THE MIDRANGE ESTIMATOR IN SYMMETRIC DISTRIBUTIONS

by

RICHARD ALLEN SUNDHEIM

B. S., Kansas State University, 1971

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1974

Approved by:

Kay O. Waller

Major Professor

TABLE OF CONTENTS

1. INTRODUCTION

When the center, θ, of a symmetric population is unknown, we have several alternatives for an estimator of θ. The almost habitual assumption that the population is normally distributed or approximately so has led to the wide use of the sample mean. If the population is in fact normally distributed, the sample mean is considered the optimal estimator since it is consistent, unbiased, sufficient, and efficient. However, the sample mean being very sensitive to extreme outliers is not the most desirable estimator when the parent population has "heavy-tails".

The alternative estimator we consider in this paper is the midrange. After defining and characterizing heavy-tailed distributions, we investigate the behavior of the midrange in a variety of heavy-tailed and light-tailed distributions. The classical properties of consistency, unbiasedness, sufficiency, and efficiency are then used to compare the relative merits of the midrange estimator with that of the sample mean and sample median.

The hypothesis put forth is that the shape of the parent distribution's tail is the determining factor in the choice of the estimator. Only when the tails are light is the sample mean preferred. When the tails are "short and heavy" the midrange is found to be a very desirable estimator while the median is preferred when sampling from populations with "long and heavy tails".

Bryson [2] contends there is evidence for placing more emphasis on the use of heavy-tailed distributions when modeling data. Therefore when the form of the distribution is unknown, rather than assuming normality, the sample information should be used to determine the tail-shape of the distribution before choosing the estimator. This is a concept of robust estimation and is briefly discussed in Chapter 5.

2. HEAVY-TAILED DISTRIBUTIONS

To define and characterize heavy-tailed distributions we consider a symmetric distribution with a density function $f(\cdot)$, distribution function $F(\cdot)$, and tail function $G(\cdot) = 1 - F(\cdot)$. The notion of heavy tails is a relative one which is used to compare the tail weights of different distributions. Intuitively, a heavy-tailed distribution assigns greater likelihood to extreme (tail) values in the range of positive density than does a light-tailed distribution. If the density function is positive over a finite range (i.e. $F(x) = 1$ for some finite x) the distribution is said to be bounded and if $F(x) < 1$ for all finite x, then the distribution is unbounded. This distinction leads us to adopt a separate characterization of heavy tails for the two cases.

2.1 Unbounded Distributions.

The amount of weight in the tails of the distribution is described by the rate at which $F(x)$ converges to one as x goes to infinity. This is equivalent to the rate at which the tail function $G(x)$ and the density function $f(x)$ converge to zero as x goes to infinity. So if $F(x)$ converges "slowly" to one, the distribution is considered to have heavy tails. For example, let

$$f_1(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \quad , \quad -\infty < x < \infty \tag{2.1}$$

$$f_2(x) = \lambda e^{-\lambda x} \quad , \quad x > 0 \ , \ \lambda > 0 \tag{2.2}$$

$$f_3(x) = \frac{1}{\pi\left(1 + x^2\right)} \quad , \qquad -\infty < x < \infty \tag{2.3}$$

represent the density functions for the standard normal, exponential, and Cauchy distributions, respectively. Since as x goes to infinity, e^x approaches infinity faster than any power of x, it follows that $f_1(x)$ approaches zero faster than $f_2(x)$ which in turn approaches zero faster than $f_3(x)$. Therefore we can conclude the Cauchy distribution has heavier tails than the other two distributions, while the normal distribution has light tails when compared to the exponential and Cauchy distributions.

To characterize heavy-tailed distributions we need some borderline distribution ϕ, such that every distribution with tails heavier than ϕ is considered to be a member of the class of heavy-tailed distributions. Bryson [2] argues that ϕ should be the exponential distribution. He does this by considering the conditional mean exceedence, CME, which is defined as

$$CME_c = E\left(X - c \mid X \geq c\right). \tag{2.4}$$

If c is sufficiently large, we obtain the following characterization of tail weights:

(i) A distribution is said to be heavy-tailed if the CME_c increases as c increases.

(ii) A distribution is said to have light tails if the CME_c decreases as c increases.

(iii) A distribution is a borderline case if the CME_c is constant as c increases.

To apply this characterization, we put (2.4) in a more tractable form:

$$CME_c = \frac{1}{1 - F(c)} \int_c^\infty (x - c)\, dF(x) \qquad (2.5)$$

Letting $y = x - c$, (2.5) becomes

$$CME_c = \frac{1}{1 - F(c)} \int_0^\infty y\, dF(y)$$

$$= \frac{-1}{1 - F(c)} \int_0^\infty y \frac{d}{dy}[1 - F(y)]\, dy \quad .$$

Integrating by parts yields

$$CME_c = \frac{1}{1 - F(c)} \int_c^\infty [1 - F(x)]\, dx$$

or

$$CME_c = \frac{1}{G(c)} \int_c^\infty G(x)\, dx \quad . \qquad (2.6)$$

To illustrate (2.6) consider the exponential distribution in (2.2).

Then

$$F(x) = 1 - e^{-\lambda x},$$

$$G(x) = e^{-\lambda x},$$

and

$$CME_c = \frac{1}{e^{-\lambda c}} \int_c^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda} \quad .$$

Therefore, the CME_c is a constant function and the exponential distribution is considered a borderline case.

The rationale behind choosing the exponential as the borderline distribution can be illustrated by considering lifetime data (in hours) of a mechanical part. The CME_c is then interpreted as the mean residual lifetime, that is, the expected lifetime remaining given that the part has already survived at least c hours. If the lifetimes follow an exponential distribution, then once a part has reached a certain age (sufficiently large c) its mean residual lifetime is unaffected by any additional aging. With heavy tails this part would improve with additional age whereas light tails produce a decreasing mean residual lifetime.

Unfortunately, for many distributions it is difficult to express $\int_c^\infty G(x)dx$ in closed form and it is often easier to compare the density function with the exponential density function as in (2.1), (2.2), and (2.3). Bryson [2] presents an alternative graphical approach which plots log G(x) against x. Heavy-tailed distributions tend to have concave-upward graphs for sufficiently large x and distributions with light tails tend to have concave-downward graphs whereas the borderline exponential graph is linear.

When we were concerned with the rate of convergence of F(x) only the extreme part of the tail was important. Intuitively, tail-weight or "tail-thickness" should be determined by a larger portion of the distribution's tail. For example, let

$$f_4(x) = \frac{1}{2} e^{-|x|} \quad , \qquad -\infty < x < \infty \tag{2.7}$$

$$f_5(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \quad , \qquad -\infty < x < \infty \tag{2.8}$$

represent the density functions for the Laplace (double exponential) and logistic distributions, respectively. As x goes to infinity, $f_4(x)$ and $f_5(x)$ approach zero at essentially the same rate as the exponential distribution. To compare the teil-weights of $f_4(x)$ and $f_5(x)$ we consider the length of the tail between the .75 and .95 quantiles. With this in mind, Crow and Siddiqui [5] suggest the following measure of tail-thickness:

$$R = \frac{\zeta(.95) - \zeta(.5)}{\zeta(.75) - \zeta(.5)} \tag{2.9}$$

where $\zeta(p)$ is the p^{th} quantile.

For unbounded distributions the larger the value of R the heavier the tail, so that the Laplace distribution (R = 3.322) has heavier tails than the logistic distribution (R = 2.680).

Another useful measure of tail-thickness is the moment coefficient of kurtosis, β_2, which is defined as

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} \tag{2.10}$$

where μ_r is the r^{th} central moment.

Traditionally, β_2 was viewed to describe the "flatness" of the density curve, however Kendall and Stuart [11] cite examples disputing this. Despite this it still seems to follow that for most of the unbounded distributions we encounter, the larger the value of β_2 the heavier the tail. The logistic distribution ($\beta_2 = 4.2$) in (2.8) is flatter than the Laplace distribution ($\beta_2 = 6.0$) in (2.7) since the maximum ordinate of $f_4(x)$ is .5 and for $f_5(x)$ is .25, yet the Laplace distribution was shown to have heavier tails by (2.9).

In order to compare the values of β_2 and R for some of the unbounded symmetric distributions we present a table from Siddiqui and Raghunandanan [16].

	β_2	R
Normal	3.00	2.439
Logistic	4.20	2.680
Laplace	6.00	3.322
t_5	9.00	2.773
t_3	∞	3.077
Cauchy	∞	6.314

For the t-distribution with three degrees of freedom, t_3, and for the Cauchy distribution, β_2 was calculated for the distribution truncated at K_1 and K_2 and then letting $K_1 \to -\infty$ and $K_2 \to \infty$. In this case R may be a preferable measure of tail-weight, however for our purposes β_2 seems quite adequate and will be the one we shall use in the remaining chapters. To characterize heavy-tailed distributions for these measures the logistic distribution, whose tails are similar to the exponential tail for extreme values of x, would appear to offer a reasonable borderline value. Hence for $\beta_2 \geq 4.2$, we describe the tails of the distribution as being long and heavy. We also use β_2 to describe the tail-weight in bounded distributions where the tails are short, but its behavior is somewhat different.

2.2 Bounded Distributions.

Technically, bounded distributions have very light and smooth tails since F(x) converges rapidly to one. However distributions such as the uniform and the U-shaped appear to have heavy tails when we consider only

the part of the tail where f(x) is positive. If these short tails are heavy, the .95 and .75 quantiles will be close together and the value of R in (2.9) will be close to 1.0. So that the smaller the value of R, the heavier the tails. This same behavior is exhibited by β_2 as can be seen in the following table:

	β_2	R
Normal	3.00	2.439
Triangular	2.40	2.335
Cosine	2.19	2.183
Parabolic	2.14	2.100
Uniform	1.80	1.800
U-shaped	1.19	1.216

The development in Chapter 4 shows that when $\beta_2 \leq 2.19$, the midrange is more efficient than the mean for estimating the center of bounded symmetric populations. We therefore conveniently describe the tails of the distribution as being short and heavy whenever $\beta_2 \leq 2.19$. In the next two chapters we investigate the effect that heavy tails has on the midrange estimator.

3. MIDRANGE AND CONSISTENCY

The sample midrange is determined by the smallest and largest observations and is therefore sensitive to outliers. This criticism, which was also leveled against the sample mean, is justified in distributions with long and heavy tails; the consequence being that the midrange is an inconsistent estimator in these distributions. In asymmetric unbounded distributions, the sample midrange depends upon the sample size and we cannot state specifically what it estimates. For this reason we consider only symmetric populations in the subsequent development. In this chapter we develop the distribution of the midrange and investigate the consistency of the midrange.

3.1 Distribution of the Midrange.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population having a continuous p.d.f. $f(x)$. Rearranging the sample in the order of their magnitude provides the n order statistics

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

The joint distribution of $X_{(1)}$ and $X_{(n)}$ is

$$f(x_{(1)},x_{(n)}) = n(n-1)\{F(x_{(n)}) - F(x_{(1)})\}^{n-2} f(x_{(1)}) f(x_{(n)}). \qquad (3.1)$$

Let M and W denote the midrange and half-range of the sample, respectively, then

$$M = \frac{X_{(1)} + X_{(n)}}{2}, \qquad W = \frac{X_{(n)} - X_{(1)}}{2}, \qquad (3.2)$$

and the inverse transformations are

$$X_{(1)} = M - W \quad , \quad X_{(n)} = M + W \quad . \tag{3.3}$$

Then the Jacobian is

$$J = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2.$$

Therefore , from (3.1), the joint distribution of M and W is

$$f(m,w) = 2n(n-1)[F(m + w) - F(m - w)]^{n-2} f(m - w) f(m + w) \quad . \tag{3.4}$$

The exact distribution of the midrange is found by

$$f(m) = \int f(m,w)\, dw \quad . \tag{3.5}$$

For bounded distributions with endpoints a and b such that $a < b$ the limits of integration are $0 \leq w \leq \frac{(b - a)}{2} - \left| m - \frac{a + b}{2} \right|$ and for unbounded distributions the limits are $0 \leq w \leq \infty$.

Most measures of central tendency, such as the mean or median (Fisz[7]), have asymptotic normal distributions where the precision of the estimate improves with increasing sample size. This is not true of the midrange as Gumbel [8] shows that the asymptotic distribution of the midrange is logistic when the parent distribution is of the exponential-type and the precision may increase, decrease, or remain invariant with increasing sample size depending on the form of the distribution. Unbounded distributions are of the exponential-type if for very large values of x the following equations

are approximately valid:

$$\frac{f(x)}{1 - F(x)} = - \frac{f'(x)}{f(x)} \quad ; \qquad \frac{f(x)}{F(x)} = \frac{f'(x)}{f(x)} \ .$$

The tails of exponential-type distributions converge at least as fast as the exponential distribution. Included in this group is the normal, logistic and Laplace distributions while the Cauchy is not of the exponential-type. Certainly precision increasing with sample size is a desirable property for an estimator to have. We now investigate the conditions under which the median, mean, and midrange attain this property.

3.2 Consistency.

A consistent estimator of θ is one which converges in probability to θ. For continuous symmetric populations Fisz [7] proves that the median is a consistent estimator of θ and if $E(X)$ exist then the mean is also consistent. In the Cauchy distribution, where $E(X)$ does not exist, Kendall and Stuart [12] show that the mean is not a consistent estimator of θ. This tends to support Robertson and Wright's [14] claim that the consistency of the sample mean depends only on the amount of weight in the tails of the parent distribution. They also report that consistency of the midrange depends not only on tail-weight but also on the smoothness of the tails. The following two theorems, useful in determining when the midrange is consistent, are due to Geffroy and are given by Robertson and Wright [14] as they apply to distributions symmetric about zero:

Theorem 1. The midrange converges in probability to zero if and only if, for all $\varepsilon > 0$

$$\lim_{x \to \infty} \frac{G(x + \varepsilon)}{G(x)} = 0$$

where G(x) is the tail function.

Theoren 2. If

$$G(x) = \int_x^\infty c\, e^{-|t|^p}\, dt \quad ,$$

then the midrange fails to converge in probability when $p \leq 1$, but converges almost surely to zero when $p > 1$.

Theorem 2 implies that the midrange is consistent in the standard normal distribution ($p = 2$), but inconsistent in the smooth-tailed logistic distribution ($p = 1$) and therefore inconsistent in all distributions with long and heavy tails. It may also be inconsistent in light-tailed distributions if the tails are not smooth, as is demonstrated by the following example found in Robertson and Wright [14]: Let $h(\cdot)$ and $H(\cdot)$ be the density and distribution function of the standardized normal population. Let $f(x)$ be a symmetric density whose values are given by:

$$f(x) = \begin{cases} h(x) + h(x+1) & ; \quad 2n \leq x < 2n+1 \\ 0 & ; \quad 2n+1 \leq x < 2n+2 \end{cases}$$

$n = 0, 1, 2, \ldots$

Then $F(x) \geq H(x)$ for all positive x so that the tail of $F(x)$ is no heavier than the tail of $H(x)$. However,

$$\limsup_{x \to \infty} \{1 - F(x + \tfrac{1}{2})\} \cdot \{1 - F(x)\}^{-1} = 1$$

so by Theorem 1, with $\varepsilon = \tfrac{1}{2}$, the midrange fails to converge in probability when sampling from $F(x)$. However, by Theorem 2, the midrange converges almost surely when sampling from $H(x)$.

3.3 Unbiasedness.

Kendall and Stuart [11] state that in every symmetric population the median is an unbiased estimator of θ while the midrange is unbiased whenever E(M) exists. They show that E(M) exists at least for the symmetric exponential-type populations. The sample mean can be found to be an unbiased estimator of θ whenever E(X) exists, by

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X).$$

For symmetric populations there are a number of consistent unbiased estimators and to choose between them we rely on their relative efficiency which is discussed in the next chapter.

4. EFFICIENCY OF THE MIDRANGE

In bounded distributions the tails are both light and smooth since for sufficiently large x, the tail function G(x) = 0. Therefore the midrange, as well as the median and mean, is a consistent unbiased estimator of θ. This is also true of unbounded light-tailed distributions. When this is the case the choice of the estimator usually rests upon its efficiency. In populations which are symmetric, bounded and heavy-tailed, the midrange is found to be more efficient than the mean and this relative efficiency increases as β_2 decreases. The uniform distribution, where the midrange is optimal, is discussed in detail. In unbounded light-tailed populations the mean is found to be the preferred estimator while in populations with long and heavy tails the median enjoys greater efficiency than either the mean or midrange.

4.1 Uniform (Rectangular) Distribution.

The two parameter uniform distribution is defined by the density function

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases} \tag{4.1}$$

or by the distribution function

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b\ . \end{cases} \tag{4.2}$$

Therefore, by symmetry,

$$\theta = E(X) = (a + b)/2 \tag{4.3}$$

Using the relation

$$\mu_r = \sum_{j=0}^{r} (-1)^j \binom{r}{j} m_{r-j} m_1^j \tag{4.4}$$

where μ_r is the r^{th} central moment and $m_r = E(X^r)$,

we can compute the variance

$$\mu_2 = \mathrm{Var}(X) = \frac{(b - a)^2}{12} \tag{4.5}$$

and from (2.10), $\beta_2 = 1.8$.

In order to find the efficiency of the midrange we need the distribution of the midrange.

4.1.1 Exact Distribution of the Midrange. To find the exact distribution of the midrange (4.1) and (4.2) are substituted into (3.4) to get

$$f(m,w) = 2n(n-1)\left[\frac{m + w - a}{b - a} - \frac{m - w - a}{b - a} \right]^{n-2} / (b - a)^2$$

$$= \frac{2^{n-1} n(n-1)\, w^{n-2}}{(b - a)^n} \tag{4.6}$$

where $a \leq m \leq b$, $0 \leq w \leq \frac{b - a}{2} - \left| m - \frac{a + b}{2} \right|$.

Integrating (4.6) with respect ot w, we get

$$f(m) = 2^{n-1}\, n \left[\frac{b-a}{2} - \left| m - \frac{a+b}{2} \right| \right]^{n-1} / (b-a)^n \quad . \tag{4.7}$$

Carlton [3] shows that in the uniform distribution the midrange has a non-normal limiting distribution and hence the concept of efficiency that we use is not strictly applicable.

4.1.2 Efficiency. The asymptotic relative efficiency, ARE, of the mean relative to the midrange is defined as

$$ARE(\bar{X},M) = \frac{Var(M)}{Var(\bar{X})} \quad . \tag{4.8}$$

By symmetry,

$$E(M) = (a + b)/2 \quad .$$

Using (4.7) and applying integration by parts twice we find

$$E(M^2) = \int_a^b m^2 f(m)\, dm$$

$$= \left(\frac{a+b}{2} \right)^2 + \frac{(b-a)^2}{2(n+1)(n+2)} \quad .$$

Therefore the variance of M is

$$Var(M) = \frac{(b-a)^2}{2(n+1)(n+2)}$$

and the variance of $\bar{X}$ is

$$Var(\bar{X}) = \frac{Var(X)}{n} = \frac{(b-a)^2}{12n} \quad . \tag{4.9}$$

Hence the efficiency in (4.8) is

$$ARE(\bar{X},M) = \frac{6n}{(n+1)(n+2)} \quad . \tag{4.10}$$

As n increases, $ARE(\bar{X},M) \to 6/n \to 0$. In other words, the midrange becomes increasingly more efficient than the mean as the sample size increases. Therefore, the mean which uses the "full sample information" by weighting all the observations equally is as an estimator of θ inferior to the midrange which weights only the two extreme observations.

To show that both the midrange and mean are more efficient than the median, $\hat{M}$, we need to find the variance of $\hat{M}$. To simplify the calculations we choose $b = -a = \frac{1}{2}$ in (4.1) and let the sample size $n = 2k + 1$ where k is an integer. The median then is the $k+1^{st}$ order statistic and therefore the distribution of $\hat{M}$ is (see Kendall and Stuart [11])

$$f(\hat{m}) = \frac{\{F(\hat{m})\}^k \{1 - F(\hat{m})\}^k}{B(k+1,\ k+1)} \tag{4.11}$$

which is a Beta distribution with both parameters equal to k+1. The variance of this Beta distribution is

$$Var(\hat{M}) = \frac{(k+1)(k+1)}{(2k+2)^2(2k+3)}$$

$$= \frac{1}{4(2k+3)} = \frac{1}{4(n+2)} \quad .$$

From (4.9), with $b - a = 1$

$$Var(\bar{X}) = \frac{1}{12n} \quad .$$

So that

$$ARE(\hat{M},\bar{X}) = \frac{Var(\hat{M})}{Var(\bar{X})}$$

$$= \frac{12n}{4(n+2)} \to 3 \qquad \text{as } n \to \infty \ .$$

Therefore the median is only 1/3 as efficient as the mean and hence it has "zero" efficiency compared to the midrange.

To show that the midrange is more efficient (i.e. smaller variance) than any other unbiased estimator we need only show that it is sufficient and that the uniform distribution is complete (see Hogg and Craig [10]). To show that M is a sufficient statistic for θ it suffices to show that $X_{(1)}$ and $X_{(n)}$ are a pair of jointly sufficient statistics for a and b. Substituting (4.1) and (4.2) into (3.1) yields

$$f(x_{(1)}, x_{(n)}) = n(n-1)(x_{(n)} - x_{(1)})^{n-2} (b - a)^{-n} \ . \qquad (4.12)$$

Hence the likelihood function, L, can be factored as

$$L(x_1, x_2, \ldots, x_n \mid a, b) = \prod_{i=1}^{n} f(x_i \mid a,b)$$

$$= (b-a)^{-n} I(x_{(1)}-a)I(b-x_{(n)}) \qquad (4.13)$$

$$= f(x_{(1)}, x_{(n)})\, k(x)$$

where I(y) is the indicator function

$$I(y) = \begin{cases} 1 & \text{if } y \geq 0 \\ 0 & \text{if } y < 0. \end{cases}$$

and

$$k(x) = (x_{(n)} - x_{(1)})^{2-n}/n(n-1)$$

is independent of a and b, so M is sufficient for θ.

Hogg and Craig [10] show that the completeness property is satisfied, and therefore the midrange is the most efficient unbiased estimator when sampling from the uniform population. When sampling from other bounded heavy-tailed distributions the midrange is not optimal but it remains more efficient than the mean.

4.2 Other Bounded Distributions.

The procedure for finding the efficiency of $\overline{X}$ compared to M in the cosine, parabolic, U-shaped and dichotomous distributions is identical to that in the uniform distribution, although the integrations necessary are often quite tedious. We therefore just summarize Rider's [13] results in the following table:

Efficiency of Mean Relative to Midrange

Sample Size	Cosine $\beta_2 = 2.19$	Parabolic $\beta_2 = 2.14$	Uniform $\beta_2 = 1.8$	U-shape $\beta_2 = 1.19$	Dichotomous $\beta_2 = 1.0$
2	1.000	1.000	1.000	1.000	1.000
3	.984	.974	.900	.760	.750
4	.971	.951	.800	.527	.500
5	.963	.932	.714	.354	.312
6	.956		.643		.188

These values indicate that for symmetric populations with $\beta_2 \leq 2.19$ (short heavy tails) the midrange is more efficient than the mean and this efficiency increases as β_2 decreases. Gumbel [8] cites four symmetric distributions from the Pearson system with $\beta_2 \geq 2.5$ where the mean is found to be more efficient than the midrange and the midrange becomes increasingly inferior as β_2 increases. We explore this further in unbounded distributions of which the normal is the most important.

4.3 Unbounded Distributions.

The mean, median and midrange were all shown in Chapter 3 to be consistent and unbiased in the normal distribution. To choose between them we again rely on their relative efficiencies which are summarized in the following table from Kendall and Stuart [11]

Sample Size	ARE($\hat{M}$,$\bar{X}$)	ARE(M,$\bar{X}$)
2	1.000	1.000
4	1.092	1.092
6	1.135	1.190
10	1.177	1.362
20	1.214	1.691
∞	1.253	∞

The mean is definitely the most reliable and the midrange becomes relatively increasingly unreliable while the median has almost attained its limiting value by the time $n = 20$.

For other unbounded distributions we refer to a paper by Broffitt [1] who considers the family of power distributions, symmetric about zero,

$$h_p(x) = \frac{1}{2\Gamma(1 + 1/p)} e^{-|x|^p} , \quad -\infty < x < \infty \tag{4.14}$$

where $p > 0$.

For this family,

$$\mathrm{Var}_p(X) = \frac{\Gamma(1 + 3/p)}{3\Gamma(1 + 1/p)} , \tag{4.15}$$

so that

$$\mathrm{Var}_p(\bar{X}) = \frac{\Gamma(1 + 3/p)}{3n\Gamma(1 + 1/p)} . \tag{4.16}$$

Kendall and Stuart [11] show the median to be asymptotically normally distributed with variance

$$\mathrm{Var}_p(\hat{M}) = \frac{1}{4n\, h_p(0)} \tag{4.17}$$

where $h_p(0)$ is the median ordinate.

Then from (4.17) and (4.14)

$$\mathrm{Var}_p(\hat{M}) = \frac{\Gamma^2(1 + 1/p)}{n} . \tag{4.18}$$

Therefore, from (4.16) and (4.18)

$$\mathrm{ARE}(\bar{X},\hat{M}) = \frac{3\Gamma^3(1 + 1/p)}{\Gamma(1 + 3/p)} , \quad 0 < p < \infty \tag{4.19}$$

Broffitt [1] states that from (4.19) it can be seen numerically that $\bar{X}$ is more efficient than the median when $p > 1.41$. The median is then preferred in the heavy-tailed logistic and Laplace ($p = 1$) distributions.

To compare the midrange and mean we consider only the light-tailed distributions since for $p \leq 1$ the midrange was found in Chapter 3 to be inconsistent. Asymptotic results for the midrange are based on $X_{(1)}$ and $X_{(n)}$ being asymptotically independent which is proved in Gumbel [8]. Broffitt [1] uses this result to find the asymptotic variance of the midrange

$$\text{Var}(M) = \frac{\pi}{[\ 12p^2\ (\ln n)^{2(1-1/p)}]} \tag{4.20}$$

and therefore from (4.16) and (4.20)

$$\text{ARE}(\bar{X},M) = \frac{nk(p)}{(\ln n)^{2(1-1/p)}} \quad , \quad 1 < p < \infty \tag{4.21}$$

where $k(p) = \dfrac{\pi^2\ \Gamma(1 + 1/p)}{4p^2\ \Gamma(1 + 3/p)}$.

For any fixed p, $\text{ARE}(\bar{X},M) \to \infty$ as $n \to \infty$ and hence the mean is more efficient than the midrange. A comparison of the mean, median, and midrange with the most efficient estimator, is given in Tiao and Lund [17] for p =1, 4/3, and 4. It demonstrates that the median fares very well in the Laplace distribution, whereas the efficiency of the mean remains constant at one-half that of the minimum variance estimator and the efficiency of the midrange falls off rapidly toward zero. For $p = 4/3$ (somewhat heavier tails than the normal) both the median and mean do fairly well, with the median slightly better for even sample sizes and the mean preferred for odd sample sizes. The efficiency of the midrange again drops rapidly. For $p = 4$ (very light tails) the median has poor efficiency while

the midrange is more efficient than the mean for $n < 13$. So for very light-tailed distributions the midrange may be preferred for small sample sizes. This can be seen by recognizing that the density in (4.14) approaches the uniform density as $p \to \infty$. Hence we can, by choosing p large enough, get arbitrarily close to the uniform density, yet the asymptotic result in (4.21) shows the unbounded tails eventually (for large enough n) yield the mean more efficient than the midrange.

We noticed that the asymptotic variances of $\overline{X}$ and $\hat{M}$ are of the order 1/n. This is not always the case with the midrange. The following results in Cramér [4] illustrate this dependence that the midrange has on the form of the distribution.

Orders of Asymptotic Variances of M

Uniform	$O(1/n^2)$
Triangular	$O(1/n)$
Normal	$O(1/\ln n)$
Laplace	$O(1)$
Cauchy	$O(n^2)$

The values for the Cauchy and Laplace distributions again point out the inconsistency of the midrange.

5. OTHER ESTIMATORS

We briefly consider best linear unbiased estimates to see how the optimal weighting of the order statistics is affected by the tail weight of the parent distribution. Then we look at an estimator having good efficiency for a wide family of distributions. This is useful when the form of the distribution is unknown, as is often the case.

5.1 Best Linear Unbiased Estimate (BLUE).

The estimator of θ that is usually chosen is some linear combination of the order statistics and it takes the form

$$\hat{\theta} = \sum_{i=1}^{n} \alpha_i X_{(i)}$$

where α_i denotes the weight given the i^{th} order statistic and $\sum_{i=1}^{n} \alpha_i = 1$. The weights should be chosen so as to guarantee that the estimator is the best one possible according to the properties of consistency, unbiasedness, sufficiency and efficiency. We have already seen that the midrange, which assigns weights $\alpha_1 = \alpha_n = .5$ and gives a zero weight to all the remaining observations, is optimal in the uniform population. However, in the normal population the optimal estimator is the sample mean which weights all observations equally. So the choice of the estimator depends on the form of the parent distribution.

Kendall and Stuart [12] use the technique of ordered least squares to find the weights guaranteeing an unbiased mininum variance estimator(BLUE).

The BLUE is found by

$$\theta^* = \frac{\underline{1}' \underline{V}^{-1} \underline{X}_{(0)}}{\underline{1}' \underline{V}^{-1} \underline{1}}$$

where

$\underline{1}$ is a column vector of ones,

$\underline{X}_{(0)}$ is a column vector of order statistics,

and $\underline{V}$ is the variance-covariance matrix of $\underline{X}_{(0)}$.

With this approach Kendall and Stuart[12] show the midrange to be the BLUE in the uniform distribution and David [6] shows $\overline{X}$ to be the BLUE in the normal distribution. For other distributions a table of weights, when one exists, should be used to determine θ^* . We present such a table from Sarhan and Greenberg [15] for samples of size 5:

β_2	Population	α_1	α_2	α_3	α_4	α_5
1.19	U-shaped	.55848	-.04486	-.02724	-.04486	.55848
1.80	Uniform	.50000	0	0	0	.50000
2.14	Parabolic	.38629	.07954	.06835	.07954	.38629
2.40	Triangular	.30608	.11885	.15014	.11885	.30608
3.00	Normal	.20000	.20000	.20000	.20000	.20000
6.00	Laplace	.01664	.22130	.52413	.22130	.01664

These values support the contention that when β_2 is small (short heavy tails) the best estimate gives more weight to the extreme observations, whereas for large values of β_2 (long heavy tails) more weight is given to the

middle observations. When tables are unavailable for a particular distribution or sample size, it may be more practical to adopt a simpler estimator which retains good properties although it may not be quite as efficient. The topic of robust estimators addresses that question.

5.2 Robust Estimators.

When sampling from a population whose distribution is unknown, it is useful to have an estimator which works well for a wide family of distributions. Although the midrange may be adopted when $\beta_2 \leq 2.19$, it is not very desirable for other distributions. The wide use of the sample mean results from assuming the unknown population has a normal distribution. Instead of weighting the ordered observations according to some assumed distribution (which may or may not be correctly assumed), Hogg [9] uses the sample information to determine the weights in the following estimator T of the center of a symmetric distribution:

$$T = \begin{cases} \overline{X}^{c}_{1/4}, & b_2 < 2.0 \\ \overline{X}, & 2.0 \leq b_2 \leq 4.0 \\ \overline{X}_{1/4}, & 4.0 < b_2 \leq 5.5 \\ m, & 5.5 < b_2 \end{cases}$$

where $\overline{X}^{c}_{1/4}$ is the mean of the [n/4] smallest and the [n/4] largest observations, $\overline{X}_{1/4}$ is the mean of the remaining interior observations, $\overline{X}$ and m are the sample mean and median, and b_2 is the sample coefficient of kurtosis:

$$b_2 = \frac{n \sum (x_i - \bar{x})^4}{\left(\sum (x_i - \bar{x})^2 \right)^2} .$$

This reasonably simple estimator makes effective use of the tail shape of the distribution provided b_2 is a good estimate of β_2. Hogg [9] shows that T has excellent asymptotic properties while his empirical studies have yielded a fine overall performance.

Another interesting robust estimator of location for symmetric populations is offered by Crow and Siddiqui [5] and extended by Siddiqui and Raghunandanan [16]. We only mention here that the efficiency of the estimator is studied in a number of different distributions, including those we have referred to in this paper. Other approaches to robust estimation are referenced in David [6].

6. CONCLUSION

We have classified symmetric heavy-tailed distributions according to the shape of their tail as follows:

(i) The tails are considered short and heavy if $\beta_2 \leq 2.19$ in which case the tail-weight increases as β_2 decreases.

(ii) The tails are considered long and heavy if $\beta_2 \geq 4.20$ in which case the tail-weight increases as β_2 increases.

When the tails of the parent distribution are long and heavy, the midrange is an inconsistent estimator and therefore undesirable. In this situation, according to the BLUE, the estimator chosen should place less weight on the extremes as the tail weight increases. The median or trimmed mean could be appropriate when weighting tables for the BLUE are unavailable.

When the tails of the parent distribution are short and heavy the midrange is a consistent and unbiased estimator having greater efficiency than the sample mean. Here the BLUE, having maximum efficiency, places more weight on the extremes as tail-weight increases. In the uniform distribution the midrange is the BLUE which indicates that the efficiency of the midrange decreases as the tail-weight deviates from that to the uniform population. Yet the midrange may be preferred due to its computational ease.

Tail-shape then is the influencing factor in determining the appropriate estimator. Only when the tails are light should the sample mean be considered. When the tails are heavy the most appropriate estimators often use less than the full sample information. Determining the tail-shape when the form of the distribution is unknown is the basis for robust

estimation procedures. These procedures, if adopted by textbooks, could reduce the almost automatic assumption of normality and use of the sample mean.

ACKNOWLEDGMENT

I would like to thank Dr. Ray A. Waller for his assistance and patience in the preparation of this report.

REFERENCES

[1] Broffitt, J.D. (1973). "An example of the large sample behavior of the midrange". Technical Report. No.26. Department of Statistics, University of Iowa.

[2] Bryson, M.C. (1974). "Heavy-tailed distributions: properties and tests". Technometrics. 16, 61-68.

[3] Carlton, A.G. (1946). "Estimating the parameters of a rectangular distribution". Annals of Mathematical Statistics. 17, 355-358.

[4] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press, Princeton.

[5] Crow, E.L. and Siddiqui, M.M. (1967). "Robust estimation of location". Journal of the American Statistical Association. 62, 353-389.

[6] David, H.A. (1970). Order Statistics. Wiley, New York.

[7] Fisz, M. (1963). Probability Theory and Mathematical Statistics. Wiley, New York.

[8] Gumbel, E.J. (1958). Statistics of Extremes. Columbia University Press, New York.

[9] Hogg, R.V. (1967). "Some observations on robust estimation". Journal of the American Statistical Association. 62, 1179-1186.

[10] Hogg, R.V. and Craig, A.T. (1970). Introduction to Mathematical Statistics. Macmillan, London.

[11] Kendall, M.G. and Stuart, A. (1969). The Advanced Theory of Statistics. Vol.1. Hafner, New York.

[12] Kendall, M.G. and Stuart, A. (1967). The Advanced Theory of Statistics. Vol.2. Hafner, New York.

[13] Rider, P.R. (1957). "The midrange of a sample as an estimator of the population midrange". Journal of the American Statistical Association. 52, 537-542.

[14] Robertson, T. and Wright, F.T. (1972). "Consistency of the midrange: distributons whose tails are both light and smooth". Technical Report. No.11. Department of Statistics, University of Iowa.

[15] Sarhan, A.E. and Greenberg, B.G.(Eds.). (1962). Contributions to Order Statistics. Wiley, New York.

[16] Siddiqui, M.M. and Raghunandanan, K. (1967). "Asymptotically robust estimators of location". Journal of the American Statistical Association. 62, 950-953.

[17] Tiao, G.C. and Lund, D.R. (1970). "The use of OLUMV estimators in inference robustness studies of the location parameter of a class of symmetric distributions". Journal of the American Statistical Association. 65, 370-386.

THE MIDRANGE ESTIMATOR IN SYMMETRIC DISTRIBUTIONS

by

RICHARD ALLEN SUNDHEIM

B.S., Kansas State University, 1971

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1974

ABSTRACT

When the center, θ, of a symmetric population is unknown, we have several alternatives for an estimator of θ. The estimators we consider in this paper are the sample mean, median, and midrange with emphasis placed on the latter. After characterizing distributions according to the shape of their tail, the behavior of the midrange with respect to the properties of consistency and efficiency is investigated in a variety of distributions.

The hypothesis put forth is that the shape of the parent distribution's tail is the determining factor in the choice of the estimator. Only when the tails are light is the sample mean preferred. When the tails are "short and heavy" the midrange is found to be a very desirable estimator while the median is preferred when sampling from populations with "long and heavy" tails. When the form of the distribution is unknown, rather than assuming normality, the sample information is used to determine the tail-shape before choosing the estimator.