BOUNDS FOR ESTIMATES OF LOCATION

by

Edward E. Leamer

Discussion Paper Number 105
10/77

Bounds for Estimates of Location

by Edward E. Leamer*
Department of Economics
University of California at Los Angeles

The traditional theory of point estimation describes a horse race among alternative estimators. Although this race does identify clear losers, it does not, in fact cannot, identify a clear winner. As a result, a practical data analyst is faced with the dilemma that his data could sensibly support many estimates. In this paper I attempt to determine the limits of this ambiguity. For each of several classes of estimators, I will report a bound for the set of estimates that may be computed from a given data set.

The first section deals with maximum likelihood estimators of the location of a normal distribution. When the variances and covariances of the observations are unknown, many estimates may be computed from a given data set by varying the covariance matrix. In fact it will be shown that <u>any</u> sample may support <u>any</u> estimate if the covariance matrix is appropriately chosen. Moreover, even stationarity is not sufficient to bound the estimates between the extreme sample observations.

The second section deals with bounds for estimates chosen to minimize the distances from the sample values to the estimate. If the distance function is monotone, the estimate must lie between the extreme sample values. If the distance function is further restricted, two types of trimming procedures may be employed to bound the estimate. The more restrictive bound makes use of the assumptions of monotonicity, symmetry, exchangeability, convexity and additivity.

---

*Conversations with Michael Ward, John Riley and Robbie Jones have contributed to this manuscript. Support from NSF grant SOC76 - 08863A01 is gratefully acknowledged. The Monte Carlo study reported in Section 3 was done by Tom Means.

Given these assumptions the estimate must lie within the range of the "folded sample," $Z_i = (X_i + X_{n-i+1})/2$, $i = 1, \ldots, n/2$, where X_i are the order statistics. This set of points is a set of weighted averages $\hat{\mu} = \Sigma X_i d_i / \Sigma d_i$ with weights $d_i > 0$ such that if $|X_i - \hat{\mu}| \geq |X_j - \hat{\mu}|$ then $d_i |X_i - \hat{\mu}| \geq d_j |X_j - \hat{\mu}|$.

The choice of point estimator within the range of the folded sample is shown in the last section not to be a matter of indifference from the standpoint of mean squared error criterion. A Monte Carlo study suggests that the median of the folded sample -- the Bickel-Hodges (1967) estimator -- is often a better estimator than the mid-point.

2.0 Bounds for Estimates of the Location of a Normal Distribution

In this section it is assumed that the error distribution is normal; thus the model can be written as

$$\underset{\sim}{x} = \underset{\sim}{1}\mu + \underset{\sim}{\varepsilon} \qquad (1)$$

where $\underset{\sim}{1}$ is an n x 1 vector of ones and $\underset{\sim}{\varepsilon}$ is normally distributed with mean zero and covariance matrix $\underset{\sim}{\Sigma}$. The maximum likelihood estimator of μ is then

$$\hat{\mu} = (\underset{\sim}{1}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{1})^{-1}(\underset{\sim}{1}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{x}).$$

In words, $\hat{\mu}$ is an average of the observations weighted by the column sums of $\underset{\sim}{\Sigma}^{-1}$.

The question now to be addressed is what set of estimates $\hat{\mu}$ can be generated from a given data set $\underset{\sim}{x}$ if $\underset{\sim}{\Sigma}$ is allowed to vary with the class of symmetric positive definite matrices. Somewhat surprising, <u>any</u> value of $\hat{\mu}$ can be generated from <u>any</u> sample provided that $\underset{\sim}{x}$ is not proportional to the vector of ones. One might have guessed that $\hat{\mu}$ would lie between the minimum and maximum observations. The fact that this isn't true in general stimulates a search for restrictions on $\underset{\sim}{\Sigma}$ that are sufficient to bound $\hat{\mu}$ within this range. Of course if $\underset{\sim}{\Sigma}$ is a diagonal matrix this bound does apply. But otherwise I have been unable to find interesting restrictions on $\underset{\sim}{\Sigma}$ which imply the bound, and the fact that several conjectures, now to be discussed, are false, is more surprising to me than the fact that any $\hat{\mu}$ is possible.

The estimate $\hat{\mu}$ may lie outside the range of the data if any of the column sums of $\underset{\sim}{\Sigma}^{-1}$ is negative. A symmetric positive definite matrix $\underset{\sim}{\Sigma}^{-1}$ can rotate the vector of ones by as much as ninety degrees, and in general the column sums of $\underset{\sim}{\Sigma}^{-1}$ can be negative. Of course, if there is only one observation there can be only one estimate of μ. But if there are two or more observations <u>any</u> number

is a maximum likelihood estimate of μ, given a suitably chosen $\underset{\sim}{\Sigma}$![1] This is proved below for two observations, but obviously generalizes since any other observation can be ignored by setting appropriate elements of $\underset{\sim}{\Sigma}$ to sufficiently small numbers.

Theorem 1. There exists a symmetric positive definite matrix $\underset{\sim}{\Sigma}$ such that $\hat{\mu} = (\underset{\sim}{1}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{1})^{-1}(\underset{\sim}{1}\underset{\sim}{\Sigma}^{-1}\underset{\sim}{x})$ for any values of $\hat{\mu}$ and the vector $\underset{\sim}{x}$, provided $\underset{\sim}{x}$ is not proportional to $\underset{\sim}{1}$.

Proof. For the 2 x 2 case write $\underset{\sim}{\Sigma}$ as

$$\underset{\sim}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and

$$\underset{\sim}{\Sigma}^{-1} \propto \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Thus $\hat{\mu} = [(\sigma_2^2 - \rho\sigma_1\sigma_2)x_1 + (\sigma_1^2 - \rho\sigma_1\sigma_2)x_2]/[\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2]$. The theorem states that for any values of $\hat{\mu}$, x_1, x_2, there exist σ_1, σ_2, ρ such that the equality just reported holds. This equality can be rewritten as $0 = (\sigma_2^2 - \rho\sigma_1\sigma_2)(x_1 - \hat{\mu}) + (\sigma_1^2 - \rho\sigma_1\sigma_2)(x_2 - \hat{\mu})$ or $(\sigma_1^2 - \rho\sigma_1\sigma_2)/(\sigma_2^2 - \rho\sigma_1\sigma_2) = (x_1 - \hat{\mu})/(\hat{\mu} - x_2)$ where the term on the right can be any real number. But the term on the left can also be any real number, say c; to see this let $r = \sigma_1/\sigma_2$ and the equation can be written as $c = (r^2 - \rho r)/(1 - \rho r)$ or $r^2 + \rho(c - 1)r - c = 0$, which can be solved for ρ in terms of r as $\rho = (c - r^2)/r(c - 1)$. [It is necessary to choose r such

---

[1] This is true also for the regression problem; i.e. for any (k x 1) vector $\hat{\underset{\sim}{\beta}}$, any (n x k) matrix $\underset{\sim}{X}$ with rank k, and any (T x 1) vector $\underset{\sim}{Y}$ which is not a linear combination of the columns of $\underset{\sim}{X}$, there exists a symmetric positive definite matrix $\underset{\sim}{\Sigma}$ such that $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\Sigma}^{-1}\underset{\sim}{Y}$.

that ρ^2 is less than one: $(c - r^2)^2/r^2(c - 1)^2 < 1$; or $c^2 - 2r^2c + r^4 < r^2c^2 - 2r^2c + r^2$, or $0 < r^2 - c^2 + r^2c^2 - r^4 = r^2(1 - r^2) + c^2(r^2 - 1) = (r^2 - c^2)(1 - r^2)$. Thus r^2 must be chosen between 1 and c^2.]

Geometrically, this result is illustrated in Figure 1. If $\underset{\sim}{\Sigma}$ is the identity matrix, the vector $\underset{\sim}{x}$ is orthogonally projected onto the vector of ones to form the "predicted $\underset{\sim}{x}$," $\underset{\sim}{1}\hat{\mu} = \underset{\sim}{1}\bar{x}$. In general, $\hat{\mu}$ is selected to minimize $(\underset{\sim}{x} - \underset{\sim}{1}\hat{\mu})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x} - \underset{\sim}{1}\hat{\mu})$. The function $f(\underset{\sim}{z}) = (\underset{\sim}{x} - \underset{\sim}{z})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x} - \underset{\sim}{z})$ defines an ellipse located at $\underset{\sim}{x}$. Minimizing $f(\underset{\sim}{z})$ subject to $\underset{\sim}{z} = \underset{\sim}{1}\hat{\mu}$ involves finding a tangency between an ellipse and the 45° line. The theorem states that any point on the 45° line is a point of tangency between the line and a suitably chosen ellipse located at $\underset{\sim}{x}$.

Figure 1 illustrates a situation in which $0 < x_2 < x_1 < \hat{\mu}$. The ellipses around the vector $\underset{\sim}{x}$ have the same shape as the usual confidence ellipses and from Figure 1 it is clear that $\hat{\mu}$ is outside the range of the data only if x_1 and x_2 are positively correlated. The relative standard errors, σ_1/σ_2, can be computed by projecting an ellipse onto each of the axes. In Figure 1 the projection onto the x_1 axis is smaller than the projection onto the x_2 axis, which means that $\sigma_1 < \sigma_2$. This may now make clear why $\hat{\mu}$ exceeds x_1. Since x_1 and x_2 are positively correlated they are likely to be on the same side of μ. Since $\sigma_1 < \sigma_2$, x_1 is likely to be closer to μ than is x_2. Hence $\hat{\mu}$ should be outside of the observed points, farther from x_2 than x_1.

This example of an estimate outside the sample range requires unequal variances. If the diagonal elements of $\underset{\sim}{\Sigma}$ are equal in the 2 x 2 case, then the mean, $(x_1 + x_2)/2$, is the only estimator. This does not generalize as will be illustrated by an example below:

False Conjecture 1. If the diagonal elements of $\underset{\sim}{\Sigma}$ are all equal, then $\hat{\mu}$ must lie between the minimum and maximum sample points.

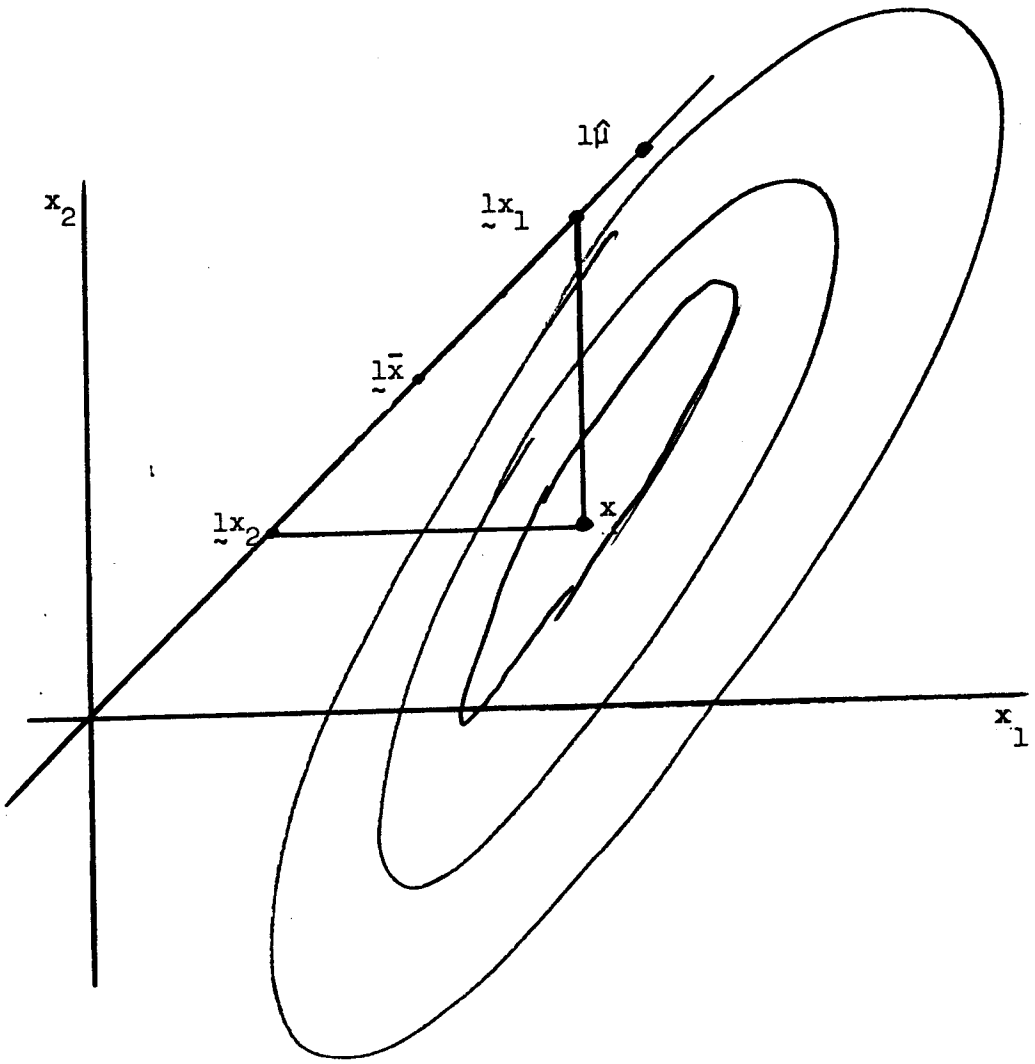

FIGURE 1 A Geometrical Illustration of Theorem 1

A stronger statement is that the process is stationary, that is that $\underset{\sim}{\Sigma}$ is a band diagonal matrix. But this too is not sufficient to restrict the $\hat{\mu}$ to the range of the observations:

False Conjecture 2. If $\underset{\sim}{\Sigma}$ is a band diagonal matrix, then $\hat{\mu}$ must lie between the minimum and maximum sample points.

Both of these conjectures can be refuted by the following example. Let

$$\underset{\sim}{\Sigma} = \begin{bmatrix} 1 & .6 & 0 \\ .6 & 1 & .6 \\ 0 & .6 & 1 \end{bmatrix}$$

$$\underset{\sim}{\Sigma}^{-1} \propto \begin{bmatrix} 1-.6^2 & -.6 & .6^2 \\ -.6 & 1 & -.6 \\ .6^2 & -.6 & 1-.6^2 \end{bmatrix}$$

$$\underset{\sim}{1}'\underset{\sim}{\Sigma}^{-1} \propto (.4, -.2, .4)$$

Suppose $x_2 = 0$, $x_1 = x_3 = 1$. Then $\hat{\mu} = 2(.4)/.6 = 4/3$.

Although stationarity is not sufficient to bound the estimate between the extreme observations, I do have two results on band diagonal matrices. First, it is well known that if $\underset{\sim}{\Sigma}$ is band diagonal the elements of $\underset{\sim}{H} = \underset{\sim}{\Sigma}^{-1}$ satisfy the symmetry conditions

$$H_{ij} = H_{ji} = H_{n+1-j,\, n+1-i} = H_{n+1-i,\, n+1-j}$$

An immediate implication of these symmetry conditions is that the weight allocated to the i^{th} observation must equal the weight allocated to the $(n + 1 - i)^{th}$ observation.

A second result, now to be presented, describes the row sums of $\underset{\sim}{\Sigma}^{-1}$ when $\underset{\sim}{\varepsilon}$ comes from an autogressive process.

Theorem 2. If the error $\underset{\sim}{\varepsilon}$ in (1) is drawn from an autoregressive process of order p

$$\varepsilon_t = \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \ldots + \phi_p\varepsilon_{t-p} + u_t$$

where u_t is a white noise normal process, and if there are enough observations n such that $n \geq 2p$, then $\hat{\mu}$ is a weighted average of the observations with weights equal to

$$(1,\ 1 - \phi_1,\ 1 - \phi_1 - \phi_2,\ \ldots\ 1 - \Sigma_i\phi_i,\ 1 - \Sigma_i\phi_i,\ \ldots,\ 1 - \phi_1 - \phi_2,\ 1 - \phi_1,\ 1).$$

Proof. The precision matrix $\underset{\sim}{H} = \underset{\sim}{\Sigma}^{-1}$ of $\underset{\sim}{\varepsilon}$ can be computed as in Galbraith and Galbraith (1974) as follows. Let $\underset{\sim}{v} = (\varepsilon_{1-p}, \varepsilon_{2-p}, \ldots, \varepsilon_{-1}, \varepsilon_o)$ and note that $\underset{\sim}{u} = (u_1, u_2, \ldots, u_n)$ can be written as

$$\underset{\sim}{u} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -\phi_1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & \\ -\phi_p & & & & \vdots \\ 0 & & & & \\ \vdots & & & & \\ 0 & \ldots & 0 & -\phi_p & 1 \end{bmatrix} \underset{\sim}{\varepsilon} + \begin{bmatrix} -\phi_p & -\phi_{p-1} & \ldots & -\phi_1 \\ 0 & -\phi_p & & -\phi_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & -\phi_p \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \underset{\sim}{v}$$

$$= \underset{\sim\sim}{A}\underset{\sim}{\varepsilon} + \underset{\sim\sim}{B}\underset{\sim}{v}$$

The random variables $\underset{\sim}{u}$ and $\underset{\sim}{v}$ can be taken to be independent with covariances $\underset{\sim}{I}$ and $\underset{\sim}{V}$ respectively. The random vector $\underset{\sim}{\varepsilon} = A^{-1}(\underset{\sim}{u} - \underset{\sim}{B}\underset{\sim}{v})$ therefore has variance $\underset{\sim}{A}^{-1}(\underset{\sim}{I} + \underset{\sim\sim\sim}{BVB'})\underset{\sim}{A}'^{-1}$ and precision

$$H = \underset{\sim}{A}'(\underset{\sim}{I} + \underset{\sim\sim\sim}{BVB'})^{-1}\underset{\sim}{A}$$

$$= \underset{\sim}{A}'\underset{\sim}{A} - \underset{\sim}{A}'\underset{\sim}{B}(\underset{\sim}{B}'\underset{\sim}{B} + \underset{\sim}{V}^{-1})^{-1}\underset{\sim}{B}'\underset{\sim}{A}.$$

But the last (n - p) columns of $\underset{\sim}{B}'\underset{\sim}{A}$ are zeroes and, regardless of $(\underset{\sim}{B}'\underset{\sim}{B} + \underset{\sim}{V}^{-1})^{-1}$, the elements of $\underset{\sim}{H}$ except for the upper left p x p block are equal to the values of $\underset{\sim}{A}'\underset{\sim}{A}$. The column sums of the last (n - p) columns, which can be computed from $\underset{\sim}{A}'\underset{\sim}{A}$, are necessarily equal to the column seems of the first n - p columns, and the whole vector of column sums is therefore computable from $\underset{\sim}{A}'\underset{\sim}{A}$ if $n \geq 2p$. The column sums of A'A are

$$\underset{\sim}{A}'\underset{\sim}{A}\underset{\sim}{1} = A' \begin{bmatrix} 1 \\ 1-\phi_1 \\ 1-\phi_1-\phi_2 \\ \vdots \\ 1-\phi_1-\phi_2-\cdots-\phi_p \\ 1-\phi_1-\phi_2-\cdots-\phi_p \\ \vdots \\ 1-\phi_1-\phi_2\cdots-\phi_p \end{bmatrix} = (1-\phi_1-\phi_2-\cdots-\phi_p) \begin{bmatrix} \vdots \\ 1-\phi_1-\phi_2-\cdots-\phi_p \\ 1-\phi_1-\phi_2-\cdots-\phi_p \\ \vdots \\ 1-\phi_1-\phi_2 \\ 1-\phi_1 \\ 1 \end{bmatrix}$$

A final result is that if the covariances are negative, then the estimate of $\hat{\mu}$ must lie within the range of the observations.

Theorem 3. If the positive definite matrix $\underset{\sim}{\Sigma}$ has only non-positive off-diagonal terms, then $\underset{\sim}{\Sigma}^{-1}\underset{\sim}{1}$ is a non-negative vector.

Proof. Scale $\underset{\sim}{\Sigma}$ so that the largest diagonal element is less than one. Write $\underset{\sim}{\Sigma}$ as $(\underset{\sim}{I} - \underset{\sim}{A})$ where $\underset{\sim}{A}$ is a positive matrix. The Hawkin- Simons conditions are satisfied, and $(\underset{\sim}{I} - \underset{\sim}{A})^{-1}$ can be written as $\underset{\sim}{I} + \underset{\sim}{A} + \underset{\sim}{A}^2 + \ldots$, which is a positive matrix.

The conclusion that seems warranted in light of the results in this section is that the set of estimates that could be supported by a given data set is surprisingly wide. Even stationarity is not sufficient to bound the estimate between the extreme observations. Independence does imply the bound, but interesting weaker conditions are hard to find.

3.0 Bounds for Robust Estimates of Location

This section deals generally with the minimization of a function of the distances from an estimate $1\hat{\mu}$ to the set of n sample values $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$. The difference between the estimate and the sample will be indicated by

$$\underset{\sim}{e}(\hat{\mu}) = (x_1 - \hat{\mu}, x_2 - \hat{\mu}, \ldots, x_n - \hat{\mu})' = \underset{\sim}{x} - 1\hat{\mu}.$$

The problem to be considered is the minimization with respect to $\underset{\sim}{e}$ of a function $W(\underset{\sim}{e};\underset{\sim}{\theta})$ subject to the constraint $\underset{\sim}{e} = \underset{\sim}{x} - 1\hat{\mu}$. The vector $\underset{\sim}{\theta}$ will index a class of distance functions, $\underset{\sim}{\theta} \in \textcircled{H}$, and the minimization just described will imply a function $\hat{\mu}(\underset{\sim}{\theta};\underset{\sim}{x})$ which is the value of $\hat{\mu}$ which minimizes $W(\underset{\sim}{e};\underset{\sim}{\theta})$, given $\underset{\sim}{x}$ and $\underset{\sim}{\theta}$. The extreme values of $\hat{\mu}(\underset{\sim}{\theta},x)$ for $\underset{\sim}{\theta} \in \textcircled{H}$ will therefore bound the estimates that may be supported by a given data set $\underset{\sim}{x}$ and a class of criterion functions $W(\underset{\sim}{e};\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \textcircled{H}$.

For example, as in the previous section, suppose the criterion function is the quadratic $W(\underset{\sim}{e};\underset{\sim}{\theta}) = \underset{\sim}{e}'\underset{\sim}{H}\underset{\sim}{e}$ where $\underset{\sim}{H}$ is a positive diagonal matrix $\underset{\sim}{H} = \text{diag}\{\theta_1, \theta_2, \ldots, \theta_n\}$, $\theta_i \geq 0$. Then of course the optimal estimate is a weighted average of the observations $\hat{\mu}(\underset{\sim}{\theta};\underset{\sim}{x}) = (\underset{\sim}{\theta}'\underset{\sim}{1})^{-1}(\underset{\sim}{\theta}'\underset{\sim}{x})$, and the extreme values of $\hat{\mu}(\underset{\sim}{\theta};\underset{\sim}{x})$ are the smallest and largest sample observations.

3.1 Properties of the Criterion Function

Several possible properties of the criterion function will now be discussed. In the special case when the criterion function is quadratic $W = \underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ these properties will restrict the matrix $\underset{\sim\sim}{H}$ in ways now to be pointed out.

(1) Monotonicity.

The function $W(\underset{\sim}{e})$ will be said to be monotone if

$$\partial W/\partial e_i \geq 0 \qquad \text{for } e_i \geq 0$$

and

$$\partial W/\partial e_i \leq 0 \qquad \text{for } e_i \leq 0.$$

In words if one of the errors e_i increases in absolute value, holding fixed the other errors, then the criterion to be minimized increases. This seems like a sensible restriction on a distance function but the quadratic form $W = \underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ need not be monotone. The vector of derivatives is then $2\underset{\sim\sim}{H}\underset{\sim}{e}$ which will have the monotonicity property if $\underset{\sim\sim}{H}$ is a positive diagonal matrix but otherwise not. It will be shown below that monotonicity is sufficient to restrict the estimate to lie within the range of the sample observations, and the failure of the maximum likelihood estimate for the location of a normal distribution to satisfy this bound is attributable to the violation of the monotonicity condition.

(2) Exchangeability

The function $W(\underset{\sim}{e})$ will be said to be exchangeable if $W(\underset{\sim}{e}) = W(\underset{\sim}{e}^*)$ where $\underset{\sim}{e}^*$ is a permutation of the vector $\underset{\sim}{e}$.

The vector of ones is an axis of symmetry for an exchangeable function. Statistically speaking, exchangeability implies that the order of the observations cannot affect the inferences. The quadratic form $\underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ is exchangeable if $\underset{\sim\sim}{H} = (\underset{\sim\sim}{I} + \underset{\sim}{1}\rho\underset{\sim}{1}')^{-1} = \underset{\sim\sim}{I} - \underset{\sim}{1}(\underset{\sim}{1}'\underset{\sim}{1} + \rho^{-1})^{-1}\underset{\sim}{1}'$, since $\underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ then depends only on the sum of the e_i and the sum of squared e_i.

(3) Additive Separability

The function $W(\underset{\sim}{e})$ will be said to be additively separable if there exists a monotone increasing function h such that

$$h(W(\underset{\sim}{e})) = \sum_i g_i(e_i)$$

Additivity may be associated with a sampling scheme which generates statistically independent observations since then the logarithm of the likelihood function is additive. The quadratic form $\underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ is additively separable if $\underset{\sim}{H}$ is a diagonal matrix.

(4) Symmetry

The function W will be said to be symmetric around the axes if

$$W(\underset{\sim}{e}) = W(|\underset{\sim}{e}|)$$

where $|\underset{\sim}{e}|$ has elements equal to the absolute value of the elements of $\underset{\sim}{e}$.

Independent sampling from a symmetric, unimodal density will imply a log-likelihood function which is additive, exchangeable and symmetric around the axes. The quadratic form $\underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$ is axis symmetric if $\underset{\sim}{H}$ is diagonal.

(5) Convexity

The function $W(\underset{\sim}{e})$ is said to be convex if for all points $\underset{\sim a}{e}$ and $\underset{\sim b}{e}$, and all c, $0 \leq c \leq 1$, $W(c\underset{\sim a}{e} + (1-c)\underset{\sim b}{e}) \leq cW(\underset{\sim a}{e}) + (1-c)W(\underset{\sim b}{e})$.

If W is twice differentiable, convexity is equivalent to the matrix of second derivatives being positive semidefinite. Estimation of the location of a normal distribution implies the criterion function $\underset{\sim}{e}'\underset{\sim\sim}{H}\underset{\sim}{e}$, which is convex. But estimating the location of a Student function does not imply a convex criterion. The negative of the logarithm of the likelihood function would be $W = \sum_i \ln(v + e_i^2)$ where v is the degrees of freedom. The second partial derivative is $\partial^2 W/\partial\mu^2 = 2(v - e_i^2)/(v + e_i^2)^2$, which is positive in the neighborhood of $e_i = 0$ but turns negative for $e_i^2 > v$.

Convexity will be shown below to have an important effect on constraining the range of supportable estimates but the property is not especially compelling. It may be pointed out that many of the criterion functions which have been used in the literature are convex. Researchers such as Forsythe (1972) and Ekblom (1974) who have used the function $\Sigma|e_i|^p$ have restricted p to exceed one. The M-estimators of Huber (1964) select $\hat{\mu}$ which is a solution to

$$\sum_i \psi(x_i - \hat{\mu}) = 0$$

where ψ is a monotone increasing function with $\psi(0) = 0$. The function ψ can be thought to be the derivative of a convex function.

3.2 Results

Three different bounds are reported in this section. The first makes use of the assumption of monotonicity, the second uses also exchangeability and the third uses all five properties. To illustrate, let X_i (i = 1, 2, 3) be the set of order statistics, and suppose that $(X_1 + X_2)/2 < X_3$ as illustrated in Figure 2. If only monotonicity is assumed, then any estimate between X_1 and X_3 is feasible. If exchangeability and monotonicity are assumed, then $\hat{\mu}$ must lie between $(X_1 + X_2)/2$ and $(X_2 + X_3)/2$. If all five properties are used, then $\hat{\mu}$ must lie between $(X_1 + X_2)/2$ and X_2.

Monotonicity

Theorem 4. If the function $W(\underset{\sim}{e})$ is monotone increasing then $\hat{\mu}$ which minimizes $W(\underset{\sim}{e}(\hat{\mu}))$ must lie between the first and last order statistics, $X_1 \leq \hat{\mu} \leq X_n$. Conversely, there exists a monotone $W(\underset{\sim}{e})$ such that any $\hat{\mu}$ in this range is a solution to the minimization problem.

Proof: Consider a point $\mu^* < X_1$. Then

$$dW/d\mu\Big|_{\mu=\mu^*} = \sum_i (\partial W/\partial e_i) < 0$$

since $\mu^* \leq X_i$ for all i. Therefore the minimum of the function must occur at a point exceeding μ^*. The converse is also simply demonstrated by letting $W(\underset{\sim}{e}) = \sum_i e_i^2/\sigma_i^2$ and choosing σ_i^2 appropriately.

Monotonicity Symmetry and Exchangeability

Definition: The anchored pairwise averages are the set of points

$$Y_{j1} = (X_1 + X_j)/2 \qquad j = 1, \ldots, n$$

$$Y_{jn} = (X_j + X_n)/2 \qquad j = 1, \ldots, n$$

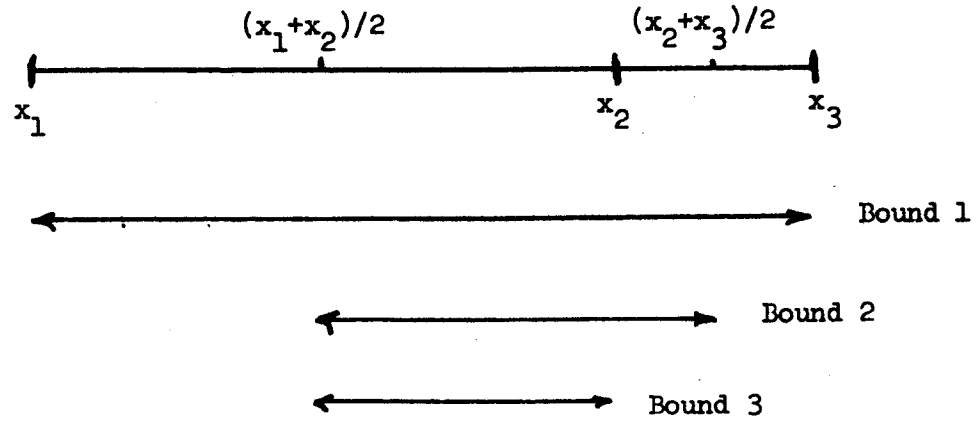

FIGURE 2 Bounds for Estimates of Location

Definition: A pairwise average anchored by the first observation, Y_{j1}, will be said to be unsupported if there is only one observation less than Y_{j1}:

$$Y_{j1} < X_k \qquad k = 2, \ldots, n$$

An average anchored by the last observation, Y_{jn}, will be said to be unsupported if

$$X_k < Y_{jn} \qquad k = 1, 2, \ldots, n-1.$$

If Y_{j1} is unsupported and if $j' > j$ then $Y_{j'1}$ is unsupported. If Y_{jn} is unsupported and $j > j'$, then $Y_{j'n}$ is unsupported.

Theorem 5. If $W(\underset{\sim}{e})$ is monotone and invariant to permutations of the element of $\underset{\sim}{e} = \{X_i - \mu\}$, and if $\underset{\sim}{e}$ has more than two elements, then μ which minimizes $W(\underset{\sim}{e}(\mu))$ must lie within the range of the "unsupported anchored averages."

Proof: Suppose that Y_{j1} is an unsupported anchored average. Then it can be shown that any location $\mu < Y_{j1}$ can be dominated by a location $\mu > Y_{j1}$. Let the new location be $\mu^* = X_j - \mu + X_1$, and let $e_i^* = X_i - \mu^*$. Notice that $e_1 = -e_j^*$ and $e_j = -e_1^*$, which by the exchangeability and symmetry assumptions, produces no change in the value of W. But for all other i we have $|e_i^*| < |e_i|$ and hence by monotonicity $W(\underset{\sim}{e}) > W(\underset{\sim}{e}^*)$. To see that $|e_i^*| < |e_i|$, note that $(e_i^*)^2 = (X_i - \mu^*)^2 = (X_i + \mu - X_j - X_1)^2 = ([X_i - \mu] - [X_j + X_1 - 2\mu])^2 = (X_1 - \mu)^2 - (X_j + X_1 - 2\mu)(2[X_i - \mu] - [X_j + X_1 - 2\mu]) = e_i^2 - 4(Y_{j1} - \mu)(X_i - Y_{j1}) < e_i^2$, since $Y_{j1} > \mu$ and $X_i > Y_{j1}$.

Monotonicity, Symmetry, Exchangeability, Convexity and Additivity

Definition. The folded sample observations are the averages of symmetrically placed order statistics.

$$Z_i = (X_i + X_{n+1-i})/2, \quad i = 1, \ldots .$$

Theorem 6. If the differentiable function $W(\underset{\sim}{e})$ is monotone, symmetric, exchangeable, additive and convex, then $\hat{\mu}$ which minimizes $W(\underset{\sim}{e}(\hat{\mu}))$ must lie within the range of the folded sample

$$\min (z_1, z_2, \ldots) \leq \max (z_1, z_2, \ldots).$$

Proof: By the additivity and exchangeability assumption, the function $W(\underset{\sim}{e})$ can be written as

$$W(\underset{\sim}{e}) = \sum_i w(e_i)$$

where by monotonicity $\partial w/\partial e_i > 0$ for $e_i > 0$ and by convexity $\partial^2 w/\partial e_i^2 > 0$

The first derivative of W with respect to μ is

$$\partial W/\partial \mu = \sum_i -(\partial w(e_i)/\partial e_i) = \sum_i g(X_i - \mu)$$

where g is a function such that (symmetry) $g(x) = -g(-x)$ and, (convexity) for $x > x'$, $g(x) < g(x')$.

Minimization of W given the constraints $\underset{\sim}{e} = \underset{\sim}{X} - \underset{\sim}{1}\mu$ implies the Lagrangian expression $\Sigma w(e_i) + \underset{\sim}{\lambda}'(\underset{\sim}{X} - \underset{\sim}{1}\mu - \underset{\sim}{e})$, where $\underset{\sim}{\lambda}$ is the vector of n Lagrange multipliers. Differentiation of this expression with respect to μ and to $\underset{\sim}{e}$, implies the equations

$$\underset{\sim}{\lambda}'\underset{\sim}{1} = 0 \qquad (2)$$

$$\{g(e_i)\} = \underset{\sim}{\lambda}$$

where $\{g(e_i)\}$ is the vector with elements equal to $g(e_i)$, the derivative of w evaluated at e_i. The function g has the properties that $g(e_i) \geq 0$ if $e_i > 0$ [monotonicity of w], $g(e_i) = -g(-e_i)$ [symmetry of w] and for $e_1 \geq e_2$, $g(e_1) \geq g(e_2)$ [convexity of w]. Hence, $\{g(e_i)\}$ is a vector in the same orthant as $\underset{\sim}{e}$ and also with elements identically ordered in absolute value. Thus $g(e_i)$ can be written as $\underset{\sim}{De}$ where $\underset{\sim}{D}$ is a diagonal matrix $\underset{\sim}{D} = \text{diag}\{d_1, d_2, \ldots, d_n\}$ with $d_i \geq 0$ and $d_i|e_i| \geq d_j|e_j|$ for $|e_i| \geq |e_j|$.

Equation (2) and $\underset{\sim}{De} = \underset{\sim}{\lambda}$ imply $\underset{\sim}{1}'\underset{\sim}{De} = 0$; and then using $\underset{\sim}{e} = \underset{\sim}{X} - \underset{\sim}{1}\mu$ we have $\underset{\sim}{1}'\underset{\sim}{D}(\underset{\sim}{X} - \underset{\sim}{1}\hat{\mu}) = 0$ which can be solved as

$$\hat{\mu} = (\underset{\sim}{1}'\underset{\sim}{D}\underset{\sim}{1})^{-1}(\underset{\sim}{1}'\underset{\sim}{DX}).$$

In words, $\hat{\mu}$ is a weighted average of the observations with the weights restricted such that if $|Y_i - \hat{\mu}| \geq |Y_j - \hat{\mu}|$ then $d_i|Y_i - \hat{\mu}| \geq d_j|Y_j - \hat{\mu}|$.

The foregoing implies that an estimate $\hat{\mu}$ is feasible if there exist $d_i \geq 0$ such that $\underset{\sim}{1}'\underset{\sim}{De} = \Sigma d_i e_i = 0$ and if $|e_i| \geq |e_j|$ then $d_i|e_i| \geq d_j|e_j|$. If $\hat{\mu}$ is less than all the elements of the folded sample, then it can be shown that $\Sigma d_i e_i > 0$ for all d_i satisfying these restrictions. Consider the elements in this summation pairwise, $d_j e_j + d_{n+1-j} e_{n+1-j}$. Since $\hat{\mu} < (X_j + X_{n+1-j})/2$ $(\hat{\mu} - X_j) < (X_{n+1-j} - \mu)$. This, together with $X_j < X_{n+1-j}$, implies $|e_j| < |e_{n+1-j}|$; thus $d_j|e_j| < d_{n-j+1}|e_{n-j+1}|$ and $0 < -d_j|e_j| + d_{n-j+1}|e_{n+1-j}|$. If e_j and e_{n-j+1} are both positive, then obviously $d_j e_j + d_{n+1-j} e_{n+1-j}$ is also positive. Otherwise the sum can be written as $d_j e_j + d_{n+1-j} e_{n+1-j} = -d_j|e_j| + d_{n+1-j}|e_{n+1-j}| > 0$.

The converse of this theorem is also true but a complete proof involves the enumeration of many cases. One case will be considered here. Suppose that $X_k < X_j$ and we would like to find d_i such that $Z_j < \mu < Z_k$, for some given μ. Suppose further that $X_k < X_j < \mu < X_{n+1-j} < X_{n-k+1}$.

Define now the constants $c_o = 0$, $c_1 = X_{n+1-j} - \mu$, $c_2 = \mu - X_j$, $c_3 = \mu - X_k$ $c_4 = X_{n-k+1} - \mu$ $c_5 = \infty$. It can be shown that $0 < c_1 < c_2 < c_3 < c_4$. Let n_i be the number of observations such that $\mu + c_{i-1} \leq X < \mu + c_i$ and let m_i be the number of observations such that $\mu - c_i < X \leq \mu - c_{i-1}$.

Assign weights to observations as follows

if	then
$\lvert e_i \rvert < c_2$	$d_i = 0$
$c_2 \leq \lvert e_i \rvert < c_4$	$d_i = \lvert e_i \rvert^{-1}$
$c_4 \leq \lvert e_i \rvert$	$d_i = (1 + \frac{n_2}{m_4}) \lvert e_i \rvert^{-1}$.

These weights do have the feature that if $|e_i| \geq |e_j|$ then $d_i|e_i| \geq d_j|e_j|$. Furthermore $\Sigma d_i e_i = n_3 + n_4 - m_3 - m_4 + (n_5 - m_5)(1 + \frac{n_2}{m_4})$. But since $n_2 + n_3 + n_4 = j - k$, $n_5 = k$, $m_3 = j - k$, and $m_5 = k - m_4$, then $\Sigma d_i e_i = j - k - n_2 - (j - k) - m_4 + m_4(1 + \frac{n_2}{m_4}) = 0$.

4. A Monte Carlo Study of the Mid-point of the Folded Sample

The bound reported in Theorem 6 implies a set of estimates over which you would be indifferent if "nothing is known" about the sampling process except that the observations come independently from a fixed symmetric unimodal distribution with the logarithm of the density function convex. In that case, any estimate in the range of the folded sample is a maximum likelihood estimate given a suitably chosen density function.

If something more is known about the sampling process, the choice of a point within the bound may not be a matter of indifference. A Monte Carlo study will illustrate this fact. For five different sampling distributions, the mean squared error of the mid-range of the folded sample is compared with the mean squared error of six other estimators. If you thought your sampling distribution were one of these, the median of the folded sample seems to be the best choice.

The following distributions were used:

(1) Uniform

$$f(x) = \begin{cases} 1 & |x| < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

(2) Standard Normal

$$f(x) = f_N(x|\mu = 0, \sigma^2 = 1)$$

(3) 5% Contaminated Normal

$$f(x) = 0.95 \times f_N(x|\mu = 0, \sigma^2 = 1) + 0.05 \times f_N(x|\mu = 0, \sigma^2 = 9)$$

(4) 10% Contaminated Normal

$$f(x) = 0.90 \times f_N(x|\mu = 0, \sigma^2 = 1) + 0.10 \times f_N(x|\mu = 0, \sigma^2 = 9)$$

(5) Double Exponential

$$f(x) = \frac{1}{2} e^{-|x|}$$

The following estimators were used where $(X_1, \ldots, X_n)$ are the order statistics

(1) Arithmetic Mean

$$\bar{x} = \sum_{i}^{n} x_i/n$$

(2) Median

$$\text{MED} = \begin{cases} X_{N/2+1} & \text{-n is odd} \\ \dfrac{X_{N/2} + X_{N/2+1}}{2} & \text{-n is even} \end{cases}$$

(3) Midrange

$$\text{MIDRNG} = \frac{(X_1 + X_n)}{2}$$

(4) 25% Trimmed Mean

$$\text{25-TM} = \frac{1}{N - 2J} \sum_{J+1}^{N-J} X_i$$

where $J = [N/4]$ and $[\cdot]$ is the greatest integer function.

(5) 25% Winsorized Mean

$$\text{25-WM} = \frac{1}{N} \left((J \cdot X_J) + \sum_{J+1}^{N-J} X_i + (J \cdot X_{N-J+1}) \right)$$

(6) Median of Folded Sample

$$\text{F-MED} = \begin{cases} Z_{M/2+1} & \text{-M is odd} \\ \dfrac{(Z_{M/2} + Z_{M/2+1})}{2} & \text{-M is even} \end{cases}$$

where $Z_1, \ldots, Z_M$ are the order statistics calculated from the folded sample.

(7) Midrange of Folded Sample

$$\text{F-MID} = \frac{(Z_1 + Z_M)}{2}$$

Relative efficiencies were calculated with respect to the last estimator (7), as follows:

$$\text{RELEFF}(i) = \frac{\text{MSE}(i)}{\text{MSE}(7)} \quad \text{where MSE}(i) = \sum_{1}^{100} (\hat{\theta} - \theta)^2 \quad i = 1, \ldots, 7$$

Hence a value less than one implies a gain in efficiency relative to the last estimator.

Table 1

(Sample Size = 10)

	Mean	Median	MDRNG	25-TM	25-WM	MED-FMDRNG	MP-FMDRNG
UNIFORM	1.023	2.264	0.465	1.658	1.157	1.136	1.000
STD-NORMAL	0.929	1.445	2.155	1.081	0.956	0.971	1.000
.05-CNORM	0.272	0.186	3.969	0.140	0.141	0.145	1.000
.10-CNORM	0.275	0.107	3.704	0.092	0.088	0.088	1.000
DBL-EXP	0.753	0.390	2.142	0.445	0.613	0.574	1.000

Table II

(Sample Size = 20)

	Mean	Median	MDRNG	25-TM	25-WM	F-MED	F-MID
UNIFORM	1.030	3.009	0.236	2.007	1.458	1.177	1.000
STD-NORMAL	0.761	1.138	2.835	0.874	0.797	0.759	1.000
.05-CNORM	0.110	0.073	3.991	0.063	0.058	0.057	1.000
.10-CNORM	0.043	0.002	4.026	0.001	0.001	0.001	1.000
DBL-EXP	0.573	0.283	2.357	0.321	0.374	0.374	1.000

The results are reported in Tables 1 and 2. Notice that the median of the folded sample out performs the midrange of the folded sample, except when the sampling distribution is uniform. The mean works relatively well for the normal, but otherwise can be substantially dominated. The median seems to trim too much, but the other trimming procedures are effective. The median of the folded sample has the further characteristic of simplicity, and overall seems the best.

REFERENCES

Andrews, D. F., et al., Robust Estimates of Locations. Princeton, New Jersey: Princeton University Press, 1972.

Bickel, P. J., and J. L. Hodges (1967), "The asymptotic theory of Galton's test and a related simple estimate of location." Annals of Mathematical Statistics, 38, 73-89.

Ekblom, H. (1974), "L_p-Methods for robust regression." Nordisk. Tidskr. Informationsbehandlung (BIT), 14, 22-32.

Forsythe, A. B. (1972), "Robust estimation of straight line regression coefficients by minimizing p^{th} power deviations." Technometrics, 14, 159-66.

Galbraith, R. F. and J. I. Galbraith (1974), "On the inverses of some patterned matrices arising in the theory of stationary time series." J. Applied Probability, 11, 63-71.

Hodges, J. L. (1967), "Efficiency in normal samples and tolerance of extreme values for some estimates of location." Fifth Berkeley Symposium on Probability and Statistics.

Hogg, Robert V. (1974), "Adaptive robust procedures: A partial review and some suggestions for future applications and theory." J. of American Statistical Association, 69, 909-927.

Huber, P. J. (1964), "Robust estimation of a location parameter." Annals of Mathematical Statistics, 35, 73-101.