimator) T with values in Θ is defined as "*Fisher-consistent*" at the parametric model iff $T(F_\theta) \equiv \theta$ for all θ (cf. [27]). This seems to come close to what Fisher had in mind with "consistency," and to his first and last definition (see, e.g., [7, 8]; cf. also [10]). Note that nothing is required for F's outside the parametric model! If T is differentiable at all F_θ with sufficiently regular IC $[T, F_\theta](x) =: \psi_\theta(x)$, it may be replaced locally (at the model) by the M-estimator $\{\psi_\theta\}$, and Fisher-consistency implies $\int \psi_\theta f_\theta \equiv 0$ for both estimators. The asymptotic variance at F_θ is $\int \psi_\theta^2 f_\theta$, and it is well-known that this is minimized for all θ among all Fisher-consistent estimators iff T is locally equivalent (including the derivative) to the maximum likelihood estimator at the model, with $\psi_\theta = f'_\theta/f_\theta$ (where the prime, naturally, denotes the derivative with respect to θ). This is, in essence, one of the basic parts of Fisher's classical theory of estimation in strict parametric models.

We now assume that the "ideal" F_θ is distorted to yield an F more or less close to F_θ (in the weak* topology). If T is reasonable, it will then tend to $T(F)$ [instead of $T(F_\theta)$], and in order to learn still as much as possible about θ, we require, in addition to Fisher-consistency and efficiency, ideally:

(i) that T be (weak*-) continuous in all F_θ;
(ii) that its breakdown point be as high as possible;
(iii) that T changes as little as possible if one goes away from any F_θ in any direction.

Usually we have to compromise (if the model, as we shall assume, allows robust estimation at all). If we start to modify T, always retaining Fisher-consistency, we may rather easily fulfill (i) and (ii) (e.g., by applying a very cautious, smooth rejection procedure based on a very robust auxiliary estimate) without changing by much the behavior at the model (including the asymptotic variance). But the main conflict typically arises between efficiency and (iii). Condition (iii) can be achieved by putting a low bound on the sup-norm of the derivative, i.e., on the gross-error-sensitivity; this will normally also imply (i) and, at least to some extent, (ii). Now the lower the bound, the larger the smallest possible

variance of a Fisher-consistent estimator (see Section 7). This optimal modification leads to the class of "admissible robust" estimators with respect to asymptotic variance (efficiency) and gross-error-sensitivity, and the M-estimates in this class constitute the first and most important step in robustifying maximum likelihood estimators.

In addition, we may also demand:

(iv) a low local-shift-sensitivity and/or
(v) a low rejection point.

Again we may ask for optimal solutions, i.e., for estimates which minimize the asymptotic variance at the model among all Fisher-consistent estimators with given bounds for (iv) and/or (v) [and perhaps (iii)]. Each additional condition may imply some loss in efficiency. However, in practice it is certainly worthwhile to pay a premium of a few percent efficiency each (compare also [2]) in order to get a bound on the gross-error-sensitivity and a finite rejection point and perhaps local-shift-sensitivity, obtaining (i) and (ii) for free.

Mathematicians may note that the terms "invariance," "group," "order," "symmetry," "unbiasedness," "continuous distribution," etc., do not appear in the general theory. In particular, there remains no problem with "asymmetric contamination," and no question as to "what to estimate"; some bias is unavoidable, except in artificial cases, and all we can do is balance bias (gross-error-sensitivity) and variance under the model against each other (sensibly dependent on the sample size). Furthermore, the stress on M-estimates is not accidental; neither L- nor R-estimates (cf. [20] or preceding) allow a proper rejection of outliers, i.e., a rejection based on the distance from the bulk of the data.

7. OPTIMAL HUBERIZING

As an example of the sort of mathematical results that can be expected in this theory, we shall give a lemma that (in all "regular" cases) supplies the structure of the optimal estimators with regard to efficiency and gross-error-sensitivity in the case that both sample space and parameter space are subsets of the real line.

Lemma (Lemma 5 of [10]): Let f be a probability density with respect to (σ-finite) λ (on R), positive on its (closed) support S; let f' be measurable on S, zero elsewhere, with $\int f'd\lambda = 0$ and $0 < \int (f'/f)^2 f d\lambda < \infty$; let $b > 0$ be some constant. Define $\mathring{\psi} := f'/f$ on S (arbitrary measurable elsewhere); define

$$\psi_\alpha^* := \min\{|\mathring{\psi} - \alpha|, b\} \cdot \operatorname{sign}(\mathring{\psi} - \alpha)$$

for any real α. Then there is an α^* such that $\int \psi_{\alpha^*}^* f d\lambda = 0$. Define $\tilde{\psi} := \psi_{\alpha^*}^*$ and $c := \int \tilde{\psi} f' d\lambda$. Then $c > 0$, and $\tilde{\psi}$ minimizes $\int \psi^2 f d\lambda / [\int \psi f' d\lambda]^2$ among all ψ with

$$\int \psi f d\lambda = 0, \quad \int \psi f' d\lambda \neq 0$$

and

$$\sup_x \left| \psi \Big/ \int \psi f' d\lambda \right| \leq k$$

where $k = b/c$. Any other solution that minimizes this integral under the given side-conditions coincides with a nonzero multiple of $\tilde{\psi}$ modulo $f d\lambda$.

The proof (see [10]), left out due to lack of space, is fairly straightforward.

The interpretation is obvious. For each parameter θ, we find a class of ψ-functions which are optimal (or admissible) with respect to gross-error-sensitivity and efficiency (or asymptotic variance) under the side condition of Fisher-consistency. If we now select pointwise a $\tilde{\psi}_\theta$ for each θ in a smooth way [e.g., by choosing a very smooth function $b(\theta)$], then the class $\{\tilde{\psi}_\theta\}$ defines an optimal M-estimator which cannot be dominated by any other (sufficiently "regular") estimator. Moreover, any sufficiently regular estimator can be replaced by an M-estimator with the same values and the same influence curves at the parametric model (notwithstanding different behavior off the parametric model) so that there is indeed no restriction involved in considering only M-estimators locally at the parametric model.

It should be noted that the lemma is valid also for such cases as the Poisson distribution, where we have no invariance, no symmetry, and no density with respect to Lebesgue measure. If we do have invariance (as for scale estimates), then a solution for a single θ (with its "correction for bias" given by α^*) generates immediately an M-estimator $\{\tilde{\psi}_\theta\}$ for all θ in a natural way; if, moreover, we do have symmetry (as for location of a symmetric distribution), then $\alpha^* \equiv 0$, and the solution simplifies further.

We repeat that the solutions for the location parameter of the normal distribution include the various Huber-estimators (see [1]) and the trimmed means. The corresponding scale solutions include Huber's scale estimators (see [15]) and trimmed variances (cf. also [25, p. 173]. See [10] for more examples).

8. SOME NUMERICAL VALUES

For those who like to look at numbers, we present a short table with some estimators of location at the standard normal distribution. There is a deeper reason for doing this. In a sense, the numerical values provide the flesh for the skeleton of a purely logical theory. It is not enough to know, e.g., that one can decrease the gross-error-sensitivity by paying with the asymptotic variance; it is the fact that one can decrease the gross-error-sensitivity tremendously (down from ∞ to 2 or 1.7, say, in the Normal model) with only a few percent loss in efficiency, which gives this robustification its full importance for statistical practice. Similarly, we can gain a low rejection point with only a marginal increase in asymptotic variance. Viewed differently, Fisher's classical theory of estimation provides only the strictly

optimal solutions for strict parametric models, which may be disastrously nonrobust; the fact of practical importance is that extremely close to these optimal solutions there are nearly optimal solutions with good to excellent robustness properties.

It may also be mentioned that the bewildering amount of Monte Carlo variances in the Princeton robustness study [1] can be quite well summarized by only 2 or, slightly better, by only 4 parameters of the estimates: one being the gross-error-sensitivity, one being something like the rejection point, and for finer description, the adaptivity (essentially change of the gross-error-sensitivity with the underlying distribution); finally, for some badly-designed estimators, some measure for the basic inefficiency of the estimator ([1], in particular Ch. 7C6).

In studying the table, we realize that all robustness concepts are meaningful and necessary. Qualitative robustness and positive breakdown point are closely connected (they were indeed treated as identical in [10], but not in [11] by a slight change of definitions); however, there are models in which, e.g., the median would not be locally robust, but would still retain its global breakdown point $\frac{1}{2}$. On the other hand, the fine distinction between qualitative robustness and finite gross-error-sensitivity is important, as the normal scores estimator shows; it is robust (as opposed to the mean which has the same influence curve at the Normal) and has a positive breakdown point; but as the infinite gross-error-sensitivity indicates, it deteriorates fast as one goes away from the Normal (faster than, e.g., the Hodges-Lehmann-estimator) and breaks down relatively early. The $\Phi(L$-type)-estimator and the similar optimal L-estimator for the logistic distribution are opposite examples of nonrobust estimators with finite gross-error-sensitivity; some trimming would barely change them locally at the Normal, yet would make them robust with positive breakdown point. The former may be compared with the Hodges-Lehmann-estimator (an R-estimator) and the $\Phi(M$-type)-estimator (an M-estimator) which have the same influence curves at the Normal. The relations between trimmed and Winsorized means as well as between trimmed means and Huber-estimators have been discussed previously. We note that trimmed means deteriorate slightly faster (and break down much earlier) than Huber's Proposal 2, and these estimators in turn deteriorate slightly faster than Huber-estimators with most robust scale. Finally, three-part descending M-estimators pay a small premium in asymptotic variance or gross-error-sensitivity, as compared with Huber-estimators, in order to be able to reject outliers completely. They share this property with estimates combined with rejection rules such as Huber-type

Some Numerical Robustness Properties at the Standard Normal Distribution[a]

Estimator	*qr*	σ^2	γ^*	δ^*	λ^*	ρ^*	*supGEM*	*difGEM*
mean	−	1.000	∞	0.00	1.00	∞	∞	∞
n.sc.	+	1.000	∞	0.24	1.00	∞	1.48	1.48
H/L	+	1.047	1.77	0.29	1.41	∞	1.29	1.29
Φ (M-type)	+	1.047	1.77	0.50	1.41	∞	1.28	1.28
Φ (L-type)	−	1.047	1.77	0.00	1.41	∞	∞	∞
median	+	1.571	1.25	0.50	∞	∞	1.74	1.74
5% Wins	+	1.014	2.13	0.05	∞	∞	1.46	1.46
5% trim	+	1.026	1.83	0.05	1.11	∞	1.30	1.30
10% trim	+	1.060	1.60	0.10	1.25	∞	1.26	1.26
6.68% trim	+	1.037	1.73	0.07	1.15	∞	1.271	1.271
H (1.5)	+	1.037	1.73	0.26	1.15	∞	1.264	1.264
A (1.5)	+	1.037	1.73	0.50	1.15	∞	1.262	1.262
H_1 (1.5)	+	1.037	1.73	0.50	1.15	∞	1.258	1.258
25A	+	1.026	1.86	0.50	1.10	6.41	1.35	1.07
A (1.686)	+	1.024	1.86	0.50	1.10	∞	1.28	1.28
H "sk" 2.71	+	1.066	2.89	0.50	∞	2.71	∞	1.10
A (2.71)	+	1.001	2.73	0.50	1.01	∞	1.52	1.52

[a] See Sec. 5 and also [10, 11] for the robustness concepts, and [15, 10, 1] for details on the estimators.

NOTE: The columns provide: qr = qualitative robustness (and weak*-continuity, which usually coincides with it) (+ = yes, − = no); σ^2 = asymptotic variance (at the standard Normal Φ); γ^* = gross-error-sensitivity; δ^* = breakdown point; λ^* = local-shift-sensitivity; ρ^* = rejection point; supGEM = supremum of the asymptotic variance in Huber's [15] symmetric 5% gross-error-model which allows $F(x) = .95\Phi(x) + .05H(x)$ with any symmetric distribution H; difGEM = asymptotic variance in the symmetric 5% diffuse gross-error-model, as the limiting case with 5% widespread contamination, e.g., of $F(x) = .95\Phi(x) + .05\Phi(x/k)$ for $k \to \infty$.

The estimators are: mean = arithmetic mean; n.sc. = normal scores estimator (R-estimator derived from the 2-sample normal scores test); H/L = Hodges-Lehmann-estimator (= median of all pairwise means of observations = R-estimator derived from the Wilcoxon test); Φ (M-type) = M-estimator with $\psi(x) = \Phi(x) - \frac{1}{2}$ (Φ = standard normal cumulative) and scale = 0.6745^{-1}·median deviation; Φ (L-type) = asymptotically equivalent L-estimator (linear function of the order statistics) which has the IC proportional to $\Phi(x) - \frac{1}{2}$ and which behaves very similarly to the asymptotically optimal L-estimator for the logistic distribution; median; 5% Wins = 5% Winsorized mean; α% trim = α% trimmed mean; H(k) = Huber's Proposal 2 with parameter k; A(k) = "Huber robust scale" = Huber-estimator with rescaled median deviation as scale estimate and parameter k; $H_1(k)$ = Huber-estimator for known (fixed) variance 1 and with parameter k; 25A = M(2.5, 4.5, 9.5)•AD = three-part descending M-estimator with a = 2.5, b = 4.5, c = 9.5 as multiples of the median deviation resp. with a' = 1.686, b' = 3.035, c' = 6.408 as multiples of the standard deviation; H "sk" k = Huber-type "skipped" mean (a mean with rejection rule) defined as the solution closest to the median of $\int \psi((x - \theta)/s)F(dx) = 0$ where s = 0.6745^{-1}·median deviation and $\psi(x) = x$ for $|x| \le k$, = 0 otherwise.

The entries are partly based on tables in [15] and [10] and on unpublished work by the author; some entries, such as qr and δ^* for H "sk" 2.71, required some special work. The last decimal given may be slightly inaccurate in some cases.

"skipped" means, which, however, are usually too sensitive to the local behavior of the underlying distribution, especially at the rejection points (compare [15]). While most estimators considered here are worst for very distant contamination, 25*A* and *H*"*sk*"*k* suffer most from very specific types of close contamination, but both are superior when the gross errors are thinly and widely spread. (The premium paid by 25*A* in the worst possible case is still unnecessarily high and could be decreased efficiently by the use of a ψ-function going back down to zero on a certain hyperbolic tangent, or first on a certain hyperbolic tangent and then on a straight line tangent to it, rather than on a simple straight line.)

[*Received May 1972. Revised October 1973.*]

REFERENCES

[1] Andrews, David F., *et al.*, *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press, 1972.

[2] Anscombe, Frank J., "Rejection of Outliers," *Technometrics*, 2 (May 1960), 123–47.

[3] ——— and Tukey, John W., "The Examination and Analysis of Residuals," *Technometrics*, 5 (May 1963), 141–60.

[4] Daniel, Cuthbert and Wood, Fred S., *Fitting Equations to Data. Computer Analysis of Multifactor Data for Scientists and Engineers*, New York: Wiley-Interscience, 1971.

[5] Feller, William, *An Introduction to Probability Theory and Its Applications, Vol. 2*, New York: John Wiley and Sons, Inc., 1966.

[6] Filippova, A.A., "Mises' Theorem on the Asymptotic Behavior of Functionals of Empirical Distribution Functions and Its Statistical Applications," *Theory of Probability and Its Applications*, 7, No. 1 (1962), 24–57.

[7] Fisher, Ronald A., "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society*, Ser. A, 222 (1922), 309–68.

[8] ———, *Statistical Methods and Scientific Inference*, London: Oliver and Boyd, Ltd., 1959.

[9] Gauss, Carl Friedrich, "Bestimmung der Genauigkeit der Beobachtungen," *Zeitschrift für Astronomie und verwandte Wissenschaften*, (März und April 1816). Reproduced in *Werke, Vol. 4*, Göttingen: Dieterichsche Universitäts-Druckerei, 1880, 109–17.

[10] Hampel, Frank R., "Contributions to the Theory of Robust Estimation," Unpublished Ph.D. thesis, University of California, Berkeley, September 1968.

[11] ———, "A General Qualitative Definition of Robustness," *Annals of Mathematical Statistics*, 42 (December 1971), 1887–96.

[12] Hodges, Joseph L., Jr., "Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, Berkeley and Los Angeles: University of California Press, 1967, 163–86.

[13] ——— and Lehmann, Erich L., "Estimates of Location Based on Rank Tests," *Annals of Mathematical Statistics*, 34 (June 1963), 598–611.

[14] Hoeffding, Wassily, "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19 (September 1948), 293–325.

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

[16] ———, "A Robust Version of the Probability Ratio Test," *Annals of Mathematical Statistics*, 36 (December 1965), 1753–8.

[17] ———, "Robust Confidence Limits," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 10, No. 3 (1968), 269–78.

[18] ———, "Robust Estimation," *Mathematical Centre Tracts*, 27, Amsterdam: Mathematisch Centrum Amsterdam, 1968, 3–25.

[19] ———, *Théorie de l'inférence statistique robuste*, Montreal: Presses de l'Université de Montréal, 1969.

[20] ———, "Robust Statistics: A Review," *Annals of Mathematical Statistics*, 43 (August 1972), 1041–67.

[21] ——— and Strassen, Volker, "Minimax Tests and the Neyman-Pearson Lemma for Capacities," *Annals of Statistics*, 1, No. 2 (1973), 251–63.

[22] Huber-Carol, Catherine, "Étude asymptotique de tests robustes," Published Ph.D. thesis, Swiss Federal Institute of Technology, Zürich, 1970.

[23] Jaeckel, Louis A., "Robust Estimates of Location," Unpublished Ph.D. thesis, University of California, Berkeley, 1969.

[24] Jeffreys, Harold, *Theory of Probability*, 3rd ed., Oxford: Clarendon Press, 1961.

[25] Johnson, Norman L. and Leone, Fred C., *Statistics and Experimental Design: In Engineering and the Physical Sciences, Vol. 1*, New York: John Wiley and Sons, Inc., 1964.

[26] Kallianpur, G., "Von Mises Functionals and Maximum Likelihood Estimation," in C.R. Rao, *et al.*, eds., *Contributions to Statistics*, Calcutta: Statistical Publishing Society Calcutta, 1963, 137–46.

[27] ——— and Rao, C.R., "On Fisher's Lower Bound to Asymptotic Variance of a Consistent Estimate," *Sankhyā*, 15, No. 4 (1955), 331–42.

[28] Knüsel, L.F., "Ueber Minimum-Distance-Schätzungen," Published Ph.D. thesis, Swiss Federal Institute of Technology, Zürich, 1969.

[29] Lehmann, Erich L., *Testing Statistical Hypotheses*, New York: John Wiley and Sons, Inc., 1959.

[30] Mallows, Colin, "Hoeffding, Tukey, Hájek, and von Mises," Unpublished talk at the Princeton Robustness Seminar, May 10, 1971.

[31] Miller, R.G., "A Trustworthy Jackknife," *Annals of Mathematical Statistics*, 35 (December 1964), 1594–1605.

[32] Sarhan, A. and Greenberg, B., eds., *Contributions to Order Statistics*, New York: John Wiley and Sons, Inc., 1962.

[33] Stigler, S.M., "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920," *Journal of the American Statistical Association*, 68 (December 1973), 872–79.

[34] Takeuchi, Kei, "Robust Estimation and Robust Parameter," Unpublished manuscript, New York, 1968.

[35] Tukey, John W., "Which Part of the Sample Contains the Information?" *Proceedings of the National Academy of Sciences*, 53 (1956), 127–34.

[36] ———, "A Survey of Sampling from Contaminated Distributions," in I. Olkin, *et al.*, eds., *Contributions to Probability and Statistics*, Palo Alto, Calif.: Stanford University Press, 1950, 448–85.

[37] ———, "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33 (March 1962), 1–67.

[38] ———, *Exploratory Data Analysis, Vol. 1*, Preliminary ed., Reading, Mass.: Addison-Wesley Publishing Co., 1970.

[39] von Mises, Richard, "On the Asymptotic Distributions of Differentiable Statistical Functions," *Annals of Mathematical Statistics*, 18 (September 1947), 309–48.

[40] ———, H. Geiringer, ed., *Mathematical Theory of Probability and Statistics*, New York: Academic Press, 1964 (published posthumously).