Z. Wahrscheinlichkeitstheorie verw. Geb. 27, 87–104 (1973)

Robust Estimation: A Condensed Partial Survey

Frank R. Hampel

Prefatory Note

This paper was written to stimulate discussion; therefore the pointed style. It was designed to be self-contained and yet to minimize overlap with Huber's (1972) long and basic survey paper, which in particular covers the technical points in more detail. The issues raised are considered basic to reasonable applications of statistics; on the other hand, they suggest and stimulate much novel research in mathematical statistics and probability theory (such as about weak*-continuous functionals serving as robustified maximum likelihood estimators, or about Choquet-capacities describing and replacing sets of probability measures). It is hoped that the paper may help in clarifying some relations between rigorous stochastic models and the world outside of mathematics, and perhaps also in improving understanding and cooperation between pure mathematicians and data analysts.

1. Why Robust Estimation?

What do those "robust estimators" intend? Should we give up our familiar and simple models, such as our beautiful analysis of variance, our powerful regression, or our high-reaching covariance matrices in multivariate statistics? The answer is no; but it may well be advantageous to modify them slightly. In fact, good practical statisticians have done such modifications all along in an informal way; we now only start to have a theory about them. Some likely advantages of such a formalization are a better intuitive insight into these modifications, improved applied methods (even routine methods, for some aspects), and the chance of having pure mathematicians contribute something to the problem. Possible disadvantages may arise along the usual transformations of a theory when it is understood less and less by more and more people. Dogmatists who insisted on the use of "optimal" or "admissible" procedures as long as mathematical theories contained no other criteria, may now be going to insist on "optimal robust" or "admissible robust" estimation or testing. Those who habitually try to lie with statistics, rather than seek for truth, may claim even more degrees of freedom for their wicked doings. In passing, those who use statistics for sanctification rather than elucidation of uncertain facts (treating it as a replacement rather than aid for thinking) might wonder about the "monolithic, authoritarian structure" (cf. Tukey, 1962) they believe statistics to be. (Furthermore, there are of course tremendous possibilities for publishing under the fashionable flag of robustness, both of very valuable and of less valuable results, keeping needy statisticians from perishing.)

Now what are the reasons for using robust procedures? There are mainly two observations which combined give an answer. Often in statistics one is using a parametric model implying a very limited set of probability distributions thought possible, such as the common model of normally distributed errors, or that of exponentially distributed observations. Classical (parametric) statistics derives results under the assumption that these models were strictly true. However, apart from some simple discrete models perhaps, such models are never exactly true. We may try to distinguish three main reasons for the deviations: (i) rounding and grouping and other "local inaccuracies"; (ii) the occurrence of "gross errors" such as blunders in measuring, wrong decimal points, errors in copying, inadvertent measurement of a member of a different population, or just "something went wrong"; (iii) the model may have been conceived only as an approximation anyway, e.g. by virtue of the central limit theorem. To point (ii) we may add the lack of fit of a simple piece of structure in a structured design, such as different behaviour of a single row in a two-way table. Even point (i), though usually least harmful, invalidates parts of statistical theory if taken seriously, e.g. the "super-efficiency" of estimators for a uniform distribution (Kempthorne, 1966), and it may cause slight inconveniences with estimators like the median. "Gross errors" may occur as clear "outliers" or as "hidden contamination" which usually cannot even be detected; their frequency depends of course greatly on the quality of the data, but some figures may be revealing. Paul Olmstead, cited by John Tukey (1962), maintains that engineering data typically involves about 10% "wild shots". Cuthbert Daniel (1968), for various types of industrial data, considers frequencies from less than 1% up to 10% and 20% as usual and cites, as rather exceptional, a set of about 3000 data points where he couldn't find anything wrong. As result of a spot check on medical data in a clinic, A.J. Porth reported 8–12% gross errors (in the meeting on medical statistics in Oberwolfach of February 1972). For routine data in exact science, Freedman (1966) could individually identify gross errors in a special case (readings of seismograms by different stations) and came up with 5% certain and an additional 2% suspected gross errors. Altogether, 5–10% wrong values in a data set seem to be the rule rather than the exception. But, as to (iii), even large high-quality samples in astronomy and geodesy, consisting of several thousand data points each, which should be the prime examples for the "normal law or error", are mildly but definitely leptocurtic (longer-tailed); there are a number of examples known, including the large ones in Romanowski and Green (1965) and the old ones by Bessel (1818). The same result (leptocurtic error distributions) was obtained by Student (1927) by studying many small samples; compare also Box and Andersen (1955) p. 2, Tukey (1960) p. 458, Jeffreys (1961) Ch. 5.7, and Huber (1972). "Normality is a myth; there never was, and never will be, a normal distribution" (Geary, 1947, p. 241, cited by Tukey, 1962, p. 20).

The other observation which in combination with the first one substantiates the need for robust procedures, is the fact that even slight and harmless-looking deviations from a strict parametric model may render classically "optimal" procedures rather inefficient and bad. The first basic reference here is Tukey's (1960) survey paper on earlier research at Princeton which had gone into a number of technical reports. A lot of empirical evidence is also contained in Andrews *et al.* (1972), where the arithmetic mean is only included marginally for reference purposes to show how bad it is. To simulate the effect of mild outliers, Tukey (1960) considered mixtures of the standard normal distribution Φ with small fractions ε of contaminating normals with threefold scale (variance 9): $F(x)=(1-\varepsilon)\,\Phi(x)+\varepsilon\,\Phi(x/3)$, with $0\leqq\varepsilon\leqq 0.1$, say. With these realistic amounts, but too mild dispersions of the contamination, the asymptotic efficiency of the mean goes already down to about 70% for $\varepsilon=0.1$, the mean being worse than the median there, while other simple estimators, such as a 6% trimmed or truncated mean (cf. next section), are more than about 95% efficient over the whole range. The case for the variance is even worse. Tukey (1960) recalls the dispute between Fisher (1920) and Eddington (*Stellar Movements*, p. 147, and footnote in Fisher, 1920, p. 762) about the use of mean deviation and standard deviation as measures of dispersion, where Fisher stressed the fact that for strictly normal data the mean deviation is only 88% efficient; however, in Tukey's model it suffices to take $\varepsilon=0.0018$(!) (corrected value by Huber) to make the mean deviation asymptotically better than the standard deviation. (And the mean deviation is not even qualitatively robust either; there are still better estimators.) This example shows clearly how dangerous and misleading mathematical theorems on optimality (or other such topics) may be, if they are applied to the real world outside of mathematics.

If we try to approximate the error distributions of "ordinary" data without clear outliers more closely than by the normal distribution, we might consider the results by Jeffreys (1961) Ch. 5.7. Jeffreys concluded from his analysis of nine long series, that the errors of careful observations made under uniform circumstances might well be described by t-distributions with about 5 to 9 degrees of freedom (with the additional qualifications that nonuniform conditions, such as different observers, cause still longer-tailed distributions, and that the occurrence of less long-tailed – more closely normal – distributions often seems to be coupled with marked correlations between the errors). But if we take, for the moment, t_5 to t_9 as our error models and use the formulas given by Fisher (1922), we find for the asymptotic efficiency of the arithmetic mean values between 80% and 93%, and for the asymptotic efficiency of the standard deviation values between 40% and 83%. Such values (and even lower ones) are thus quite likely to hold true while the efficiency of both estimates is believed to be 100% by the unsuspecting statistician.

We should also remember that we never know the exact distribution of ordinary data; and even if we did, or as far as we do, there remain serious questions about how to handle the excess knowledge of details. After all, a statistical model has to be simple (where "simple", of course, has a relative meaning, depending on state and standards of the subject matter field); Ockham's razor is an essential tool for the progress of science.

For mathematical statisticians, there are purely mathematical and aesthetical reasons to consider robust estimators. In functional analysis and its applications (e.g. in physics), functionals are often required to be continuous or even differentiable, in the same vein in which ordinary functions on the real line are usually required to be at least continuous (or piecewise continuous, in some applications). For decades, however, mathematical statisticians have been bravely struggling with functionals on the space of probability distributions, such as the arithmetic mean, which are *nowhere continuous* (in any statistically reasonable topology, such as weak*-topology) and which are only defined on a dense subset (with dense complement) of the whole space. Apparently they did not even clearly recognize this situation which has caused them troubles in their proofs on more than one occasion. The theory of robust estimation (cf. Hampel, 1968) approximates these functionals by other functionals which are continuous and even differentiable, as far as desired.

If we look briefly to some related fields, we can recognize parallel developments there. Robustness may be viewed as a set of stability requirements, analogous to stability of ordinary differential equations, for example. Only quite recently, numerical analysts have become dissatisfied with such possibilities as obtaining a negative variance with the familiar formula $\{\sum X_i^2 - (\sum X_i)^2/n\}/(n-1)$ and a correctly operating computer; so they started to investigate numerical stability. And in parts of prediction and control theory, engineers resorted to "sub-optimal solutions" after finding the "optimal" solutions offered to them inappropriate in practice (oral remark by Th. Gasser; cf. also Åström, 1970, p. 184).

The question arises how statisticians could get along until now with such non-robust estimators as the arithmetic mean. There are several partial answers (cf. Hampel, 1968, p. 60 f.): small avoidable losses by the "inadmissible use of admissible estimators" will have gone entirely unnoticed, and (in hopefully not too few cases) large losses may have been prevented by the use of the median rather than the mean in view of plainly non-normal data, or by a transformation, or by the use of a formal rejection procedure, or, perhaps most frequently, by the habit of "throwing away" any strays before taking the mean. The latter procedure (better replaced by: "setting them aside and treating them separately") amounts to using an informal, vague rejection rule; and any halfways reasonable rejection procedure, even though usually not good, will at least prevent the worst (cf. Andrews *et al.*, 1972, p. 243 f.). However, there is an increasing danger in that more and more data is automatically processed on the computer without being looked at by a competent statistician.

Indeed, some users of least squares programs have recognized this danger. Feeling the urgent practical need to suppress unwanted outliers in their case, and being left in the lurch by statistics as they knew it, they robustified their program on their own, using intuition, trial and error, and came up with a remarkably good solution (see Merrill and Schweppe, 1971).

Sometimes it is objected that results of mathematical statistics, like the Gauss-Markov theorem, yield the arithmetic mean as optimal even if nothing is known about normality. In such cases, the main fallacy lies in assuming that we always want to estimate the expected value of the observations, however wrong they be. (We rather often want to estimate roughly some value in the bulk of the data, or perhaps the expectation of a normal distribution close to the actual distribution in a metric leading to the weak*-topology.) And even if we, for the Gauss-Markov theorem in least squares theory, impose symmetry and existence of moments on the distribution of errors (e.g. as in a lightly truncated Cauchy distribution), then linearity is far too severe a restriction, and unbiasedness is of doubtful value anyway.

Another objection may be raised based on the wellknown "robustness" of the t-test (and of the analysis of variance, for that matter). As tests will not be discussed in this paper, it may suffice to remark that the t-test is only moderately robust (as long as the tails of the error distribution are not too long) with regard to its *level*, but not with regard to its *power* and the corresponding length of confidence intervals; furthermore, that robustness studies based on Edgeworth expansions, closeness of higher moments and other similarly restricted alternatives, though of limited value if interpreted correctly (especially for "good" data), are missing the main effects of actual contamination.

Some may find it surprising that nonparametric techniques basically have nothing to do with robust techniques in parametric models. Yet the former, while having a proper realm of application of their own, often also have good robustness properties and may then be used as robust techniques. Unfortunately, they don't seem to be very appropriate for complex designs, so that specifically robust methods still have to be developed anew. In some sense and in analogy with computers, robust statistics may be regarded as "third-generation statistics" after parametric statistics and nonparametric statistics (Tukey, 1970/71, Vol. I, Ch. 6G).

It should be stressed that there are situations where robust methods are not applicable (though much rarer than naive belief would have it): e.g. if the population members to be measured show large variability in the measured characteristic (apart from measurement error), and if, *specifically*, the expected value of this characteristic is to be estimated (in a *nonparametric* sense). An example may be a distribution of incomes, which is practically unbounded on one side. Apart from trying to establish a suitable parametric model (or at least a "smoothed" model, or truncating the distribution if permitted), there seems to be no remedy but to pay special attention and great care to the tails of the distribution and otherwise hope and pray. (Cf. also Bahadur and Savage, 1956.)

As the main aim of robust estimation, we can consider building in safeguards against unsuspectedly large amounts of gross errors, putting a bound on the influence of hidden contamination and questionable outliers, isolating clear outliers for separate treatment (if desired), and still being nearly optimal at the strict parametric model. Robust estimation considers *both* small, barely detectable, and large, conspicuous contamination of the model – as long as one still wants to retain it. The interest is concentrated on the behaviour of the *bulk* of the data, and on routine methods for determining it; but these methods may also be helpful in exposing more clearly the deviating behaviour of parts of the data.

After treating the classical "random errors" of "equally good" observations (mainly due to small measurement errors or to inherent variability of the material) and, to some extent, controlling "systematic errors" (e.g. by randomization or trend-free designs), statistics finally starts to deal explicitly with "gross errors". At the same time, the "rigorous models" of classical statistical theory are being superseded by "approximate models". The conditions are favourable: the mathematical tools have further improved; and owing to the development of computers and Monte Carlo methods, statistics is now in part an experimental science like experimental physics (as evidenced by the book of Andrews *et al.*, 1972), much more so than in the days of Student (1908). But we are only at the beginning. We know a lot about point estimation of a single location parameter (especially, as most important, near the normal distribution), and we still discover more and more open questions there; we badly need more techniques for finding confidence intervals, to make more estimators practically usable; and we only dimly see the contours of a general theory of robustness in linear and locally linear designs, including robustness of design. Much more has to be done.

2. Some Examples of Robust Estimators

Before briefly mentioning some robust estimators which are either well-known or otherwise of current interest, we may remember that large parts of classical statistics are not robust (to be more refined: in varying degree not robust). Non-robust (sensitive to small changes of the model) are, e.g.: the arithmetic mean and the method of least squares; standard deviation, mean deviation and range; covariance and correlation. Robust are: the same methods *with* sensible "looking at the data", setting aside outliers etc. Robust (under some restrictions which will not be mentioned) are also estimators combined with reasonable formal rejection procedures (cf. Anscombe, 1960; Ferguson, 1961; Grubbs, 1969; for the latter), and the estimators to follow. For more details cf. the literature (e.g. Andrews *et al.*, 1972; Huber, 1964; Huber, 1972; Hampel, 1972).

The median $\check{x}$ is well-known and in some sense the "most robust" estimator of location, tolerating up to one-half of the sample of totally wrong values and being least affected by single gross errors. The α-trimmed mean $(0<\alpha<\frac{1}{2})$ deletes (roughly) αn observations on each side of the ordered sample before taking the mean. It can be defined as the functional $\int_{\alpha}^{1-\alpha} F^{-1}(t)\,dt/(1-2\alpha)$, where F may be the empirical cumulative distribution function. It has a long history (cf. Anonymous, 1821; Stigler, 1973; and Eisenhart, 1971, about Mendeleev) and (together with its one-sided variant) is also naively and sensibly used by laymen very distant from professional statistics. It is sharply distinguished from rejection procedures, as it does already what "Winsorizing" was supposed to do. The α-Winsorized mean $\int_{\alpha}^{1-\alpha} F^{-1}(t)\,dt+\alpha\{F^{-1}(\alpha+0)+F^{-1}(1-\alpha-0)\}$ $(0<\alpha<\frac{1}{2})$ "brings in" αn observations on each side to the first unaffected order statistic, but with some surprising side effects (cf. Hampel, 1968). A one-sided Winsorized mean (though not under this name) is well-known and commonly used in life-testing (cf. e.g. Feller, 1966, p. 41, or Johnson and Leone, 1964, p. 160f.); here the pressure of early termination has brought about the involuntary use of a slightly inefficient, but pleasantly robust statistic. The estimator derived from the nonparametric Wilcoxon-(Mann-Whitney-)test, called Hodges-Lehmann-estimator, can be defined as the median of the means of all pairs of observations (with or without diagonal). Of special importance are Huber's M-estimators which should look rather familiar as slight generalization of maximum likelihood estimators. In general, they are given by a family of functions $\{\psi_\theta(x)\}$, with θ in the parameter space and x in the sample space, and define the estimate θ by the implicit relation $\int \psi_\theta(x)\,F(dx)=0$ (together with side conditions if necessary). For estimates of location, $\psi_\theta(x)=\psi(x-\theta)$, and specifically $\psi(x)=x$ for $|x|\leqq k$, $\psi(x)=k\,\mathrm{sign}(x)$ otherwise $(0<k<\infty$; e.g. $k=1.5)$ defines the prototype of the Huber-estimator $H_1(k)$ which puts a bound on the influence of any observation. It may be combined with any robust scale estimate in order to be also scale-invariant. The author proposed to let the defining ψ-function smoothly and gently go to zero in the tails, as in $\psi_{abc}(x)$ which is symmetric and linear from 0 to a, constant till b and linearly descending to zero till c from where on it stays zero (e.g. "25A" with $a=2.5$, $b=4.5$, and $c=9.5$ multiples of the "median deviation", see below). This combines the advantages of "Huberizing" and of rejecting outliers. (Of course, the computer program should also report the outliers separately.) There are many other estimates of location (hundreds of them have been tested recently), e.g. Tukey's skipping procedures (which reject observations according to their relative distance from the quartiles) and the "adaptive" estimators (which try to estimate the shape of the underlying distribution in one way or another).

The interquartile range is a simple robust scale estimate, while the precise counterpart of the median is the slightly different "median deviation" med $(|x_i - \acute{x}|)$, with $\acute{x} = \mathrm{med}(x_i)$ (cf. Hampel, 1968 (p. 83), 1972). It is the limiting case of Huber's scale estimators which have parabolas bounded from above and later also from below as defining ψ-functions. Other robust scale estimators are trimmed and Winsorized variances.

Correlation may be estimated by the Spearman or Kendall rank correlation, or by the quadrant correlation (cf. e.g. Quenouille, 1959, or Sachs, 1972, p. 312f.) which in time series analysis is known as the result of "hard limiting" (cf. Thomas, 1969, p. 298ff.; Huber, 1972). Some other estimators are those obtained by "smooth limiting".

3. The Cushny and Peebles Data Revisited

The data by Cushny and Peebles (1905) on the prolongation of sleep by two soporific drugs were used by Student (1908) as the first illustration of his famous test, were then cited by Fisher ((1925) 1970) and thenceforth copied in numerous books as example of (univariate) normally distributed data; the original bivariate sample is even one of two samples given and treated as multivariate normal in Anderson (1958). A glance at a simple scatter diagram reveals that there is one point way off the pattern set by the rest, and two more are slightly suspect. But rarely are the differences analyzed in another way than with mean and t-test; e.g. Pfanzagl (1968, p. 131f.) uses only the sign test, and Tukey (1970/71, Vol. I, Ch. 6D, E) applies his skipping procedures, remarking cooly that we do expect fairly frequent stray values in data of this kind.

For a brief first look, we neglect to go beyond the data, to the way they were generated; we ignore the medical and biochemical knowledge and other pertinent information needed for a thorough analysis; within the data, we disregard the bivariate structure, including choice of model and potential special meaning of zeros; we pass over the test against zero difference, where the results of t-test and sign test are very similar (the two-sided levels of significance reached are 0.3% resp. 0.4%); we only try to estimate the difference between the two drugs by various methods, including confidence intervals on the customary 95%-level where possible.

The ordered sample (of differences) is 0.0; 0.8; 1.0; 1.2; 1.3; 1.3; 1.4; 1.8; 2.4; 4.6. Some central values are:

mean:		1.58
10%	(10%-trimmed mean):	1.40
20%	(20%-trimmed mean):	1.33
H/L	(Hodges-Lehmann-estimator):	1.32
50%	(median):	1.30
P15	(a variant of Huber's H(1.5); see Andrews *et al.*, 1972, also for other estimators):	1.38

trimean:	1.35
(rejecting estimators:) 25A:	1.29
mean with 4.6 rejected:	1.24
$c=t=s$-skipped trimean:	1.25
$c=t=s$-skipped mean:	1.26

(the latter ones having 4.6, 2.4, 0.0 rejected). We note that all estimators save the mean, even though of very different types, range from 1.24 to 1.40, leaving a clear gap up to 1.58. Even estimators which are highly correlated with the mean in the case of strictly normal samples, like 10%, P15, 25A and H/L (all being about 95% efficient), are now clearly separated from it. Those estimators which never reject an outlier, no matter how bad it is, range from 1.30 to 1.40, the median being most resistant and farthest away; the others, including the very efficient 25A, range from 1.24 to 1.29. If we wanted to select any single point estimate, we could take 1.29, which just happens to nearly coincide with the median.

Confidence intervals can presently be constructed for mean and median (and other estimates derived from rank tests which, however, become better usable only for slightly larger sample sizes), and approximate ones (with a robustified t-test) for 10% and 20% (cf. Tukey and McLaughlin, 1963, and Huber, 1970b). They are (on the 95%-level): mean: [0.70, 2.46]; 10%: [0.85, 1.95]; 20%: [0.87, 1.79]; median: [0.8, 2.4]. We discover that the interval for 20% is only about half as long as that for the mean (derived from the t-test). Moreover, the interval for the median, derived from the sign test, often shunned because of its inefficiency, is still on both ends slightly narrower than that of the mean. In contemplating the length of the confidence intervals, we infer that (for most purposes) the point estimates should be rounded to full tenths (or at most half tenths), yielding 1.3 (possibly 1.2 or 1.4); that the differences among the robust estimators are negligible; that the mean 1.6, however, is clearly somewhat worse, being by about one estimated standard error or more away from the rest.

4. Historical Notes; The "Princeton Robustness Year" 1970/71

After Gauss (1821) had introduced the normal distribution to suit the arithmetic mean (of "equally accurate" observations, to be sure) and had (since about 1800) developed his statistical theories mainly under the criterion of mathematical simplicity and elegance, little progress was made towards robustifying the classical models for about a century and a half. There were, of course, analyses with only partial or no regard to the theory of least squares; some first rejection procedures for outliers were introduced (Peirce, 1852; Chauvenet, 1863; for current references cf. Anscombe, 1960; Ferguson, 1961; Grubbs, 1969); and it was soon discovered that even samples of purely "good" observations tend to be indeed *approximately* normal, but also clearly longer-tailed. But apparently only a few noted statisticians, such as Newcomb (1886), Student (1927) and Jeffreys (1932, 1939) took the latter fact serious, and tried to do something about it. (For the beginnings of robust testing, cf. e.g. Pearson, 1931; Box, 1953; Box and Andersen, 1955.)

Perhaps some words should be said about the place of Fisher, in particular since his emphasis on strict model optimality, though very valuable for theory, has entailed rather untenable and illogical attitudes towards practice (by mathematicians) and within practice. Fisher (1922, p. 355 and before) had pointed out that the method of moments (in particular mean and standard deviation) is more than 80% efficient only in a small region of Pearson curves; and he was the only statistician queried by Tukey (1960) to anticipate large effects in the latter's contamination model. He mentioned the possible use of the sample kurtosis for deciding whether to use the standard deviation or the mean deviation (Fisher, 1920, p. 770) and envisaged the future possibility of something like adaptive estimation (Fisher, 1936, p. 250). He sometimes carefully checked parametric models in real applications (cf. e.g. Fisher, Thornton and Mackenzie, 1922); he provided tests for departures from normality (Fisher, 1930) and applied them (with somewhat fragmentary comments) to real data (Fisher, (1925) 1970, p. 52f.). Two crucial references seem to be Fisher (1922) p. 314 and p. 322f. After stressing its existence, Fisher (loc. cit. p. 314) *explicitly ignored* the "question of specification" (of a parametric model) in his discussion, leaving it to the "practical statistician". Later (loc. cit. p. 322f.) he made short remarks about rejection of outliers, nonnormal distributions in reality, the importance of collecting large samples for determining (better: estimating) the true error curves, and assured observers of no criticism if (in view of nonnormal data) they used other estimators than the mean. He did not suggest any alternative model to be used, but (loc. cit. p. 314) he did expect progress in the area of specification, quite likely in connection with a change of viewpoint, and quite possibly faster than any progress has come (cf. Tukey, 1960, p. 472f.).

It seems surprising that (Neo-)Bayesians have not contributed more to robustness or even the problem of outliers (for some different types of work done, see e.g. de Finetti, 1961; Box and Tiao, 1962; Gebhardt, 1966). They have made some valid theoretical points, and undoubtedly much work will have escaped this author; but some parts of their work are apt to leave the unfortunate impression of formal derivation of complicated integrals under unwarranted assumptions (which are then often left unsolved), or of the substitution of simple models by equally strict but more complicated ones. The surprise stems from the fact that there are indeed many questions in applying statistics which reasonable practical statisticians would think most suitable for formalization by a Bayesian or similar theory, but somehow they don't seem to find the attention they deserve by Bayesians. A simple example may be given by the statement: "I expect about 10% gross errors, which could be anywhere, in this sort of data."

The short history of robust estimation starts with the investigations of Tukey and his co-workers about the non-robustness of some classical estimates and the robustness of some simple substitutes for them (cf. Tukey, 1960). The first theoretical breakthrough came when Huber (1964) considered a rather full class of neighbouring alternatives which dwells on the crucial junction between being already sufficiently general and being still elegantly manageable. In a second basic approach, Huber (1965, 1968a) derived robust confidence intervals from robust tests, for which he is working out a general theory (cf. Huber and Strassen, 1973). There are other papers containing memorable ideas, e.g. by Anscombe (1960), Tukey (1962), Anscombe and Tukey (1963), Hodges (1967), Tukey (1970/71) (this selection is of course subjective and incomplete). The author (Hampel, 1968) studied the various stability aspects of robustness: the qualitative local behaviour, the "breakdown" aspect and the theory of infinitesimal properties (cf. also Hampel, 1971, 1972). So far, most successful studies consider only point estimation of a single parameter, as a necessary but transient first step; comparatively little is known yet about interval estimation (cf. Tukey and McLaughlin, 1963, and Huber, 1970b) and about estimation in more general designs (where Huber, 1968b, 1970a, appears to have pointed the way for linear designs); but work is going on. A recent survey (with additional references) is given by Huber (1972).

On the more empirical side, there was a number of computer studies of the finite sample behaviour of various robust estimators for different underlying error distributions, using either numerical integration or Monte Carlo methods. A particularly large study was conducted at Princeton in 1971 (Andrews *et al.*, 1972). It was an outgrowth of a year-long research seminar about robustness by Bickel, Huber, Tukey and the author which at first delineated various methods of attack for robust regression, but then concentrated on finding a common and communicable basis with regard to the simplest but fundamental case of robust estimation, that of estimating a single location parameter. Proposals were made and collected from the literature and from a number of statisticians, and finally about 70 estimators yielded their Monte Carlo variances and percentage points under about 20 different sampling distributions (mostly for sample size $n=20$, but in some situations also for $n=5$, 10, and 40). Several interesting side studies, such as of finite-sample "influence" or "sensitivity curves" and of "breakdown" bounds, helped in rounding the picture. The detailed results are too numerous to be reviewed here (cf. Andrews *et al.*, 1972); they are in good agreement with theory (as far as existent) and supply the necessary quantitative details. Some results are that estimators which in effect "bring in" outlying observations closer to the bulk of the sample, such as Huber-estimators, trimmed means and the Hodges-Lehmann-estimator, are doing quite well, but still lose unnecessarily much efficiency compared with descending M-estimators which eventually "throw out" (reject) very distant values completely if such distant outliers may be present. This type of M-estimator lets the influence of outliers go to zero smoothly with the distance; it allows for a region of doubt, so to speak, and therefore avoids the unpleasant jumpy features of classical "hard" rejection procedures, which treat observations as "completely good" up to a certain distance, and as "completely bad" (to be thrown out) just beyond.

A comparison of the practice of applying a (hard) rejection procedure and taking the mean of the remaining observations with other robust estimators under equal conditions seemed desirable. Now Tukey continued the Monte Carlo studies in detail with more and more sets of 75 estimators each under a few situations, and one of these sets, filled with suggestions by Bickel, Huber and the author, eventually contained also several classical rules for the rejection of outliers. The first results were not unexpected: if the "contamination" by "gross errors" was rather distant and clearly separated from the rest of the sample, these "hard" rejection rules did about as well as the best other robust estimators, namely smoothly rejecting (descending) M-estimators; but when the underlying distribution carried just somewhat more mass in the "flanks" than the normal distribution, making the – artificial – distinction between "good" and "bad" observations much more difficult, then hard rejection rules were considerably inferior to many other robust estimators. They also showed great differences among themselves; as a very tentative and preliminary result, rejection rules based on the 4th moment, on the maximum studentized residual and perhaps also on the Shapiro-Wilk-test appear to be relatively good, while some others may fail rather badly. Still, the fact remains that reasonable formal rejection procedures, though usually not very good, can often at least prevent the worst; they and, even more so, judicious subjective rejection procedures may already be far superior to formal application of classically optimal nonrobust methods.

5. Some Concepts in Robustness Theories

This short chapter can only give some hints about the meaning of some concepts; details can be found in the literature. We also repeat the warning that theory, any theory, can at most provide *guidance*, but not the answer in real life; hence also robustness theories should not be taken literally in applications.

The *gross-error-model* (Huber, 1964) is described by $F(x)=(1-\varepsilon)\,\Phi(x)+\varepsilon H(x)$ $(0<\varepsilon<1)$ where Φ is (e.g.) the standard normal cumulative and H is either symmetric (for mathematical convenience) or completely arbitrary. (By shift, etc., this produces a sort of neighbourhood of the parametric model.) The interpretation is that an observation is "good" with probability $1-\varepsilon$, but may be anything with the small probability ε. For fixed ε, there is (usually) a least favourable distribution among this set, which is hardest to estimate; and its maximum likelihood estimator has an even smaller asymptotic variance for all other distributions allowed (saddelpoint property). In the case of the Normal, the least favourable distribution is normal in the middle and exponential in the tails, and its maximum likelihood estimator is the Huber-estimator (with k depending on ε).

Confidence intervals can be derived from *censored likelihood ratio tests* (Huber 1965, 1968a) which put bounds from above and below on the likelihood ratio. Exact finite sample solutions for the Normal case are again the Huber-estimators. For the asymptotic evaluation of the bounds, cf. Huber-Carol (1970). For the relation to *capacities*, which had been introduced into statistics by Strassen (1964), cf. Huber (1968a, 1969), Huber and Strassen (1973).

Qualitative robustness (small change of the behaviour with small change of the model) can essentially be described by continuity of the estimator with respect to the Prohorov metric. The *Prohorov distance* takes care of a small fraction of arbitrary gross errors, of small rounding and grouping effects, and of the type of deviations occurring in the central limit theorem. There are close relations to the "total-variation-model" and the gross-error-model (cf. Hampel, 1968, p. 41 ff.).

The *breakdown point* is a weak and simple, but important global measure of robustness. It tells us the fraction of gross errors needed until the estimator becomes completely unreliable and totally disastrous. (E.g. for the α-trimmed mean, it is equal to α; more than αn free outliers on one side can enforce any value of the estimate whatsoever.) While usually a breakdown point $\frac{1}{2}$ is desirable and also possible, estimators with breakdown points near zero are obviously too risky, except for very good data.

The richest information can be gained by considering infinitesimal changes and extrapolating them back to finite samples. The central tool is the *influence* *curve* IC of an estimator (functional) T at a distribution F, defined on the sample space by $IC_{T,F}(x)=\lim_{\varepsilon\downarrow 0}[T\{(1-\varepsilon)F+\varepsilon\,\delta_x\}-T(F)]/\varepsilon$, where δ_x is the point mass 1 in x (Hampel, 1968). Roughly speaking, it measures the change of the estimate caused by an additional observation in x. There are close ties to derivatives of functionals (such as in von Mises, 1947), to the jackknife and to frequently occurring expressions in mathematical statistics; in particular, the IC of a maximum likelihood estimator resp. of an M-estimator is nothing but a certain multiple of the log likelihood derivative resp. the defining ψ-function. Thus it is possible to see at a glance local robustness properties of maximum likelihood (and other) estimators, and to construct new estimators with predetermined properties (as successfully done in Andrews *et al.*, 1972). A maximum likelihood estimator may be robustified in the following ways (in decreasing order of importance): by putting a bound on $|IC|$ *("gross-error-sensitivity")*, by putting a bound on the absolute value of the slope *("local-shift-sensitivity")*, by setting the $IC\equiv 0$ outside a certain point *("rejection point")*, by going down to zero (in the location case) on a certain hyperbolic tangent (determining the "*change-of-variance-sensitivity*"), and by putting (in the same case) a (stricter) bound from below on the slope (determining the "*change-of-bias-sensitivity*"). This means roughly: limited effect of any fixed amount of unknown contamination; limited effect of "wiggling" (rounding, grouping etc.); no effect of clear outliers; and a certain stability of asymptotic variance and bias. Under such side conditions, the estimator should still estimate the correct quantity *if* the model were true – meaning $T(F_\theta)\equiv\theta$ *("Fisher-consistency")* – and be as efficient as possible under the strict model. In this framework, there remains no question of "what to estimate", no "question of bias" (cf. Hampel, 1968, p. 84f.); we estimate a quantity which coincides with the parameter (e.g. the expectation) if the model is true and which changes very little if the model is not quite true (and which breaks down as late as possible). By robustifying an estimator, we may exchange an arbitrarily small loss in efficiency against a tremendous gain in quantitative robustness properties. The optimality problem of maximizing the efficiency under a given bound on the gross-error-sensitivity ("optimal Huberizing") is basically solved and leads again to the Huber-estimators in the Normal case. – Among the first uses of the influence curve was the surprising clarification of the effects of rejecting, trimming, Winsorizing and Huberizing. The IC provided also a simple four-parameter summary of the mass of data in Andrews *et al.* (1972; cf. p. 248ff.). More uses (e.g. in multivariate statistics) may be anticipated.

6. Open Problems and Recent Developments

As there are many more things which we don't know (but can already vaguely point to) than things we do know, we can only briefly touch upon some likely or important areas of future research; keeping in mind that weights and even the questions themselves may drastically change within a short time.

There are many open questions even with a single parameter. An outstanding problem is Studentizing (confidence intervals; finite-sample distributions), in particular of M-estimators. (Formulas for asymptotic variances are known, and their values can be estimated, but the problem lies with small samples.) A small point to be discussed is the selection of a sequence of different functionals for different sample sizes. A topic which has received much recent attention are "adaptive" estimators which try to estimate the true underlying distribution in order to become "asymptotically everywhere efficient". They seem to be more nonparametric than robust in spirit, though some "mild" form of adaptivity has already proved useful in Andrews *et al.* (1972). There are still open regularity conditions and missing theorems about Fréchet-derivatives, optimal robustifications and similar points. A special topic is sufficiency; in one sense or other the robust estimators are "nearly as" sufficient for the parametric model as the classical ones, but a satisfactory quantitative interpretation (which one, and why?), together with theorems about minimal loss of sufficiency under robustness restrictions, are still waiting to be written down. Robustification of Bayes estimators (or even the full aposteriori distribution) might consist of two parts: the "classical" one with regard to the model and a new one with respect to the apriori distribution. One approach for the latter problem (before the question was posed) is already given by the theory of "restricted Bayes solutions" (Hodges and Lehmann, 1952; cf. also Lehmann, 1959, p. 14), but its suitability remains to be checked. Those for whom the separation of the problem is too simple or not "Bayesian" enough may put a robust apriori distribution on a full Prohorov neighbourhood of the whole parametric model and solve for the estimator.

The problem of estimating a single variance is largely solved, although for practical use some constants and tables have to be worked out or expanded (cf. also Johnson and Leone, 1964, p. 172f.). A number of possibilities are known for estimating covariance and correlation matrices, as used in multivariate statistics and time series analysis, but their properties have to be studied in more detail and compared; the final choice will vary with the purpose. A general question is: should outliers only be "brought in" ("Huberized") or even "thrown out" (rejected)? (And do we want "smooth" or "hard" rejecting?) Another feature of interest is whether there exists a random transformation of the data such that the (vectorvalued) estimate is a simple bijective function of the classical estimate of the *transformed* sample (as in Huber's "proposal 2"); and if so, whether this transformation can be achieved in each coordinate separately. Obviously, a scalar factor as bijective function ensures positive semidefiniteness of the covariance matrix. (Perhaps a random "weight" function for the size of the observations—usually $1/n$—could also be used.) Along these lines, the author (Zürich robustness seminar, summer 1972) suggested the following affine invariant procedure: given a ψ-function as used for location estimators, find a matrix A such that with $y=A^{-1}x$ and $z=\psi(\|y\|)\cdot y/\|y\|$ the covariance matrix of the z_i is a fixed multiple of the identity matrix; then AA^T is the estimated covariance matrix of the x_i. The solution (existence? uniqueness?) hopefully can be found iteratively (convergence?), by transforming the z_i to spherical symmetry (as outlined by Gnanadesikan and Kettenring, 1972) and starting, e.g., with median deviations and raw quadrant correlations. The ψ-function may be bounded monotone ("bringing in" outliers) or returning smoothly to zero (rejecting outliers); the quantitative form may depend on the dimension. A similar method, which is a hard rejection procedure, has already been tried out in some cases by Gnanadesikan and Kettenring (1972). Huber (Zürich robustness seminar, summer 1972) suggested and studied another approach which is not affine invariant but on the other hand still simpler; namely to transform each coordinate separately and to use the correlations of the transformed variables. It contains smooth and hard limiting as special cases. There are surprisingly close ties to the theory for a location parameter.

The first step in robustifying the linear model may be taken by Huberizing the residuals (cf. Huber, 1968b, 1970a, 1972, 1973a); other approaches (cf. e.g. Bickel, 1971) may be asymptotically equivalent. Historically and practically, the first step was hard rejecting (cf. Anscombe, 1960, Daniel, 1960, Kruskal *et al.*, 1960). The second step, combining the initial alternatives, would be smooth rejecting (Hampel, Princeton robustness seminar; cf. also Anscombe, 1967). But then there comes a novel feature. The total influence (of an observation on the fitted model) is determined not only by the size of the residual, but also by the "influence of position (in the factor space)" (in a multiplicative way), and the question arises whether and when (and how, in practice) to Huberize or even to reject certain positions, as a possible third and fourth step. An extreme example would be simple linear regression with one x_i far away from the rest (e.g. astronomical observations, a single one from antiquity). This x_i may be known to be physically impossible (so that one would reject it), or it may be the most important point in factor space (so that one *has* to keep and tentatively trust it), or it may be more or less doubtful (so that one can restrict the influence of position to an arbitrary degree, shifting the emphasis of the fit towards the densely occupied part of the factor space in the frequent case where the model is only approximative). It should be noted that merely restricting the influence of residuals, without regard to position, would *not* suffice. Eventually such considerations may also lead to a quantitative theory about "robustness of design" (Hampel, Princeton robustness seminar, 1970; cf. also Huber, 1973b).

There are still many open problems in time series analysis. Of minor importance may be χ^2-related estimation in contigency tables and similar designs, which is already qualitatively robust; but sometimes it may still be desirable to cut out or down the influence of a few single cells. Another, quite different type of topic is possibly that of bias in complex designs, in particular bias caused by overfitting (of many parameters) and bias caused by ill-fitting approximate models (leading to new aspects with regard to "optimal" designs, see Huber, 1973b). But all these questions are still more or less within conventional framework. Some very important questions of data analysis are still on a deeper level. One is basically pattern recognition (in the presence of random errors); in particular, fitting of a suitable model to the data (rather than the other way around) and identification of simple substructures of a structure which behave like outliers (Daniel, 1968). (Here belong also breakdown aspects with regard to substructures.) Another one are the "semi-systematic fluctuations" of certain time series (cf. Student, 1927; Jeffreys, 1961, Ch. 5.6; Mandelbrot and Wallis, 1968; cf. also Daniel and Wood, 1971, p. 59); there may be similar slow trends in space, and both may be related to the "semi-systematic inhomogeneity" of the target population or "reference set" (cf. also de Finetti, 1964, p. 115f.) and the question of the range of validity (and attainable accuracy) of a statistical model. A third, somewhat related, broad area which seems to underly many uses of data analysis (perhaps even more so in the future) is that of smoothing (including "resistant" smoothing, as discussed by

Tukey, Princeton robustness seminar, 1971). Our theories are still a long way away from the "high art of data analysis", as exemplified by the book of Daniel and Wood (1971). We do well to remember this; we do well to remember that statistical theories, at best, provide a partial aid by clarifying specific aspects of the situation at hand (and that, in our case, even robust routine methods do not obviate the need for careful thinking if more than a routine answer is wanted); we do well to remember how little we always know.

7. Some Summarizing Theses

1. (a) Robust methods, in one form or another (and be it a glance at the data), are necessary; those who still don't use them are either careless or ignorant. (b) On the other hand, some robust methods, even if very vague, subjective and inefficient, were always applied by good practical statisticians; this use, long frowned upon by dogmatists, is now also justified by theory.

2. (a) The theorems of mathematical statistics, sharpened by logic and correspondingly narrow-pointed, can and should provide guidance and illumination of special aspects; if taken superficially for unwarranted deductions in real life, they can be totally misleading. (b) In practice, it is much more important to prevent the worst in every conceivable aspect, including those not covered by any current theory, than to optimize in one or a few directions, completely ignoring everything else.

3. (a) Those who don't trust their statistical intuition, or who worry even about small losses of efficiency, say, from 5% to 50%(!), may use good robust estimators as far as currently known; they may remember that these estimators help to control and expose contamination, not to interpret it. (b) Those who try to devise systematically new robust methods may be warned of doing so without the utmost aid by robustness theory presently attainable; there were and are too many fallacies and traps for naive intuition.

4. (a) The statistical theory of robust estimation gives us in many cases already a good intuitive insight into what we can do, and how we can do it; in particular, we can understand old methods more deeply and can develop greatly improved new ones. (b) The practical approach is to act as if the parametric model would hold, but with methods which are hardly affected by small perturbations of the model and which are still safe under large contamination; the results may be used for routine purposes if this makes sense, or they may be analyzed further. (c) The mathematical theory of robust estimation, slowly coming into existence, can provide many new problems to work on and a surprising new outlook at many old results of mathematical statistics; and perhaps some topics will become more elegant, with the aid of some increasingly used tools of functional analysis and other parts of pure mathematics.

5. (a) Statistical practice, much more than practice of other fields, may contain a large share of incompetent, nonsensical and greatly inadequate work which goes unnoticed; and even if the consequences of a blunder become apparent, it may still be ascribed to some error of first or second kind. (b) Statistics, much more than other fields, is also liable to misuse up to outright fraud which it may even be very hard to expose; attempts to confine the misuse by prescribing a limited

canonical set of procedures for all circumstances have the side-effect of letting statistics altogether appear unreasonable and foolish. (c) While misuse and incompetent use seemingly cannot be eliminated, it may still be of value to discuss more frequently and openly examples and higher standards of good data analysis (as opposed to those of pure mathematics). Statistics and common sense should be compatible, after all. (d) The history of robust estimation may provide another incentive to discuss these questions.

References

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. New York: Wiley 1958

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

Anonymous: Dissertation sur la recherche du milieu le plus probable. Ann. Math. Pures et Appl., **12**, 181–204 (1821)

Anscombe, F.J.: Rejection of outliers. Technometrics, **2**, 123–147 (1960)

Anscombe, F.J.: Topics in the investigation of linear relations fitted by the method of least squares. J. Roy. Statist. Soc. Ser. B, **29**, 1–52 (1967)

Anscombe, F.J., Tukey, J.W.: The examination and analysis of residuals. Technometrics, **5**, 141–160 (1963)

Åström, K.J.: Introduction to Stochastic Control Theory. New York and London: Academic Press 1970

Bahadur, R.R., Savage, L.J.: The nonexistence of certain statistical procedures in nonparametric problems. Ann. Math. Statist., **27**, 1115–1122 (1956)

Bessel, F.W.: Fundamenta Astronomiae. Königsberg: Nicolovius (1818)

Bickel, P.J.: On some analogues to linear combinations of order statistics in the linear model. Manuscript. (1971)

Box, G.E.P.: Non-normality and tests on variances. Biometrika, **40**, 318–335 (1953)

Box, G.E.P., Andersen, S.L.: Permutation theory in the derivation of robust criteria and the study of departures from assumption. J. Roy. Statist. Soc. Ser. B, **17**, 1–34 (1955)

Box, G.E.P., Tiao, G.C.: A further look at robustness via Bayes' theorem. Biometrika, **49**, 419–432 (1962)

Chauvenet, W.: Manual of Spherical and Practical Astronomy. Philadelphia: Lippincott (1863)

Cushny, A.R., Peebles, A.R.: The action of optical isomers. II. Hyoscines. Journal of Physiology, **32**, 501–510 (1905)

Daniel, C.: Locating outliers in factorial experiments. Technometrics, **2**, 149–156 (1960)

Daniel, C.: Statistical aids in the planning of experimental work. Course at the University of California, Berkeley (1968)

Daniel, C., Wood, F.S.: Fitting Equations to Data. Computer Analysis of Multifactor Data for Scientists and Engineers. New York: Wiley-Interscience (1971)

De Finetti, B.: The Bayesian approach to the rejection of outliers. Proc. 4th Berkeley Sympos. Math. Statist. Probab., **1**, 199–210. Berkeley and Los Angeles: University of California Press (1961)

De Finetti, B.: Foresight: its logical laws, its subjective sources. (French original 1937.) Translation in: H.E. Kyburg and H.E. Smokler (ed.): Studies in Subjective Probability. New York: Wiley 1964

Eisenhart, C.: Mendeleev's expressed preference (1895) of the $33\frac{1}{3}\%$-trimmed mean. Memorandum, Nat. Bureau of Standards, Washington D.C. 1971

Feller, W.: An Introduction to Probability Theory and Its Applications. Vol. II. New York: Wiley 1966

Ferguson, T.S.: On the rejection of outliers. Proc. 4th Berkeley Sympos. Math. Statist. Probab., **1**, 253–287. Berkeley and Los Angeles: University of California Press (1961)

Fisher, R.A.: A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error. Monthly Not. Roy. Astron. Soc., **80**, 758–770 (1920)

Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. London Ser. A, **222**, 309–368 (1922)

Fisher, R.A.: Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd ((1925) 1970)

Fisher, R.A.: The moments of the distribution for normal samples of measures of departure from normality. Proc. Roy. Soc. London, Ser. A, **130**, 16–28 (1930)
Fisher, R.A.: Uncertain inference. Proc. Amer. Acad. Arts and Sciences, **71**, 245–258 (1936)
Fisher, R.A., Thornton, H.G., Mackenzie, W.A.: The accuracy of the plating method of estimating the density of bacterial populations. Ann. Appl. Biology, **9**, 325–359 (1922)
Freedman, H.W.: The "little variable factor". A statistical discussion of the reading of seismograms. Bull. Seismol. Soc. of America, **56**, 593–604 (1966)
Gauss, C.F.: Göttingische gelehrte Anzeigen (1821). Reprinted in: Werke, **4**, 98. Göttingen: Dieterich 1880
Geary, R.C.: Testing for normality. Biometrika, **34**, 209–242 (1947)
Gebhardt, F.: On the effect of stragglers on the risk of some mean estimators in small samples. Ann. Math. Statist., **37**, 441–450 (1966)
Gnanadesikan, R., Kettenring, J.R.: Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, **28**, 81–124 (1972)
Grubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics, **11**, 1–21 (1969)
Hampel, F.R.: Contributions to the theory of robust estimation. Unpublished dissertation, Berkeley: University of California (1968)
Hampel, F.R.: A general qualitative definition of robustness. Ann. Math. Statist., **42**, 1887–1896 (1971)
Hampel, F.R.: The influence curve and its role in robust estimation. Manuscript (1972)
Hodges, J.L.: Efficiency in normal samples and tolerance of extreme values for some estimates of location. Proc. Fifth Berkeley Sympos. Math. Statist. Probab., 1965/1966, **1**, 163–186, Berkeley and Los Angeles: Univ. Cal. Press (1967)
Hodges, J.L., Lehmann, E.L.: The use of previous experience in reaching statistical decisions. Ann. Math. Statist., **23**, 396–407 (1952)
Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Statist., **35**, 73–101 (1964)
Huber, P.J.: A robust version of the probability ratio test. Ann. Math. Statist., **36**, 1753–1758 (1965)
Huber, P.J.: Robust confidence limits. Z. Wahrscheinlichkeitstheorie und verw. Gebiete, **10**, 269–278 (1968a)
Huber, P.J.: Robust estimation. Mathematical Centre Tracts, **27**, 3–25 (Amsterdam) (1968b)
Huber, P.J.: Théorie de l'inférence statistique robuste. Montréal: Presses de l'Université de Montréal (1969)
Huber, P.J.: Robust regression. Unpublished manuscript (1970a)
Huber, P.J.: Studentizing robust estimates. In: Nonparametric Techniques in Statistical Inference (ed. M.L. Puri), 453–463. Cambridge: Cambridge Univ. Press (1970b)
Huber, P.J.: Robust statistics: a review. Ann. Math. Statist., **43**, 1041–1067 (1972)
Huber, P.J.: Robust regression. To appear in: Ann. Statist., **1**, No. 5, 1973a
Huber, P.J.: Robustness and designs. To appear in: Proceedings of the Internat. Sympos. on Statistical Design and Linear Models. Fort Collins: Colorado State Univ. 1973b
Huber, P.J., Strassen, V.: The Neyman-Pearson lemma for capacities. Ann. Statist., **1**, 251–263 (1973)
Huber-Carol, C.: Etude asymptotique de tests robustes. Dissertation, ETH Zürich (1970)
Jeffreys, H.: An alternative to the rejection of outliers. Proc. Roy. Soc. London Ser. A, **137**, 78–87 (1932)
Jeffreys, H.: Theory of Probability. Oxford: Clarendon Press, 1939, 1948, 1961
Johnson, N.L., Leone, F.C.: Statistics and Experimental Design: In Engineering and the Physical Sciences. Vol. I. New York: Wiley 1964
Kempthorne, O.: Some aspects of experimental inference. J. Amer. Statist. Assoc., **61**, 11–34 (1966)
Kruskal, W., Ferguson, T.S., Tukey, J.W., Gumbel, E.J.: Discussion of the papers of Messrs. Anscombe and Daniel. Technometrics, **2**, 157–166 (1960)
Lehmann, E.L.: Testing Statistical Hypotheses. New York: Wiley 1959
Mandelbrot, B., Wallis, J.R.: Noah, Joseph, and operational hydrology. Water Resources Research, **4**, 909–918 (1968)
Merrill, H.M., Schweppe, F.C.: Bad data suppression in power system static state estimation. IEEE Trans., **PAS-90**, 2718–2725 (1971)
Newcomb, S.: A generalized theory of the combination of observations so as to obtain the best result. Amer. J. Math., **8**, 343–366 (1886)
Pearson, E.S.: The analysis of variance in cases of non-normal variation. Biometrika, **23**, 114–133 (1931)

Peirce, B.: Criterion for the rejection of doubtful observations. Astronom. J., **2**, 161–163 (1852)
Pfanzagl, J.: Allgemeine Methodenlehre der Statistik. Vol. II. Berlin: de Gruyter 1968 (Sammlung Göschen)
Quenouille, M.H.: Rapid Statistical Calculations. London: Griffin (1959)
Romanowski, M., Green, E.: Practical applications of the modified normal distribution. Bull. Géodésique, **76**, 1–20 (1965)
Sachs, L.: Statistische Auswertungsmethoden. (3rd. Ed.) Berlin-Heidelberg-New York: Springer 1972
Stigler, S.M.: Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. Submitted to J. Amer. Statist. Assoc. (1973)
Strassen, V.: Messfehler und Information. Z. Wahrscheinlichkeitstheorie und verw. Gebiete, **2**, 273–305 (1964)
"Student": The probable error of a mean. Biometrika, **6**, 1–25 (1908)
"Student": Errors of routine analysis. Biometrika, **19**, 151–164 (1927)
Thomas, J.B.: An Introduction to Statistical Communication Theory. New York: Wiley 1969
Tukey, J.W.: A survey of sampling from contaminated distributions. In: Contrib. to Prob. and Statist. (ed. Olkin), 448–485. Stanford: Stanford University Press (1960)
Tukey, J.W.: The future of data analysis. Ann. Math. Statist., **33**, 1–67 (1962)
Tukey, J.W.: Exploratory Data Analysis. 3 Vol. Limited Preliminary Edition. Reading-Menlo Park-London-Don Mills: Addison-Wesley (1970/71)
Tukey, J.W., McLaughlin, D.H.: Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization I. Sankhyā Ser. A, **25**, 331–352 (1963)
von Mises, R.: On the asymptotic distributions of differentiable statistical functions. Ann. Math. Statist., **18**, 309–348 (1947)

Frank R. Hampel
Seminar für Angewandte Mathematik
der Universität
Freiestrasse 36
CH-8032 Zürich
Switzerland

(Received April 3, 1973)