ROBUST ESTIMATION OF A LOCATION PARAMETER[1]

By Peter J. Huber[2]

University of California, Berkeley

1. Introduction and summary. This paper contains a new approach toward a theory of robust estimation; it treats in detail the asymptotic theory of estimating a location parameter for contaminated normal distributions, and exhibits estimators—intermediaries between sample mean and sample median—that are asymptotically most robust (in a sense to be specified) among all translation invariant estimators. For the general background, see Tukey (1960) (p. 448 ff.)

Let $x_1, \cdots, x_n$ be independent random variables with common distribution function $F(t - \xi)$. The problem is to estimate the location parameter ξ, but with the complication that the prototype distribution $F(t)$ is only approximately known. I shall primarily be concerned with the model of indeterminacy $F = (1 - \epsilon)\Phi + \epsilon H$, where $0 \leqq \epsilon < 1$ is a known number, $\Phi(t) = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{t} \exp(-\frac{1}{2}s^2)\, ds$ is the standard normal cumulative and H is an unknown contaminating distribution. This model arises for instance if the observations are assumed to be normal with variance 1, but a fraction ϵ of them is affected by gross errors. Later on, I shall also consider other models of indeterminacy, e.g., $\sup_t |F(t) - \Phi(t)| \leqq \epsilon$.

Some inconvenience is caused by the fact that location and scale parameters are not uniquely determined: in general, for fixed ϵ, there will be several values of ξ and σ such that $\sup_t|F(t) - \Phi((t - \xi)/\sigma)| \leqq \epsilon$, and similarly for the contaminated case. Although this inherent and unavoidable indeterminacy is small if ϵ is small and is rather irrelevant for practical purposes, it poses awkward problems for the theory, especially for optimality questions. To remove this difficulty, one may either (i) restrict attention to symmetric distributions, and estimate the location of the center of symmetry (this works for ξ but not for σ); or (ii) one may define the parameter to be estimated in terms of the estimator itself, namely by its asymptotic value for sample size $n \to \infty$; or (iii) one may define the parameters by arbitrarily chosen functionals of the distribution (e.g., by the expectation, or the median of F). All three possibilities have unsatisfactory aspects, and I shall usually choose the variant which is mathematically most convenient.

It is interesting to look back to the very origin of the theory of estimation, namely to Gauss and his theory of least squares. Gauss was fully aware that his main reason for assuming an underlying normal distribution and a quadratic loss function was mathematical, i.e., computational, convenience. In later times,

Received 4 June 1963.

[1] This research was performed while the author held an Adolph C. and Mary Sprague Miller Fellowship.

[2] Now at Cornell University.

this was often forgotten, partly because of the central limit theorem. However, if one wants to be honest, the central limit theorem can at most explain why many distributions occurring in practice are approximately normal. The stress is on the word "approximately."

This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): What happens if the true distribution deviates slightly from the assumed normal one? As is now well known, the sample mean then may have a catastrophically bad performance: seemingly quite mild deviations may already explode its variance. Tukey and others proposed several more robust substitutes—trimmed means, Winsorized means, etc.—and explored their performance for a few typical violations of normality. A general theory of robust estimation is still lacking; it is hoped that the present paper will furnish the first few steps toward such a theory.

At the core of the method of least squares lies the idea to minimize the sum of the squared "errors," that is, to adjust the unknown parameters such that the sum of the squares of the differences between observed and computed values is minimized. In the simplest case, with which we are concerned here, namely the estimation of a location parameter, one has to minimize the expression $\sum_i (x_i - T)^2$; this is of course achieved by the sample mean $T = \sum_i x_i/n$. I should like to emphasize that no loss function is involved here; I am only describing how the least squares estimator is defined, and neither the underlying family of distributions nor the true value of the parameter to be estimated enters so far.

It is quite natural to ask whether one can obtain more robustness by minimizing another function of the errors than the sum of their squares. We shall therefore concentrate our attention to estimators that can be defined by a minimum principle of the form (for a location parameter):

(M) $\quad T = T_n(x_1, \cdots, x_n)$ *minimizes* $\sum_i \rho(x_i - T)$, *where* ρ *is a non-constant function.*

Of course, this definition generalizes at once to more general least squares type problems, where several parameters have to be determined.

This class of estimators contains in particular (i) the sample mean ($\rho(t) = t^2$), (ii) the sample median ($\rho(t) = |t|$), and more generally, (iii) all maximum likelihood estimators ($\rho(t) = -\log f(t)$, where f is the assumed density of the untranslated distribution).

These (M)-estimators, as I shall call them for short, have rather pleasant asymptotic properties; sufficient conditions for asymptotic normality and an explicit expression for their asymptotic variance will be given.

How should one judge the robustness of an estimator $T_n(x) = T_n(x_1, \cdots, x_n)$? Since ill effects from contamination are mainly felt for large sample sizes, it seems that one should primarily optimize large sample robustness properties. Therefore, a convenient measure of robustness for asymptotically normal estimators seems

to be the supremum of the asymptotic variance ($n \to \infty$) when F ranges over some suitable set of underlying distributions, in particular over the set of all $F = (1 - \epsilon)\Phi + \epsilon H$ for fixed ϵ and symmetric H.

On second thought, it turns out that the asymptotic variance is not only easier to handle, but that even for moderate values of n it is a better measure of performance than the actual variance, because (i) the actual variance of an estimator depends very much on the behavior of the tails of H, and the supremum of the actual variance is infinite for any estimator whose value is always contained in the convex hull of the observations. (ii) If an estimator is asymptotically normal, then the important central part of its distribution and confidence intervals for moderate confidence levels can better be approximated in terms of the asymptotic variance than in terms of the actual variance.

If we adopt this measure of robustness, and if we restrict attention to (M)-estimators, then it will be shown that the most robust estimator is uniquely determined and corresponds to the following ρ: $\rho(t) = \frac{1}{2}t^2$ for $|t| < k$, $\rho(t) = k|t| - \frac{1}{2}k^2$ for $|t| \geqq k$, with k depending on ϵ. This estimator is most robust even among all translation invariant estimators. Sample mean ($k = \infty$) and sample median ($k = 0$) are limiting cases corresponding to $\epsilon = 0$ and $\epsilon = 1$, respectively, and the estimator is closely related and asymptotically equivalent to Winsorizing.

I recall the definition of Winsorizing: assume that the observations have been ordered, $x_1 \leqq x_2 \leqq \cdots \leqq x_n$, then the statistic $T = n^{-1}(gx_{g+1} + x_{g+1} + x_{g+2} + \cdots + x_{n-h} + hx_{n-h})$ is called the *Winsorized mean*, obtained by *Winsorizing* the g leftmost and the h rightmost observations. The above most robust (M)-estimators can be described by the same formula, except that in the first and in the last summand, the factors x_{g+1} and x_{n-h} have to be replaced by some numbers u, v satisfying $x_g \leqq u \leqq x_{g+1}$ and $x_{n-h} \leqq v \leqq x_{n-h+1}$, respectively; g, h, u and v depend on the sample.

In fact, this (M)-estimator is the maximum likelihood estimator corresponding to a unique least favorable distribution F_0 with density $f_0(t) = (1 - \epsilon)(2\pi)^{-\frac{1}{2}}e^{-\rho(t)}$. This f_0 behaves like a normal density for small t, like an exponential density for large t. At least for me, this was rather surprising—I would have expected an f_0 with much heavier tails.

This result is a particular case of a more general one that can be stated roughly as follows: Assume that F belongs to some convex set C of distribution functions. Then the most robust (M)-estimator for the set C coincides with the maximum likelihood estimator for the unique $F_0 \varepsilon C$ which has the smallest Fisher information number $I(F) = \int (f'/f)^2 f\, dt$ among all $F \varepsilon C$.

Miscellaneous related problems will also be treated: the case of non-symmetric contaminating distributions; the most robust estimator for the model of indeterminacy $\sup_t |F(t) - \Phi(t)| \leqq \epsilon$; robust estimation of a scale parameter; how to estimate location, if scale and ϵ are unknown; numerical computation of the estimators; more general estimators, e.g., minimizing $\sum_{i<j} \rho(x_i - T, x_j - T)$, where ρ is a function of two arguments.

Questions of small sample size theory will not be touched in this paper.

2. Asymptotic normality of (M)-estimators for convex ρ. It will be assumed throughout this section that ρ is a continuous convex real-valued function of a real variable t, tending to $+\infty$ as $t \to \pm\infty$.

Let $x_1, \cdots, x_n$ be independent identically distributed random variables with common distribution function F. Let $[T_n(x)]$ be the set of all those ξ for which $Q(\xi) = \sum_{i=1}^n \rho(x_i - \xi)$ reaches its infimum $Q_{\inf}$. Obviously, $[T_n(x)]$ is invariant under translations: $[T_n(x + c)] = [T_n(x)] + c$. By $T_n(x)$ we shall denote any representation of the set valued function $(x_1, \cdots, x_n) \to [T_n(x)]$ by a single-valued function $(x_1, \cdots, x_n) \to T_n(x)\ \varepsilon\ [T_n(x)]$, e.g., $T_n(x) =$ midpoint of $[T_n(x)]$.

LEMMA 1. *$Q(\xi)$ is a convex function of ξ, and $[T_n(x)]$ is non-empty, convex and compact. If ρ is strictly convex, then $[T_n(x)]$ is reduced to a single point.*

PROOF. (Strict) convexity of Q follows immediately from (strict) convexity of ρ. The sets $\{\xi \mid Q(\xi) \leqq Q_{\inf} + m^{-1}\}$ form a decreasing sequence of non-empty convex compact sets as $m \to \infty$, hence their intersection $[T_n(x)]$ is non-empty convex compact. If Q is strictly convex, and if ξ', ξ'' were distinct points of $[T_n(x)]$, then we would have $Q_{\inf} = Q(\frac{1}{2}\xi' + \frac{1}{2}\xi'') < \frac{1}{2}Q(\xi') + \frac{1}{2}Q(\xi'') = Q_{\inf}$, which is a contradiction.

Let $\psi = \rho'$ be the derivative of ρ, normalized such that $\psi(t) = \frac{1}{2}\psi(t - 0) + \frac{1}{2}\psi(t + 0)$. ψ is monotone increasing and strictly negative (positive) for large negative (positive) values of t.

If ψ is continuous, $T_n(x)$ may equivalently be defined by the equation $\sum_{i=1}^n \psi(x_i - T_n(x)) = 0$.

Define $\lambda(\xi) = \int \psi(t - \xi)F(dt) = E\psi(t - \xi)$.

LEMMA 2. *If there is a ξ_0 such that $\lambda(\xi_0)$ exists and is finite, then $\lambda(\xi)$ exists for all ξ (possibly $\lambda(\xi) = \pm\infty$), is monotone decreasing and strictly positive (negative) for large negative (positive) values of ξ.*

PROOF. Split ψ into its positive and negative part $\psi = \psi^+ - \psi^-$. Then $\lambda(\xi) = \int \psi^+(t - \xi)F(dt) - \int \psi^-(t - \xi)F(dt)$. For $\xi = \xi_0$ both integrals exist and are finite. For $\xi \geqq \xi_0$ the first integral is bounded $0 \leqq \int \psi^+(t - \xi)F(dt) \leqq \int \psi^+(t - \xi_0)F(dt)$, and similarly, for $\xi \leqq \xi_0$, the second integral is bounded $0 \leqq \int \psi^-(t - \xi)F(dt) \leqq \int \psi^-(t - \xi_0)F(dt)$. Hence, at least one of the two integrals is finite, thus $\lambda(\xi)$ exists everywhere. $\lambda(\xi)$ is monotone decreasing in ξ since $\psi(t - \xi)$ is. Now we want to show that $\lambda(\xi)$ is strictly negative for large positive values of ξ, and strictly positive for large negative ξ; of course, it suffices to prove the first of these two assertions. Let $\epsilon > 0$, and let M be such that $\int_M^\infty \psi^+(t - \xi_0)F(dt) < \epsilon$; for sufficiently large ξ we have $\int_{-\infty}^M \psi^+(t - \xi)F(dt) = 0$, hence $\int_{-\infty}^{+\infty} \psi^+(t - \xi)F(dt) < \epsilon$, which implies that $\int \psi^+(t - \xi)F(dt) \to 0$ for $\xi \to \infty$. Since ψ takes upon strictly negative values, there is a $\delta > 0$ such that $\int \psi^-(t - \xi)F(dt) > \delta$ for sufficiently large ξ, thus $\lambda(\xi)$ is strictly negative for sufficiently large values of ξ.

LEMMA 3. ("Consistency of T_n"). *Assume that there is a c such that $\lambda(\xi) > 0$ for $\xi < c$ and $\lambda(\xi) < 0$ for $\xi > c$. Then $T_n \to c$ almost sure and in probability.*

PROOF. Let $\epsilon > 0$. Then, by the law of large numbers,

$$(1/n)\sum_{i=1}^{n}\psi(x_i - c - \epsilon) \to \lambda(c + \epsilon) < 0,$$

$$(1/n)\sum_{i=1}^{n}\psi(x_i - c + \epsilon) \to \lambda(c - \epsilon) > 0,$$

a.s. and in probability. Hence, by monotonicity of ψ, for a.a. sample sequences, $c - \epsilon < [T_n(x)] < c + \epsilon$ holds from some n on, and similarly, $P[c - \epsilon < [T_n(x)] < c + \epsilon] \to 1$.

REMARK. The assumption of Lemma 3 could have been replaced—in view of Lemma 2—by the assumption that $\lambda(\xi)$ exists and is finite for some ξ_0, and that it does not identically vanish on a nondegenerate interval.

LEMMA 4. ("Asymptotic normality"). *Assume that* (i) $\lambda(c) = 0$, (ii) $\lambda(\xi)$ *is differentiable at* $\xi = c$, *and* $\lambda'(c) < 0$, (iii) $\int \psi^2(t - \xi)F(dt)$ *is finite and continuous at* $\xi = c$. *Then* $n^{\frac{1}{2}}(T_n(x) - c)$ *is asymptotically normal with asymptotic mean* 0 *and asymptotic variance* $V(\psi, F) = \int \psi^2(t - c)F(dt)/(\lambda'(c))^2$.

PROOF. Without loss of generality assume $c = 0$. We have to show that for every fixed real number g, $P[n^{\frac{1}{2}}T_n < g\sigma] \to \Phi(g)$, where $\sigma = V(\psi, F)^{\frac{1}{2}}$. Since

$$[T_n < g\sigma n^{-\frac{1}{2}}] \subset \left[\sum_{i=1}^{n}\psi(x_i - g\sigma n^{-\frac{1}{2}}) < 0\right] \subset [T_n \leqq g\sigma n^{-\frac{1}{2}}],$$

it suffices to show that

$$p_n = P\left[\sum_{i=1}^{n}\psi(x_i - g\sigma n^{-\frac{1}{2}}) < 0\right] \to \Phi(g).$$

Let $s^2 = \int (\psi(t - g\sigma n^{-\frac{1}{2}}) - \lambda(g\sigma n^{-\frac{1}{2}}))^2F(dt)$, then the $y_i = (\psi(x_i - g\sigma n^{-\frac{1}{2}}) - \lambda(g\sigma n^{-\frac{1}{2}}))/s$ are independent random variables with mean 0 and variance 1. We have

$$p_n = P\left[n^{-\frac{1}{2}}\sum_{i=1}^{n} y_i < -n^{\frac{1}{2}}\lambda(g\sigma n^{-\frac{1}{2}})/s\right]$$

and $-n^{\frac{1}{2}}\lambda(g\sigma n^{-\frac{1}{2}})/s \to g$. We shall see presently that $n^{-\frac{1}{2}}\sum y_i$ is asymptotically normal with mean 0 and variance 1, hence $p_n \to \Phi(g)$.

There is a slight complication: although the y_i are independent identically distributed summands, they are different for different values of n, and the usual formulations of normal convergence theorems do not apply. But by the normal convergence criterion, as given in Loève (1960), p. 295, the distribution of $\sum n^{-\frac{1}{2}}y_i$ converges toward the standard normal iff for every $\epsilon > 0$, as $n \to \infty$, $\int_E y_1^2F(dx_1) \to 0$, the integration being extended over the set $E = [|y_1| \geqq n^{\frac{1}{2}}\epsilon]$. Since $\lambda(g\sigma n^{-\frac{1}{2}}) \to 0$, this holds iff for every $\epsilon > 0$, $\int_{E'} \psi^2(x_1 - n^{-\frac{1}{2}}g\sigma)F(dx_1) \to 0$, the integration being extended over the set $E' = [|\psi(x_1 - n^{-\frac{1}{2}}g\sigma)| \geqq n^{\frac{1}{2}}\epsilon]$. Now, let $\eta > 0$ be given and let n_0 be such that $|n^{-\frac{1}{2}}g\sigma| < \eta$ for $n \geqq n_0$. Then, since ψ is monotone, we have $\psi^2(x_1 - n^{-\frac{1}{2}}g\sigma) \leqq u^2(x_1)$ for $n \geqq n_0$ with

$u(x_1) = \max\{|\psi(x_1 - \eta)|, |\psi(x_1 + \eta)|\}$. Since $E\psi^2(x_1 \pm \eta)$ exists, we have $\int_{E''} u^2(x_1)\, F(dx_1) \to 0$, the integration being extended over the set $E'' = [|u| \geqq n^{\frac{1}{2}}\epsilon] \supset E'$. This proves the assertion.

REMARK. Lemma 4 admits a slight generalization, which is sometimes useful: If λ has different left and right derivatives at c, then the asymptotic distribution of $n^{\frac{1}{2}}(T_n(x) - c)$ is pieced together from two half-normal distributions, with variances determined by the one-sided derivatives of λ respectively in the same way as in Lemma 4.

For the remainder of this section, assume for simplicity $c = 0$. By formally interchanging the order of integration and differentiation one obtains $\lambda'(0) = [(d/d\xi)\int\psi(t - \xi)F(dt)]_{\xi=0} = -\int\psi'(t)F(dt) = -E\psi'$. This makes sense either when ψ is sufficiently regular or when F is sufficiently regular so that ψ' may be interpreted as a Schwartz distribution. Moreover, if F has an absolutely continuous density f, we find by partial integration $\lambda'(0) = \int\psi(t)f'(t)\,dt$. The Schwarz inequality now yields the inequality

$$V(\psi, F) = \frac{E\psi^2}{(E\psi')^2} = \frac{\int\psi^2 f\,dt}{[\int\psi(f'/f)f\,dt]^2} \geqq \frac{1}{\int(f'/f)^2 f\,dt}. \tag{1}$$

We have strict inequality unless $\psi = -pf'/f$, for some constant p, that is, unless $f(t) = \text{const.}\exp(-\rho(t)/p)$ and then the (M)-estimator is the maximum likelihood estimator.

3. The case of non-convex ρ. If ρ is not convex, the estimators T_n will, in general, no longer converge toward some constant c. Apparently, one has to impose not only local but also some global conditions on ρ in order to have consistency. Compare also Wald (1949).

Asymptotic normality is easier to handle. The following is a simple proof of asymptotic normality under somewhat too stringent regularity conditions, but without assuming monotonicity of ψ.

LEMMA 5. *Assume that* (i) $T_n(x) \to 0$ *in probability*; (ii) ψ *is continuous and has a uniformly continuous derivative* ψ'; (iii) $E\psi^2 < \infty$; (iv) $0 < E\psi' < \infty$. *Then* $n^{\frac{1}{2}}T_n$ *is asymptotically normal, with asymptotic mean* 0 *and asymptotic variance* $E\psi^2/(E\psi')^2$.

PROOF. T_n satisfies $\sum\psi(x_i - T_n) = 0$, hence, by the mean value theorem, there is some ϑ, $0 \leqq \vartheta \leqq 1$, such that $\sum\psi(x_i) - T_n\sum\psi'(x_i - \vartheta T_n) = 0$, or $n^{\frac{1}{2}}T_n = n^{-\frac{1}{2}}\sum\psi(x_i)/(n^{-1}\sum\psi'(x_i - \vartheta T_n))$. The numerator is asymptotically normal with mean 0 and variance $E\psi^2$, the denominator tends in probability toward the constant $E\psi'$, hence $n^{\frac{1}{2}}T_n$ is asymptotically normal with mean and variance as given above (see Cramér (1946), 20.6).

4. Examples. In all examples below we shall choose the origin such that $E\psi(x) = 0$, that is, $T_n \to 0$.

(i) $\rho(t) = t^2$. The corresponding estimator is the sample mean $T_n = \sum x_i/n$, $T_n \to Ex = 0$, and $n^{\frac{1}{2}}T_n$ is asymptotically normal with mean 0 and variance $Ex^2 = \int t^2 F(dt)$.

(ii) $\rho(t) = |t|$. The corresponding estimator is the sample median, and $T_n \to \mu = 0$, where μ is the median of F. If F has a non-zero derivative at the origin, then $n^{\frac{1}{2}}T_n$ is asymptotically normal with asymptotic mean 0 and asymptotic variance $1/(2F'(0))^2$.

(iii) $\rho(t) = \frac{1}{2}t^2$ for $|t| \leqq k$, $\rho(t) = k|t| - \frac{1}{2}k^2$ for $|t| > k$. The corresponding estimator is closely related to Winsorizing, since it has the following property: if we put for short $t_0 = T_n(x)$, then all observations x_i such that $|x_i - t_0| > k$ may be replaced by $t_0 \pm k$, whichever is nearer, without changing the estimate $T_n(x)$, and t_0 equals the arithmetic mean of the modified set of observations. *Proof*: since ρ is convex, it suffices to look at the first derivative of $Q(\xi) = \sum \rho(x_i - \xi)$, which still vanishes at $\xi = t_0$ if it did before the change:

$$Q'(t_0) = \sum_{|x_i - t_0| \leqq k} (x_i - t_0) + \sum_{x_i > t_0 + k} ((t_0 + k) - t_0) + \sum_{x_i < t_0 - k} ((t_0 - k) - t_0).$$

The asymptotic variance of $n^{\frac{1}{2}}T_n$ is

$$V = \left(\int_{-k}^{+k} t^2 F(dT) + k^2 \int_{-\infty}^{-k} F(dt) + k^2 \int_{k}^{\infty} F(dt)\right) \Big/ \left(\int_{-k}^{+k} F(dt)\right)^2.$$

(iv) $\rho(t) = \frac{1}{2}t^2$ for $|t| \leqq k$, $\rho(t) = \frac{1}{2}k^2$ for $|t| > k$. The corresponding estimator is a trimmed mean: let $t_0 = T_n(x)$, and assume for simplicity that for no observation $|x_i - t_0| = k$. Then $T_n(x)$ equals the sample mean of those observations for which $|x_i - t_0| < k$, and remains unchanged if some or all of the x_i for which $|x_i - t_0| > k$ are removed. *Proof*: Compute the derivative of $Q(\xi) = \sum \rho(x_i - \xi)$ at t_0

$$Q'(t_0) = \sum_{|x_i - t_0| < k} (x_i - t_0).$$

Thus, if $Q(\xi)$ attained its infimum at t_0 before the outliers were removed, it must still be stationary there when the outliers are excluded. But removing an outlier decreases $Q(t_0)$ by $\frac{1}{2}k^2$ and cannot decrease $Q(\xi)$ by more than that for any ξ, hence Q still attains its infimum at t_0.

Assume that T_n is consistent $T_n \to 0$, then the asymptotic variance of $n^{\frac{1}{2}}T_n$ is, formally,

$$V = \int_{-k}^{+k} t^2 F(dt) \Big/ \left(\int_{-k}^{+k} F(dt) - kF'(k) - kF'(-k)\right)^2.$$

(However, one should be reminded that we did not prove the formula in this case, ρ not being convex. I conjecture that it is valid if and only if the expression inside the parentheses of the denominator is strictly positive.)

A glance at the asymptotic variances computed in these four cases shows plainly that

(i) The sample mean is very sensitive to the tails of F.

(ii) The sample median is very sensitive to the behavior of F at its median, and neglects its behavior elsewhere.

(iii) "Winsorizing" avoids these shortcomings and seems to be practically foolproof. Apparently, this is connected with the fact that the corresponding ψ is monotone, bounded and absolutely continuous.

(iv) The "trimming procedure" is rather sensitive to the behavior of F at the rejection points $\pm k$; a high density at these points will play havoc with the estimate. This shortcoming seems to be common also to other rejection procedures; here one might avoid it by smoothing ρ at $\pm k$.

5. Minimax questions; a special case. At first glance, it might seem absurd to look for asymptotic minimax solutions, since, asymptotically, one could do better by estimating the true underlying distribution (see also Stein (1956), Hájek (1962)). However, this seems to require an exorbitant number of observations. On the other hand, it is hoped—and preliminary numerical experiments seem to substantiate this hope—that (M)-estimators approach their asymptotic behavior rather fast, provided ψ is bounded. So an asymptotic minimax theory should be useful in those frequent cases where the sample size is perhaps large enough to indicate deviations from the assumed model but not yet large enough to establish their nature. In the case of the contaminated normal distribution $F = (1 - \epsilon)\Phi + \epsilon H$, this means that asymptotic minimax theory would be appropriate whenever the sample size n is fairly large, but ϵn, the average number of outliers, is still rather small. We shall discuss this point in a later section.

First we shall treat a special case that can be solved explicitly by a direct verification of the saddlepoint property.

Let C be the set of all distributions of the form $F = (1 - \epsilon)G + \epsilon H$, where $0 \leqq \epsilon < 1$ is a fixed number, G is a fixed and H is a variable distribution function. Assume that G has a convex support and a twice continuously differentiable density g such that $-\log g$ is convex on the support of G. Let T_n be an (M)-estimator belonging to a certain ρ, let $\psi = \rho'$ be the derivative of ρ, and let $c = c(F)$ be such that $\int \psi(t - c)F(dt) = 0$. The asymptotic variance of $n^{\frac{1}{2}}(T_n - c)$ will be

$$V(\psi, F) = E_F\psi^2(t - c)/(E_F\psi'(t - c))^2,$$

provided ψ is "nice." For the moment we shall not bother about c—it might be interpreted as the bias of T_n—and shall try to minimize the supremum $\sup_F V(\psi, F)$ of the asymptotic variance only for those pairs (ψ, F) for which $c(F) = 0$.

THEOREM 1. *The asymptotic variance $V(\psi, F)$ has a saddlepoint: there is an $F_0 = (1 - \epsilon)G + \epsilon H_0$ and a ψ_0 such that*

$$\sup_F V(\psi_0, F) = V(\psi_0, F_0) = \inf_\psi V(\psi, F_0)$$

where F ranges over those distributions in C for which $E_F\psi_0 = 0$. Let $t_0 < t_1$ be the

endpoints of the interval where $|g'/g| \leqq k$ *(either or both of these endpoints may be at infinity), and* k *is related to* ϵ *by*

$$(1 - \epsilon)^{-1} = \int_{t_0}^{t_1} g(t)\, dt + (g(t_0) + g(t_1))/k.$$

Then the density f_0 *of* F_0 *is given explicitly by*

$$\begin{aligned} f_0(t) &= (1 - \epsilon)g(t_0)e^{k(t-t_0)} && \text{for } t \leqq t_0, \\ &= (1 - \epsilon)g(t) && \text{for } t_0 < t < t_1, \\ &= (1 - \epsilon)g(t_1)e^{-k(t-t_1)} && \text{for } t \geqq t_1. \end{aligned}$$

$\psi_0 = -f_0'/f_0$ *is monotone and bounded and corresponds to the maximum likelihood estimator of the location parameter when* F_0 *is the underlying distribution.*

REMARK. The statement of this theorem is unsatisfactory insofar as the class over which H ranges depends on ψ_0. This could be avoided by restricting G to be symmetric, and letting H range over all symmetric distributions. However, I preferred to give the stronger statement above.

PROOF. It is easy to check that F_0 has total mass 1, hence also $H_0 = (F_0 - (1 - \epsilon)G)/\epsilon$ has total mass 1, and it remains to check that its density h_0 is nonnegative. But

$$\begin{aligned} \epsilon h_0(t) &= (1 - \epsilon)[g(t_0)e^{k(t-t_0)} - g(t)] && \text{for } t \leqq t_0, \\ &= 0 && \text{for } t_0 < t < t_1, \\ &= (1 - \epsilon)[g(t_1)e^{-k(t-t_1)} - g(t)] && \text{for } t \geqq t_1. \end{aligned}$$

Because the function $-\log g(t)$ is convex, it lies above its tangents at t_0 and t_1, i.e., $-\log g(t) \geqq -\log g(t_0) - k(t - t_0)$, thus $g(t) \leqq g(t_0)e^{k(t-t_0)}$, etc., which implies non-negativity of h_0.

$\psi_0 = -f_0'/f_0$ is bounded and monotone, so we have

$$V(\psi_0, F_0) = \frac{E_{F_0}\psi_0^2}{(E_{F_0}\psi_0')^2} = \frac{(1 - \epsilon)E_G\psi_0^2 + \epsilon k^2}{((1 - \epsilon)E_G\psi_0')^2}.$$

The right side is an obvious upper bound for $V(\psi_0, F)$, provided $E_F\psi_0 = 0$, so we have $V(\psi_0, F) \leqq V(\psi_0, F_0)$.

Note that λ_F has left and right derivatives, so the asymptotic distribution of $n^{\frac{1}{2}} T_n$ must be either normal or pieced together from two half normal distributions, and the latter case can only occur if F puts positive mass on the points t_0, t_1.

The inequality $V(\psi_0, F_0) \leqq V(\psi, F_0)$ follows directly from inequality (1) of Section 2, noticing that $V(\psi_0, F_0) = (\int (f_0'/f_0)^2 f_0\, dt)^{-1}$.

More generally, using results of LeCam (1953), (1958), one may prove that the maximum likelihood estimator for F_0 is indeed efficient among all translation

invariant estimators, in the sense that for any translation invariant measurable estimator $T_n = T_n(x_1, \cdots, x_n)$ one has

$$V(\psi_0, F_0) \leqq \lim_{c\to\infty} \lim\inf_{n\to\infty} \int \min(nT_n^2, c^2) F_0(dx_1) \cdots F_0(dx_n).$$

In other words, one may strengthen the theorem: the (M)-estimator corresponding to ψ_0 minimizes the maximal asymptotic variance not only among (M)-estimators, but even among all translation invariant estimators. (Instead of translation invariance, any weaker condition still excluding superefficient estimators would suffice in this context.)

6. The contaminated normal distribution. The assumptions of the preceding section are satisfied if $G = \Phi$ is the standard normal cumulative with density $\varphi(t) = (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}t^2\}$. Because of the importance of this case, we summarize the results in a

COROLLARY. *Define the estimator* $T_n = T_n(x_1, \cdots, x_n)$ *by the property that it minimizes* $\sum_i \rho(x_i - T_n)$, *where* $\rho(t) = \frac{1}{2}t^2$ *for* $|t| < k$, $\rho(t) = k|t| - \frac{1}{2}k^2$ *for* $|t| \geqq k$. *Let* $\psi = \rho'$. *For symmetric* H, $n^{\frac{1}{2}}T_n$ *is asymptotically normal with asymptotic mean* 0, *and attains the maximal asymptotic variance* $\sup_F V(\psi, F) = \{(1-\epsilon)E_\Phi\psi^2 + \epsilon k^2\}/\{(1-\epsilon)E_\Phi\psi'\}^2$, *whenever* H *puts all its mass outside the interval* $[-k, +k]$. *If* ϵ *and* k *are related by* $(1-\epsilon)^{-1} = \int_{-k}^{+k} \varphi(t)\,dt + 2\varphi(k)/k$, *then* T_n *minimizes the maximal asymptotic variance among all translation invariant estimators, the maximum being taken over the set of all symmetric* ϵ*-contaminated distributions* $F = (1-\epsilon)\Phi + \epsilon H$. *There is a unique asymptotically least favorable* F_0 *having the density* $f_0(t) = (1-\epsilon)(2\pi)^{-\frac{1}{2}}e^{-\rho(t)}$

Table I gives the value of $\sup_F V(\psi, F)$ in dependence of k and ϵ. Notice that $E_\Phi\psi' = P_\Phi[|x| < k]$ is the probability that an observation falls inside the interval $(-k, +k)$ if Φ is the true underlying distribution. $\epsilon_{\min}$ is the value of ϵ for which the corresponding k yields the asymptotic minimax solution.

It can be seen from the table that the performance of T_n is not very sensitive to the choice of k—any value between 1.0 and 2.0 will give acceptable results for all $\epsilon \leqq 0.2$. The limiting case $k = 0$ is the sample median.

7. The question of bias. Assume that we have the situation of Section 5, the prototype distribution being $F = (1-\epsilon)G + \epsilon H$, with a symmetric G. What happens if the contaminating distribution H is not symmetric? (Suggested interpretation: most observations are distributed according to $G(x - \vartheta)$, but a small fraction ϵ of them is distributed according to $H(x)$, and is unrelated to the parameter ϑ to be estimated. This might happen if the observations are made with an instrument that jams with probability ϵ.)

Let ψ be any of the minimax procedures described in Theorem 1, or any other sufficiently regular bounded monotone skew symmetric function. Let $c = c(F)$ be such that $\lambda_F(c) = E_F\psi(t - c) = 0$; for reasons of symmetry, we have $c(G) = 0$. Thus, $c(F)$ may be considered as the bias caused by the contamination ϵH.

The mean value theorem implies that $\lambda_F(0) + c(F)\lambda_F'(\vartheta c) = 0$ for some ϑ, $0 < \vartheta < 1$, if λ_F is differentiable. Hence

$$c(F) = -\epsilon\lambda_H(0)/[(1-\epsilon)\lambda_G'(\vartheta c) + \epsilon\lambda_H'(\vartheta c)].$$

Thus, since ψ and λ are monotone,

$$|c(F)| \leqq [\epsilon/(1-\epsilon)] \sup |\psi|/(-\lambda_G'(\vartheta c)). \tag{2}$$

If $\lambda_G'(c)$ is continuous at $c = 0$, and if ϵ and hence c are small, we may replace ϑc by 0 to obtain an approximation to the right side of (2).

The upper bound of the asymptotic variance calculated above for symmetric H will have to be increased if also asymmetric H are admitted. But under mild regularity conditions this correction will be of the order $o(c)$ and may thus be neglected in comparison with the bias c, if the latter is small.

In the particular case where $G = \Phi$ is the standard normal cumulative, and ψ is the minimax solution for ϵ (minimizing the maximal asymptotic variance for symmetric H), we obtain the following approximate upper bounds for the asymptotic bias c:

$$|c| \leqq 0.17 \qquad \text{for } \epsilon = 0.1, \quad k = 1.14,$$

$$|c| \leqq 0.021 \qquad \text{for } \epsilon = 0.01, \quad k = 1.95,$$

$$|c| \leqq 0.0027 \qquad \text{for } \epsilon = 0.001, \quad k = 2.64.$$

These bounds will be reached by c whenever H puts all its mass to the right of $k + c$, or to the left of $-k - c$.

It does not seem that one can make the bias much smaller without doing serious damage to the variance of the estimate. The sample median gives the smallest bias, namely $|c| \leqq \epsilon(1-\epsilon)^{-1}(2\varphi(0))^{-1} \approx 1.25\epsilon(1-\epsilon)^{-1}$. It is easy to see that this is indeed the best possible bound for translation invariant estimators: let

$$f(t) = (1-\epsilon)\varphi(0) \qquad \text{for } |t| < c,$$

$$= (1-\epsilon)\varphi(|t| - c) \qquad \text{for } |t| \geqq c,$$

f and ϵ being related by $\int f(t)\,dt = 1$, i.e., $c = \epsilon(1-\epsilon)^{-1}(2\varphi(0))^{-1}$. Then both $c(t-c)$ and $f(t+c)$ may be considered as ϵ-contaminated standard normal densities, one having the contamination to the right, and the other to the left of the origin. No translation invariant estimator can have a bias of absolute value less than c for both of these densities simultaneously, since the difference of the two biases must be $2c$.

As a consequence of these considerations, if somebody wants to control the asymptotic bias and keep it below $\frac{1}{2}n^{-\frac{1}{2}}$, that is, below one half of the asymptotic standard deviation of T_n, he should not increase the sample size above $n = 9$, 600 and 35,000, respectively, for the above-mentioned values of ϵ and k.

Of course, such sample sizes are not large enough to estimate H to a sufficient

TABLE I

Upper bounds for the asymptotic variance of $n^{\frac{1}{2}}T_n$ for $F = (1 - \epsilon)\Phi + \epsilon H$, H symmetric

k	$E_\Phi\psi'$	$E_\Phi\psi^2$	$\epsilon_{\min}$	$\epsilon = 0$	0.001	0.002	0.005	0.010	0.020	0.050	0.100	0.200	0.500
0.0	0.0000	0.0000	1.0000	1.571	1.574	1.577	1.587	1.603	1.636	1.741	1.939	2.454	6.283
0.1	0.0797	0.0095	0.8753	1.492	1.495	1.498	1.508	1.523	1.556	1.658	1.853	2.358	6.137
0.2	0.1585	0.0358	0.7542	1.423	1.426	1.429	1.438	1.454	1.485	1.586	1.778	2.276	6.030
0.3	0.2358	0.0758	0.6401	1.362	1.365	1.368	1.377	1.393	1.424	1.524	1.714	2.209	5.962
0.4	0.3108	0.1265	0.5354	1.309	1.312	1.315	1.324	1.339	1.370	1.470	1.659	2.154	5.930
0.5	0.3829	0.1851	0.4417	1.263	1.266	1.269	1.278	1.293	1.324	1.423	1.613	2.111	5.935
0.6	0.4515	0.2491	0.3599	1.222	1.225	1.228	1.237	1.252	1.284	1.384	1.576	2.079	5.976
0.7	0.5161	0.3160	0.2899	1.187	1.190	1.193	1.202	1.217	1.249	1.351	1.546	2.058	6.053
0.8	0.5763	0.3840	0.2311	1.156	1.159	1.162	1.172	1.188	1.220	1.324	1.523	2.047	6.166
0.9	0.6319	0.4511	0.1825	1.130	1.133	1.136	1.146	1.162	1.195	1.302	1.506	2.046	6.317
1.0	0.6827	0.5161	0.1428	1.107	1.111	1.114	1.124	1.140	1.175	1.284	1.495	2.055	6.506
1.1	0.7287	0.5777	0.1109	1.088	1.091	1.095	1.105	1.122	1.158	1.272	1.490	2.072	6.734
1.2	0.7699	0.6352	0.0855	1.072	1.075	1.079	1.089	1.107	1.144	1.263	1.491	2.099	7.003
1.3	0.8064	0.6880	0.0655	1.058	1.062	1.065	1.077	1.095	1.134	1.258	1.497	2.135	7.314
1.4	0.8385	0.7358	0.0498	1.047	1.050	1.054	1.066	1.086	1.126	1.256	1.507	2.180	7.669
1.5	0.8664	0.7785	0.0376	1.037	1.041	1.045	1.057	1.078	1.121	1.258	1.522	2.233	8.069
1.6	0.8904	0.8160	0.0282	1.029	1.034	1.038	1.051	1.073	1.118	1.262	1.542	2.296	8.517
1.7	0.9109	0.8487	0.0211	1.023	1.027	1.032	1.046	1.069	1.116	1.270	1.567	2.367	9.012
1.8	0.9281	0.8767	0.0156	1.018	1.023	1.027	1.042	1.066	1.117	1.280	1.595	2.448	9.558
1.9	0.9426	0.9006	0.0115	1.014	1.019	1.024	1.039	1.065	1.119	1.292	1.628	2.537	10.154

2.0	0.9545	0.9205	0.0084	1.010	1.016	1.021	1.038	1.065	1.122	1.307	1.665	2.635	10.802
2.1	0.9643	0.9371	0.0061	1.008	1.014	1.019	1.037	1.066	1.137	1.324	1.705	2.742	11.501
2.2	0.9722	0.9507	0.0044	1.006	1.012	1.018	1.037	1.068	1.133	1.343	1.750	2.858	12.253
2.3	0.9786	0.9617	0.0032	1.004	1.011	1.017	1.037	1.071	1.140	1.363	1.798	2.982	13.058
2.4	0.9836	0.9705	0.0023	1.003	1.010	1.017	1.038	1.074	1.148	1.386	1.850	3.115	13.914
2.5	0.9876	0.9776	0.0016	1.002	1.010	1.017	1.040	1.078	1.156	1.410	1.905	3.255	14.821
2.6	0.9907	0.9831	0.0011	1.002	1.010	1.018	1.042	1.082	1.166	1.436	1.963	3.405	15.779
2.7	0.9931	0.9873	0.0008	1.001	1.010	1.018	1.044	1.087	1.176	1.463	2.025	3.562	16.787
2.8	0.9949	0.9906	0.0005	1.001	1.010	1.019	1.046	1.092	1.186	1.492	2.090	3.726	17.843
2.9	0.9963	0.9931	0.0004	1.001	1.010	1.020	1.048	1.097	1.198	1.523	2.158	3.899	18.947
3.0	0.9973	0.9950	0.0003	1.000	1.011	1.021	1.051	1.103	1.209	1.554	2.229	4.078	20.098

T_n is defined by $\Sigma\psi(x_i - T_n) = 0$, where $\psi(t) = t$ for $|t| < k$, $\psi(t) = k$ sign (t) for $|t| \geqq k$.

degree of accuracy, since, on the average, less than $\epsilon n = 1$, 6 or 35 observations, respectively, would come from H, and one would not even know which ones. This may be taken as a justification for using minimax procedures, in the sense that sample sizes reasonable for a given amount of unknown contamination will not allow to determine the nature of this contamination, except in rather extreme cases.

8. Minimax theory. The most robust estimator constructed in Section 5 coincided with the maximum likelihood estimator for some least favorable distribution. The present section shows that this is quite a general feature of most robust (M)-estimators.

Consider the following game against Nature:

Nature chooses a distribution F from some set C of distributions on the real line.

The statistician chooses a function ψ from some set Ψ of functions.

The payoff to the statistician is

$$K(\psi, F) = \left(\int \psi' \, dF\right)^2 \Big/ \int \psi^2 \, dF$$

where ψ' denotes the derivative (in measure) of ψ; it shall be tacitly understood that the statistician chooses not only ψ, but also a particular representative of ψ'.

On purpose, we shall remain rather vague about the set Ψ and shall change it according to convenience. For instance, if F has an absolutely continuous density f, the payoff function may be transformed by partial integration into $K(\psi, F) = (\int \psi f' dt)^2 / \int \psi^2 f \, dt$, which makes sense for a broader class of functions than the original definition and allows to use a larger class Ψ.

If C and Ψ are restricted in a suitable way, for instance to symmetric distributions and sufficiently regular skew symmetric monotone functions ψ, we will have $K(\psi, F) =$ (asymptotic variance of $n^{\frac{1}{2}}T_n)^{-1}$, hence $K(\psi, F)$ is a measure of the asymptotic efficiency of the estimator. For the moment, we shall not bother about this aspect and shall just try to establish properties of the abstractly defined game.

The reason for preferring the utility function $K(\psi, F)$ over the loss function $V(\psi, F) = K(\psi, F)^{-1}$ used in earlier sections is that K has more convenient convexity properties.

The following convexity inequality is hardly new, but since it will be used several times, it is given the status of a lemma.

LEMMA 6. *Let* $v_1 > 0$, $v_2 > 0$, $0 \leqq \alpha \leqq 1$. *Then*

$$\frac{(\alpha u_1 + (1 - \alpha)u_2)^2}{\alpha v_1 + (1 - \alpha)v_2} \leqq \alpha \frac{u_1^2}{v_1} + (1 - \alpha)\frac{u_2^2}{v_2}.$$

PROOF. Put $v = \alpha v_1 + (1 - \alpha)v_2$, $\beta = \alpha v_1/v$. Then

$$\frac{(\alpha u_1 + (1 - \alpha)u_2)^2}{\alpha v_1 + (1 - \alpha)v_2} = v\left\{\beta\frac{u_1}{v_1} + (1 - \beta)\frac{u_2}{v_2}\right\}^2$$

$$\leqq v\left\{\beta\left(\frac{u_1}{v_1}\right)^2 + (1 - \beta)\left(\frac{u_2}{v_2}\right)^2\right\} = \alpha\frac{u_1^2}{v_1} + (1 - \alpha)\frac{u_2^2}{v_2}.$$

In particular, if we put $u_i = \int\psi'\,dF_i$, $v_i = \int\psi^2\,dF_i$, $(i = 1, 2)$, we obtain that $K(\psi, \cdot)$ is a convex function: $K(\psi, \alpha F_1 + (1 - \alpha)F_2) \leqq \alpha K(\psi, F_1) + (1 - \alpha)K(\psi, F_2)$. Hence, if μ is a mixed strategy of Nature, that is, a probability distribution on C, it follows that $K(\psi, F_\mu) \leqq \int K(\psi, F)\,d\mu$, where $F_\mu = \int F\,d\mu$. In other words, every mixed strategy of Nature is dominated by a pure one, provided C is convex in the sense that it contains all averages F_μ of elements of C. However, we shall not use this result, so we do not give it a formal proof.

THEOREM 2. *Let C be a convex set of distribution functions such that every $F \in C$ has an absolutely continuous density f satisfying $I(F) = \int(f'/f)^2 f\,dt < \infty$.* (i) *If there is an $F_0 \in C$ such that $I(F_0) \leqq I(F)$ for all $F \in C$, and if $\psi_0 = -f_0'/f_0$ is contained in Ψ, then (ψ_0, F_0) is a saddlepoint of the game, that is,*

$$K(\psi, F_0) \leqq K(\psi_0, F_0) = I(F_0) \leqq K(\psi_0, F)$$

for all $\psi \in \Psi$ and all $F \in C$. (ii) *Conversely, if (ψ_0, F_0) is a saddlepoint and Ψ contains a nonzero multiple of $-f_0'/f_0$, then $I(F_0) \leqq I(F)$ for all $F \in C$, F_0 is uniquely determined, and ψ_0 is $[F_0]$-equivalent to a multiple $-f_0'/f_0$.* (iii) *Necessary and sufficient for F_0 to minimize $I(F)$ is that $\int(-2\psi_0 g' - \psi_0^2 g)\,dt \geqq 0$ for all functions $g = f_1 - f_0$, $F_1 \in C$.*

PROOF. The inequality $K(\psi, F_0) = (\int\psi(f_0'/f_0)f_0\,dt)^2/\int\psi^2 f_0\,dt \leqq \int(f_0'/f_0)^2 f_0\,dt = I(F_0) = K(\psi_0, F_0)$ is the Schwarz inequality (1) and holds independently of the particular properties of F_0. Assume now that $I(F_0) \leqq I(F)$ for all $F \in C$. Put for short $F_\epsilon = (1 - \epsilon)F_0 + \epsilon F_1$, with $F_1 \in C$, and let $J(\epsilon) = K(\psi_0, F_\epsilon)$, $I(\epsilon) = I(F_\epsilon)$. Both I and J are convex functions of ϵ; this follows immediately from Lemma 6. Thus, in order to prove $K(\psi_0, F_0) \leqq K(\psi_0, F)$ it suffices to show that for all $F_1 \in C$ we have $J'(0) \geqq 0$. An explicit calculation yields

$$J'(0) = \int(-2\psi_0 g' - \psi_0^2 g)\,dt,$$

with $g = f_1 - f_0$. On the other hand, we have also

$$I'(0) = \int(-2\psi_0 g' - \psi_0^2 g)\,dt,$$

provided we are allowed to interchange differentiation and integration. But for fixed t, $(f_\epsilon'(t))^2/f_\epsilon(t)$ is convex in ϵ by Lemma 6, hence the difference quotient satisfies the inequalities

$$-2\psi_0 g' - \psi_0^2 g \leqq \{(f_\epsilon')^2/f_\epsilon - (f_0')^2/f_0\}/\epsilon \leqq (f_1')^2/f_1 - (f_0')^2/f_0.$$

The right-hand term is integrable by assumption, the middle term tends decreasingly toward the left-hand term as $\epsilon \downarrow 0$, and has a nonnegative integral. One concludes by the monotone convergence theorem that the interchange of integration and differentiation is legitimate, and that the integral of the left-hand term is $I'(0) \geqq 0$. Hence, we have also $J'(0) \geqq 0$. Conversely, if F_0 is such that $\int(-2\psi_0 g' - \psi_0^2 g)\,dt \geqq 0$ for all $F_1 \,\varepsilon\, C$, then $I'(0) \geqq 0$, and F_0 minimizes $I(F)$. This proves (i) and (iii).

If (ψ_0, F_0) is a saddlepoint, then

$$I(F_0) = K(-f_0'/f_0, F_0) \leqq K(\psi_0, F_0) \leqq K(\psi_0, F) \leqq K(-f'/f, F) = I(F),$$

for all $F \,\varepsilon\, C$, provided $-f_0'/f_0$ (or a nonzero multiple of it) is in Ψ. This proves the first part of (ii).

Since $I(F)$ is a convex function of F, the set of those F that minimize $I(F)$ is convex. Assume that $F_1 \neq F_0$ also minimizes $I(F)$. Without loss of generality, we may assume that F_0 and F_1 are absolutely continuous with respect to each other (otherwise replace them, say, by $0.9F_0 + 0.1F_1$ and $0.1F_0 + 0.9F_1$, respectively, which are still distinct and still minimize $I(F)$). Let $\psi_i = -f_i'/f_i$, $(i = 0, 1)$, then both pairs (ψ_i, F_i) have the saddlepoint property, hence

$$K(\psi_1, F_0) \leqq K(\psi_0, F_0) \leqq K(\psi_0, F_1) \leqq K(\psi_1, F_1) \leqq K(\psi_1, F_0),$$

and we have equality signs throughout. But the first equality can only hold if $\psi_1 = -f_1'/f_1$ is $[F_0]$-equivalent to a multiple of $-f_0'/f_0$. Integrating this relation, we obtain $f_1(t) = a(f_0(t))^p$ for some constants a and p. Let $f = \frac{1}{2}f_0 + \frac{1}{2}f_1$, then, for the same reason, f must satisfy $f = \frac{1}{2}f_0 + \frac{1}{2}af_0^p = bf_0^q$ for some constants b and q. Taking the logarithmic derivative, we obtain

$$\frac{f'}{f} = \frac{(1 + apf_0^{p-1})f_0'}{(1 + af_0^{p-1})f_0} = q\frac{f_0'}{f_0}.$$

On the set where $f_0'/f_0 \neq 0$, we have $1 - q = a(q - p)f_0^{p-1}$. Since f_0 is not constant on this set, this implies $p = q = 1$. Hence f_0 and f_1 must be equal. This terminates the proof.

Theorem 2 does not answer the question whether such an F_0 exists, and it does not tell us what happens for distributions F that do not possess an absolutely continuous density. To settle these questions, we have to generalize a little bit further.

From now on, let C be a vaguely compact convex set of substochastic distributions on the real line. (Under the vague topology we understand the weakest topology such that the maps $F \to \int \psi\, dF$ are continuous for all continuous functions ψ with compact support.) We no longer require that the elements of C have a density, and we do not exclude the possibility that some might have a total mass <1. This latter case is not so far-fetched as it might seem. It corresponds loosely to the usual practice of screening the data and excluding extremely wild observations from further processing: If we formalize this by assuming that

Nature chooses a probability distribution on the extended real line and that the statistician excludes infinite-valued observations, we end up with substochastic distributions on the ordinary real line. Theorem 2 remains valid for substochastic distributions.

On the other hand, vague compactness would be unduly restrictive if we did not admit substochastic distributions.

First we shall extend the definition of the Fisher information number, essentially by putting it equal to ∞ whenever the classical expression $\int (f'/f)^2 f\,dt$ does not work. However, a more devious approach is more useful, where the above statement turns up as the conclusion of a theorem.

DEFINITION. *Let*

$$I(F) = \sup_\psi \left(\int \psi'\,dF \right)^2 \Big/ \int \psi^2\,dF$$

where ψ ranges over the set of continuous and continuously differentiable functions with compact support satisfying $\int \psi^2\,dF \neq 0$.

THEOREM 3. *$I(F) < \infty$ if and only if F has an absolutely continuous density f such that $\int (f'/f)^2 f\,dt < \infty$, f' being the derivative in measure of f, and then $I(F) = \int (f'/f)^2 f\,dt$.*

PROOF. If F has an absolutely continuous density f, then we may conclude by the Schwarz inequality (1) that $I(F) \leqq \int (f'/f)^2 f\,dt$, which proves the "if" part of the theorem. The "only if" part is more involved. Assume that $I(F) < \infty$.

(i) First we have to show that F has a density f. Assume that it has not, then there would be a set N of Lebesgue measure $\lambda(N) = 0$ and of F-measure $F(N) > 0$. The idea now is to approximate the indicator function of N by a continuous ψ' such that ψ has a compact support. The actual construction of ψ, ψ' is somewhat tedious: let $\eta > 0$ be given. Let U be an open set containing N such that $\lambda(U) < \eta F(N)$. Let V be a compact set contained in U such that $F(V) > (1 - \eta)F(N)$, and let V_1 be an open subset of U containing V, having a compact closure. Then let g be a continuous function, $0 \leqq g \leqq 1$, equal to 1 on V, to 0 outside V_1 and let ψ' be a continuous function with compact support, equal to g on V_1, and satisfying $-\eta F(N) \leqq \psi' \leqq 0$ outside V_1, such that $\int \psi'\,dt = 0$. Let ψ be the indefinite integral of ψ', normed such that it vanishes outside a compact, then $|\psi| < \eta F(N)$ and $\int \psi'\,dF > (1 - 2\eta)F(N)$; hence $K(\psi, F) > ((1 - 2\eta)/\eta)^2$ can be made arbitrarily large. This leads to a contradiction, and thus F must have a density.

For the following parts of the proof we shall freely use measurable functions for ψ and ψ', leaving it understood that one should approximate them by continuous functions with compact support in a similar way as above, in order to complete the proofs.

(ii) Now we show that f is essentially bounded (modulo Lebesgue measure). If this would not be the case, the sets $A_m = \{t \mid f(t) \geqq m\}$ would have strictly positive Lebesgue measure $\lambda(A_m) > 0$ for all m. Let ψ' be the indicator function of A_m, then $\int \psi' f\,dt \geqq m\lambda(A_m)$. We may choose ψ such that $|\psi| \leqq \lambda(A_m)$, then

$\int \psi^2 f\, dt \leqq (\lambda(A_m))^2$, thus $K(\psi, F) \geqq m^2$ can be made arbitrarily large; this leads to a contradiction.

(iii) Then we show that f can be chosen to be continuous. If not, then there would be three numbers t_0, c, and ϵ such that for all $\delta > 0$, the sets $A_+ = \{t \mid f(t) > c + \epsilon, |t - t_0| < \delta\}$ and $A_- = \{t \mid f(t) < c - \epsilon, |t - t_0| < \delta\}$ would have strictly positive Lebesgue measure simultaneously. Replace A_+ and A_- by smaller sets B_+ and B_-, respectively, such that $\lambda(B_+) = \lambda(B_-) = p > 0$. Put $\psi'(t) = +1$ on B_+, $= -1$ on B_-, $= 0$ elsewhere, and norm its indefinite integral ψ such that it vanishes outside the interval $[t_0 - \delta, t_0 + \delta]$. Then $|\psi| \leqq p$. Since f is essentially bounded, say by M, it follows that $\int \psi^2 f\, dt \leqq 2p^2 M\delta$, $\int \psi' f\, dt = \int \psi'(f - c)\, dt \geqq 2p\epsilon$. Thus $K(\psi, F) \geqq 4\epsilon^2/(2M\delta)$, which can be made arbitrarily large; this leads to a contradiction.

(iv) Thus we can choose a continuous version of f; we have to show that it is absolutely continuous. Assume that it is not, then there would be a $c > 0$ such that for every $\epsilon > 0$ there exists a finite collection (a_i, b_i) of disjoint intervals, of total length $< \epsilon$, such that $|\sum_i f(b_i) - f(a_i)| > c$. Now let ψ' be defined by

$$\begin{aligned} \psi'(t) &= \delta^{-1} \quad \text{if} \quad |t - b_i| < \tfrac{1}{2}\delta \quad \text{for some} \quad i, \\ &= -\delta^{-1} \quad \text{if} \quad |t - a_i| < \tfrac{1}{2}\delta \quad \text{for some} \quad i, \\ &= 0 \quad \text{elsewhere;} \end{aligned}$$

$\delta > 0$ has to be chosen so small that this definition makes sense. We may choose the indefinite integral ψ of ψ' such that $|\psi| \leqq 1$, and that $\psi = 0$ outside the union of the intervals $(a_i - \frac{1}{2}\delta, b_i + \frac{1}{2}\delta)$, which will have a total length $< \epsilon$ if δ is chosen sufficiently small. Then $\int \psi^2 f\, dt \leqq \epsilon M$, and $|\int \psi' f\, dt| > c$ by continuity of f, if δ is chosen sufficiently small. Hence $K(\psi, F) \geqq c^2/(\epsilon M)$ can be made arbitrarily large, which leads to a contradiction.

(v) In a way very similar to (iv) one shows that f is of bounded variation. This implies in particular that $\int |f'|\, dt < \infty$.

(vi) By the Schwarz inequality, we have

$$K(\psi, F) = \left(\int \psi f'\, dt\right)^2 \Big/ \int \psi^2 f\, dt \leqq \int (f'/f)^2 f\, dt.$$

Now approximate $-f'/f$ by a continuous, continuously differentiable function ψ with compact support in the sense that the integrals $\int |(-f'/f) - \psi|^2 f\, dt$ and $\int |(-f'/f) - \psi|\, |f'|\, dt$ are made arbitrarily small. Then both $\int \psi^2 f\, dt$ and $\int \psi f' dt$ approach $\int (f'/f)^2 f\, dt$, which shows that $\sup_\psi K(\psi, F) = \int (f'/f)^2 f\, dt$. This terminates the proof.

THEOREM 4. *Assume that C is vaguely compact and convex. If $\inf_{F \in C} I(F) = a < \infty$, then there exists a unique $F_0 \in C$ such that $I(F_0) = a$.*

PROOF. For $b > a$, the set $K_b = \{F \in C \mid I(F) \leqq b\}$ is nonempty and convex. Being the intersection of the vaguely closed sets $K_{b,\psi} = \{F \in C \mid (\int \psi'\, dF)^2 \leqq b \int \psi^2\, dF\}$, it is vaguely closed and hence compact. By the finite intersection property, $K_a = \bigcap_{b > a} K_b$ is nonempty, convex and compact, and Theorem 2 implies that it is reduced to a single element F_0.

THEOREM 5. *Assume that C is vaguely compact and convex, and that the set $C_1 = \{F \in C \mid I(F) < \infty\}$ is dense in C. Let F_0 be the unique distribution minimizing $I(F)$. If F_0 happens to be such that $\psi_0 = -f_0'/f_0$ is in Ψ, is differentiable in measure and ψ_0' is either continuous or nonnegative upper semi-continuous, and vanishes at infinity, then (ψ_0, F_0) is not only a saddlepoint with respect to C_1, but also with respect to C.*

PROOF. Assume that $F \in C$, and let (F_ν) be a net of elements of C_1 converging toward F. We have to prove that $K(\psi_0, F) \geqq a$. If ψ_0' is continuous and vanishes at infinity, then $\int \psi_0' \, dF_\nu \to \int \psi_0' \, dF$. If $\psi_0' \geqq 0$ is upper semi-continuous and vanishes at infinity, then $\int \psi_0' \, dF = \inf_\chi \int \chi dF = \inf_\chi \limsup_\nu \int \chi \, dF_\nu \geqq \limsup_\nu \int \psi_0' \, dF$, where χ ranges over the continuous functions $\chi \geqq \psi_0'$ vanishing at infinity. On the other hand, we have

$$\int \psi_0^2 \, dF = \sup_\chi \int \chi \, dF = \sup_\chi \liminf_\nu \int \chi \, dF_\nu \leqq \liminf_\nu \int \psi_0^2 \, dF_\nu,$$

where χ ranges over the set of continuous functions $0 \leqq \chi \leqq \psi_0^2$ vanishing at infinity. It follows that $K(\psi_0, F) \geqq \limsup_\nu K(\psi_0, F_\nu) \geqq a$.

REMARK. This theorem solves the abstract game; however, for the original statistical problem some further considerations are needed. Let $\lambda_F(\xi) = \int \psi_0(t - \xi) \, dF$. According to Lemma 4, the asymptotic variance of $n^{\frac{1}{2}} T_n$ is $V(\psi_0, F) = \int \psi_0^2 \, dF / (\lambda_F'(0))^2$, provided ψ_0 is bounded and monotone increasing, $\lambda_F(0) = 0$ and $\lambda_F'(0) < 0$ exists. If ψ_0' is uniformly continuous, then $-\lambda_F'(0) = \int \psi_0' \, dF$, thus $V(\psi_0, F) = K(\psi_0, F)^{-1}$ and (ψ_0, F_0) is also a saddlepoint of $V(\psi, F)$ if it is one of $K(\psi, F)$ and if $\lambda_F(0) = 0$ for all $F \in C$. If ψ_0' is discontinuous, then $-\lambda_F'(0) = \int \psi_0' \, dF$ still holds for $F \in C_1$. Under the assumptions (i) ψ_0 is continuous, bounded and monotone increasing, (ii) for all $F \in C$, $\lambda_F(0) = 0$ and $\lambda_F'(0) < 0$ exists, (iii) for every $F \in C$, there is a net $F_\nu \to F$, $F_\nu \in C_1$, such that $\lim \int \psi_0' \, dF_\nu \leqq -\lambda_F'(0)$, we may still conclude that (ψ_0, F_0) is a saddlepoint of $V(\psi, F)$. The proof is a simplified version of that of Theorem 5.

9. The minimax solution for ϵ-normal distributions. We say that a distribution F is ϵ-normal with approximate mean ξ and approximate variance σ^2 if it satisfies $\sup_t |F(t) - \Phi((t - \xi)/\sigma)| \leqq \epsilon$. Of course, ξ and σ are in general not uniquely determined, but the indeterminacy is small for small ϵ. We shall now determine the saddlepoint of the game, when Nature chooses an F from the set C of symmetric ϵ-standard-normal distributions:

$$C = \{F \mid \sup_t |F(t) - \Phi(t)| \leqq \epsilon, F(t) + F(-t) = 1\}.$$

This set C is vaguely closed in the set of all substochastic distributions, hence it is compact; it is convex and $C_1 = \{F \in C \mid I(F) < \infty\}$ is dense in C. So the theory of the preceding section can be applied.

The saddlepoint (ψ_0, F_0) was found by some kind of enlightened guesswork, using variational techniques; instead of going through the cumbersome heuristic arguments, I shall state the solution explicitly and shall then verify it.

The solution is valid only for values of ϵ between 0 and (approximately) 0.03, but it seems that this range is just the range important in practice.

F_0 has the symmetric density

$$\begin{aligned} f_0(t) = f_0(-t) &= \varphi(a)(\cos \tfrac{1}{2}ca)^{-2}(\cos \tfrac{1}{2}ct)^2 && \text{for } 0 \leqq t \leqq a \\ &= \varphi(t) && \text{for } a < t < b \\ &= \varphi(b)e^{-b(t-b)} && \text{for } t \geqq b \end{aligned}$$

where $\varphi(t) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}t^2}$, and a, b, c are three constants determined through the relations

(i) $c \tan(\frac{1}{2}ca) = a \qquad (0 \leqq ca < \pi)$,
(ii) $\int_0^a f_0\, dt = \int_0^a \varphi\, dt - \epsilon$,
(iii) $\int_b^\infty f_0\, dt = \int_b^\infty \varphi\, dt + \epsilon$.

It is easy to check that the F_0 thus defined belongs to C, provided (i), (ii), and (iii) can be satisfied with $0 \leqq a \leqq b$; this is the case for values of ϵ between 0 and (approximately) 0.03. For numerical results, see Table II.

Putting $\psi_0 = -f_0'/f_0$, we have

$$\begin{aligned} \psi_0(t) = -\psi_0(-t) &= c \tan \tfrac{1}{2}ct, && \text{for } 0 \leqq t \leqq a \\ &= t && \text{for } a < t < b \\ &= b && \text{for } t \geqq b, \end{aligned}$$

which is continuous and has a piecewise continuous bounded derivative.

We have to prove that (ψ_0, F_0) is a saddlepoint. In view of Theorem 2, we have to show that $J'(0) = \int(-2\psi_0 g' - \psi_0^2 g)\, dt \geqq 0$ for all functions $g = f_1 - f_0$, $F_1 \varepsilon C_1$. By partial integration, the condition becomes

$$J'(0) = \int (2\psi_0' - \psi_0^2)g\, dt \geqq 0 \qquad \text{for all } g = f_1 - f_0.$$

Let $\hat{g}(t) = \int_0^t g(s)\, ds = F_1(t) - F_0(t)$, then $F_1 \varepsilon C$ implies the necessary feasibility condition $\hat{g}(t) \geqq 0$ for $a \leqq t \leqq b$, since in this interval $F_0(t) = \Phi(t) - \epsilon$.

We have

$$J'(0) = \int_0^a c^2 g\, dt + \int_a^b (2 - t^2)g\, dt + \int_b^\infty (-b)^2 g\, dt,$$

TABLE II

ϵ	a	b	c
0.001	0.65	2.44	1.37
0.002	0.75	2.23	1.35
0.005	0.91	1.95	1.32
0.01	1.06	1.72	1.29
0.02	1.24	1.49	1.26
0.03	1.34	1.36	1.23

thus, if we transform the middle integral by partial integration,

$$J'(0) = c^2\hat{g}(a) + (2 - b^2)\hat{g}(b) - (2 - a^2)\hat{g}(a) + \int_a^b 2t\hat{g}(t)\,dt$$
$$+ b^2\hat{g}(b) + b^2(1 - F_1(\infty)) = (c^2 + a^2 - 2)\hat{g}(a)$$
$$+ 2\hat{g}(b) + \int_a^b 2t\hat{g}(t)\,dt + b^2(1 - F_1(\infty)).$$

We shall show presently that $c^2 + a^2 - 2 \geqq 0$, hence all summands are nonnegative and it follows that $J'(0) \geqq 0$. To prove that remaining inequality, notice that in the interval $(0, a)$, ψ_0' increases monotonely from $\psi_0'(0) = \frac{1}{2}c^2$ to $\psi_0'(a) = \frac{1}{2}(c^2 + a^2)$. Since $\int_0^a \psi_0'\,dt = \psi_0(a) = a$, it follows from the mean value theorem that $\psi_0'(t) = 1$ for some $0 \leqq t \leqq a$, hence $\psi_0'(a) = \frac{1}{2}(c^2 + a^2) \geqq 1$, which establishes the inequality in question.

This proves the saddlepoint property for distributions from C_1. To prove it for the whole of C, one has to check the points (i), (ii) and (iii) of the remark at the end of Section 8. (i) and (ii) are immediate. Instead of doing (iii), it is probably easier to go back to Theorem 5, to show that $-\lambda_F'(0) = \int \psi_0'\,dF$ unless F puts a positive mass on $\{\pm a, \pm b\}$; and that in this latter case one may decrease $-\lambda_F'(0)$ by shifting this mass away from $\{\pm a, \pm b\}$.

10. Estimation of a scale parameter. The theory of estimating a scale parameter is less satisfactory than that of estimating a location parameter. Perhaps the main source of trouble is that there is no natural "canonical" parameter to be estimated. In the case of a location parameter, it was convenient to restrict attention to symmetric distributions; then there is a natural location parameter, namely the location of the center of symmetry, and we could separate difficulties by optimizing the estimator for symmetric distributions (where we know what we are estimating) and then investigate the properties of this optimal estimator for nonstandard conditions, e.g., for nonsymmetric distributions. In the case of scale parameters, we meet, typically, highly asymmetric distributions, and the above device to ensure unicity of the parameter to be estimated fails. Moreover, it becomes questionable, whether one should minimize bias or variance of the estimator.

So we shall just go ahead and shall construct estimators that are invariant under scale transformations and that estimate their own asymptotic values as accurately as possible. Of course, one has to check afterward in a few typical cases what these estimators really do estimate.

The problem of estimating a scale parameter for the random variable X can be reduced to that of estimating a location parameter for the random variable $Y = \log X^2$. This amounts to estimating the parameter $\tau = \log \sigma^2$, where σ is a scale parameter for X. The change of parameters here is made for technical reasons, but it might also be justified on purely philosophical grounds. Compare also Tukey (1960), especially p. 461 ff.

We shall again be concerned with the contaminated normal case, that is, we shall assume that the distribution of Y is of the form $F = (1 - \epsilon)G + \epsilon H$, where G is the distribution of Y corresponding to a standard normal distribution for X, ϵ is a known number and H is an unknown contaminating distribution.

If the distribution of X is the standard normal, then the distribution function of $Y = \log X^2$ is $G(u) = P[Y < u] = P[X^2 < e^u] = \Phi(e^{(\frac{1}{2})u}) - \Phi(-e^{(\frac{1}{2})u})$, having the density $g(u) = G'(u) = (2\pi)^{-(\frac{1}{2})} \exp(-\frac{1}{2}e^u + \frac{1}{2}u)$.

Then, $-\log g$ is a convex function, and $-g'(u)/g(u) = \frac{1}{2}(e^u - 1)$ is monotone and differentiable. Hence, we may apply the theory of Section 5, especially Theorem 1. To avoid confusion later on, we replace the letters ψ, k of Theorem 1 by χ, c and define

$$\begin{aligned} \chi_0(u) &= \tfrac{1}{2}(e^u - 1) && \text{for } \tfrac{1}{2}|e^u - 1| < c, \\ &= c \operatorname{sign}(e^u - 1) && \text{for } \tfrac{1}{2}|e^u - 1| \geqq c. \end{aligned}$$

Then, if c is chosen appropriately, the estimator T_n defined by $\sum_i \chi_0(y_i - T_n) = 0$ will minimize the maximal asymptotic variance among all estimators of $\tau = \log \sigma^2$ that are invariant under a change of scale of the x_i (resulting in a translation of the $y_i = \log x_i^2$), the maximum being taken over those H for which $E_F\chi_0 = 0$.

The minimax solution shows a different behavior for $c < \frac{1}{2}$ than for $c \geqq \frac{1}{2}$; for the following, we shall assume that $c \geqq \frac{1}{2}$; thus, with $q^2 = 2c + 1 \geqq 2$,

$$\begin{aligned} \chi_0(u) &= \tfrac{1}{2}(e^u - 1) && \text{for } e^u < q^2 \\ &= \tfrac{1}{2}(q^2 - 1) && \text{for } e^u \geqq q^2. \end{aligned}$$

Unfortunately, the condition $E_F\chi_0 = 0$ now becomes rather restrictive: it means that only those H are admitted for competition in $\sup_F V(\chi_0, F)$ which put all their mass on the set $[|x| > q]$, and for this class of distributions we have identically $V(\chi_0, F) = \sup_F V(\chi_0, F)$. Nevertheless, it seems intuitively plausible that an estimator behaving satisfactorily against this type of contamination should not fare too badly against other types.

The parameters q and ϵ are related by

$$(1 - \epsilon)^{-1} = \int_{-\infty}^{t_1} g(t)\, dt + g(t_1)/c$$

where $t_1 = \log q^2$. Hence

$$(1 - \epsilon)^{-1} = \int_{-q}^{+q} \varphi(t)\, dt + 2q(q^2 - 1)^{-1}\varphi(q). \tag{3}$$

It will be convenient to express the results again in terms of the x_i and of the estimate $S_n = e^{\frac{1}{2}T_n}$ of the scale parameter of the x_i. If we introduce the function ψ_0 (already used in Section 6), displaying explicitly the parameter q occurring in it

$$\begin{aligned} \psi_0(q, t) &= t && \text{for } |t| < q \\ &= q \operatorname{sign}(t) && \text{for } |t| \geqq q, \end{aligned} \tag{4}$$

then S_n satisfies

$$\sum_{i=1}^{n} \psi_0^2(q, x_i/S_n) = n,$$

as an elementary calculation shows, and may be defined by this equation.

Summarizing, we may describe the properties of S_n as follows. Let C be the class of all distributions of the form $F = (1 - \epsilon)\phi + \epsilon H$, where ϵ is given by (3), and H puts all its mass outside the interval $[-q, +q]$. The problem is to estimate the variance σ^2 of the normal component of F, where F has been obtained by a scale transformation from some (unknown) element of C. Then, $\log S_n^2$ is an estimate of $\log \sigma^2$, invariant under scale transformations, asymptotically unbiased and asymptotically normal; it minimizes the maximal asymptotic variance among all scale invariant estimators of $\log \sigma^2$, and has a constant asymptotic variance over the whole of C.

In other words, S_n is gauged for the class C and the class of distributions that can be generated from C by scale transformations and has there about the most pleasant asymptotic properties one can reasonably expect. The flaw is that S_n may have quite a considerable bias for distributions not in C, for instance for the normal distribution Φ itself.

In most cases it will be preferable to gauge the bias such that it is zero at Φ. In order to do this, let S_∞ be the asymptotic value of S_n under Φ; then define $S_n' = S_n/S_\infty$. Obviously, S_n' has the same pleasant properties as S_n, except that it is biased for the class C, but asymptotically unbiased for Φ.

We have, by the definition of S_n,

$$1 = n^{-1} \sum \psi_0^2(q, x_i/S_n) = n^{-1} S_\infty^{-2} \sum \psi_0^2(qS_\infty, x_i/S_n'),$$

and, going over to the limit,

$$1 = E_\Phi \psi_0^2(q, t/S_\infty) = S_\infty^{-2} E_\Phi \psi_0^2(qS_\infty, t).$$

Put $q' = qS_\infty$, then S_n' can be defined by the equation

$$n^{-1} \sum_i \psi_0^2(q', x_i/S_n') = E_\Phi \psi_0^2(q', t). \tag{5}$$

Table III gives some numerical results. $\epsilon_{\min}$ is the value of ϵ for which the corresponding q' yields the asymptotic minimax solution.

11. Estimation of a location parameter, if scale and ϵ are unknown. Now consider the problem of estimating a location parameter if neither the scale

TABLE III

q'	S_∞	$q = q'/S_\infty$	$\epsilon_{\min}$
1.1	0.760	1.45	0.182
1.5	0.882	1.70	0.074
2.0	0.960	2.08	0.019
2.5	0.989	2.53	0.0038
3.0	0.997	3.01	0.0006

σ of the normal component nor the amount ϵ of contamination are known. We are now not interested in estimating scale or ϵ, but only in estimating location by some estimator T_n, and in addition, in estimating the (asymptotic) variance of T_n.

The following are three heuristic proposals on how one can apply the theory of (M)-estimators to this problem. I conjecture that the third proposal is asymptotically minimax, in the sense that independently of scale σ and ϵ, the supremum over symmetric H of the asymptotic variance of T_n has the least possible value. The first two proposals are asymptotically minimax if k is chosen in relation to the true ϵ.

All three proposals are asymptotically equivalent to Winsorizing; they differ among each other in the determination of the number of observations to be Winsorized. From now on I shall omit the indices from ψ_0, T_n and S_n; $\psi = \psi_0$ is defined by (4) in the preceding section.

PROPOSAL 1. Choose beforehand a fixed number k (either by intuition or by using previous experience). Then determine T and S such that

$$n^{-1}\sum \psi(k, (x_i - T)/S) = 0$$
$$n^{-1}\sum \psi'(k, (x_i - T)/S) = E_\Phi \psi'(k, x).$$

The second equation can only be fulfilled approximately, since $\sum \psi'(k, (x_i - T)/S)$ is an integer, namely the number of observations contained in the interval $[T - kS, T + kS]$. Hence this proposal is similar (and asymptotically equivalent) to Winsorizing a fixed, predetermined number of observations; it is asymmetric in the sense that it tends to Winsorize more observations on the side with the heavier contamination. k might be compared to a kind of insurance against ill effects from contamination: if one insures against high contamination by choosing a small k, one has to pay by losing some efficiency if the actual contamination is low, and vice versa; compare also Anscombe (1960), especially p. 127.

PROPOSAL 2. Choose beforehand a fixed number k. Then determine T and S such that

$$n^{-1}\sum \psi(k, (x_i - T)/S) = 0$$
$$n^{-1}\sum \psi^2(k, (x_i - T)/S) = E_\Phi \psi^2(k, x).$$

This proposal is of course suggested by the most robust estimator of scale as determined in Equation (5) of the preceding section. It corresponds to Winsorizing a variable number of observations: slightly more if F has heavier tails, slightly less if F has lighter tails, and if H is asymmetric, more on the side with heavier contamination. Certainly for large n, but probably also for small n, this proposal is better than the preceding one, since it is less sensitive to a "wrong" choice of k (i.e., to a choice not adapted to the true value of ϵ). Moreover, the S of Proposal 2 will be more accurate in general ($\log S$ having the smaller asymptotic variance), which also improves accuracy of T.

For both proposals, T and S can be determined relatively easily by iterative procedures (see below), and the (asymptotic) variance of T might be estimated by

$$\frac{n}{n-1}\frac{\sum\psi^2(k,(x_i-T)/S)}{[\sum\psi'(k,(x_i-T)/S)]^2}S^2.$$

The correction term $n/(n-1)$ is suggested by the classical theory ($k=\infty$); one will have to revise it in the light of later knowledge, but it seems to agree quite well with exploratory Monte Carlo computations for sample sizes 5 and 10.

PROPOSAL 3. Determine k such that the estimated (asymptotic) variance of T is minimized, i.e., minimize

$$\sum\psi^2(k,x_i-T)/[\sum\psi'(k,x_i-T)]^2$$

subject to the side condition (determining T)

$$\sum\psi(k,x_i-T)=0.$$

To avoid trouble at $k=0$, one should safeguard by requiring that enough observations (say more than $n^{\frac{1}{2}}$) are contained in the interval $[T-k,T+k]$. This proposal seems to give the best asymptotic properties of all three, but numerical computation of k and T apparently is a very difficult task.

For practical applications, I personally would favor Proposal 2, since it seems to fit best into the framework of conventional least squares techniques. For nonlinear least squares problems, which have to be solved by iterative methods anyway, this proposal probably could be incorporated into existing computer programs with an only marginal increase of the amount of computation.

First we have to show that Proposal 2 constitutes a legitimate definition of T and S; that is, we have to show existence and uniqueness of the solution.

Let $x_1,\cdots,x_n$ be a fixed sample of size n and consider the equations

$$\begin{aligned}\sum_{i=1}^{n}\psi(k,(x_i-T)/S)&=0\\ \sum_{i=1}^{n}\psi^2(k,(x_i-T)/S)&=n\beta.\end{aligned}\tag{6}$$

Proposal 2 corresponds to $\beta=E_\Phi\psi^2(k,x)$, and for this choice we have, obviously, $\beta<k^2$ whenever $k>0$. I do not want to exclude the possibility that small sample size considerations might lead to a different choice of β.

PROPOSITION. *Assume that $\beta<k^2$ and that the sample is such that $x_{i_1}=x_{i_2}=\cdots=x_{i_m}$, $i_1<i_2<\cdots<i_m$ implies $m<n(1-\beta/k^2)$. Then the system* (6) *has a unique solution T, S.*

PROOF. First we remark that for any pair T, S satisfying the second equation of (6), the number m_2 of observations satisfying $|(x_i-T)/S|<k$ must be greater than or equal to $n(1-\beta/k^2)$, since $(n-m_2)k^2\leqq\sum\psi^2(k,(x_i-T)/S)$. Since ties of high multiplicity between the x_i are forbidden, it follows that the

right side of this last inequality will exceed $n\beta$ for small S. Continuity of ψ now implies that for every t there is an $S_t > 0$ such that $\sum \psi^2(k, (x_i - t)/S_t) = n\beta$. Since $m_2 \geqq n(1 - \beta/k^2)$, this S_t is uniquely determined, and is a continuous function of t; for large t it behaves asymptotically like $\beta^{-\frac{1}{2}}|t|$. Hence, $n^{-1}\sum \psi(k, (x_i - t)/S_t)$ is a continuous function of t, tending to $\mp\beta^{\frac{1}{2}}$ for $t \to \pm\infty$. Thus, by continuity, there will be a T such that $\sum \psi(k, (x_i - T)/S_T) = 0$, which proves the existence of a solution.

This solution T, S is unique: Assume that T_1, S_1 and T_2, S_2 are two distinct solutions. If $S_1 = S_2$, we would necessarily have $T_1 = T_2$, since $\sum \psi(k, (x_i - t)/S)$ is strictly monotone in t so long as $m_2 \geqq 1$. Hence we may assume $S_1 \neq S_2$; put $c = (T_2 - T_1)/(S_2 - S_1)$. Let S vary from S_1 to S_2, and put $T(S) = T_1 + c(S - S_1)$. Consider the function

$$R(S) = \tfrac{1}{2}\sum(\psi^2(k, (x_i - T(S))/S) - \beta) + c\sum \psi(k, (x_i - T(S))/S)$$

which vanishes at S_1 and S_2 : $R(S_1) = R(S_2) = 0$. A straightforward calculation shows that the derivative of this function is

$$R'(S) = -S^{-1}\sum{}'(c + (x_i - T(S))/S)^2,$$

where the primed summation sign denotes that the summation is extended only over those indices for which $|(x_i - T(S))/S| < k$. Thus, $R'(S)$ has the same sign for $S_1 \leqq S \leqq S_2$, and since $R(S_1) = R(S_2) = 0$, it must identically vanish. This implies $cS + x_i - T(S) = 0$ or $x_i = T_1 - cS_1$ for all i such that $|(x_i - T(S))/S| < k$. In particular, this holds for $S = S_1$, hence there would be a tie of multiplicity $m_2 \geqq n(1 - \beta/k^2)$ between the x_i; but this has expressly been excluded. This terminates the proof.

Let m_1, m_2, m_3 be the number of observations satisfying $x_i \leqq T - kS$, $T - kS < x_i < T + kS$ and $T + kS \leqq x_i$, respectively. Then (6) may be written

$$\begin{aligned} \sum{}'x_i - m_2T + (m_3 - m_1)kS &= 0 \\ \sum{}'(x_i - T)^2 + (m_1 + m_3)k^2S^2 - n\beta S^2 &= 0. \end{aligned}$$

By determining T from the first equation and inserting it in the second, one obtains the equivalent system

$$(6') \qquad \begin{aligned} T &= \bar{x}' + kS(m_3 - m_1)/m_2 \\ S^2 &= \sum{}'(x_i - \bar{x}')^2/(n\beta - (m_1 + m_3 + (m_3 - m_1)^2/m_2)k^2) \end{aligned}$$

where $\bar{x}' = m_2^{-1}\sum{}' x_i$ is the trimmed mean.

This may be used to compute T and S by iterations: start with some initial values T_0, S_0. Compute

$$\begin{aligned} \bar{x}' &= m_2^{-1}\sum{}' x_i \\ S_1^2 &= \sum{}'(x_i - \bar{x}')^2/(n\beta - (m_1 + m_3 + (m_3 - m_1)^2/m_2)k^2) \\ T_1 &= \bar{x}' + kS_1(m_3 - m_1)/m_2, \end{aligned}$$

where $\sum'$ denotes that the summation is extended over those indices for which $|(x_i - T_0)/S_0| < k$, and m_1, m_2, m_3 are the numbers of observations satisfying $x_i \leqq T_0 - kS_0$, $T_0 - kS_0 < x_i < T_0 + kS_0$ and $T_0 + kS_0 \leqq x_i$, respectively. If $T_1 = T_0$ and $S_1 = S_0$, one has found the solution, otherwise do the same again, with T_1, S_1 in place of T_0, S_0, etc. For automatic computation, it might be wise to safeguard against too small values of m_2 and of the denominator in the expression defining S_1^2.

Since, for a fixed sample, the values of T_1 and S_1 are uniquely determined by the pair of integers m_1, m_3, the process must either stop after a finite number of steps, or it will repeat itself periodically. I do not know whether and when the latter case occurs.

Ordinarily, the procedure converges rather fast, as an exploratory study with Monte Carlo methods shows. For $k = 1.5$, sample sizes up to 100 and distributions that do not have too heavy tails, the stationary value is on the average reached after 1-2 steps, starting from sample mean and sample variance. For distributions with heavier tails (Cauchy), somewhat more steps are needed (about 2 for sample size 10, 4 for sample size 100).

Table IV gives sharp upper bounds for the asymptotic variance of $n^{\frac{1}{2}}T$, if T and S are computed according to Proposal 2, and $F = (1 - \epsilon)\Phi + \epsilon H$, H symmetric. These upper bounds can be determined as follows. The asymptotic variance of $n^{\frac{1}{2}}T$ is $E_F\psi^2(kS, x)/(E_F\psi'(kS, x))^2$, where S satisfies $E_F\psi^2(kS, x) = S^2\beta(k)$, with $\beta(k) = E_\Phi\psi^2(k, x)$. Hence, we have, with $q = kS$,

$$q^2\beta(k)/k^2 = (1 - \epsilon)\beta(q) + \epsilon E_H\psi^2(q, x) \leqq (1 - \epsilon)\beta(q) + \epsilon q^2$$

or

$$\beta(q)/q^2 \geqq (\beta(k)/k^2 - \epsilon)/(1 - \epsilon). \tag{7}$$

Since $\psi^2(q, x)/q^2$ is monotone decreasing for increasing q, $\beta(q)/q^2$ is also monotone decreasing, and equality in (7) determines a sharp upper bound $q_{\max}$ for q.

TABLE IV

Sharp upper bounds for the asymptotic variance of $n^{\frac{1}{2}}T$, if T and S are computed according to Proposal 2, for symmetric contamination

	$\epsilon = 0.000$	0.010	0.020	0.050	0.100	0.200
$k = 1.0$	1.107	1.138	1.172	1.276	1.490	2.15
1.1	1.088	1.120	1.155	1.265	1.495	2.23
1.2	1.072	1.105	1.140	1.258	1.506	2.35
1.3	1.058	1.093	1.131	1.256	1.526	2.50
1.4	1.047	1.084	1.124	1.257	1.557	2.70
1.5	1.037	1.077	1.120	1.264	1.592	2.97
1.6	1.029	1.072	1.117	1.275	1.639	3.36
1.7	1.023	1.068	1.117	1.288	1.697	3.92
1.8	1.018	1.066	1.118	1.306	1.769	>4
1.9	1.014	1.065	1.121	1.324	1.859	>4
2.0	1.010	1.065	1.127	1.353	2.029	>4

Now we have to find a sharp upper bound for the asymptotic variance

$$\frac{E_F \psi^2(q, x)}{(E_F \psi'(q, x))^2} = \frac{q^2\beta(k)/k^2}{((1 - \epsilon)\alpha(q) + \epsilon E_H \psi'(q, x))^2}$$

where $\alpha(q) = E_\Phi\psi'(q, x)$. This is solved if we can find a sharp lower bound for $(1 - \epsilon)\alpha(q)/q + \epsilon E_H\psi'(q, x)/q$. But $\alpha(q)/q$ is monotone decreasing for increasing q (notice that $\alpha(q)/q$ is the average of the normal density in the interval $(0, q)$), hence this expression is bounded from below by $(1 - \epsilon)\alpha(q_{\max})/q_{\max}$. Hence it follows that the asymptotic variance of $n^{\frac{1}{2}}T$ is bounded from above by

$$\frac{q^2_{\max}\beta(k)/k^2}{((1 - \epsilon)\alpha(q_{\max}))^2} = \frac{(1 - \epsilon)\beta(q_{\max}) + \epsilon q^2_{\max}}{((1 - \epsilon)\alpha(q_{\max}))^2}.$$

This bound is sharp: it is attained for contaminating distributions H that put all their mass outside the interval $[-q_{\max}, q_{\max}]$. Hence, one may determine the upper bound $q_{\max}$ from (7) and enter Table 1 with $k = q_{\max}$ and ϵ to determine the upper bound for the variance. This was done, using linear interpolation, to obtain Table 4.

12. A generalization. (M)-estimators $T_n = T_n(x_1, \cdots, x_n)$ have been defined by the property that they minimize an expression of the form $\sum_{i=1}^n \rho(x_i - T_n)$. More generally, one may define (M_r)-estimators by the property that they minimize an expression of the form $\sum_I \rho_r(x_I - T_n)$ where ρ_r is a symmetric function of r arguments, the summation is extended over the $\binom{n}{r}$ subsets $I = \{i_1, \cdots, i_r\} \subset \{1, 2, \cdots, n\}$ containing r distinct elements, and $(x_I - T_n)$ stands short for $(x_{i_1} - T_n, \cdots, x_{i_r} - T_n)$.

Let $\psi_r(x_I - T) = -(\partial/\partial T)\rho_r(x_I - T)$, and assume that $E\psi_r = 0$. Then define $\psi(x_1) = \int\psi_r(x_1, \cdots, x_r)F(dx_2) \cdots F(dx_r)$.

It follows from Hoeffding's work on U-statistics (1948), that the suitably normed sum $\sum_I\psi_r(x_I - T)$ behaves asymptotically like $\sum_i \psi(x_i - T)$. In particular, one shows in much the same way as in Section 2, and under analogous regularity conditions, that $T_n \to 0$ in probability, and that $n^{\frac{1}{2}}T_n$ is asymptotically normal, with asymptotic mean 0 and asymptotic variance $V(\psi_r, F) = E\psi^2/(E\psi')^2$.

Since the amount of computation needed for determining the value of an (M_r)-estimator increases with the rth power of the sample size, they will hardly be of practical use for $r > 2$.

Example. Take $r = 2$, and $\rho_2(t_1, t_2) = |t_1 + t_2|$. It is easy to see that $\sum_I \rho_2(x_I - T)$ is minimized by $T(x) = \{$sample median of the $z_{ij} = \frac{1}{2}(x_i + x_j)$, $i < j\}$, i.e., by the estimator proposed by Hodges and Lehmann (1963). If the underlying distribution has a symmetric density f, we end up with an asymptotic variance $V(\psi_2, F) = 1/(12(\int f^2\,dt)^2)$. In the contaminated normal case, $F = (1 - \epsilon)\Phi + \epsilon H$, H symmetric, we have the sharp upper bound $V(\psi_2, F) <$

$(1 - \epsilon)^{-4}\pi/3$ for this particular ψ_2 ; the bound is approached when H spreads its mass toward infinity. Numerically:

ϵ =	0.000	0.001	0.002	0.005	0.010	0.020	0.050	0.100	0.200	0.500
$V \leqq$	1.047	1.051	1.056	1.068	1.090	1.135	1.286	1.596	2.557	16.76

The reader is invited to compare this with Tables 1 and 4; apparently, the Hodges-Lehmann estimator and the estimators proposed in the preceding sections of the present paper are close competitors.

For the conventional model of contamination $F(t) = (1 - \epsilon)\Phi(t) + \epsilon\Phi(t/3)$, these estimators have practically equivalent performances also:

	ϵ = 0.0	0.01	0.02	0.05	0.10
Hodges-Lehmann:	V = 1.047	1.07	1.09	1.17	1.31
Proposal 2, k = 1.5:	V = 1.037	1.06	1.08	1.16	1.30

I would like to express my thanks for a number of stimulating discussions with Professor E. L. Lehmann on robust estimation and with Professor L. LeCam on asymptotic efficiency.

REFERENCES

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

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

HÁJEK, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist.* **33** 1124–1147.

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

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.* **1** 277–330.

LECAM, L. (1958). Les propriétés asymptotiques des solutions de Bayes. *Publications de l'Institut de Statistique de l'Université de Paris.* **7** 17–35.

LOÈVE, M. (1960). *Probability Theory*, (2nd ed.). Van Nostrand, New York.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. and Prob.* **I** 187–195.

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

WALD, A. (1949). Note on the consistency of the M. L. estimate. *Ann. Math. Statist.* **20** 595–601.