The Annals of Mathematical Statistics
1972, Vol. 43, No. 4, 1041-1067

THE 1972 WALD LECTURE

ROBUST STATISTICS: A REVIEW[1,2]

BY PETER J. HUBER

Swiss Federal Institute of Technology, Zürich

This is a selective review on robust statistics, centering on estimates of location, but extending into other estimation and testing problems. After some historical remarks, several possible concepts of robustness are critically reviewed. Three important classes of estimates are singled out and some basic heuristic tools for assessing properties of robust estimates (or test statistics) are discussed: influence curve, jackknifing. Then we give some asymptotic and finite sample minimax results for estimation and testing. The material is complemented by miscellaneous remarks on: computational aspects; other estimates; scale, regression, time series and other estimation problems; some tentative practical recommendations.

TABLE OF CONTENTS

1. Introduction. This review neither claims to be exhaustive nor attempts to be impartial. I have tried to give a reasonably coherent, but not too technical account of one line of development—the one to which I have contributed myself. The presentation centers around the simplest, best known and most important special case: that of estimating one single location parameter. We have now attained a sufficiently strong foothold there, both with regard to rigorous theory and to intuitive insight, that we may confidently transfer ideas and methods to other, more complicated situations, both in estimation and testing. This present review considerably revises and updates an earlier survey (Huber (1968b)).

I gratefully acknowledge a number of comments and suggestions from several colleagues who have read the manuscript, in particular from C. L. Mallows and J. W. Tukey.

Received October 13, 1971; revised November 22, 1971.

[1] Work performed while holding a NSF Senior Foreign Scientist Fellowship at Princeton University; partially supported by ONR Grant N00014-67-0151-0017.

[2] Invited Wald Lecture; presented in part at The IMS Annual Meeting at Hanover, New Hampshire, August 28-September 1, 1972.

2. Historical remarks, or: the dogma of normality. The dogma that measurement errors should be distributed according to the normal law is still widespread among users of the method of least squares; I hope that the following historical remarks will help to clarify some of the issues. I am indebted to Churchill Eisenhart for drawing my attention to two crucial nineteenth century references.

The theory of estimation originated with problems where almost all of the statistical variability is due to measurement errors. This situation should be clearly distinguished from the opposite case where the data shows a large internal variability and where good reasons can be advanced for the use of the sample mean, or of the sample median, as estimates of the corresponding population parameters. But in the first case, statistical variability is just a nuisance to get rid of, and one is mainly interested in finding that combination of the observations which lies on the average nearest to the true value.

It is illuminating to witness how the normal, or Gaussian, distribution was introduced by Gauss himself. I quote Gauss (1821):

> The author of the present treatise, who in the year 1797 first investigated this problem according to the principles of the theory of probability, soon realized that it was impossible to determine the most probable value of the unknown quantity, unless the function representing the probability of the errors is known. But since it is not, there is no other recourse than to assume such a function in a hypothetical fashion. It seemed most natural to him to take the opposite approach and to look for that function which must be taken as a base in order that for the simplest of all cases a rule is obtained which is generally accepted as a good one, namely that the arithmetic mean of several observations of equal accuracy for one and the same quantity should be considered the most accurate value. This implied that the probability of an error x must be assumed proportional to an exponential expression of the form e^{-hhxx}, and that then just the same method which he had found by other considerations already a few years earlier, would become necessary in general. This method, which afterwards, in particular since 1801, he had almost daily opportunity to use in diverse astronomical computations, and which in the meantime also Legendre had happened upon, now is in general use under the name method of least squares.

Note that Gauss here introduces the normal distribution to suit the sample mean. It is amusing to observe how the use of the arithmetic mean became almost sacred over the years—I believe mostly because one misunderstood the

Gauss-Markov theorem ("the best *linear unbiased* estimate of the expected value is the sample mean") and the Central Limit theorem ("the sum of *many small* independent elementary errors is *approximately* normal"), in conjunction with the theorem that for independent identically distributed normal observations the sample mean is indeed best in almost every conceivable sense. I have italicized the crucial words in the above paraphrases of the theorems; for instance, there is no reason, except mathematical convenience, to impose linearity or unbiasedness, and one might argue from sad experience that the model should also allow for a *few gross* elementary errors occurring with low probability.

Moreover, one can hardly claim that the sample mean was universally accepted, as Gauss did. There is a charming contemporary paper (Anonymous[3] 1821), which first states that good reasons can be advanced for the use of the sample mean in the case of inherent statistical variability of the data, as opposed to mere measurement errors, but which then continues (page 189):

> Though, even in this case [the case of inherent statistical variability], the popular method [the sample mean], has neither generally been followed nor has it been used without some restrictions. For example, there are certain provinces of France where, to determine the mean yield of a property of land, there is a custom to observe this yield during twenty consecutive years, to remove the strongest and the weakest yield and then to take one eighteenth of the sum of the others.

The author then continues to remark that a considerable arbitrariness is involved here: why should not one exclude the two greatest and the two smallest observations? But nevertheless he does not believe that all observations should enter with the same weight into the determination of the mean.

Bessel (Bessel and Baeyer (1838) page 67) states that he never rejected an observation for internal reasons, i.e., because it deviated too much from the majority of the observations, and that he gave them all the same weight. He states: "We believe that only through a firm adherence to this rule we have been able to remove arbitrariness from our results."

It seems to me that this kind of discussion borders on dogmatism; a more rational action would have been to look at actual error distributions in large samples, to check whether they were compatible with a normal distribution and, if not, to develop a different theory of estimation.

Actually, Bessel himself ((1818) page 19 ff.) had made such a comparison. He notes that all three of his test samples show a slightly higher frequency of large errors than the normal distribution would predict, but he discards this discrepancy as very small and fails to recognize its potential significance (the

[3] According to Czuber (1891), page 227, the author is Svanberg.

sample mean is a poor estimate of location for longer-tailed distributions). Compare also Bessel (1838).

Around the middle of the century, Peirce (1852) and Chauvenet (1863) developed procedures for detecting and rejecting grossly erroneous observations or "outliers"; incidentally, these procedures seem to be among the earliest examples of nontrivial statistical tests. For a recent survey of rejection procedures, see Grubbs (1969). The statistics of choice for detecting and identifying outliers are the sample skewness and kurtosis (Ferguson (1961)). However, the traditional philosophy behind rejection procedures is highly objectionable (cf. Anscombe (1960)). First, the separation of the observations into normal and grossly erroneous ones is artificial and does not make sense for distributions which just carry somewhat more mass in the flanks. Second, there will be statistical errors of both kinds (false rejections and false retentions), therefore the retained observations do not form a sample from a normal population. This has the unfortunate consequence that the performance of a composite procedure (rejection plus estimation based on the retained observations) cannot be deduced in a simple way from the performance of the two parts.

The first serious attempt to deal directly with somewhat longer-tailed distributions seems to be due to Newcomb (1886). He states flatly: "In practice, large errors are more frequent than this equation [the normal law] would indicate them to be." He suggests that the square exponent of the normal density should be replaced by a less rapidly increasing function. "The management of such an exponent might, however, prove inconvenient, and I shall adopt a law of error founded on the very probable hypothesis that we are dealing with a mixture of observations having various measures of precision." Thus, he adopts an error distribution with density

$$\frac{1}{(2\pi)^{-\frac{1}{2}}}\left\{\frac{p_1}{\sigma_1}e^{-(x^2/2\sigma_1{}^2)}+\cdots+\frac{p_m}{\sigma_m}e^{-(x^2/2\sigma_m{}^2)}\right\} \tag{2.1}$$

and proposes to use the Bayes estimate for a uniform prior distribution (i.e., what is now called the Pitman estimate). Roughly, this amounts to giving lesser weights to more extreme observations.

There was little, if any, progress beyond Newcomb in the following sixty years, even though the situation had been quite clearly recognized by eminent statisticians like Student (1927) and Jeffreys (1932).

The estimates proposed by Newcomb and Jeffreys were excessively laborious; according to the latter "each approximation took about 6 hours' work, using a Marchant calculating machine and the tables of Milne-Thomson and Comrie" ((1932) page 85).

Moreover, hardly anybody realized how bad the classical estimates could be in slightly nonnormal situations. E. S. Pearson (1931) may have been the first to note the high sensitivity to deviations from normality of some standard procedures (tests for equality of variances); incidentally, in connection with the

same test problems, G.E.P. Box later coined the term "robustness" (Box (1953)).

In the late forties, the emergence of distribution free procedures brought some relief for testing problems. The turning point for estimation came at the same time, when Tukey and the Statistical Research Group at Princeton began to propagandize the problem, to emphasize the shortcomings of the classical estimates and—most important of all—to establish properties of several really practicable alternatives to them. In particular, the α-trimmed mean was rediscovered and investigated—the old French custom of removing a fixed fraction α of extreme observations from each end of the sample before taking the mean. Unfortunately, most of the material was disseminated only in technical reports, which are almost inaccessible now. A survey paper was later published by Tukey (1960); compare also Tukey (1962).

Hodges and Lehmann (1963) noticed that estimates of location could be derived from the Wilcoxon and other rank tests; that confidence intervals and asymptotic variances could be computed from the power functions of these tests, and that these estimates never have much lower but sometimes infinitely higher efficiencies than the sample mean.

So far one was mostly concerned either with parametric families (as for example Newcomb's mixture of normals, or the t-family), or with the set of, say, all symmetric continuous distributions, as alternatives to the normal law. Huber (1964) then proposed an intermediate approach, the use of small, but rather full neighborhoods (e.g., the set of all distribution functions differing less than ε from a normal one), and he solved some corresponding asymptotic minimax problems. The minimax approach is also able to yield exact, fixed sample size results (Huber (1965), (1968a)).

Hampel (1968) recognized and sorted out the stability aspect of robustness, in close analogy to the stability of a mechanical structure (say of a bridge): (i) the qualitative aspect: a small perturbation should have small effects; (ii) the breakdown aspect: how big can the perturbation be before everything breaks down; (iii) the infinitesimal aspect: the effects of infinitesimal perturbations.

3. What is a robust procedure? By 1960 one had recognized that

—one never has a very accurate knowledge of the true underlying distribution;

—the performance of some of the classical tests or estimates is very unstable under small changes of the underlying distribution;

—some alternative tests or estimates (like the Wilcoxon instead of the t-test, or the α-trimmed mean instead of the mean) lose very little efficiency for an exactly normal law, but show a much better and more stable performance under deviations from it.

While for years one had been concerned mostly with what was later called "robustness of validity" (that the actual confidence levels should be close to, or at least on the safe side of the nominal levels), one realized now that

"robustness of performance" (stability of power, or of the length of confidence intervals) was at least as important and usually brought for free a satisfactory robustness of validity (but not vice versa).

From the beginning, "robustness" has been a rather vague concept; for example, Box and Anderson (1955) had introduced the notion as follows: Procedures are required which are 'robust' (insensitive to changes in extraneous factors not under test) as well as powerful (sensitive to specific factors under test).

But if one wants to choose in a rational fashion between different robust competitors to a classical procedure, one has to make precise the goals one wants to achieve. Unfortunately, a consensus has not been reached; although the goals rarely are stated in an explicit fashion, one can discern at least five or six conflicting ones, and I do not think that all of them should be called by the same name "robust."

To fix the idea, let us consider the problem of estimating a location parameter θ from a large number of independent observations $X_1, \cdots, X_n$, distributed according to $P(X_i < x) = F((x - \theta)/\sigma)$, where the shape F is not exactly known. For most good estimators $n^{\frac{1}{2}}(T_n - \theta)$ will be asymptotically normal, and one will have to judge estimators in terms of their asymptotic variance $\sigma_F^2(T)$; or their absolute efficiency $1/(I(F)\sigma_F^2(T))$, where $I(F)$ is the Fisher information; or their relative efficiency $\sigma_F^2(T')/\sigma_F^2(T)$.

According to the first goal, a robust estimator should possess

(i) a high absolute efficiency for *all* suitably smooth shapes F.

While this goal can be achieved asymptotically for large sample sizes, the convergence seems to be much too slow for practical purposes. Compare Hájek and Šidák ((1967) page 264 ff.), van Eeden (1970); Takeuchi (1971).

Thus, one modifies the requirements to one of the following:

(ii) a high efficiency *relative* to the sample mean (and some other selected estimates), and this for all F (cf. Bickel (1965));

(iii) a high absolute efficiency over a strategically selected *finite* set $\{F_i\}$ of shapes (e.g., the normal, logistic, double exponential, Cauchy and rectangular shapes), cf. Birnbaum and Laska (1967), Crow and Siddiqui (1967), Siddiqui and Raghunandanan (1967), Hogg (1967), Birnbaum and Miké (1970).

A variant of (iii) is

(iii') a high absolute efficiency over a strategically selected *parametric* family of shapes, cf. Box and Tiao (1962, 1964a, b).

(iv) A small asymptotic variance over some neighborhood of one shape, in particular of the normal one (Huber (1964)).

Neither of these goals guarantees the qualitative stability requirement (the convergence in (i) and (iv) need not be uniform):

(v) the distribution of the estimate should change little under arbitrary small variations of the underlying distribution F, and this uniformly in the sample size n (Hampel (1968), (1971)).

Personally, I think that the local goals (iv) and (v) are the important ones. With Anscombe (1960) I am inclined to view robustness as a kind of insurance problem: I am willing to pay a premium (a loss of efficiency of, say, 5 to 10% at the ideal model) to safeguard against ill effects caused by small deviations from it; although I am happy if the procedure performs well also under large deviations, I do not really care—inferences based upon a grossly wrong statistical model may have little physical significance.

Moreover, we often have quite a good idea of the approximate shape of the true underlying distribution (say from looking at histograms and probability plots of related previous samples) so that it should suffice to consider a neighborhood of only one shape. On the other hand, we need a rather full set to include all conceivable nasty nearby possibilities for the true shape.

While (iii) is very attractive for empirical small sample studies *after* one has proposed some (parametric families of) estimates, it is dangerous as an independent goal for optimization and might lead to rather unstable "nonrobust" estimates.

For *finite* sample sizes, the appropriate criteria are much more difficult to put down. Despite its seductive simplicity, the ordinary variance is *not* an adequate measure of performance of a robust estimate; it is much too sensitive to the irrelevant extreme tail behavior of the estimate. To see this, it suffices to consider an extremely long-tailed symmetric distribution for the observations, say with $P(|X| > x) \sim c/\log |x|$. Then an estimate whose value is always contained in the convex hull of the sample cannot have a finite moment of any order for any sample size. In practice, where all random variables are bounded, this means that the variance, although finite, may be much larger than the middle, almost normal part of the distribution of the estimate would suggest. It is preferable to look at selected quantiles (suggested minimal set: 0.25, 0.1, 0.025, 0.005, 0.001)—they will also indicate how fast the limiting (normal) law is approached—and if one needs a single number, one may take the variance of the best normal fit to the central part of the distribution.

For finite sample sizes one is almost forced to fall back to (iii) (with sophisticated Monte Carlo techniques to obtain accurate quantiles, cf. Andrews *et al.* (1972)). We have already noted under (i) that it may not be possible to achieve a satisfactory overall performance. As it then might be too pessimistic to minimize the maximum loss of efficiency over $\{F_i\}$, I propose to give the model distribution a preferential treatment and limit the loss there.

It may sound provocative if I maintain that the notions "nonparametric" and "distribution-free" have little relation to robustness. The sample mean and the sample median are *the* nonparametric estimates of the true mean and median respectively. But it is rather the exception than the rule that the user knows precisely which functional of the distribution he wants to estimate. Ultimately, he might select the functional with the better efficiency or robustness properties at and near the assumed model.

Many of the so-called "distribution-free" procedures derived from rank order statistics have good robustness properties. However, this seems to be a fortunate accident: distribution-freeness stabilizes the level, but not necessarily the power of a test; the performance of estimates derived from tests is intimately linked to the power of these tests.

One sometimes forgets that robustness also should include insensitivity to grouping effects and the like, and that some estimates commonly accepted as robust, like the sample median, are not robust in this sense. For a recent discussion of some effects of granularity, cf. Noether (1967).

There are other aspects of robustness about which little is known, for instance insensitivity towards deviations from independence (cf. Hφyland (1968), Gastwirth and Rubin (1969)).

Deviations from the assumption that the observations in the sample are identically distributed seem to be relatively harmless, at least for procedures which are symmetric in the observations: since such a sample will behave very much like one with identically distributed observations from the average distribution, we can expect that any procedure which does well for random mixtures (say of the type (2.1)) will also do well for a corresponding deterministic mixture, and vice versa.

4. Three methods for constructing estimates. As before, let $X_1, \cdots, X_n$ be independent random variables with common distribution $P(X_i < x) = F((x - \theta)/\sigma)$. We shall assume that F has a density f and that the scale $\sigma = 1$ is known; we shall not bother about regularity conditions. Since all estimates will be translation invariant, it suffices to consider their behavior for $\theta = 0$.

4.1. *Maximum likelihood type estimators (M-estimators).* Let ρ be a real valued function of a real parameter, with derivative $\psi = \rho'$. Define a statistic $T_n = T_n(X_1, \cdots, X_n)$ either by

$$\sum_{i \leqq n} \rho(X_i - T_n) = \inf_t \sum_{i \leqq n} \rho(X_i - t)$$

or by

$$\sum_{i \leqq n} \psi(X_i - T_n) = 0 .$$

Under quite general conditions, T_n converges to $T(F)$, defined by

$$\int \psi(x - T(F))F(dx) = 0 ,$$

and $n^{\frac{1}{2}}(T_n - T(F))$ is asymptotically normal with asymptotic mean 0 and asymptotic variance

$$\sigma_M{}^2(F) = \int \Omega_F(x)^2 F(dx) ,$$

where

$$\Omega_F(x) = \frac{\psi(x - T(F))}{\int \psi'(x - T(F))F(dx)} .$$

If we choose

$$\psi_0(x) = -f_0'(x)/f_0(x)\,, \tag{4.1}$$

for $\psi(x)$, then T_n is the maximum likelihood estimator of θ for the true underlying distribution F_0 and will under suitable regularity conditions be asymptotically efficient for $F = F_0$ (Huber (1964), (1967)).

We obtain a scale invariant version of this estimate if we replace the defining equations by

$$\sum_{i\leqq n} \psi\left(\frac{X_i - T_n}{S_n}\right) = 0$$

and by

$$\int \psi\left(\frac{x - T(F)}{S(F)}\right)F(dx) = 0$$

respectively, where $S_n = S(F_n)$ is any robust estimate of scale, e.g. the interquartile range (compare also Section 12.1). If F is symmetric, then T_n and S_n are asymptotically independent, and the variance of T_n can be expressed as before, with

$$\Omega_F(x) = \frac{\psi(x/S(F))S(F)}{\int \psi'(x/S(F))F(dx)}\,.$$

4.2. *Linear combinations of order statistics (L-estimates).* Let $X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$ be the ordered sample; put

$$T_n = \sum_{i\leqq n} a_i X_{(i)}$$

where the weights are generated by $a_i = \int_{(i-1)/n}^{i/n} J(t)\,dt$ from some function J satisfying $\int_0^1 J(t)\,dt = 1$.

Under quite general (but not yet entirely satisfactory) regularity conditions, T_n converges to

$$T(F) = \int J(t)F^{-1}(t)\,dt$$

and $n^{\frac{1}{2}}(T_n - T(F))$ is asymptotically normal with asymptotic mean 0 and asymptotic variance

$$\sigma_L^2(F) = \int_0^1 U(t)^2\,dt - \left(\int_0^1 U(t)\,dt\right)^2\,,$$

where U is an indefinite integral of

$$U'(t) = \frac{J(t)}{f(F^{-1}(t))}\,.$$

We may also write

$$\sigma_L^2(F) = \int \Omega_F(x)^2 F(dx)$$

where

$$\Omega_F(x) = U(F(x)) - \int_0^1 U(t)\,dt\,.$$

If we choose

$$J(t) = \frac{1}{I(F_0)} \psi_0'(F_0^{-1}(t)) ,$$

where ψ_0' is the derivative of (4.1) and

$$I(F_0) = \int \psi_0(x)^2 F_0(dx)$$

is Fisher's information, then T_n is asymptotically efficient for F_0 (Jung (1955); the best results to date are those of Chernoff, Gastwirth and Johns (1967); compare also Bickel (1967)).

4.3. *Estimates derived from rank tests (R-estimates).* Consider a 2-sample rank test for shift: let $Y_1, \cdots, Y_n$ and $Z_1, \cdots, Z_n$ be two independent samples with distributions $F(x)$ and $F(x - \Delta)$ respectively. Form the combined sample of size $N = 2n$ and take as test statistic for testing $\Delta = 0$ against $\Delta > 0$

$$W(Y_1, \cdots, Y_n; Z_1, \cdots, Z_n) = \sum_{i \leq n} J\left(\frac{i}{N+1}\right) V_i$$

where $V_i = 1$ if the ith smallest entry in the combined sample is a Y, and $V_i = 0$ otherwise.

One can derive estimates of location from such tests: determine $T_n(X_1, \cdots, X_n)$ such that

$$W(X_1 - T_n, \cdots, X_n - T_n; -(X_1 - T_n), \cdots, -(X_n - T_n)) = 0 . \tag{4.2}$$

The asymptotic behavior of T_n can be determined from the asymptotic power of the rank test; it turns out that T_n tends to the solution $T(F)$ of

$$\int J\left(\frac{F(x) + 1 - F(2T(F) - x)}{2}\right) F(dx) = 0 ,$$

and that $n^{\frac{1}{2}}(T_n - T(F))$ is asymptotically normal with asymptotic mean 0 and asymptotic variance

$$\sigma_R^2(F) = \int \Omega_F(x)^2 F(dx) ,$$

where $\Omega_F(x)$ is defined as follows. Standardize the location parameter so that $T(F) = 0$, let $U(x)$ be an indefinite integral of

$$U'(x) = J'\left(\frac{F(x) + 1 - F(-x)}{2}\right) f(-x) ,$$

then, assuming $J(1 - t) = -J(t)$,

$$\Omega_F(x) = \frac{U(x) - \int U(x) f(x)\, dx}{\int U'(x) f(x)\, dx} .$$

In particular, if F is symmetric, then

$$\Omega_F(x) = \frac{J(F(x))}{\int J'(F(x)) f(x)^2\, dx} .$$

For symmetric F_0 we obtain a locally most powerful rank test with

$$J(t) = \psi_0(F_0^{-1}(t))$$

and then the estimate is asymptotically efficient for F_0. For asymmetric F_0 one cannot in general reach full efficiency with R-estimates. (Chernoff and Savage (1958), Hodges and Lehmann (1963), Hájek and Šidák (1967).)

5. Stability aspects of the preceding estimates. The three estimates considered in the preceding section can be written as functionals of the empirical distribution function $T_n = T(F_n)$ (either exactly or at least approximately). In particular, the L-estimate corresponds to

$$T(F) = \int J(t)F^{-1}(t)\,dt\,;$$

the M-estimate is defined by the implicit formula

$$\int \psi(x - T(F))F(dx) = 0$$

and the R-estimate by

$$\int J\left(\frac{F(x) + 1 - F(2T(F) - x)}{2}\right)F(dx) = 0\,.$$

In a somewhat simplified form, Hampel's (1968), (1971) ideas on stability can be described as follows.

A basic requirement is that a small change in F_n (either small changes affecting most or all observations, like rounding or grouping, or large changes affecting a few of them) should cause only small changes in $T_n = T(F_n)$.

This means that T should be continuous for the Prohorov metric d_p in the space of probability distributions:

$$d_p(F, G) = \inf\{\varepsilon | F(A) \leqq G(A^\varepsilon) + \varepsilon \quad \text{for all measurable sets} \quad A\}\,,$$

where A^ε denotes the closed ε-neighborhood of the set A.

For instance, L-estimates cannot be continuous unless $J(t) = 0$ for $t \notin [\alpha, 1-\alpha]$ for some $0 < \alpha < \frac{1}{2}$.

But then up to a fraction α of the observations can be grossly erroneous before anything catastrophic happens ("breakdown point").

If the true underlying distribution F is sufficiently smooth, M-, L- and R-estimates possess a von Mises derivative (von Mises (1947), Filippova (1962)). That is,

$$\lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)F + \varepsilon G) - T(F)}{\varepsilon} = \int \Omega_F(x)G(dx)$$

for some function Ω_F. (Note that this formula yields the value $\Omega_F(x)$ itself, if we take G to be the distribution function of a point mass 1 at x.) In most cases, a kind of Taylor expansion is valid:

$$n^{\frac{1}{2}}[T(F_n) - T(F) - \int \Omega_F(x)F_n(dx)] \to 0$$

in probability. Since

$$\int \Omega_F(x)F_n(dx) = \frac{1}{n}\sum \Omega_F(x_i)\,,$$

$\Omega_F(x)$ describes the influence of an observation with value x toward the estimate $T(F_n)$, and Ω_F has therefore been called "influence curve" by Hampel. Moreover, it follows that $n^{\frac{1}{2}}(T(F_n) - T(F))$ is asymptotically normal and that its asymptotic variance can be written as $\int \Omega_F(x)^2F(dx)$, as we already did systematically in the preceding section.

Note in particular that the influence function of the α-trimmed mean for a symmetric F is

$$\begin{aligned}
\Omega_F(x) &= \frac{1}{1-2\alpha}F^{-1}(\alpha) && \text{for} \quad x \leqq F^{-1}(\alpha) \\
&= \frac{1}{1-2\alpha}x && \text{for} \quad F^{-1}(\alpha) < x < F^{-1}(1-\alpha) \\
&= \frac{1}{1-2\alpha}F^{-1}(1-\alpha) && \text{for} \quad x \geqq F^{-1}(1-\alpha)
\end{aligned}$$

(for asymmetric F, a constant must be added so that $\int \Omega_F(x)F(dx) = 0$). In other words, and contrary to naive intuition, α-trimming does what α-Winsorizing was *supposed to* do, namely to reduce the influence of extreme observations to that of suitable order statistics.

The M-estimate given by

$$\psi(x) = \max(-k, \min(k, x))$$

has the same influence function as the α-trimmed mean and hence the same asymptotic behavior at a given symmetric F if k and α are related through $F(-k) = \alpha$.

If the true underlying distribution F is replaced by $(1-\varepsilon)F + \varepsilon H$, then

$$\varepsilon \int \Omega_F(x)H(dx)$$

describes the bias introduced by the small perturbation εH.

In my opinion, Hampel's influence function is the most important single heuristic tool for constructing robust estimates with specified properties. One will strive for influence functions which are bounded (to limit the influence of any single "bad" observation), which are reasonably continuous in x (to achieve insensitivity against roundoff and grouping effects) and which are reasonably continuous as a function of F (to stabilize the asymptotic variance of the estimate under small changes of F). At the same time, one will try to have an influence function roughly proportional to $-(\log f_0(x))'$, to achieve a high efficiency at the model distribution F_0.

6. Jackknifing. Let $T_n = T_n(X_1, \cdots, X_n)$, $n \geqq n_0$, be a sequence of estimates

such that we may reasonably consider T_{n-1} and T_n to be the "same" estimators despite the different sample sizes.

By definition, the *jackknifed pseudo-value* T_{ni}^* is

$$T_{ni}^* = nT_n(X_1, \cdots, X_n) - (n-1)T_{n-1}(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n) .$$

For instance, if T_n is the sample mean, then $T_{ni}^* = X_i$.

Originally Quenouille (1956) had proposed this device to reduce bias: if the bias of T_n has an asymptotic expansion

$$E(T_n) = \frac{a_1}{n} + \frac{a_2}{n^2} + O\left(\frac{1}{n^3}\right),$$

then the corresponding expansion for the jackknifed estimate

$$T_n^* = \frac{1}{n} \sum T_{ni}^*$$

lacks the $O(1/n)$ term.

Tukey (1958) noted that the pseudo-values often can be treated as if they were independent; in particular,

$$\frac{1}{n(n-1)} \sum_1^n (T_{ni}^* - T_n^*)^2$$

usually is a reliable estimate of the variance of both the original estimate T_n and of the jackknifed version T_n^*. Thus, once one has a computer program calculating the estimate, one is also able to assess the variability, which is a tremendous advantage.

Moreover,

$$\Omega_n(X_i) = T_{ni}^* - T_n^*$$

is a finite sample version of the influence function $\Omega_F(X_i)$ considered in the preceding section. This allows a simple empirical assessment of robustness of a complicated estimator. In particular, if the pseudo-sample contains outliers, then something should be done about the estimator T_n to improve its robustness.

For further information, cf. Miller (1964), (1968). It is hardly worthwhile to write down precise regularity conditions under which the jackknife has these useful properties—more work might be needed to check them than to devise a more specific and better (in particular computationally faster) variance estimate. Typically, the jackknife fails if the von Mises derivative $\Omega_F(x)$ is not continuous as function of F (as in the case of the median, where $\Omega_F(x) = \text{sign}\,(x)/(2F'(0))$).

7. Studentizing. As a rule, jackknifing gives a usable estimate for the variance of a robust estimate. However, it is often possible to find a simpler and sometimes better direct estimate. For instance, a glance at the influence function of the α-trimmed mean shows that its variance can be estimated through the α-Winsorized sample variance: assume $\alpha = g/n$ and put

$$s_n^2 = \frac{1}{n(n-1)} \cdot \frac{1}{(1-2\alpha)^2} \cdot \sum_i (X_i^W - \bar{X}_i^W)^2$$

where $\{X_i^W\}$ denotes the α-Winsorized sample (i.e., where the g extreme values on each side have been replaced by $X_{(g+1)}$ and $X_{(n-g)}$ respectively). (Cf. Tukey and McLaughlin (1963).)

For large n, $(T_n - \theta)/s_n$ then is asymptotically normal $N(0, 1)$. For not-so-large n, one may expect that it is approximately t-distributed. The proper number of degrees of freedom depends on the true underlying distribution; for normal observations $n - 2g - 1$ is a feasible approximation, but for longer tailed distributions, this number has to be decreased (the same remark applies to the classical t-statistic, $g = 0$), cf. Huber (1970).

For R-estimates, the parent rank tests furnish confidence intervals in a most direct fashion (Lehmann (1963c)).

8. Asymptotic minimax results. Let $\mathscr{C}$ be a convex compact set of distributions F on the (extended) real line. The problem is to find a sequence T_n of estimators of location which have a small asymptotic variance over the whole of $\mathscr{C}$; more precisely, the supremum over $\mathscr{C}$ of the asymptotic variance should be least possible. We shall restrict attention to symmetric distributions and to translation invariant estimates.

Let F_0 be the distribution in $\mathscr{C}$ having the smallest Fisher information $I(F) = \int (f'/f)^2 f\,dx$; there is one and usually only one such F_0 and in many interesting cases F_0 can be determined explicitly through variational methods (cf. Huber (1964)).

Thus, for any sequence T_n, the asymptotic variance of $n^{\frac{1}{2}}T_n$ under F_0 will at best be $1/I(F_0)$; our goal is to find a T_n such that the asymptotic variance does not exceed $1/I(F_0)$ for *any* $F \in \mathscr{C}$. More precisely, we shall have to require that for every $c < \infty$ there is an n_0, such that for $n \geqq n_0$

$$\sup_{F \in \mathscr{C}} E_F(\min(nT_n^2, c^2)) \leqq 1/I(F_0)\,.$$

In particular, this sequence T_n must be asymptotically efficient at F_0, and we shall therefore have a closer look at the estimates determined in Section 4.

Consider first the behavior of $\sigma_*^2(F)$ under infinitesimal variations of F, where the star stands for M, L or R. Let $F_\gamma = (1-\gamma)F_0 + \gamma F_1$, with $F_1 \in \mathscr{C}$, $0 \leqq \gamma \leqq 1$, then $F_\gamma \in \mathscr{C}$ because of convexity. Explicit computation yields that

$$\frac{d}{d\gamma}(1/\sigma_*^2(F_\gamma)) = \frac{d}{d\gamma} I(F_\gamma) \geqq 0 \qquad \text{for} \quad \gamma = 0\,.$$

In the case of M-estimators, $1/\sigma_M^2(F)$ is a convex function of F, hence $\sigma_M^2(F)$ has a global maximum at F_0, and the sequence of maximum likelihood estimates for F_0 solves the problem.

At least in the following important special case, also $\sigma_L^2(F)$ and $\sigma_R^2(F)$ have a global maximum at F_0 (Jaeckel (1971a)).

Assume that $\mathscr{E}$ is the set of all ε-contaminated normal distributions, i.e., the set of all distributions of the form $F = (1 - \varepsilon)\Phi + \varepsilon H$, where $0 \leqq \varepsilon < 1$ is a fixed number, Φ is the standard normal cumulative, and H varies over the set of all symmetric probability distributions. Then, the least favorable F_0 has the density

$$f_0(x) = \frac{1 - \varepsilon}{(2\pi)^{\frac{1}{2}}} e^{-\rho_0(x)}, \tag{8.1}$$

where

$$\begin{aligned} \rho_0(x) &= \tfrac{1}{2}x^2 && \text{for} \quad |x| < k \\ &= k|x| - \tfrac{1}{2}k^2 && \text{for} \quad |x| \geqq k, \end{aligned} \tag{8.2}$$

with k depending on ε through

$$\frac{\varepsilon}{1 - \varepsilon} = \frac{2}{k}\phi(k) - 2\Phi(-k), \tag{8.3}$$

where $\phi(x) = \Phi'(x) = (2\pi)^{-\frac{1}{2}}e^{-(x^2/2)}$. Thus

$$\begin{aligned} \psi_0(x) &= x && \text{for} \quad |x| < k \\ &= k \operatorname{sign}(x) && \text{for} \quad |x| \geqq k. \end{aligned} \tag{8.4}$$

The maximum likelihood estimate for this F_0 was treated in Huber (1964). The best linear combination of order statistics for F_0 is the α-trimmed mean, with $\alpha = F_0(-k)$. The corresponding R-estimate does not seem to allow a simple description.

The fact that ψ_0' is discontinuous sometimes causes trouble (e.g., in connection with grouped data and with Studentizing), and one might prefer to smooth ψ_0 near $\pm k$.

Note in this connection that the $\psi(x) = (1 - e^{-x})/(1 + e^{-x})$ corresponding to the logistic distribution $F(x) = 1/(1 + e^{-x})$ behaves much like a smooth version of ψ_0. The asymptotically efficient R-estimate for the logistic distribution is the so-called Hodges-Lehmann estimate—the median of the pairwise means $\frac{1}{2}(X_i + X_j)$.

9. Finite sample minimax results. According to the Neyman-Pearson lemma, the most powerful tests of a simple hypothesis P_0 against a simple alternative P_1 are given by likelihood ratio tests: form

$$h(\mathbf{X}) = \prod_{i \leqq n} p_1(X_i)/p_0(X_i)$$

and reject P_0 is $h(\mathbf{X}) > c$ (p_j is a density of P_j).

What happens if the P_j are only approximately known? Clearly, likelihood ratio tests may fail to be robust: a single factor $p_1(X_i)/p_0(X_i)$ equal (or almost equal) to 0 or ∞ might determine the outcome of the test.

If the uncertainty is formalized either in terms of ε-contamination, Prohorov distance, or total variation, e.g., if we replace P_j by the composite hypothesis

$\mathscr{P} = \{Q \mid \|Q - P_j\| \leqq \varepsilon\}$, then it turns out that there is a least favorable pair $(Q_0, Q_1) \in \mathscr{P}_0 \times \mathscr{P}_1$, such that the Neyman-Pearson tests of any level α between Q_0 and Q_1 coincide with maximin tests of the same level and same minimum power between $\mathscr{P}_0$ and $\mathscr{P}_1$. The likelihood ratio

$$\frac{q_1(x)}{q_0(x)} = \min\left(c'', \max\left(c', \frac{p_1(x)}{p_0(x)}\right)\right), \qquad c' < c''$$

is a censored version of $p_1(x)/p_0(x)$ (Huber (1965)).

The existence of such a least favorable pair is essentially equivalent to the assumption that each of the hypotheses $\mathscr{P}_j$ consists of the set of all probability measures majorized by some alternating capacity of order 2 (Strassen (1964), Huber and Strassen (1971)).

This robustized version of the Neyman-Pearson lemma can be used to build a finite sample minimax theory for estimates.

Assume that the measurement errors $\Delta_i = X_i - \theta$ are independent random variables whose distribution functions F_i lie anywhere within δ of some model distribution G:

$$\sup_x |F_i(x) - G(x)| \leqq \delta . \tag{9.1}$$

The logarithm of the density of G should be concave, *but there are no symmetry assumptions*; the idea works also for some other neighborhoods. To fix the idea, assume that $G = \Phi$ is the normal cumulative.

Let $a > 0$ be a fixed number; the goodness of any estimate $T = T(X_1, \cdots, X_n)$ of θ shall be assessed by the least α for which one can guarantee

$$P\{T < \theta - a\} \leqq \alpha$$
$$P\{T > \theta + a\} \leqq \alpha$$

for all θ and for all distributions satisfying (9.1)—the smaller α, the better the estimate.

The corresponding minimax solution T^0 can be described explicitly as follows. Let $\psi_0(x)$ be defined by (4.1), where k depends on δ and a (but not on the sample size n) through the relation

$$e^{-2ak}\Phi(a - k) - \Phi(-a - k) = (1 + e^{-2ak})\delta .$$

Let T^* and T^{**} be the smallest and the largest solution T of

$$\sum_{i \leqq n} \psi_0(X_i - T) = 0$$

respectively. Then put $T^0 = T^*$ or $T^0 = T^{**}$ at random with equal probability (Huber (1968a)).

The idea behind this result is very simple: one first constructs a maximin test (with level α and power $1 - \alpha$) between $\theta - a$ and $\theta + a$, then one derives an estimate from this test in the manner of Hodges and Lehmann (1963).

Note that this family of estimates happens to coincide with the asymptotically

minimax M-estimates for symmetric contamination, determined in the preceding section.

The actual value of the diffidence level α is difficult to determine, but asymptotic approximations (for $n \to \infty$, $\delta = O(n^{-\frac{1}{2}})$, $a = O(n^{-\frac{1}{2}})$) are available (Huber-Carol (1970)).

10. Criticisms and complements.

10.1. One might question the appropriateness of a minimax theory, especially of an asymptotic one, since minimax methods generally are too pessimistic. I think there are two answers: first, it seems that sample sizes reasonable for a given amount of not necessarily symmetric contamination will not allow to determine the nature of this contamination to a sufficient degree of accuracy (cf. Huber (1964) page 82 ff.). Second, one might check some actual error distributions in extremely large samples. Romanowski and Green (1965) have collected some quite impressive examples; it turns out that their largest sample behaves very much like the least favorable F_0 for the 2%-contaminated normal distribution (it lies between the slightly different curves for the least favorable F_0 for location (8.1) and the least favorable one for scale (12.1)). In this case, the 5%-trimmed mean would seem to be a very nearly efficient estimate. For their smaller samples, the conclusions are less definitive, but also these suggest trimming rates between 1% and 10%. In any case, safeguarding against the family of least favorable distributions (8.1) might not be overly pessimistic.

10.2. There is the usual objection against any asymptotic theory: one never knows whether it is applicable for any given finite sample size. Direct calculations are not very manageable except for rather small sample sizes (e.g., Tukey and McLaughlin (1963), Anscombe and Barron (1966), Crow and Siddiqui (1967), Gastwirth and Cohen (1970)); in addition, several Monte Carlo studies have been reported (e.g., Leone, Jayachandran and Eisenstat (1967), Dixon and Tukey (1968)); a very comprehensive one has just been completed at Princeton (Andrews *et al.* (1972)). While such studies mathematically do not prove anything about the applicability of the asymptotic theory, they furnish convincing evidence that the better among the more "rigid" (non-adaptive) procedures approach their asymptotic behavior rather fast, i.e., confidence levels between 1% and 5% derived from asymptotics seem to be sufficiently reliable for sample size 20 and larger.

Then there is the more basic problem whether asymptotic optimality bears any resemblance to finite sample optimality—after all, if, say, 1% of the observations are outliers, it makes quite a difference, whether the sample size is 5 (and 19 out of 20 samples are free from outliers) or whether it is 1000. But the answer to this is affirmative (cf. the end of Section 9).

10.3. The asymptotic minimax approach of Section 8 restricted attention to symmetric shapes. If one does not do so, the bias caused by the unknown small asymmetries of the true F would ultimately take precedence over the statistical

errors, leading to the sample median as the unique asymptotic minimax estimate (cf. Huber (1964) page 83). But presumably the practicing statisticians would then conclude that the sample size is unreasonably large. The finite sample approach of Section 9 does not make any symmetry assumptions.

10.4. *Adaptive estimates.* It is very tempting to devise estimates which adapt themselves to each particular sample. We already remarked in Section 3 that this seems to need exorbitant sample sizes. A very moderate type of adaptation—tailoring one parameter, say the trimming proportion α to the sample—might however work (Jaeckel (1971b)). For sample size 20, Jaeckel's procedure performs well for heavy tailed distributions, but the loss of efficiency at the normal model is rather high; the performance for not-so-long-tailed distributions is comparable to that of the 15% trimmed mean. A very promising approach is that of Takeuchi (1971); but there are some doubts whether it is robust against extremely long tails.

10.5. *Computational aspects.* The trimmed mean is probably easiest to compute; most of the work, $O(n \log n)$ operations, is spent for ordering the sample. M-estimates do not need much more work; my favorites are scale invariant versions of (8.4) (e.g. proposal 2 of Huber (1964) page 96 ff.). These are iterative procedures, but both small sample experiments and asymptotic theory show that it is hardly worthwhile to go beyond the first step of the iteration if one starts with the median and the suitably scaled interquartile range as preliminary estimates for location and scale.

The straightforward Hodges-Lehmann estimate—the median of the pairwise means $(X_i + X_j)/2$—is prohibitively expensive for all but the smallest samples, since it needs $O(n^2)$ operations. However, the Wilcoxon statistic W occurring in (4.2) is asymptotically linear in T_n, and one or two applications of the *regula falsi* to (4.2) should give an entirely adequate approximation to the estimate. (Jurečková (1969), Koul (1969), Jaeckel (1969).)

10.6. *Monte Carlo methods.* Small and intermediate sample properties ($n = 5$ to 50) of robust procedures almost always have to be determined by empirical sampling. Straightforward simulation is easy and wasteful, but sometimes drastic savings are possible. For instance, for a normal sample $\mathbf{X} = (X_1, \cdots, X_n)$, the sample mean $\bar{X}$ and the residual vector $\mathbf{Z} = (X_i - \bar{X})$ are independent, and the approximate distribution of the translation invariant estimate $T(\mathbf{X}) = T(\mathbf{Z}) + \bar{X}$ can be found efficiently by convoluting the empirical distribution of $T(Z)$ with the known normal distribution of $\bar{X}$. (Hodges (1967), Andrews *et al.* (1972).)

11. A few other selected estimates.

11.1. The so-called "quick-and-dirty" methods are estimates based on a few selected order statistics. One proposal (Gastwirth (1966)) for instance takes the $33\frac{1}{3}$, 50 and $66\frac{2}{3}$ percentiles with weights 0.3, 0.4 and 0.3 respectively and

achieves an efficiency of approximately 80% or better, simultaneously for the Cauchy, double-exponential, logistic and normal distributions. Of course, some care is needed if one works with grouped data—one should "degroup" the observations near the quantiles in question. Compare also Crow and Siddiqui (1967).

11.2. It may be desirable to cut to zero the influence of extremely outlying observations. As we have seen, the trimmed mean does not do so. The simplest way to achieve this is to use an M-estimate with a ψ vanishing outside of some finite interval (cf. the formula for the influence curve $\Omega_F(x)$ in Section 4.1), and Hampel recently has proposed some extremely promising estimates of this type (see Section 13). The so-called skipping procedures of Tukey also eliminate the influence of extreme outliers, but show a somewhat poorer performance (perhaps because their influence curves are rather jumpy): remove all observations from the sample whose distance to the nearest quartile exceeds the interquartile range (or $\frac{3}{2}$ times the interquartile range, etc.); repeat this until the sample does not change anymore. Then use any reasonable estimate on the remainder.

11.3. Bickel and Hodges (1967) have investigated a simplified version of the Hodges-Lehmann estimate, namely the median of the pairwise means

$$\tfrac{1}{2}(X_{(1)} + X_{(n)}), \cdots, \tfrac{1}{2}(X_{(k)} + X_{(n-k+1)}), \cdots$$

of symmetric order statistics. This estimate has a very good performance, as it seems, but its asymptotic distribution is *not* normal (it can be represented in terms of the time which Brownian motion spends above some curve).

11.4. Not every intuitively appealing estimate lives up to expectations. For instance, the "shorth" (the mean of the shortest half of the sample) behaved very poorly in an empirical sampling study ($n = 20$); a closer scrutiny reveals that even its rate of convergence is of the wrong order: $n^{\frac{1}{3}}(T_n - \theta)$ has a nondegenerate limiting distribution. For this kind of investigation, weak convergence (the "invariance principle") is an almost indispensable tool (Pyke and Shorack (1968), Billingsley (1968), Chernoff (1964), Andrews *et al.* (1972)).

12. Other estimation problems.

12.1. The scale parameter problem can be reduced to that of a location parameter by taking logarithms. However, some difficulties arise because the resulting distributions tend to be highly asymmetric, and it is not clear what one is estimating if the underlying distribution is only approximately known. But since one has taken logarithms, this uncertainty only acts as an unknown additive constant, and it makes sense to try to minimize the maximum of the asymptotic variance over a suitable neighborhood of some model distribution. Also here the asymptotically efficient M- and L-estimates for the least favorable distributions (=minimizing the Fisher information) seem to have good robustness properties.

For example, in the ε-contaminated normal case, the least favorable F_0 has the density

$$f_0(x) = \frac{1-\varepsilon}{(2\pi)^{\frac{1}{2}}} e^{-(x^2/2)} \qquad \text{for} \quad |x| < q\,, \tag{12.1}$$

$$= \frac{1-\varepsilon}{(2\pi)^{\frac{1}{2}}} e^{-(q^2/2)} \left(\frac{q}{|x|}\right)^{(q^2)} \qquad \text{for} \quad |x| \geqq q\,,$$

where ε and $q \geqq 2^{\frac{1}{2}}$ are related through

$$\frac{\varepsilon}{1-\varepsilon} = \frac{2q}{q^2-1}\,\phi(q) - 2\Phi(-q)\,. \tag{12.2}$$

The corresponding maximum likelihood estimate was treated by Huber (1964); another asymptotically efficient estimate of σ^2 for the distribution $F_0(x/\sigma)$ is the suitably scaled α-trimmed variance, where $\alpha = F_0(-q)$.

12.2. The higher dimensional location parameter problem can be treated in very much the same way as the one-dimensional problem, if one assumes that the error distribution is spherically symmetric. In particular, one can determine a least favorable F_0 just as in the one-dimensional case; it is somewhat surprising that $-\log f_0(x)$ fails to be convex even in the simplest contaminated normal case. The maximum likelihood approach works well, compare Gentleman (1965), Huber (1967).

Much less is known about affinely invariant procedures. A higher dimensional analogue to trimming has been called "peeling" by Tukey; it consists of deleting extreme points of the convex hull of the sample and to repeat this operation either a fixed number of times, or until a fixed percentage of the points has been removed. Not much is known about the behavior of such a procedure, but it might have quite intriguing properties, cf. Rényi and Sulanke (1963), (1964) and Carnal (1970).

A multivariate version of the Hodges-Lehmann estimate has been considered by Bickel (1964).

12.3. Relatively little is known about robust estimation in the general case, where there is neither translation nor scale invariance. It is fairly clear that a modified maximum likelihood estimate should have good robustness properties: put

$$\psi(x, \theta) = \min\left(c_2(\theta), \max\left(c_1(\theta), \frac{\partial}{\partial\theta} \log f(x, \theta)\right)\right) + b(\theta)\,,$$

where $f(x, \theta)$ is the probability density of the assumed family of distributions, and define an estimator T_n of θ by

$$\sum_{i \leqq n} \psi(X_i, T_n) = 0\,.$$

The difficulty resides in the proper choice of the truncating functions $c_1(\theta) < c_2(\theta)$

and of the function $b(\theta)$ regulating the bias. A serious conceptual difficulty is caused by the fact that one does not quite know what one is estimating; perhaps one should define the parameter to be estimated in terms of the estimator. One approach to this problem has been proposed by Hampel (1968), another one has been tentatively explored by Huber-Carol (1970).

12.4. *Regression and analysis of variance problems.* Consider the general linear regression problem

$$X_i = \sum_{j=1}^{p} c_{ij}\theta_j + U_i\,, \qquad 1 \leqq i \leqq n\,,$$

where the X_i are observed, the θ_j are to be estimated, the c_{ij} are known coefficients, and the U_i are independent random errors whose distribution functions F_i are approximately equal, but only approximately known.

The classical least squares method is to minimize

$$\sum_i (X_i - \sum_j c_{ij}\hat{\theta}_j)^2\,,$$

which generalizes at once to minimizing

$$\sum_i \rho_0(X_i - \sum_k c_{ik}\hat{\theta}_k)$$

with ρ_0 as in (8.3), for example. The procedure can be made scale invariant, if one determines $(\hat{\theta}_1, \cdots, \hat{\theta}_p)$ and $\hat{\sigma}$ from the $p + 1$ equations

$$\sum_i \psi_0((X_i - \sum_k c_{ik}\hat{\theta}_k)/\hat{\sigma})c_{ij} = 0\,, \qquad j = 1, \cdots, p$$

$$\frac{1}{n-p}\sum_i \psi_0((X_i - \sum_k c_{ik}\hat{\theta}_k)/\hat{\sigma})^2 = \beta\,,$$

where $\beta = E_\Phi(\psi_0(X)^2)$.

The asymptotic theory for these estimates is relatively straightforward if one assumes that p stays fixed as n tends to infinity (Relles (1968)), but some difficulties arise in the more realistic case where p/n is small but not entirely negligible. I conjecture that for symmetric error distributions the following statements are true.

Let ε be the maximum diagonal element of the projection matrix $\Gamma = C(C^TC)^{-1}C^T$, where $C = (c_{ij})$; since $\operatorname{tr}(\Gamma) = p$, we have $p/n \leqq \varepsilon$.

Then all estimates of the form $\hat{\alpha} = \sum a_j\hat{\theta}_j$ are asymptotically normal, iff $\varepsilon \to 0$. If the errors U_i have the common distribution F, then $\hat{\sigma}$ tends to σ_F determined by $E_F\psi_0(U_i/\sigma_F)^2 = \beta$, and the asymptotic covariance matrix of the $\hat{\theta}$ is given by

$$\frac{E_F\psi_0(U_i/\sigma_F)^2}{(E_F\psi_0'(U_i/\sigma_F))^2} \cdot \sigma_F{}^2 \cdot (C^TC)^{-1}\,.$$

If $p/n \approx \varepsilon$ is not entirely negligible, I would tentatively recommend to estimate this covariance matrix by the expression

$$\frac{n\beta}{m}\left(1 + \frac{p(n-m)}{nm}\right)\hat{\sigma}^2 W^{-1}$$

where $m = \sum_i \psi_0'((X_i - \sum_k c_{ik}\hat{\theta}_k)/\hat{\sigma})$ is the number of residuals falling into the middle, linear part of ψ_0, and the matrix W is determined by

$$W_{jr} = \sum_i \psi_0'((X_i - \sum_k c_{ik}\hat{\theta}_k)/\hat{\sigma})c_{ij}\,c_{ir}\,.$$

Thus, the customary methods of linear regression and analysis of variance can be carried over in a relatively straightforward way. Compare also Anscombe (1967).

Also other robust estimates can be generalized to the regression problem. A little consideration shows that it suffices in principle to robustize the special case $p = 1$, the simple straight line regression. If one can do that, one may attack the general problem iteratively by estimating one parameter at a time, keeping the others fixed at trial values.

Lehmann and his students have attacked several regression and analysis of variance problems with the aid of rank tests (Lehmann (1963a, b), plus a number of papers by different authors in subsequent volumes of the *Annals of Mathematical Statistics*). To illustrate the basic idea, consider the straight line regression problem (Adichie (1967)):

$$X_i = \alpha + \beta c_i + U_i\,,$$

where α and β are to be estimated.

Every test of the hypothesis $\beta = 0$ furnishes an estimate of β: apply the test of the pseudo-observations $X_i' = X_i - \hat{\beta}c_i$, and adjust the value of $\hat{\beta}$ in such a way that the test is least able to reject the hypothesis. The asymptotic power of these tests then can be used in a more or less straightforward way to compute the asymptotic variances and covariances, and hence the asymptotic efficiencies of these estimates.

Recently, Bickel (1971) has been able to extend the L-approach to the general regression problem.

On purpose, I have described the regression problem in terms of the classical least squares theory, where the matrix $C = (c_{ij})$ is thought to derive from a fixed and rigorous mathematical model. In statistics, it is more customary to treat the coefficients c_{ij} as "independent variables," possibly also subject to errors. Next to nothing is known about how to robustize regression procedures with respect to errors in the c_{ij}.

12.5. *Time series problems.* Stationary time series seem to pose different and in some sense unique robustness problems. Consider spectrum analysis of an observed stochastic process $X_1, \cdots, X_N$. Against which deviations from what model should one safeguard, especially if one does *not* have a specific parametric model? In my (admittedly limited) experience, isolated "outliers" (e.g., isolated malfunctions of the recording apparatus affecting a single X_i) are quite rare. What occurs are "bumps" and "quakes"—the first being a local shift in the mean value, extending over several consecutive X_i, the second being a local shift of variance. Both cases correspond to grafting a short piece of an extraneous process onto the process under consideration. Also here, sample skewness and kurtosis are excellent for detecting the presence of such accidents and one

might blot them out by smooth data windows. This would be a close analogue to rejection of outliers.

The following procedure is related to the one-step M-estimates mentioned in Section 10.5. Subtract first a (robust) average and trend, i.e., filter out the lowest frequencies, then apply a "smooth limiter": replace the process X_t by

$$\tilde{X}_t = 2\Phi(X_t/c) - 1$$

where Φ is the normal cumulative and c is some constant. The case $c = 0$ corresponds to "hard limiting"

$$\tilde{X}_t = \text{sign}\,(X_t)\,.$$

For a Gaussian process, the covariance functions of the two processes are related by

$$\tilde{R}(t) = \frac{2}{\pi} \text{ arc sin } \frac{R(t)}{1 + c^2}$$

(assuming $R(0) = 1$), cf. Thomas (1969) page 298 ff. But some care is needed—a strong low frequency component might mask a weak high frequency component, if one applies a limiter.

13. Concluding remarks. "Which estimate do you recommend for practical use?" This question is frequently asked but does not have a simple answer. The variety of situations (and also attitudes of statisticians!) occurring in practical applications will always demand a variety of tools. And the ever increasing number of good robust estimates makes the choice progressively harder. Nevertheless, in the light of the large empirical study we have just completed at Princeton (Andrews *et al.* (1972)), I shall venture some conditional suggestions. We looked at some 65 estimates of location under some 30 different situations (mostly symmetric distributions and mostly sample size 20).

In a poorly specified and presumably long-tailed situation, I might use Gastwirth's estimate (Section 11.1), which is very simple, or Hampel's estimate 12A (see below), which seems to admit a better estimate of its own variability (through its influence function, cf. Sections 5 and 7).

In a more tightly specified approximately normal situation, I might not want to sacrifice more than 5 to 10% efficiency at the normal model. Then the simplest good estimate is an α-trimmed mean (with $\alpha = 0.1$ or 0.15).

If one objects to the relatively poor breakdown point (Section 5) of the trimmed mean, a one-step M-estimate (Section 10.5) is an attractive alternative, e.g. the following (identified as P15 in Table I):

(i) take the median $T_0 = \text{med}\,\{X_i\}$ as a preliminary estimate of location;
(ii) take $S_0 = \text{med}\,\{|X_i - T_0|\}$ as a very robust estimate of scale;
(iii) estimate location by

$$T = T_0 + \frac{\sum \psi((X_i - T_0)/S_0)}{\sum \psi'((X_i - T_0)/S_0)} S_0$$

TABLE I

Monte Carlo variances of $n^{\frac{1}{2}}T_n$ for selected estimates and distributions, sample size $n = 20$

		$N(0,1)$		$(n-p)N(0,1)$ plus $pN(0,9)$, $n=20$					$18N(0,1)$ plus $2N(0,100)$	Cauchy		$(1-\varepsilon)N(0,1)$ $+\varepsilon N/U^*)$		
		$n=\infty$	$n=20$	$p=1$	$p=2$	$p=3$	$p=5$	$p=10$		$n=\infty$	$n=20$	$\varepsilon=0.1$	$\varepsilon=0.25$	$\varepsilon=1$
	mean	**1.000**	**1.00**	1.40	1.80	2.20	3.00	5.00	10.90	∞	—	—	—	—
	$\alpha=0.05$	1.026	1.02	**1.16**	1.39	1.64	2.27	4.45	2.90	8.77	24.0	1.47	3.84	35.9
trimmed	$\alpha=0.1$	1.060	1.06	1.17	**1.31**	1.47	1.93	3.98	1.46	4.77	7.3	1.26	1.81	13.6
mean	$\alpha=0.15$	1.100	1.10	1.19	1.32	**1.44**	1.80	3.56	1.43	3.48	4.6	1.26	1.64	9.3
	$\alpha=0.25$	1.195	1.20	1.27	1.41	1.50	1.79	3.13	1.47	2.55	3.1	1.33	1.64	6.6
	median	1.571	1.50	1.52	1.70	1.75	2.16	3.37	1.80	2.48	2.9	1.64	1.94	6.6
	$k=2.0$	1.010	1.01	1.17	1.41	1.66	2.30	4.56	1.78	6.74	9.3	1.30	2.17	18.4
Huber (1964)	$k=1.5$	1.037	1.04	**1.16**	1.32	1.49	1.96	4.09	1.50	4.44	5.7	1.24	1.74	11.4
prop. 2	$k=1.0$	1.107	1.11	1.21	1.34	**1.44**	**1.78**	3.40	1.43	3.02	3.7	1.26	1.62	7.5
	$k=0.7$	1.187	1.20	1.27	1.42	1.49	1.79	3.13	1.47	2.52	3.0	1.33	1.64	6.6
Hodges-Lehmann		1.047	1.06	1.18	1.35	1.50	1.88	3.62	1.52	3.29	4.2	1.26	1.70	8.4
Gastwirth (1966)		1.28	1.23	1.30	1.45	1.52	1.82	**3.12**	1.50	2.50	3.1	1.36	1.67	6.6
Jaeckel (1969)		1.000	1.10	1.21	1.37	1.47	1.82	3.54	1.45	2.55	3.5	1.27	1.63	7.2
Hogg (1967)		1.000	1.06	1.28	1.56	1.79	2.42	4.83	1.79	2.48	4.4	1.42	1.90	9.4
Takeuchi (1969)		1.000	1.05	1.19	1.38	1.53	2.02	4.06	1.32	2.00	3.5	1.22	1.60	7.6
A15		1.037	1.05	1.17	1.33	1.47	1.91	3.78	1.49	3.77	4.5	1.24	1.69	8.8
P15		1.037	1.05	1.17	1.33	1.47	1.91	3.81	1.49	3.77	4.5	1.24	1.70	8.8
Hampel 25A		1.025	1.05	**1.16**	1.32	1.49	1.94	3.97	**1.26**		3.7	**1.19**	1.59	8.0
Hampel 12A		1.166	1.20	1.26	1.40	1.47	**1.78**	3.24	1.32		2.7	1.30	**1.56**	6.2
Max Likelihood Cauchy			1.72	1.66	1.84	1.84	2.14	3.24	1.71	**2.00**	**2.3**	1.75	1.96	**5.8**

* N/U denotes the distribution of the quotient of a normal (0, 1) variable divided by a uniform (0, 1) variable.

with $\psi(x) = \max(-k, \min(k, x))$, where $k = 1.5/\Phi^{-1}(\frac{3}{4}) \doteq 2.22$ for P15.

Iteration of (iii), i.e. solving $\sum \psi((X_i - T)/S_0) = 0$ for T by Newton's method improves the estimate only very slightly (estimate A15).

Hampel's recent proposals (Section 11.2) give even better performances for long tails: let (i), (ii), with iterative (iii) as above, but put

$$\begin{aligned} \psi(x) = -\psi(-x) &= x && \text{for } 0 \leqq x < a \\ &= a && \text{for } a \leqq x < b \\ &= \frac{c-x}{c-b}a && \text{for } b \leqq x < c \\ &= 0 && \text{for } x \geqq c \end{aligned}$$

with, say, $a = 2.5$, $b = 4.5$, $c = 9.5$ (estimate 25A) or $a = 1.2$, $b = 3.5$, $c = 8.0$ (estimate 12A). These estimates (whose one-step versions have not yet been investigated) certainly look very promising and will be investigated further.

REFERENCES

ADICHIE, J. N. (1967). Estimates of regression parameters based on rank-tests. *Ann. Math. Statist.* **38** 894–904.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., and TUKEY, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press.

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

ANSCOMBE, F. (1960). Rejection of outliers. *Technometrics* **2** 123–147.

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

ANSCOMBE, F. and BARRON, B. A. (1966). Treatment of outliers in samples of size three. *J. Res. Nat. Bur. Standards Sect. B* **70** 141–147.

BESSEL, F. W. (1818). *Fundamenta Astronomiae.* Königsberg.

BESSEL, F. W. (1938). Untersuchungen über die Wahrscheinlichkeit von Beobachtungsfehlern. *Astronom. Nachr.* **15** 369–404.

BESSEL, F. W. and BAEYER (1838). *Gradmessung in Ostpreussen.* Berlin.

BICKEL, P. J. (1964). On some alternative estimates for shift in the p-variate one sample problem. *Ann. Math. Statist.* **35** 1079–1090.

BICKEL, P. J. (1965). On some robust estimates of location. *Ann. Math. Statist.* **36** 847–858.

BICKEL, P. J. (1967). Some contributions to the theory of order statistics. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 575–591.

BICKEL, P. J. (1971). On some analogues to linear combinations of order statistics in the linear model. (Unpublished manuscript).

BICKEL, P. J. and HODGES, J. L. JR. (1967). The asymptotic theory of Galton's test and a related simple estimate of location. *Ann. Math. Statist.* **38** 73–89.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BIRNBAUM, A. and LASKA, E. (1967). Optimal robustness: A general method with applications to linear estimators of location. *J. Amer. Statist. Assoc.* **62** 1230–1240.

BIRNBAUM, A. and MIKÉ, V. (1970). Asymptotically robust estimators of location. *J. Amer. Statist. Assoc.* **65** 1265–1282.

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

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

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

BOX, G. E. P. and TIAO, G. C. (1964a). A Bayesian approach to the importance of assumptions applied to the comparison of variances. *Biometrika* **51** 153–167.

BOX, G. E. P. and TIAO, G. C. (1964b). A note on criterion robustness and inference robustness. *Biometrika* **51** 169–173.

CARNAL, H. (1970). Die konvexe Hülle von n rotationssymmetrisch verteilten Punkten. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **15** 168–176.

CHAUVENET, W. (1863). *Manual of Spherical and Practical Astronomy.* Philadelphia.

CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41.

CHERNOFF, H., GASTWIRTH, J. L. and JOHNS, M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.* **38** 52–72.

CHERNOFF, H. and SAVAGE, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29** 972–994.

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

CZUBER, E. (1891). *Theorie der Beobachtungsfehler.* Leipzig.

DIXON, W. J. and TUKEY, J. W. (1968). Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization 2). *Technometrics* **10** 83–98.

FERGUSON, T. S. (1961). On the rejection of outliers. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 253–287.

FILIPPOVA, A. A. (1962). Mises' theorem of the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theor. Probability Appl.* **7** 24–57.

GASTWIRTH, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61**, 929–948.

GASTWIRTH, J. L. and COHEN, M. L. (1970). Small sample behavior of some robust linear estimators of location. *J. Amer. Statist. Assoc.* **65** 946–973.

GASTWIRTH, J. L. and RUBIN, H. (1969). The behavior of robust estimators on dependent data. Purdue University, Dept. of Statistics, Mimeograph Series No. 197.

GAUSS, C. F. (1821). Göttingische gelehrte Anzeigen, pages 321–327 (reprinted in *Werke* Bd. **4** page 98).

GENTLEMAN, W. M. (1965). Robust estimation of multivariate location by minimizing pth power deviations. Ph. D. Dissertation, Princeton Univ.

GRUBBS, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11** 1–21.

HÁJEK, J. and ŠIDÁK Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph. D. Dissertation, Univ. of California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HODGES, J. L. JR. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 163–186.

HODGES, J. L. and LEHMANN E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

HOGG, R. V. (1967). Some observations on robust estimation. *J. Amer. Statist. Assoc.* **62** 1179–1186.

HØYLAND, A. (1968). Robustness of the Wilcoxon estimate of location against a certain dependence. *Ann. Math. Statist.* **39** 1196–1201.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36** 1753–1758.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 221–233.

HUBER, P. J. (1968a). Robust confidence limits. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** 269–278.

HUBER, P. J. (1968b). Robust estimation. *Mathematical Centre Tracts* **27** 3–25. (Amsterdam)

HUBER, P. J. (1970). Studentizing robust estimates, in *Nonparametric Techniques in Statistical Inference*, ed. M. L. Puri. Cambridge Univ. Press, 453–463.

HUBER, P. J. and STRASSEN, V. (1971). The Neyman-Pearson lemma for capacities. (Unpublished manuscript).

HUBER-CAROL, C. (1970). Etude asymptotique de tests robustes. Ph. D. Dissertation, Swiss Federal Institute of Technology.

JAECKEL, L. A. (1969). Robust estimates of location. Ph. D. Dissertation, Univ. of California, Berkeley.

JAECKEL, L. A. (1971a). Robust estimates of location: Symmetry and asymmetric contamination. *Ann. Math. Statist.* **42** 1020–1034.

JAECKEL, L. A. (1971b). Some flexible estimates of location. *Ann. Math. Statist.* **42** 1540–1552.

JEFFREYS, H. (1932). An alternative to the rejection of outliers. *Proc. Roy. Soc. Ser. A* **137** 78–87.

JUNG, J. (1955). On linear estimates defined by continuous weight function. *Ark. Mat.* **3** 199–209.

JUREČKOVÁ, J. (1969). Asymptotic linearity of a rank statistic in regression parameter. *Ann. Math. Statist.* **40** 1889–1900.

KOUL, H. L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40** 1950–1979.

LEHMANN, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34** 957–966.

LEHMANN, E. L. (1963b). Asymptotically nonparametric inference: an alternative approach to linear models. *Ann. Math. Statist.* **34** 1494–1506.

LEHMANN, E. L. (1963c). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34** 1507–1512.

LEONE, F. C., JAYACHANDRAN, T. and EISENSTAT, S. (1967). A study of robust estimators. *Technometrics* **9** 652–660.

MILLER, R. G. JR. (1964). A trustworthy jackknife. *Ann. Math. Statist.* **35** 1594–1605.

MILLER, R. G. JR. (1968). Jackknifing variances. *Ann. Math. Statist.* **39** 567–582.

NEWCOMB, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *Amer. J. Math.* **8** 343–366.

NOETHER, G. E. (1967). Wilcoxon confidence intervals for location parameters in the descrete case. *J. Amer. Statist. Assoc.* **62** 184–188.

PEARSON, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* **23** 114–133.

PEIRCE, B. (1852). Criterion for the rejection of doubtful observations. *Astronom. J.* **2** 161–163.

PYKE, R. and SHORACK, G. (1968). Weak convergence of a two sample empirical process and a new approach to Chernoff-Savage theorems. *Ann. Math. Statist.* **39** 755–771.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

RELLES, D. A. (1968). Robust regression by modified least squares. Ph. D. Dissertation, Yale Univ.

RÉNYI, A. and SULANKE, R. (1963; 1964). Über die konvexe Hülle von n zufällig gewählten Punkten. I, II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **2** 75–84, **3** 138–147.

ROMANOWSKI, M. and GREEN, E. (1965). Practical applications of the modified normal distribution. *Bull. Géodésique* **76** 1–20.

SIDDIQUI, M. M. and RAGHUNANDANAN K. (1967). Asymptotically robust estimators of location. *J. Amer. Statist. Assoc.* **62** 950–953.

STRASSEN, V. (1964). Messfehler und Information. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **2** 273–305.

"STUDENT" (1927). Errors of routine analysis. *Biometrika* **19** 151–164.

TAKEUCHI, K. (1971). A uniformly asymptotically efficient estimator of a location parameter. *J. Amer. Statist. Assoc.* **66** 292–301.

THOMAS, J. B. (1969). *An Introduction to Statistical Communication Theory*. Wiley, New York, 298 ff.

TUKEY, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford Univ. Press.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.

TUKEY, J. W. and MCLAUGHLIN, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhyā Ser. A* **25** 331–352.

VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.* **41** 172–181.

VON MISES, R. (1947). On the asymptotic distributions of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.