

Robustness of the Simultaneous Estimators of Location and Scale From Approximating a
Histogram by a Normal Density Curve

A.S. Hedayat and Guoqin Su¹

January 28, 2012

¹A.S. Hedayat is a Distinguished Professor, Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL 60607 (e-mail: hedayat@uic.edu). Guoqin Su is an expert statistician, Novartis Pharmaceuticals, East Hanover, NJ 07936 (e-mail: guoqinsu@yahoo.com). This work was partially supported by National Science Foundation (NSF) Grants DMS-0603761, DMS-0904125, and Mathematical Sciences Center at Tsinghua University, Beijing, China. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the NSF.

Abstract

The robust properties of the simultaneous estimators of location and scale parameters (μ^*, σ^*) proposed by Brown and Hwang (1993) are studied. As a pair of simultaneous M -estimators of location and scale, their asymptotic efficiencies (0.650 for μ^* and 0.541 for σ^*) are higher than those for Median (0.637) and Median Absolute Deviation (0.368) under the normal distribution. Simulation indicates that the distributions of $\sqrt{n}\mu^*$ and $\sqrt{n-1}\sigma^*$ are much flatter than those based on the sample mean and sample standard deviation under the normal distribution when the sample size is small.

Key Words: Robust estimation, Gross-error sensitivity, Asymptotic efficiency.

1. INTRODUCTION

The most exciting feature of statistics is its wide applicability in physics, chemistry, engineering, environmental science, public health science, etc. The sample mean and sample standard deviation are widely used in applications by non-statisticians. When the underlying population distribution is normal or Gaussian, the sample mean and sample variance are the uniformly minimum variance unbiased estimators for the population mean and population variance respectively. The central limit theorem is often used to justify the assumption of normality when using the sample mean and sample standard deviation. But it is inevitable that real data contain gross errors. Five to ten percent unusual values in a data set seem to be the rule rather than the exception (Hampel, 1973). The distribution of such data is no longer normal.

It has long been known that the sample mean and sample standard deviation are not robust. As an example, the confidence interval for the population mean based on student's t statistic can be extremely conservative when the underlying population distribution has heavier tails than the normal distribution. However, their non-robustness does not discourage people from using them. One reason is that they are simple to compute and easy for non-statisticians to understand.

Robust methods have existed in the literature for more than forty years. There was and is no shortage of important and exciting research on robustness (Stigler, 2010). But robust methods have not yet seen widespread application. Stahel (1991) listed three obstacles to every day use of robust estimators of mean:

1. A large number of classes of procedures have been proposed, most of which rely on a class of (ψ -, ρ -, score) functions and one or more tuning constants. Often, an auxiliary estimator of scale or other nuisance parameter is also needed.
2. A method for estimating its variability as a basis for tests or confidence intervals is needed, since a point estimate alone is rarely very useful.
3. There is a lack of easy-to-use computer programs.

The first obstacle is the main obstacle for the lack of widespread application of robust procedures. The variability of an estimator usually depends on the scale or other nuisance parameters, which is usually not estimated simultaneously. The Princeton robustness study (Andrews, et al., 1972) considered 68 estimators of location for symmetric distributions. Even for the same robust procedure, there may be disagreement about its performance among researchers (Stigler, 1977). Often researchers adjust the tuning constant and the estimator of scale to improve a robust procedure. As an example, the Tukey biweight estimator of location is an M -estimator with a tuning constant c and an estimator s for the scale in the estimating equation. Mosteller and Tukey (1977, P. 205) recommended a value of c such that (cs) is between 4σ and 6σ in the normal case. Kafadar (1982) chose $c = 6$ while Du Mond and Lenth (1987) selected $c = 4.69$. Gross (1976) used the MAD (median absolute deviation) for s and Kafadar (1982) used an adjusted MAD, which itself contains a tuning constant, while Stigler (1977) adjusted the estimator of scale on each iteration. Many choices for the tuning constant and the estimator of scale make it difficult to convince people, especially non-statisticians, to use the procedure. For widespread application of robust procedures, more work should be done on the following (Stahel, 1991):

1. developing and studying robust procedures which are motivated by heuristic arguments and are easy for practitioners to understand and to accept.
2. guidance for the choices of tuning constants and estimates of nuisance parameters.

A robust location estimate without an estimate of scale is not satisfactory in applications.

Brown and Hwang (1993) proposed a graphical procedure which provides simultaneous estimators of the location and the scale parameters for a normal distribution. Their graphical procedure is based on approximating a sample histogram by a normal density curve. For a given random sample, the usual sample histogram is rescaled so that it encloses an area of one. The normal density curve which best approximates the sample histogram is the normal curve that minimizes the integrated squared deviation between the histogram and the normal curve.

This best approximating normal density curve provides a visually satisfying fit. Further, it converges to a limiting normal density curve as the bin width of the sample histogram goes to zero. Therefore, this limiting normal density curve provides a normal density curve that is a good fit for any histogram drawn from the data, so long as the bin width is not too large. What is more interesting is that the mean and the standard deviation of the limiting normal density curve can be regarded as estimators of the location and the scale parameters respectively for the normal population distribution. These estimators of the location and the scale parameters have the following three virtues:

1. They are simultaneous estimators of the location and the scale and do not need any tuning constants.
2. The normal density curve based on the estimators of the location and the scale is usually a good fit for histograms drawn for the sample, so long as the bin width is not too large.
3. They are highly robust estimators of the location and the scale.

In applications, it is very common for practitioners to draw a sample histogram. The first virtue gives practitioners a clear-cut way of attaching a normal density curve to the sample histogram. The second virtue guarantees the *visual* satisfaction of the attached normal density curve. The third virtue guarantees some *theoretic* good properties of the attached normal density curve. The robustness of the simultaneous estimators were pointed out by Brown and Hwang (1993). They compared the performance of the estimators of the location with the sample mean by simulating samples of sizes 20, 30, and 50 from the standard Cauchy distribution. For simultaneous estimators of the location and the scale, we will compute the following measures of robustness when the population distribution is normal:

1. their efficiencies and influence functions, and
2. their quantiles. These quantiles are compared with the asymptotic quantiles for the purpose of constructing confidence intervals using the asymptotic quantiles.

In applications, relying on the central limit theorem, the normality distribution is the default distribution when the underlying data are continuous. Moreover, data transformations may be considered if lack of normality of the data is suspected. In addition to point estimates, it is important to provide confidence intervals to reflect the precision of estimates. The asymptotic variances may be used to construct confidence intervals for the location μ and the scale σ using

the normal or χ^2 approximation when the sample size is large. The quantiles will provide important information about the sample size for which these approximations are appropriate.

2. APPROXIMATING A HISTOGRAM BY A NORMAL DENSITY CURVE

Let x_1, \dots, x_n be a random sample from an unknown distribution and $\xi_0 < \xi_1 < \dots < \xi_M$ the $M + 1$ endpoints of M bins which cover the sample range. Denote by b_j the width of the j -th bin $(\xi_{j-1}, \xi_j]$ and by n_j the number of data points in the j -th bin. A scaled histogram $h(t)$ given by

$$h(t) = \frac{n_j}{nb_j}, \text{ if } \xi_{j-1} < t \leq \xi_j, \quad j = 1, \dots, M, \quad (1)$$

has the property that the area under $h(t)$ is one.

Remark 1 The scaled histogram $h(t)$ defined here does not require that all bins are of equal width and therefore is an extension of the scaled histogram defined by Brown and Hwang (1993).

Let $g(t)$ denote a probability density function satisfying the condition:

$$\int_{-\infty}^{\infty} g(t)^2 dt < \infty, \quad (2)$$

and define

$$D(\mu, \sigma) = \int [\phi_{\mu, \sigma}(t) - g(t)]^2 dt, \quad (3)$$

where $\phi_{\mu, \sigma}(t)$ denotes the normal density with mean μ and standard deviation σ . Condition (2) assures that $D(\mu, \sigma)$ is finite for all μ and σ .

Theorem 1 *When $g(t)$ is a scaled histogram (1), any values (μ, σ) that minimize $D(\mu, \sigma)$ satisfy:*

$$\sum_{j=1}^M \frac{n_j}{b_j} [\phi_{\mu, \sigma}(\xi_j) - \phi_{\mu, \sigma}(\xi_{j-1})] = 0, \quad (4)$$

$$4\sigma\sqrt{\pi} \sum_{j=1}^M \frac{n_j}{nb_j} [(\xi_j - \mu) \phi_{\mu, \sigma}(\xi_j) - (\xi_{j-1} - \mu) \phi_{\mu, \sigma}(\xi_{j-1})] = 1. \quad (5)$$

As $b = b(M) := \max(b_1, \dots, b_M) \rightarrow 0$, there is a convergent subsequence of (μ_b, σ_b) . Let (μ^*, σ^*)

denote the limit of such a subsequence. Then (μ^*, σ^*) satisfy

$$\sum_{i=1}^n (x_i - \mu^*) \phi_{\mu^*, \sigma^*}(x_i) = 0, \quad (6)$$

$$\frac{4\sigma^* \sqrt{\pi}}{n} \sum_{i=1}^n \left(1 - \frac{(x_i - \mu^*)^2}{\sigma^{*2}}\right) \phi_{\mu^*, \sigma^*}(x_i) = 1. \quad (7)$$

If the bins are of equal length, Theorem 1 reduces to Corollary 3.1 and Theorem 4.1 given by Brown and Hwang (1993). Theorem 1 can be proved similarly. The values (μ_b, σ_b) depend on the bin widths b_j 's. However, as the maximum bin width b decreases, the convergence limit (μ^*, σ^*) only depends on the data. Once μ^* and σ^* are computed directly from the data, the corresponding normal density curve is expected to provide a satisfactory fit to the histogram as long as the maximum bin width b is not too large.

Remark 2 (A) Equations (6) and (7) imply that $x_{(1)} \leq \mu^* \leq x_{(n)}$ and $\sigma^* < 2(x_{(n)} - x_{(1)})$, where $x_{(1)} = \min\{x_1, \dots, x_n\}$ and $x_{(n)} = \max\{x_1, \dots, x_n\}$. This means that the solution (μ^*, σ^*) to Equations (6) and (7) cannot stray off to infinity (Brown and Hwang, 1993).

(B) When there is a unique solution to Equations (6) and (7), there is a unique convergence limit for (μ_b, σ_b) as the maximum bin width b goes to 0.

(C) Theoretically, the solution to Equations (6) and (7) may not be unique. Equations (6) and (7) can be re-written as follows:

$$\mu^* = \frac{\sum_{i=1}^n x_i \exp\left\{-\frac{(x_i - \mu^*)^2}{2\sigma^{*2}}\right\}}{\sum_{i=1}^n \exp\left\{-\frac{(x_i - \mu^*)^2}{2\sigma^{*2}}\right\}}, \quad (8)$$

$$\sigma^{*2} = \frac{\sum_{i=1}^n (x_i - \mu^*)^2 \exp\left\{-\frac{(x_i - \mu^*)^2}{2\sigma^{*2}}\right\}}{\sum_{i=1}^n \exp\left\{-\frac{(x_i - \mu^*)^2}{2\sigma^{*2}}\right\} - \frac{n}{\sqrt{8}}}. \quad (9)$$

These equations can be solved by inserting old values on the right side to get new values on the left side iteratively until the old values and new values are identical. Practically, there is no convergence problem using the above equations to iteratively solve μ^* and σ^* .

(D) Equations (6) and (7) in fact define a pair of *simultaneous M-estimators of location and scale* (see Section 3.) with the following ψ - and χ - functions:

$$\psi(x; \mu, \sigma) = (x - \mu) \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad (10)$$

$$\chi(x; \mu, \sigma) = 2\sqrt{2} \left[1 - \frac{(x - \mu)^2}{\sigma^2} \right] \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] - 1. \quad (11)$$

(E) When σ^* is replaced by the sample standard deviation, Equation (6) defines an M -estimator of location called *mean likelihood estimate* and denoted by MEL in the Princeton Study (Andrews et al., 1972, 2C3(d)).

3. INFLUENCE FUNCTIONS, EFFICIENCIES, AND GROSS-ERROR SENSITIVITY

It is pointed out in Remark 2 (D) that Equations (6) and (7) define a pair of simultaneous M -estimators of location and scale with the corresponding ψ - and χ - functions given in (10–11) respectively. M -estimators (or maximum likelihood type estimators) were first introduced by Huber (1964). They form a very large class of estimators, which includes the maximum likelihood estimators. We are going to introduce briefly the M -estimators and a few properties (see Hampel, *et. al.* (1986), and Huber and Ronchetti (2009) for more detailed information).

Definition 1 Let x_1, \dots, x_n be a random sample from a population with density

$$\frac{1}{\sigma} f \left(\frac{x - \mu}{\sigma} \right), \quad \sigma > 0, \quad -\infty < \mu < +\infty, \quad (12)$$

where μ is called the location parameter and σ the scale parameter. Any pair of statistics (T_n, S_n) determined by two implicit equations

$$\sum_{i=1}^n \psi \left(\frac{x_i - T_n}{S_n} \right) = 0 \quad (13)$$

$$\sum_{i=1}^n \chi \left(\frac{x_i - T_n}{S_n} \right) = 0, \quad (14)$$

is called simultaneous M -estimators of location and scale.

The distribution associated with the density (12) is denoted by $F_{\mu, \sigma}$. Let $F(x)$ be a cumulative distribution function. Define the functionals $T(F)$ and $S(F)$ to be the solutions of following two equations:

$$\int \psi \left(\frac{x - T(F)}{S(F)} \right) dF(x) = 0 \quad (15)$$

$$\int \chi \left(\frac{x - T(F)}{S(F)} \right) dF(x) = 0. \quad (16)$$

If F_n denotes the empirical distribution function, then $T_n = T(F_n)$ and $S_n = S(F_n)$. The concepts in robust statistics are more convenient and more often to be defined in terms of the functionals $T(F)$ and $S(F)$. If

$$T(F_{\mu,\sigma}) = \mu \quad \text{and} \quad S(F_{\mu,\sigma}) = \sigma \quad \text{for all } \mu, \sigma, \quad (17)$$

then the functionals $T(F_{\mu,\sigma})$ and $S(F_{\mu,\sigma})$ are said to be *Fisher consistent*. This means that the estimators $\{T_n, S_n, n \geq 1\}$ asymptotically measures the right quantities.

Definition 2 *The influence function of the functional $T(F)$ at a distribution F is given by*

$$IF(x; T, F) = \lim_{h \downarrow 0} \frac{T((1-h)F + h\delta_x) - T(F)}{h} \quad (18)$$

The influence function of the functional $S(F)$ is defined similarly and denoted as $IF(x; S, F)$.

The interpretation of influence functions is that they describe the effects of an infinitesimal contamination at the point x on the estimators, standardized by the mass of the contamination.

Definition 3 *The gross-error sensitivity of estimators $T(F)$ and $S(F)$ at a distribution F is given by*

$$\gamma_u^*(T, S, F) = \sup_x \sqrt{IF(x; T, F)^2 + IF(x; S, F)^2} \quad (19)$$

The gross-error sensitivity measures the worst influence which a small amount of contamination can have on the values of the estimators. Finite gross-error sensitivity is a desirable feature.

Theorem 2 *Under the normality assumption, the simultaneous M -estimators μ^* and σ^* are Fisher consistent. Further, their influence functions are*

$$IF(x; \mu^*, N(\mu, \sigma^2)) = 2\sqrt{2}(x - \mu) \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}, \quad (20)$$

$$IF(x; \sigma^*, N(\mu, \sigma^2)) = \frac{2\sigma}{3} \left[1 - 2\sqrt{2} \left(1 - \frac{(x - \mu)^2}{\sigma^2} \right) \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \right], \quad (21)$$

The gross-error sensitivity $\gamma_u^ = 2.01\sigma$, the asymptotic covariance matrix of $\sqrt{n}\mu^*$ and $\sqrt{n}\sigma^*$*

is

$$V(\mu^*, \sigma^*, N(\mu, \sigma^2)) = \sigma^2 \begin{bmatrix} \frac{8}{3\sqrt{3}} & 0 \\ 0 & \frac{4(16-3\sqrt{3})}{27\sqrt{3}} \end{bmatrix}. \quad (22)$$

Thus, the asymptotic efficiency is 0.65 for the location and 0.541 for the scale.

See Appendix A for the proof. Laubscher, *et. al.* (1994) provided the asymptotic variances for μ^* and σ^* in their letter to the editor, although they did not provide a detailed derivation. The asymptotic efficiency for the location is slightly higher than that (.637) of the sample median (Hampel, *et. al.*, 1986). However, the asymptotic efficiency for the scale is much higher than that (0.368) of the MAD (47% more efficient). Hence, as a pair of simultaneous M -estimators of location and scale, (μ^*, σ^*) is more efficient than (Med, MAD).

4. PERCENTAGE POINTS OF (μ^*, σ^*) FOR THE NORMAL DISTRIBUTION

Theorem 3 *The joint probability density of $\frac{\mu^*(X_1, \dots, X_n) - \mu}{\sigma}$ and $\frac{\sigma^*(X_1, \dots, X_n)}{\sigma}$ does not depend on μ and σ when the data X_1, \dots, X_n is a random sample from a population with density (12),*

The theorem follows from the fact that $\mu^*([X_1 - c]/d, \dots, [X_n - c]/d) = [\mu^*(X_1, \dots, X_n) - c]/d$ and $\sigma^*([X_1 - c]/d, \dots, [X_n - c]/d) = \sigma^*(X_1, \dots, X_n)/|d|$, where c and $d(\neq 0)$ are constant, and the joint distribution of $\frac{X_1 - \mu}{\sigma}, \dots, \frac{X_n - \mu}{\sigma}$ does not depend on μ and σ when the data X_1, \dots, X_n is a random sample from a population with density (12).

For constructing the confidence intervals for μ , we will use $\frac{\mu^*(X_1, \dots, X_n) - \mu}{\sigma}$ and estimate σ by σ^* . It's obvious that the probability density of $\frac{\mu^*(X_1, \dots, X_n) - \mu}{\sigma^*}$ does not depend on μ and σ under a population with density (12). The quantiles of $\frac{\mu^*(X_1, \dots, X_n) - \mu}{\sigma^*}$ and $\frac{\sigma^*(X_1, \dots, X_n)}{\sigma}$ will be used to construct confidence intervals for μ and σ respectively. Unfortunately, their theoretical values are very difficult to derive even under the normal distribution. We will obtain these quantiles by simulation when X_1, \dots, X_n is a random sample from the normal population $N(\mu, \sigma^2)$. In fact, we will simulate the quantiles of $\sqrt{n} \frac{\mu^*(X_1, \dots, X_n) - \mu}{\sigma^*(X_1, \dots, X_n)}$ and $\sqrt{n-1} \frac{\sigma^*(X_1, \dots, X_n)}{\sigma}$ when the sample is from the normal distribution $N(\mu, \sigma^2)$ to compare with distributions of sample mean and standard deviation. Since these distributions do not depend on the parameters μ and σ , we

only need to simulate the quantiles of $\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)}$ and $\sqrt{n-1} \sigma^*(X_1, \dots, X_n)$ when the sample is from the standard normal distribution.

5. MONTE CARLO SWINDLES

A Monte Carlo swindle represents any mathematical/statistical trick which is used in a Monte Carlo simulation either to reduce the effort or to improve the precision, or both. Monte Carlo swindles were greatly used in the robust estimation of location in the Princeton robustness study (Andrews, et al., 1972). Gross (1973) and Simon (1976) made more systematic studies of Monte Carlo swindles. The idea of Monte Carlo swindles is explained here.

Assume that the problem is to estimate $\theta = E(X)$, which is difficult to calculate analytically. Usually, X does not have an explicit expression. As an example, X is an M -estimator of location and is obtained by solving an equation iteratively. The parameter θ could be anything, but it must be expressed as the expectation of some random variable. A simple method for estimating θ is to generate N independent observations X_1, \dots, X_N and estimate θ by $\bar{X} = \sum_{i=1}^N X_i/N$. Clearly, $E(\bar{X}) = \theta$ and $\text{Var}(\bar{X}) = \text{Var}(X)/N$. Next, assume that $\theta = E(X)$ can be rewritten as

$$\theta = E(X) = E(E(X|W)) = E(q(W)), \quad (23)$$

where $q(W) = E(X|W)$. Then

$$\text{Var}(X) = \text{Var}(E(X|W)) + E(\text{Var}(X|W)) = \text{Var}(q(W)) + E(\text{Var}(X|W)) \geq \text{Var}(q(W)). \quad (24)$$

If $q(W_1), \dots, q(W_N)$ is a random sample of $q(W)$, averaging X_1, \dots, X_N to estimate θ will be inferior to averaging $q(W_1), \dots, q(W_N)$. The usual situation in which this swindle can be applied is that $E(X|W)$ can be computed with the same degree of difficulty as X .

We will simulate the bias of $\sigma^*(X_1, \dots, X_n)$, the variances of $\sqrt{n} \mu^*(X_1, \dots, X_n)$ and $\sqrt{n} \sigma^*(X_1, \dots, X_n)$, the quantiles of $\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)}$ and $\sqrt{n-1} \sigma^*(X_1, \dots, X_n)$ under the standard normal distribution.

Theorem 4 *Let X_1, \dots, X_n be a random sample of size n from the standard normal distribution and denote the sample mean and sample standard deviation as \bar{X} and S respectively. The sample is standardized using the mean and standard deviation as $R_k = \frac{X_k - \bar{X}}{S}, k = 1, \dots, n$.*

(i) $\mu^*(X_1, \dots, X_n)$ is an unbiased estimator of the population mean.

(ii) The following equalities hold:

$$\text{Var}(\sqrt{n}\mu^*(X_1, \dots, X_n)) = 1 + nE(\mu^{*2}(R_1, \dots, R_n)) \quad (25)$$

$$E\sigma^*(X_1, \dots, X_n) = \frac{\sqrt{2\pi} E\sigma^*(R_1, \dots, R_n)}{B(\frac{n-1}{2}, 0.5)\sqrt{n-1}} \quad (26)$$

$$\text{Var}(\sqrt{n}\sigma^*(X_1, \dots, X_n)) = n \left[E\sigma^{*2}(R_1, \dots, R_n) - \frac{2\pi(E\sigma^*(R_1, \dots, R_n))^2}{B^2(\frac{n-1}{2}, 0.5)(n-1)} \right] \quad (27)$$

$$\begin{aligned} \Pr\left(\sqrt{n}\frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)} \leq x\right) &= .5[Et(x\sigma^*(R_1, \dots, R_n) - \sqrt{n}\mu^*(R_1, \dots, R_n); n-1) \\ &\quad + Et(x\sigma^*(R_1, \dots, R_n) + \sqrt{n}\mu^*(R_1, \dots, R_n); n-1)] \end{aligned} \quad (28)$$

$$\Pr(\sqrt{n-1}\sigma^*(X_1, \dots, X_n) \leq x) = E\left(\chi^2\left(\frac{x^2}{\sigma^{*2}(R_1, \dots, R_n)}; n-1\right)\right), \quad (29)$$

where $B(p, q)$ denotes the beta function, $t(x; n)$ the cumulative distribution function of a t distribution with n degrees of freedom, and $\chi^2(x; n)$ the cumulative distribution function of a χ^2 distribution with n degrees of freedom.

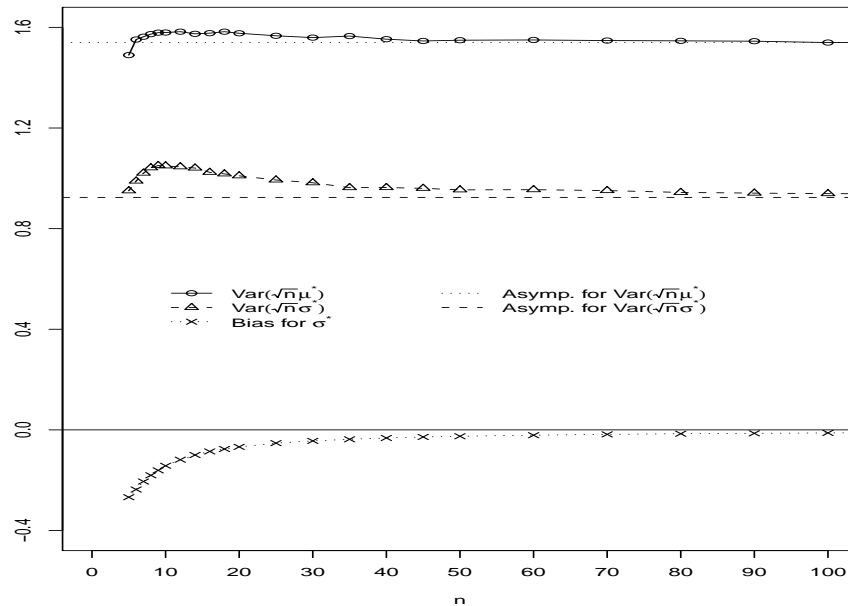
See Appendix B for the proof. Parameters of X_1, \dots, X_n on the left side can be considered as expectations conditional on R_1, \dots, R_n similar to Equation (23). Therefore, they can be estimated in simulation by generating random samples of R_1, \dots, R_n and the right side to reduce the variance according to Equation (24).

6. SIMULATION RESULTS

For each sample sizes of 5 to 10, 12 to 20 by 2, 25 to 50 by 5, 60 to 100 by 10, 200, 300, 400, 500, 1000, we generated 20,000 random samples from the standard normal distribution to estimate:

- Variances of $\sqrt{n}\mu^*$ and $\sqrt{n}\sigma^*$ and bias of σ^* (Figure 1),
- Quantiles of $\sqrt{n}\mu^*(X_1, \dots, X_n)/\sigma^*(X_1, \dots, X_n)$ (Figure 2),
- Quantiles of $\sqrt{n-1}\sigma^*(X_1, \dots, X_n)$ (Figure 3)

Figure 1: Variances of $\sqrt{n}\mu^*$ and $\sqrt{n}\sigma^*$ and bias of σ^* for the standard normal distribution



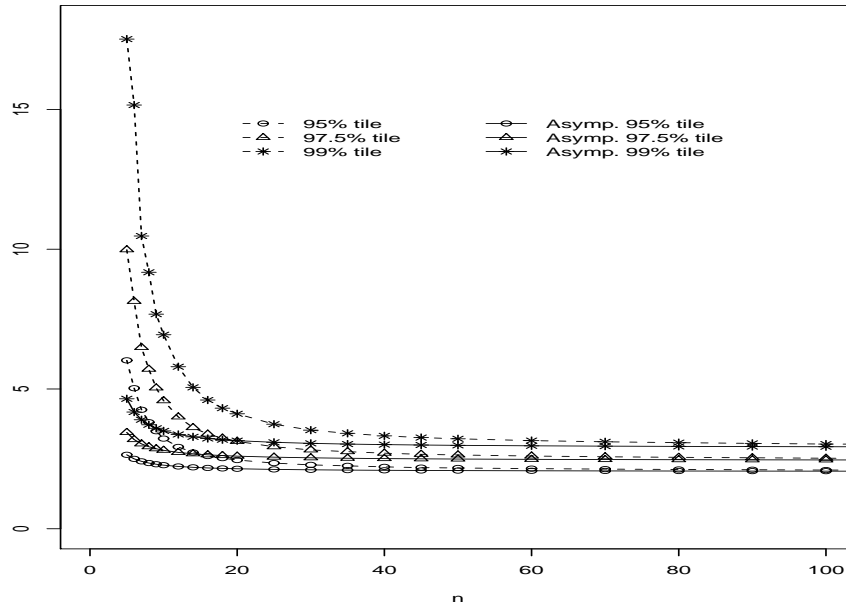
for the standard normal distribution. The results for sample sizes of 200, 300, 400, 500, and 1000 are similar to those for sample size of 100 and are omitted from these plots for better visual illustration. In the simulation, each random sample was standardized first using the mean and standard deviation. The right sides of Equations (25) to (29) were then applied to the standardized samples to reduce the simulation variance. Quantiles were obtained by linear interpolating several well-chosen cumulative probabilities. The simulation was carried out using R2.6.2 on an IBM compatible personal computer.

From Figure 1, it can be seen that

- Variances of $\sqrt{n}\mu^*$ and $\sqrt{n}\sigma^*$ decrease and approach the asymptotic values (Equation 22) in general as the sample size increases.
- When the sample size is small, variance of $\sqrt{n}\mu^*$ is only at most 2.6% higher than its asymptotic variance while the variance of $\sqrt{n}\sigma^*$ may be 13.7% higher than its asymptotic variance.
- σ^* underestimates the standard deviation, especially when the sample size is small. However, the bias becomes smaller and smaller as the sample size increases.

Figure 1 confirms the asymptotic variances for μ^* and σ^* in Equation (22) and that σ^* is an

Figure 2: Quantiles of $\sqrt{n} \mu^* / \sigma^*$ for the standard normal distribution

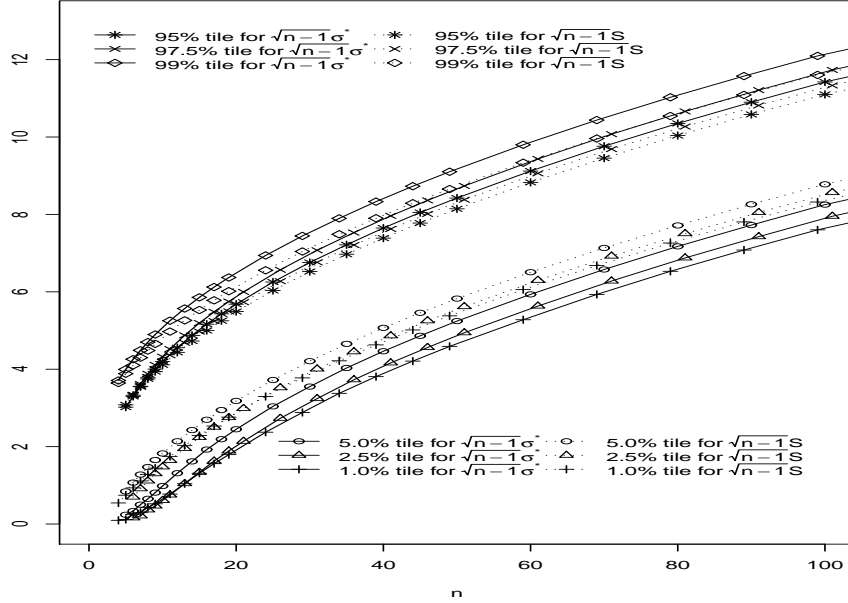


asymptotically unbiased estimator for σ .

The asymptotic quantiles in Figure 2 were obtained by multiplying the corresponding quantiles of the t distribution with $n - 1$ degrees of freedom and the asymptotic standard deviation of $\sqrt{n} \mu^*$ (1.24). When the sample size is large (≥ 60), the simulated quantiles are close to the asymptotic quantiles, which indicates that the distribution of $\sqrt{n} \mu^* / \sigma^*$ may be approximated by a t distribution with an asymptotic standard deviation factor of 1.24. When the sample size is small, the simulated quantiles are much larger than the asymptotic quantiles, which means that the distribution is much flatter than its asymptotic distribution. This may be due to the fact that σ^* underestimates the standard deviation when the sample size is small.

Quantiles of $\sqrt{n-1} \sigma^*(X_1, \dots, X_n)$ are presented in Figure 3 (solid curves). For comparison, quantiles of $\sqrt{n-1} S$ (square root of χ^2 distribution with $n - 1$ degrees of freedom) are also presented in Figure 3 (dotted curves). Similar to the quantiles of $\sqrt{n-1} S$ under normality, these quantiles for $\sqrt{n-1} \sigma^*$ increase as the sample size n increases. The distribution of $\sqrt{n-1} \sigma^*$ is flatter (longer tail) than the distribution of $\sqrt{n-1} S$ since its quantiles are smaller at the lower tails (1%, 2.5%, and 5%) and larger at the upper tails (95%, 97.5%, and 99%) compared to those for $\sqrt{n-1} S$.

Figure 3: Quantiles of $\sqrt{n-1}\sigma^*$ for the standard normal distribution



Acknowledgment: We are greatly indebted to the Editor, Associate Editor, and the Referees whose constructive suggestions and comments helped us to prepare a far better version of the original submission.

APPENDICES

A PROOF OF THEOREM 2

Assume that $\psi-$ and $\chi-$ are given as in (10) and (11), denote $T(F)$ and $S(F)$, the solutions of Equations (15) and (16), by $\mu^*(F)$ and $\sigma^*(F)$ (or μ^* and σ^* if there is no confusion about the underlying distribution), respectively. For *Fisher consistency*, we will show that when F is the normal distribution $N(\mu, \sigma^2)$, $\mu^*(F) = \mu$ and $\sigma^*(F) = \sigma$. First, we have

$$\int (x - \mu^*(F)) \exp \left\{ -\frac{(x - \mu^*(F))^2}{2\sigma^{*(F)2}} \right\} dF(x) = 0, \quad (30)$$

$$2\sqrt{2} \int \left(1 - \frac{[x - \mu^*(F)]^2}{\sigma^{*(F)2}} \right) \exp \left\{ -\frac{[x - \mu^*(F)]^2}{2\sigma^{*(F)2}} \right\} dF(x) = 1. \quad (31)$$

Since the normal density function is of the type in (12), we first assume that F has the density (12).

$$\begin{aligned} \int (x - \mu^*) \exp \left\{ -\frac{(x - \mu^*)^2}{2\sigma^{*2}} \right\} \frac{1}{|a|\sigma} f \left(\frac{x - (a\mu + b)}{|a|\sigma} \right) &= 0, \\ 2\sqrt{2} \int \left(1 - \frac{(x - \mu^*)^2}{\sigma^{*2}} \right) \exp \left\{ -\frac{(x - \mu^*)^2}{2\sigma^{*2}} \right\} \frac{1}{|a|\sigma} f \left(\frac{x - (a\mu + b)}{|a|\sigma} \right) &= 1, \end{aligned}$$

are equivalent to

$$\begin{aligned} \int \left(x - \frac{\mu^* - b}{a} \right) \exp \left\{ -\frac{\left[\frac{x - \mu^* - b}{a} \right]^2}{2(\sigma^*/|a|)^2} \right\} \frac{1}{\sigma} f \left(\frac{x - \mu}{\sigma} \right) &= 0, \\ 2\sqrt{2} \int \left(1 - \frac{\left[\frac{x - \mu^* - b}{a} \right]^2}{(\sigma^*/|a|)^2} \right) \exp \left\{ -\frac{\left[\frac{x - \mu^* - b}{a} \right]^2}{2(\sigma^*/|a|)^2} \right\} \frac{1}{\sigma} f \left(\frac{x - \mu}{\sigma} \right) &= 1. \end{aligned}$$

This means that

$$\frac{\mu^*(F_{a\mu+b,|a|\sigma}) - b}{a} = \mu^*(F_{\mu,\sigma}) \quad \text{and} \quad \frac{\sigma^*(F_{a\mu+b,|a|\sigma})}{|a|} = \sigma^*(F_{\mu,\sigma}),$$

hence, $\mu^*(F_{\mu,\sigma})$ and $\sigma^*(F_{\mu,\sigma})$ are equivariant estimators. Now assume that the underlying distribution is normal $N(\mu, \sigma^2)$. After simple algebra, we have

$$\begin{aligned} &\exp \left(-\frac{(x - \mu^*)^2}{2\sigma^{*2}} \right) \cdot \exp \left(-\frac{(x - \mu)^2}{2\sigma^2} \right) \\ &= \exp \left(-\frac{\sigma^{*2} + \sigma^2}{2\sigma^{*2}\sigma^2} \cdot \left[x - \frac{\mu\sigma^{*2} + \mu^*\sigma^2}{\sigma^{*2} + \sigma^2} \right]^2 \right) \cdot \exp \left(-\frac{(\mu^* - \mu)^2}{2(\sigma^2 + \sigma^{*2})} \right). \end{aligned} \quad (32)$$

Substituting $dF(x)$ by $\phi_{\mu,\sigma}(x) dx$ in Equation (30) and using (32), we obtain

$$\mu^* = \frac{\mu\sigma^{*2} + \mu^*\sigma^2}{\sigma^{*2} + \sigma^2},$$

hence, $\mu^* = \mu$. Substituting $dF(x)$ by $\phi_{\mu,\sigma}(x) dx$ in Equation (31) and using (32), we obtain

$$2\sqrt{2} \frac{\sigma^{*3}}{(\sigma^2 + \sigma^{*2})^{3/2}} = 1,$$

that is, $\sigma^* = \sigma$.

Next, we are going to find the influence functions, the gross-error sensitivity, and the asymptotic covariance matrix under the normal distribution $N(\mu, \sigma^2)$. First, we recall a few facts for

these robust concepts (Hampel, *et. al.* [1986]; Huber and Ronchetti [2009]):

- The influence functions of the functional $T(F)$ and $S(F)$ under a distribution F are

$$IF(x; T, F_{\mu, \sigma}) = \frac{\psi\left(\frac{x-T}{S}\right) S}{\int \psi'\left(\frac{t-T}{S}\right) \frac{1}{\sigma} f\left(\frac{t-\mu}{\sigma}\right) dt} \quad (33)$$

$$IF(x; S, F_{\mu, \sigma}) = \frac{\chi\left(\frac{x-T}{S}\right) S}{\int \chi'\left(\frac{t-T}{S}\right) \frac{t-T}{S} \frac{1}{\sigma} f\left(\frac{t-\mu}{\sigma}\right) dt} \quad (34)$$

- It is usually true that the asymptotic covariance matrix of $\sqrt{n}T(F_n)$ and $\sqrt{n}S(F_n)$ is

$$V(T, S, F) = \begin{bmatrix} \int IF(x; T, F)^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx & 0 \\ 0 & \int IF(x; S, F)^2 \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx \end{bmatrix}, \quad (35)$$

Let ψ - and χ - be the functions given in (10–11) respectively, and $F_{\mu, \sigma}$ be the Normal distribution. $T(F) = \mu$ and $S(F) = \sigma$, we have

$$\begin{aligned} \int \psi'\left(\frac{t-\mu}{\sigma}\right) \phi_{\mu, \sigma}(t) dt &= \int \left[1 - \frac{(t-\mu)^2}{\sigma^2}\right] \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \\ &= \int \left[1 - \frac{(t-\mu)^2}{\sigma^2}\right] \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{\sigma^2}\right] dt = \frac{1}{2\sqrt{2}} \end{aligned}$$

Plugging this into (33) leads to (20).

$$\begin{aligned} &\int \chi'\left(\frac{t-\mu}{\sigma}\right) \frac{t-\mu}{\sigma} \phi_{\mu, \sigma}(t) dt \\ &= -2\sqrt{2} \int \left[3 - \left(\frac{t-\mu}{\sigma}\right)^2\right] \left(\frac{t-\mu}{\sigma}\right)^2 \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \\ &= \frac{-2\sqrt{2}}{\sigma^4} \int [3\sigma^2 - (t-\mu)^2] (t-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{\sigma^2}\right] dt = -\frac{3}{2} \end{aligned}$$

Plugging this into (34) leads to (21). It is obvious that we only need to evaluate the gross-error sensitivity and the asymptotic covariance under the standard normal distribution.

$$\begin{aligned} &\sqrt{IF(x; \mu^*, N(0, 1))^2 + IF(x; \sigma^*, N(0, 1))^2} \\ &= \sqrt{8x^2 \exp(-x^2) + \frac{4}{9} \left[1 - 2\sqrt{2}(1-x^2) \exp(-x^2/2)\right]^2} \end{aligned}$$

reaches a maximum of 2.01 at $x^2 = 1.888$. Hence $\gamma_u^* = 2.01\sigma$. The asymptotic covariance matrix

(22) follows from:

$$\begin{aligned} & \int IF(x; \mu^*, N(0, 1))^2 \phi_{0,1}(x) dx = \int 8x^2 \exp(-x^2) \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = \frac{8}{3\sqrt{3}} \\ & \int IF(x; \sigma^*, N(0, 1))^2 \phi_{0,1}(x) dx \\ &= \frac{4}{9} \int \left[1 - 2\sqrt{2}(1-x^2) \exp(-\frac{x^2}{2}) \right]^2 \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = \frac{4(16-3\sqrt{3})}{27\sqrt{3}} \end{aligned}$$

The asymptotic efficiency is $\frac{3\sqrt{3}}{8} = 0.65$ for the location μ^* and $0.5 \times \frac{27\sqrt{3}}{4(16-3\sqrt{3})} = 0.541$ for the scale σ^* .

B PROOF OF THEOREM 4

- (i) Since $\mu^*(-X_1, \dots, -X_n) = -\mu^*(X_1, \dots, X_n)$ and $\{-X_1, \dots, -X_n\}$ have the joint probability distribution as $\{X_1, \dots, X_n\}$, the distribution of $\mu^*(X_1, \dots, X_n)$ is symmetric about 0. Hence, $E\mu^*(X_1, \dots, X_n) = 0$ and $\mu^*(X_1, \dots, X_n)$ is an unbiased estimator of the population mean.
- (ii) Under the normality distribution, $\{R_1, \dots, R_n\}$, \bar{X} , and S are independent following Basu's theorem (Lehmann, 1983). Equation (25) follows from

$$\begin{aligned} \text{Var}(\sqrt{n}\mu^*(X_1, \dots, X_n)) &= nE([\bar{X} + S\mu^*(R_1, \dots, R_n)]^2) \\ &= n\left(E(\bar{X}^2) + 2E[S\bar{X}\mu^*(R_1, \dots, R_n)] + E[S^2\mu^{*2}(R_1, \dots, R_n)]\right) \\ &= 1 + nE[\mu^{*2}(R_1, \dots, R_n)]. \end{aligned}$$

Assume that $\{x_{i1}, \dots, x_{in}\}$ ($i = 1, \dots, N$) are N random samples of size n from the standard normal distribution. Let \bar{x}_i and s_i be the sample mean and sample standard deviation of the i -th random sample $\{x_{i1}, \dots, x_{in}\}$ and $r_{ik} = \frac{x_{ik} - \bar{x}_i}{s_i}$, $k = 1, \dots, n$, $i = 1, \dots, N$. Then, for estimating $\text{Var}(\sqrt{n}\mu^*(X_1, \dots, X_n))$,

$$1 + \frac{n}{N} \sum_{i=1}^N \mu^{*2}(r_{i1}, \dots, r_{in})$$

is a better estimator than

$$\frac{n}{N} \sum_{i=1}^N \mu^{*2}(x_{i1}, \dots, x_{in}).$$

Equation (26) follows from

$$\begin{aligned}
\mathbb{E}(\sigma^*(X_1, \dots, X_n)) &= \mathbb{E}(S\sigma^*(R_1, \dots, R_n)) \\
&= \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})\sqrt{n-1}} \mathbb{E}(\sigma^*(R_1, \dots, R_n)) \\
&= \frac{\sqrt{2\pi} \mathbb{E}\sigma^*(R_1, \dots, R_n)}{B(\frac{n-1}{2}, 0.5)\sqrt{n-1}}.
\end{aligned}$$

Equation (27) can be derived similarly.

$$\begin{aligned}
&\Pr\left(\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)} \leq x\right) \\
&= \mathbb{E}\left[\Pr\left(\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)} \leq x \mid R_1, \dots, R_n\right)\right] \\
&= \mathbb{E}\left[\Pr\left(\frac{S\mu^*(R_1, \dots, R_n) + \bar{X}}{S\sigma^*(R_1, \dots, R_n)} \leq \frac{x}{\sqrt{n}} \mid R_1, \dots, R_n\right)\right] \\
&= \mathbb{E}\left[\Pr\left(\frac{\sqrt{n}\bar{X}}{S} \leq x\sigma^*(R_1, \dots, R_n) - \sqrt{n}\mu^*(R_1, \dots, R_n) \mid R_1, \dots, R_n\right)\right] \\
&= \mathbb{E}\left[t(x\sigma^*(R_1, \dots, R_n) - \sqrt{n}\mu^*(R_1, \dots, R_n); n-1)\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Pr\left(\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)} \geq -x\right) \\
&= \mathbb{E}\left[t(x\sigma^*(R_1, \dots, R_n) + \sqrt{n}\mu^*(R_1, \dots, R_n); n-1)\right].
\end{aligned}$$

Since the distribution of $\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)}$ is symmetric about 0, the cumulative probability $\Pr\left(\sqrt{n} \frac{\mu^*(X_1, \dots, X_n)}{\sigma^*(X_1, \dots, X_n)} \leq x\right)$ is

$$\begin{aligned}
&.5 [\mathbb{E}(t(x\sigma^*(R_1, \dots, R_n) - \sqrt{n}\mu^*(R_1, \dots, R_n); n-1)) \\
&+ \mathbb{E}(t(x\sigma^*(R_1, \dots, R_n) + \sqrt{n}\mu^*(R_1, \dots, R_n); n-1))],
\end{aligned}$$

and Equation (28) follows. Equation (29) follows from

$$\begin{aligned}
&\Pr(\sqrt{n-1}\sigma^*(X_1, \dots, X_n) \leq x) \\
&= \mathbb{E}\left[\Pr(\sqrt{n-1}S\sigma^*(R_1, \dots, R_n) \leq x \mid R_1, \dots, R_n)\right] \\
&= \mathbb{E}\left[\Pr\left((n-1)S^2 \leq \frac{x^2}{\sigma^{*2}(R_1, \dots, R_n)} \mid R_1, \dots, R_n\right)\right]
\end{aligned}$$

$$= \text{E} \left[\chi^2 \left(\frac{x^2}{\sigma^{*2}(R_1, \dots, R_n)}; n-1 \right) \right].$$

REFERENCES

- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, NJ: Princeton University Press.
- Brown, L.D., and Hwang, J.T.G. (1993), "How to Approximate a Histogram by a Normal Density", *American Statistician*, 47, 251–255.
- Du Mond, C.E., and Lenth, R.V. (1987), "A Robust Confidence Interval for Location", *Technometrics*, 29, 211–219.
- Gross, A.M. (1973), "A Monte Carlo Swindle for Estimators of Location", *Applied Statistics*, 22, 347–353.
- Gross, A.M. (1976), "Confidence-Interval Robustness with Long-Tailed Symmetric Distribution", *Journal of the American Statistician Association*, 71, 409–416.
- Hampel, F.R. (1973), "Robust estimation: A condensed partial survey", *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 27, 87–104.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics. The Approach Based on Influence*, New York, NY: Wiley.
- Huber, P.J. (1964), "Robust estimation of a location parameter", *Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics*, 2nd ed., Hoboken, NJ: John Wiley & Sons.
- Kafadar, K. (1982), "A Biweight Approach to the One-Sample Problem", *Journal of the American Statistician Association*, 77, 416–424.
- Laubsche, N.F., Hjort, N. L., Jones, M. C., Brown, L.D., and Hwang, J.T.G. (1994), "Letters to the Editor", *American Statistician*, 77, 351–354.
- Lehmann, E.L. (1983), *Theory of Point Estimation*, New York, NY: Springer-Verlag
- Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison-Wesley Publishing Company.

- Simon, G. (1976), "Computer Simulation Swindles with Applications to estimates of Location and Dispersion", *Applied Statistics*, 25, 266-275.
- Stahel, W.A. (1991), "Research Directions in Robust Statistics", in *Directions in Robust Statistics and Diagnostics*, ed. Stahel, W., and Weisberg, S., New York, NY: Springer-Verlag, 243–278.
- Stigler, S.M. (1977), "Do Robust Estimators Work With Real Data?", *Annals of Statistics*. 5, 1044-1078.
- Stigler, S.M. (2010), "The Changing History of Robustness", *American Statistician*. 64, 277-281.