Comparing the Sample Mean and Median in Measurement Error Models

by

Anirban DasGupta*
Purdue University

Technical Report #93–35

Department of Statistics
Purdue University

July 1993

*Research supported by NSF grant DMS 93-077-27.

△

COMPARING THE SAMPLE MEAN AND MEDIAN IN MEASUREMENT ERROR MODELS

by

Anirban DasGupta
Department of Statistics
Purdue University
1399 Mathematical Sciences Bldg.
West Lafayette, IN 47907-1399

ABSTRACT

The sample mean and median are compared in a measurement error model with the asymptotic relative efficiency as the basis for comparison. The model treats the observable X_i as $X_i = \theta + Z_i + e_i$, where θ is an unknown location, Z_i has a fixed distribution F which is unimodal and symmetric about zero and e_i is independent of Z_i and has a distribution G, also unimodal and symmetric, but supported on a bounded interval $[-m, m]$. An explicit formula is derived for the minimum relative efficiency over all such G of the mean with respect to the median; the formula depends on F and m. As an application of the results, it is proved that if an independent measurement error of magnitude at most 3.8σ gets added to a variate with $F = N(\theta, \sigma^2)$ distribution, then the mean is more efficient than the median provided the measurement error has a unimodal symmetric distribution G, regardless of the choice of G. For the logistic case, the same result is established if the error is at most 1.105σ in magnitude. Our results apply to the large class of problems that can be formulated as measurement error problems.

Key words: Mean, median, measurement error, asymptotic relative efficiency, unimodal, symmetric.

1. Introduction. There is a large body of literature on asymptotic comparison of the sample mean and the median as estimators of a location parameter. Many of these results demonstrate that if the population is heavy tailed, then the median is a better choice than the mean. For instance, while the sample median is only about 63% efficient in comparison to the mean for a normal population, it is about 63% more efficient than the mean if the population is a t distribution with three degrees of freedom. See Bickel (1965) and Lehmann (1983) for lucid and readable accounts. The purpose of this article is to consider a very practical and natural model within the domain of estimating a location parameter. The type of model we consider says that the observable $X_i = \theta + U_i$, where θ is an unknown parameter and U_i, ideally, should have a known distribution, say $N(0,1)$. However, due to the process generating the observations, an independent error of measurement e_i is made, so that $U_i = Z_i + e_i$, where $Z_i \sim F$ (known) and e_i has a distribution G, which may or may not be completely known. Under such an error of measurement model, we compare the mean and the median as estimators of the parameter θ. We will use the usual concept of asymptotic relative efficiency (ARE) as the basis for comparison. Formally, the following structure is considered:

$$X_i = \theta + Z_i + e_i,$$

where $Z_i \sim F, e_i \sim G, Z_i, e_i$ are independent; let $\mathcal{F}$ denote the family

$$\mathcal{F} = \{G\colon\ G \text{ is unimodal and symmetric about zero, supported on } [-m, m], m > 0 \text{ (specified)}\}. \tag{1.1}$$

We let G belong to $\mathcal{F}$, but otherwise let it be arbitrary. The 'ideal' sampling distribution F is assumed to be unimodal and symmetric about zero as well, but F may have unbounded support. We will assume F is absolutely continuous with density f, is otherwise arbitrary subject to these restrictions, but is specified. The main result gives an explicit formula for the minimum asymptotic relative efficiency of the mean with respect to the median. This formula depends on F and m, but not on G. Indeed, the minimum is over all G in the family $\mathcal{F}$. In particular, it will follow from this result as a corollary, that if an independent error of measurement gets added to a $N(\theta, \sigma^2)$ variable, then the sample mean is more efficient than the median if the error of measurement is at most 3.8σ in magnitude, and has a symmetric unimodal distribution in the interval $[-3.8\sigma, 3.8\sigma]$. Notice that in contrast to usual results on robustness, this result says something positive about the sample mean. The same phenomenon is proved for the logistic case if the error is at most 1.105σ in magnitude. The error of measurement model we consider here seems to be very applicable to many realistic situations. For a general account of the rich literature on such models, see Fuller (1987).

2. Notation, Preliminaries. Since Z and e are independent, under $\theta = 0$ the observable X is distributed as the convolution $H = F * G$ when $e \sim G$. Since F is absolutely continuous, so is H. Let $h(\cdot)$ denote the density of H. We will derive a formula for

$$e = \sup_{G \in \mathcal{F}} 4h^2(0)\sigma^2(h), \tag{2.1}$$

where $\sigma^2(h)$ denotes the variance of H. If F does not have a finite variance, then e is automatically infinite. It is clear that the reciprocal of e is the infimum of the efficiency of the mean with respect to the median. First we derive two simple formulas for $h(0)$ and $\sigma^2(h)$. The following result of Khintchine (1938) will be needed.

Theorem 2.1 (Khintchine). Let $e \sim G$, where G is symmetric and unimodal about zero, with possibly unbounded support. Then e can be represented as $e \stackrel{\mathcal{L}}{=} U.V$, where $U \sim U[-1,1], V \geq 0$, and U, V are independent. If supp $(G) \subseteq [-m, m]$ for some $m > 0$, then $0 \leq V \leq m$ with probability 1.

Definition. If $V \sim \mu$, we will call μ the mixing distribution corresponding to G.

Lemma 2.1. Let $\sigma^2 = \int z^2 dF(z)$. Then,

$$\sigma^2(h) = \sigma^2 + \frac{\int_0^m v^2 d\mu}{3},$$

where μ is the mixing distribution corresponding to G.

Proof: Obvious.

Lemma 2.2.

$$h(0) = \frac{1}{2} \cdot \int_0^m \frac{2F(v) - 1}{v} d\mu(v),$$

where μ is as in Lemma 2.1.

Proof: Since h is the density of the convolution $H = F * G$, and F is symmetric, it follows

$$\begin{aligned} h(0) &= \int f(e)dG(e) \\ &= \frac{1}{2}\int_0^m \int_{-1}^1 f(uv)dud\mu(v) \\ &= \frac{1}{2}\int_0^m \int_{-v}^v f(t)dt\frac{d\mu(v)}{v} = \frac{1}{2}\int_0^m \frac{2F(v)-1}{v}d\mu(v). \end{aligned}$$

Lemma 2.3. For given F and G, the ARE of the median with respect to the mean equals

$$e(F,G) = \left(\int_0^m \frac{2F(v)-1}{v}d\mu(v)\right)^2 \left(\sigma^2 + \frac{1}{3}\int_0^m v^2 d\mu(v)\right), \qquad (2.2)$$

where μ is the mixing distribution corresponding to G.

Proof: Immediate from Lemma 2.1 and Lemma 2.2.

Discussion. The supremum of the ARE therefore equals the supremum of the expression in (2.2) over all possible distributions μ on $[0, m]$. We will prove that under an assumption on the density f of F, the supremum is attained at a distribution supported on the set $\{0, m\}$. The probability p assigned to the point m depends on m and F. The value of p will be obtained by careful calculus. The reduction to distributions supported on $\{0, m\}$ will require some moment theory and some analysis. The assumption needed on f is the following:

Assumption A: f is twice differentiable and $\frac{f'(z)}{z}$ is nondecreasing for $z > 0$.

We will first show that Assumption A is frequently satisfied. In particular, the $N(0, \sigma^2)$ density satisfies it.

Theorem 2.2. Let f be any normal scale mixture density

$$f(z) = \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} d\nu(\sigma^2),$$

where ν is a probability measure on $\mathbb{R}^+$. Then f satisfies Assumption A.

Proof: In fact, f is infinitely differentiable. Also, by familiar use of the Dominated convergence theorem,

$$\frac{f'(z)}{z} = -\int \frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{z^2}{2\sigma^2}} d\nu(\sigma^2),$$

from which the result follows immediately.

We now introduce some notation for future use: Γ will denote the mixing distribution corresponding to the symmetric unimodal law F; also, define

$$\left.\begin{aligned} G(z) &= \int_0^z d\Gamma(t) \\ R(z) &= \int_z^\infty \tfrac{1}{t}\, d\Gamma(t) \\ \psi(z) &= \tfrac{2F(z)-1}{z} \\ g(z) &= G'(z) \end{aligned}\right\} \tag{2.3}$$

Note that by virtue of Assumption A, $G(z)$ is twice differentiable. Furthermore, it is known that $g(z) = -zf'(z)$ (see Dharmadhikari and Joag-Dev (1988)).

3. Main results. The first theorem in this section achieves the stated reduction to two point distributions supported on $\{0, m\}$. For this result, we will only need Assumption A on the density f. The second theorem builds up on the first and gives a completely explicit formula for the supremum of $e(F, G)$, defined in (2.2). For this theorem, we will need to make another assumption on the density f, again frequently satisfied, and in particular satisfied by any $N(0, \sigma^2)$ density. We will state this assumption when we need it.

Theorem 3.1. Assume F is symmetric and unimodal and that Assumption A holds. Then,

$$\sup_{G \in \mathcal{F}} e(F, G)$$
$$= \sup_{0 \le p \le 1} \left[\left(\sigma^2 + \frac{p}{3} m^2\right) \left\{ p \frac{2F(m)-1}{m} + (1-p) \cdot 2f(0) \right\}^2 \right]$$

For the proof of Theorem 3.1, we will need the following lemma.

Lemma 3.2. If Assumption A holds, then $\frac{\psi(z)-\psi(0)}{z^2}$ is nondecreasing for $z > 0$.

Proof: Clearly, the assertion holds if

$$d_1(z) = z\psi'(z) - 2\psi(z) + 2\psi[0] \ge 0 \quad \forall\, z > 0$$
$$\Leftrightarrow d_2(z) = z^2\psi'(z) - 2z\psi(z) + 2\psi(0)z \ge 0 \quad \forall\, z > 0 \tag{3.1}$$

(3.1) will be proved by showing that

$$\left.\begin{aligned} &\text{i.} \;\; d_2(0) = 0 \\ &\text{ii.} \;\; d_2'(z) \ge 0. \end{aligned}\right\} \tag{3.2}$$

Towards this end, first observe the following elementary facts:

a $\psi(z) = \frac{G(z)}{z} + R(z)$ (3.3)

b $\psi'(z)$ exists, is everywhere finite, and equals $\psi'(z) = -\frac{G(z)}{z^2}$; furthermore $\psi'(0+) = 0$

c $\psi''(z)$ exists, is everywhere finite, and equals $\psi''(z) = -\frac{g(z)}{z^2} + \frac{2G(z)}{z^3}$; furthermore, $\lim_{z \to 0+} z^2\psi''(z) = 0$

d Under Assumption A, $\frac{g(z)}{z^2}$ is nonincreasing.

Of these, **a** follows on writing F as a mixture of symmetric uniform CDF's with Γ as the mixing distribution; **b** follows on using the fact that $R(z)$ is differentiable with $R'(z) = -\frac{g(z)}{z}$ and then an application of L'Hospital's rule gives $\psi'(0+) = 0$; **c** is immediate from **b** and another application of L'Hospital's rule; **d** is a consequence of the previously stated fact $g(z) = -zf'(z)$.

Returning now to (3.2), **i** is evident and **ii** follows because

$$
\begin{aligned}
d_2'(z) &= z^2\psi''(z) - 2\psi(z) + 2\psi(0) \qquad \text{(calculus)}\\
&= -g(z) + 2\frac{G(z)}{z} - 2\frac{G(z)}{z} - 2R(z) + 2\psi(0) \text{ (from 3.3))}\\
&= 2 \cdot \int_0^\infty \frac{g(t)}{t} dt - 2\int_z^\infty \frac{g(t)}{t} dt - g(z) \text{ (from (2.3))}\\
&= 2 \cdot \int_0^z \frac{g(t)}{t} dt - g(z)\\
&= 2 \cdot \int_0^z t \cdot \frac{g(t)}{t^2} dt - g(z)\\
&\geq 2 \cdot \frac{g(z)}{z^2} \cdot \int_0^z t dt - g(z) \text{ (from (3.3))}\\
&= 0.
\end{aligned}
$$

The proof of Lemma 3.2 is now complete.

Theorem 3.3. Let $\mathcal{P}$ denote the class of all probability distributions on the interval $[0, m]$. Then, $\sup_{\mu \in \mathcal{P}} \int_0^m \psi d\mu$ subject to $\int_0^m z^2 d\mu = c$ (specified) is attained at a distribution μ_0 supported on the two point set $\{0, m\}$; furthermore, this is true for any c in the interval $0 \leq c \leq m^2$.

Proof: Consider the quadratic

$$Q(z) = az^2 + b,$$

with

$$\left.\begin{aligned} a &= \tfrac{\psi(m) - \psi(0)}{m^2},\\ b &= \psi(0) \end{aligned}\right\} \tag{3.4}$$

It follows from Lemma 3.2 that

$$\frac{\psi(z)-\psi(0)}{z^2} \le \frac{\psi(m)-\psi(0)}{m^2} = a, \text{ for all } 0 \le z \le m$$

and therefore,

$$\psi(z) \le az^2 + b, \text{ for } 0 \le z \le m, \text{ with equality at } z = 0, m.$$

Consequently, $\int_0^m \psi d\mu \le ac + b$ for any μ such that $\int_0^m z^2 d\mu = c$. If μ_0 is now chosen such that it is supported on $\{0, m\}$ and $\int_0^m z^2 d\mu_0 = c^2$ (such a probability measure μ_0 clearly exists), then evidently the equality $\int_0^m \psi d\mu_0 = ac + b$ holds and Theorem 3.3 follows.

Proof of Theorem 3.1: Immediate from Theorem 3.3 on identifying p as $\mu_0\{m\}, 1-p$ as $\mu_0\{0\}$, and on using formula (2.2) for $e(F, G)$.

Discussion. Theorem 3.1 reveals the use of moment methods for reducing an original infinite dimensional problem to a problem of maximizing a cubic polynomial in the interval $[0, 1]$. Even without any further mathematics, this is an easy numerical exercise in any given situation. However, we will actually show that the maximum value of the cubic in Theorem 3.1 can be obtained in closed form under an additional assumption on the density f. The closed form formula will require careful analytic arguments, but is worth obtaining because of its clean nature. The additional assumption we now make on f is the following:
Assumption B: Let $f(x) = p(x^2), x > 0$. Assume $x^2 p(x) \to 0$ as $x \to \infty$ and that there is at most one root x_0 of the equation

$$-\frac{p'(x)}{p(x)} = \frac{2x-5}{(x-1)(x-4)} \tag{3.5}$$

for $x > 4$; moreover, if such a root x_0 exists, then $4p(0) - (x_0^2 - 5x_0 + 4)p(x_0) \ge 0$.

Remark: Assumption B does not seem to be related to any property like monotone likelihood ratio of f. However, it is satisfied if f is a normal or a Double exponential density, and in other common situations. In the $N(0, 1)$ case, the only solution of (3.5) for $x > 4$ is $x_0 = 7$, and $4p(0) - (x_0^2 - 5x_0 + 4)p(x_0) = 1.3789 > 0$.

The following theorem gives a closed form formula for $\sup_{G \in \mathcal{F}} e(F, G)$ if Assumption B also holds. Some notation is needed.

Define

$$\left.\begin{aligned} \alpha &= \alpha(m) = 2f(0) - \tfrac{2F(m)-1}{m} \\ \beta &= 2f(0) \\ \xi &= \xi(m) = 3 \cdot \tfrac{2f(0) - \frac{2F(m)-1}{m}}{m^2 f(0)} \end{aligned}\right\} \tag{3.6}$$

Theorem 3.4. Assume F is symmetric and unimodal and that Assumptions A and B hold. Assume $\sigma^2 = \int_{-\infty}^{\infty} x^2 dF(x) = 1$. Then, $\sup_{G \in \mathcal{F}} e(F, G) = \frac{\beta^2}{\xi}(1 + \frac{\xi-1}{3})^3$ if (3.5) has a root for $x > 4$ and is $\psi^2(m) \cdot (1 + \frac{m^2}{3})$ otherwise. For a general $\sigma^2 \neq 1$, the assertion holds with $m\sigma$ in place of m.

Proof: We will describe the main steps and skip the more mechanical details.

Step 1. On expansion, the cubic

$$\begin{aligned} &\left(1 + \frac{p}{3} m^2\right) \left\{ p \frac{2F(m)-1}{m} + (1-p) 2f(0) \right\}^2 \\ &= p^3 \cdot m^2 \alpha^2 + p^2 \cdot [3\alpha^2 - 2\alpha\beta m^2] + p \cdot [\beta^2 m^2 - 6\alpha\beta] + 3\beta^2 \end{aligned} \tag{3.7}$$

Step 2. The derivative at $p = 0$ of the cubic (3.7) equals $\beta^2 m^2 - 6\alpha\beta$; algebra, elementary calculus, and the fact that $f(m) \leq f(0)$ gives $\beta^2 m^2 - 6\alpha\beta \geq 0$, implying the cubic (3.7) has a nonnegative slope at $p = 0$.

Step 3. To identify other possible values of p at which the derivative is 0, we set

$$3\alpha^2 m^2 \cdot p^2 + 2[3\alpha^2 - 2m^2\alpha\beta]p + m^2\beta^2 - 6\alpha\beta = 0; \tag{3.8}$$

on algebra, the roots of (3.8) are

$$\left.\begin{aligned} &p = \tfrac{\beta}{\alpha} \\ \text{and} \quad &p = \tfrac{m^2\beta - 6\alpha}{3\alpha m^2} = \tfrac{\beta}{3\alpha} - \tfrac{2}{m^2} \end{aligned}\right\} \tag{3.9}$$

of which $\frac{\beta}{\alpha} > 1$; hence the only possible root of (3.8) in $(0, 1]$ is $\frac{\beta}{3\alpha} - \frac{2}{m^2}$.

Step 4. From Step 2, it follows that $\frac{\beta}{3\alpha} - \frac{2}{m^2} \geq 0$. Again, on algebra and calculus, $\frac{\beta}{3\alpha} - \frac{2}{m^2} \leq 1$ if

$$(m^4 - 5m^2 + 4) f(m) \leq 4f(0), \tag{3.10}$$

failing which the derivative has no zero in $(0, 1)$. In this case, Step 1 therefore implies that the cubic is nondecreasing in the entire interval $[0, 1]$; hence the maximum is at $p = 1$. This gives the second assertion in the Theorem already. On the other hand, if the inequality

(3.10) holds, then the derivative has a zero in $(0,1]$ and because of Step 1, this therefore is the unique maxima of the cubic.

Step 5. This step shows that Assumption B precisely implies (3.10) for any $m > 0$. On using the definition of p, i.e., $f(x) = p(x^2)$, inequality (3.10) is equivalent to

$$r(y) = 4p(0) - (y^2 - 5y + 4)p(y) \geq 0. \tag{3.11}$$

It is easy to verify that $r(0) = 0$ and $r'(y) \geq 0$ for $y \leq 1$; thus for $y \leq 1$, (3.11) holds. For y between 1 and 4, it is clear that $r(y) \geq 4p(0) \geq 0$; finally, Assumption B insures that $r(y) \to 4p(0) \geq 0$ as $y \to \infty$. Thus (3.11) can be violated only if there is a finite minima of $r(y)$ at which $r(y) < 0$. Assumption B makes this exactly impossible. Thus (3.10) holds for all m and from Step 4 we therefore know that the second root $p = \frac{\beta}{3\alpha} - \frac{2}{m^2}$ is the maxima of our cubic in the interval $[0,1]$.

Step 6. Substitution of $p = \frac{\beta}{3\alpha} - \frac{2}{m^2}$ in Theorem 3.1 gives $\sup_{G \in \mathcal{F}} e(F,G) = \frac{\beta^2}{\xi}\left(1 + \frac{\xi-1}{3}\right)^3$ on rearrangement of terms and algebra.

4. Two examples. In this final section, the results of Section 3 are applied to the specific cases $F = N(0,1)$ and $F = L(0,1)$, where $L(0,1)$ stands for the standard logistic density

$$L(0,1)(x) = \frac{e^{-x}}{(1+e^{-x})^2}, -\infty < x < \infty. \tag{4.1}$$

Recall that the ARE of the median in comparison to the mean is $\frac{2}{\pi}$ in the normal and $\frac{\pi^2}{12}$ in the logistic case. One would therefore expect that up to a certain value of m, $\sup_{G \in \mathcal{F}} e(F,G)$ continues to be less than or equal to 1. We will identify this threshold value in each of the two cases.

Example 1. Let F be the $N(0,1)$ distribution. The threshold value of m such that $\sup_{G \in \mathcal{F}} e(F,G) = 1$ is easily obtained from Theorem 3.4 as follows: first by ignoring the fact that ξ depends on m, solve the equation

$$\frac{\beta^2}{\xi}\left(1 + \frac{\xi - 1}{3}\right)^3 = 1. \tag{4.2}$$

The solution is $\xi = .2791473$. Now use the definition of ξ in (3.6) to solve $\xi = .2791473$ in the variable m. The solution is $m = 3.7939$ (approximately). This is interesting. This says that if an independent measurement error gets added to a $N(0, \sigma^2)$ variate, then the mean is more efficient than the median if the error is at most 3.8σ in magnitude and has a symmetric unimodal distribution, regardless of what this distribution exactly is.

The following table gives the infimum of the efficiency of the mean in comparison to the median for selected values of m:

Table 1

m	.5	1	1.5	2	3	3.8	4	5
inf ARE	1.56988	1.56033	1.52419	1.44994	1.20980	.99852	.94925	.73637

This says that for measurement errors up to 4σ in magnitude, the mean comes out well in comparison to the median, provided of course our other model assumptions are valid.

Example 2. Let F be the logistic density defined in (4.1). The solution to (4.2) in this case is $\xi = .4456611$ and the threshold value of m such that $\sup_{G \in \mathcal{F}} e(F, G) \leq 1$ is $m = 1.105$. Notice the expected reduction in this threshold value in comparison to the normal case. Table 2 below is the analog of Table 1 for this example:

Table 2

m	.25	.5	.75	1	1.105	1.5
inf ARE	1.04787	1.04002	1.02703	1.00909	1	.95967

Thus, in the logistic case, if the measurement error is up to roughly 1.5σ in magnitude, then the mean comes out quite well in comparison to the median provided our other model assumptions are valid.

5. Summary. The measurement error model we describe here is quite natural and seems to be relevant in practical cases. The finding in the normal example is particularly positive for the sample mean. The minimax problem for this structure, as studied in other cases in Huber (1973) and others, appears to be an interesting and worthwhile problem.

References

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

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity and applications*, Academic Press, New York.

Huber, P.J. (1973). *Robust Statistics*, John Wiley, New York.

Fuller, W. (1987). *Measurement Error Models*, John Wiley, New York.

Khintchine, A.Y. (1938). On unimodal distributions, Izv. Nauchno - Isled. *Inst. Mat. Mech. Tomsk. Gos. Univ.*, **2**, 1–7.

Lehmann, E.L. (1983). *Theory of Point Estimation*, John Wiley, New York.