International Journal of Business and Statistical Analysis
ISSN (2384-4663)
Int. J. Bus. Stat. Ana. 2, No. 1 (Jan-2015)

Extended Mean with Distribution-Free Variance

A. M. Kandil[1] and T. Hamza[1]

[1]*Department of Statistics, Faculty of Commerce, Benha University*

Received 30 October 2014, Revised 18 November 2014, Accepted 19 November 2014, Published 1 January 2015

Abstract: Location estimation is one of the basic activities in statistical data analysis so considerable effort has been put into the development of procedures for the robust estimation of measures of location. Because the distribution-free variance of most of existing measures is difficult to obtain in closed form, these measures work under strong modelling assumptions. We propose a robust location measure in which the expectation of a lower order statistics is replaced by the expectation of a larger order statistics. The main attraction of this measure is that its distribution-free variance is obtained in closed form. Comparisons with some of the best location estimators, mean, Hodges-Lehmann estimator, Huber's M-estimator and median are given based on Monte Carlo simulations. Computationally, the new estimator has an explicit expression and requires no iteration.

Keywords: Huber's M-estimators; L-statistics; mean; median; order Statistics.

1. Introduction

Robust alternative to the arithmetic mean for estimating location have a history going back at least to Laplace, see Stigler (1973). Using normal contamination model, Tukey (1960) dramatically demonstrated how little efficient the mean can become when contamination increases, also showed that alternative location estimators such as the median can achieve higher asymptotic efficiency than the mean. As a result, considerable effort has been put into the development of procedures for the robust estimation of measures of location where the statisticians have felt aware of the need for robust procedures, in the sense of procedures that remain good when the assumed model does not quite fit; see, for example, Staudte and Sheather (1990) and Huber (1981) for a comprehensive accounts of these developments.

Estimators that avoid such pitfalls do exist. Two good choices are Huber's M-estimator and Hodges-Lehmann estimator. The former minimizes $\sum_i p(x_i - \theta)$ when p is chosen to be quadratic for small values and linear for larger one where θ is the location parameter; see, Huber (1981) and the latter is the median of $0.5n(n+1)$ pairwise averages $0.5(X_i + X_j)$ $(1 \leq i < j \leq n)$; see, Staudte and Sheather (1990). These estimators maintain good efficiency from small to large sample sizes for a wide range of symmetric distributions but exact variances and distribution-free of these variances are not directly available for these estimators. Consequently, robust location estimators which have distribution- free variance are desirable.

In this article, we propose a robust location measure in the class of symmetric distributions in which the expectation of a lower order statistics is replaced by the expectation of a larger order statistics. This measure has a number of advantages: its exact variance and distribution-free of this variance are available in closed form and requires no iteration.

In Section 2 we define the proposed estimator. In Section 3 we obtain exact and distribution-free variances of the proposed estimator. Robustness criteria, simulation study and comparisons with other methods are given in Section 5.

2. Extended mean

Let $X_1, \dots, X_n$ be an iid sample of size n from a continuous symmetric distribution with cumulative distribution function $F_X(.)$, density function $f_X(.)$ and quantile function $x(F)$, $0 < F < 1$ and let $X_{1:n}, \dots, X_{n:n}$ denote the corresponding order statistics. The population mean is defined as

Email: abdelfattah.kandil@gmail.com, tamersaad_5@yahoo.com

$$\mu = E(X) = E(X_{1:1}) = \int_0^1 x(F)dF$$

We define population extended mean (EM) in terms of higher order statistics as

$$\mu^t = E(X_{t+1:2t+1}), \quad t = 0,1,2,\ldots$$

as the expectation of $X_{r:s}$, $r = 1,2,\ldots s$, can be written as

$$E(X_{r:s}) = \frac{s!}{(r-1)!\,(s-r)!}\int_0^1 x(f)F^{r-1}(1-F)^{s-r}dF$$

see; for example, David (1981), we may re-express μ^t as

$$\mu^t = \frac{\Gamma(2t+2)}{\Gamma^2(t+1)}\int_0^1 x(f)F^t(1-F)^t dF, \quad t = 1,2,\ldots.$$

Important special case is $t = 0$, ordinary population mean μ, and we would emphasis that μ^t is weighted by the cumulative distribution functions F^t and $(1-F)^t$ and this will give different weights for observations $\hat{F}^t(1-\hat{F})^t/n$ while μ is weighted by 1 which gives equal weights for all the observations $(1/n)$.

We study EM in details for values of $t = 1$ and 2 only where the distribution-free variances of sample EM are obtained in simple form and could be computationally handle easily. In this case we have the following expressions for population EM

$$\mu^1 = E(X_{2:3}) = 6\int_0^1 x(F)dF$$

and

$$\mu^2 = E(X_{3:5}) = 30\int_0^1 x(F)F^2(1-F)^2 dF$$

Observe that μ^t may be defined even if $\mu = \mu^0$ is not defined, for example, Cauchy distribution μ^0 does not exist while μ^1 is defined.

2.1 Sample extended mean

We now consider estimators of population EM which are linear function of order statistics $X_{1:n},\ldots,X_{n:n}$ of a random sample $X_1, X_2, \ldots, X_n$ of size n from the population. Let us define the random variable

$$Z_{tn} = \sum_{i=1}^{n}(i-1)^{(t)}(n-i)^{(t)}X_{i:n}$$

From Sillito (1951) and Downton (1966) it is straightforward to prove that

$$E(Z_{tn}) = \sum_{i=1}^{n}(i-1)^{(t)}(n-i)^{(t)}E(X_{i:n}) = t!^2\binom{n}{2t+1}E(X_{t+1:2t+1})$$

where $n^{(r)} = n(n-1)\ldots(n-r+1)$.

We could rewrite $E(Z_{tn})$ as

$$E(X_{t+1:2t+1}) = \sum_{i=1}^{n} w_i(t,n)E(X_{i:n})$$

where

$$w_i(t,n) = \frac{(2t+1)!\,(i-1)^{(t)}(n-i)^{(t)}}{(t!)^2 n^{(2t+1)}}$$

and $\sum_i w_i(t,n) = 1$.

From $E(X_{t+1:2t+1})$ we could estimate $E(X_{t+1:2t+1})$ as

$$\hat{E}(X_{t+1:2t+1}) = \sum_{i=1}^{n} w_i(t,n)x_{i:n}$$

which is an unbiased estimator of $E(X_{t+1:2t+1})$.

Hence, we define the sample EM using μ^t and $\hat{E}(X_{t+1:2t+1})$ to be

$$\hat{\mu}^t = E(X_{t+1:2t+1}) = \sum_{i=1}^{n} w_i(t,n)x_{i:n}$$

Which is clearly an unbiased estimator of μ^t for fixed t.

In particular, an unbiased estimator of μ^1 and μ^2 are

$$\hat{\mu}^1 = \sum_{i=1}^{n} w_i(1,n)x_{i:n} \text{ and } \hat{\mu}^2 = \sum_{i=1}^{n} w_i(2,n)x_{i:n}$$

The weights $w_i(t,n)$ are plotted in Figure 1 for $t = 0,1,2$ and 3 and sample size $n = 25$. From this Figure we see that when t = 0 the weights are all equal, t > 0 the weights decrease from a maximum for median value to zero for 2t extreme sample values.

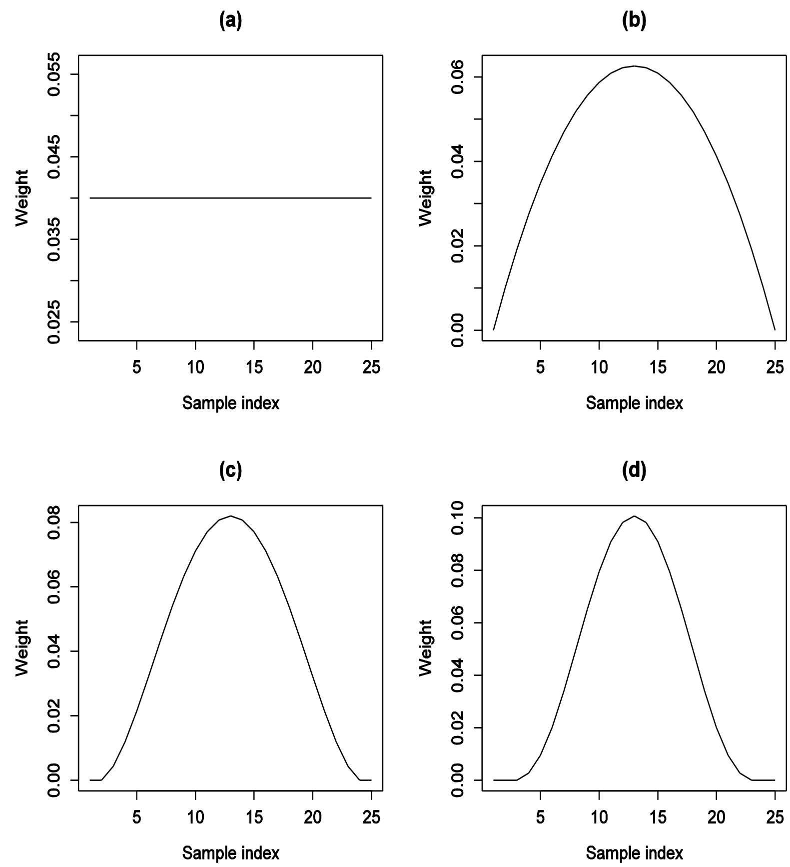

Figure1: The weights $w_i(t,n)$ for (a) $t=0$, (b) $t=1$, (c) $t=2$, (d) $t=3$ of sample size $n=25$.

3 Exact variance of sample extended mean

To calculate the exact variance of sample EM we use the following well-known expressions for the first and second moments of order statistics

$$E(X_{i:n}) = \frac{n!}{(i-1)!\,(n-i)!}\int_0^1 x(F)F^{i-1}(1-F)^{n-i}dF,$$

$$E(X^2{}_{i:n}) = \frac{n!}{(i-1)!\,(n-i)!}\int_0^1 x^2(F)F^{i-1}(1-F)^{n-i}dF,$$

and

$$E(X_{i:n}X_{j:n}) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\int_0^1\int_0^G x(F)x(G)F^{i-1}(G-F)^{j-i-1}(1-G)^{n-j}dFDG$$

To evaluate the exact variance of sample EM we note that sample EM is linear combinations of order statistics and their particular form allows us to use the method of Elamir and Seheult (2004) and Downton (1966) to obtain the exact variances in terms of expectations, variances and covariances of order statistics from samples of fixed size which do not depend on the actual sample size n as

$$var(\hat{\mu}^t) = E(\hat{\mu}^t)^2 - (E(\hat{\mu}^t))^2$$

where

$$E(\hat{\mu}^t)^2 = \sum_{r=0}^{t}\sum_{s=0}^{t}\binom{t}{r}\binom{t}{s} t^{(t-r)} t^{(t-s)} (t+r)!\,(t+s)!\binom{n}{2t+r+s+2} E(X^2_{t+r+1:2t+r+s+1}) + 2\sum_{r=0}^{t}\sum_{s=0}^{t}(-1)^{r+s}\binom{t}{r}\binom{t}{s}(n-t-r-1)^{(t-r)}(n-t-s-1)^{(t-s)}(t+r)!\,(t+s)!\binom{n}{2t+r+s+2} E\left(X_{t+r+1,t+r+2:2t+r+s+2}\right) / t!^4\binom{n}{2t+1}^2$$

and

$$(E(\hat{\mu}^t))^2 = E^2(X_{t+1:2t+1})$$

It should be noted that the computational effort required to obtain the variances and covariances is considerably less than that required using the complete sample. In particular, simple expressions of asymptotic variances of $\hat{\mu}^1$ could be obtain from normal distribution as

$$(n-1.18)(n-1.59)/(n-1)(n-2)$$

4 Distribution-free unbiased estimator

Expression for exact variance of sample EM is only useful when we know the exact form of the population distribution of X as we have seen in Section 3. In this section, we show how to find distribution-free unbiased estimator of variance of sample EM from a general class of symmetric distributions. To obtain an unbiased estimator for the variance of $\hat{\mu}^t$ we must find an unbiased estimators of $E(X^2_{t+r+1:2t+r+s+1})$ and $E\left(X_{t+r+1,t+r+2:2t+r+s+2}\right)$.

First, from Elamir and Seheult (2004) we find an unbiased estimators of $E(X^2_{t+r+1:2t+r+s+1})$ and $E\left(X_{t+r+1,t+r+2:2t+r+s+2}\right)$ from any continuous distribution are given by

$$\hat{E}(X^2_{t+r+1:2t+r+s+1}) = \sum_{i=1}^{n}\frac{(i-1)^{(t+r)}(n-i)^{(t+s)}}{(t+r)!\,(t+s)!\binom{n}{2t+r+s+1}} x^2_{i:n}$$

and

$$\hat{E}\left(X_{t+r+1,t+r+2:2t+r+s+2}\right) = \sum_{1<i<j}\sum_{<n}\frac{(i-1)^{(t+r)}(n-j)^{(t+s)}}{(t+r)!\,(t+s)!\binom{n}{2t+r+s+2}} x_{i:n}x_{j:n}$$

But biased estimator of $E^2(X_{t+1:2t+1})$ is given by

$$\hat{E}^2(X_{t+1:2t+1}) = \left[\sum_{i=1}^{n}\frac{(i-1)^{(t)}(n-i)^{(t)}(2t+1)!}{t!^2\, n^{(2t+1)}} x_{i:n}\right]^2$$

Second to obtain an unbiased estimator of $E^2(X_{t+1:2t+1})$ it could re-express it in terms of cross product of order statistics as follows. From David (1981) it found that

$$E(X_{r:r})E(X_{s:s}) = \sum_{j=0}^{s-1}(-1)^j\frac{r!\,s!}{(r+1+j)!\,(s-1-j)!}E\left(X_{r,r+1:r+1+j}\right) + \sum_{j=0}^{r-1}(-1)^j\frac{r!\,s!}{(s+1+j)!\,(r-1-j)!}E\left(X_{s,s+1:s+1+j}\right)$$

and

$$E(X_{i:n}) = \sum_{j=i}^{n}(-1)^{j-i}\binom{n}{j}\binom{j-1}{i-1}E(X_{j:j})$$

From Sillito (1966) we could write $E(X_{i:n})E(X_{s:m})$ as

$$E^2(X_{t+1:2t+1}) = \sum_{j=t+1}^{2t+1}\sum_{s=t+1}^{2t+1}(-1)^{j-t-1}(-1)^{s-t-1}\binom{2t+1}{j}\binom{j-1}{t}\binom{2t+1}{s}\binom{s-1}{t}\left[\sum_{m=0}^{s-1}\frac{(-1)^m j!\,s!\,E(X_{j,j+1:j+1+m})}{(j+1+m)!\,(s-1-m)!} + \sum_{m=0}^{j-1}\frac{(-1)^m j!\,s!\,E(X_{s,s+1:s+1+m})}{(s+1+m)!\,(j-1-m)!}\right]$$

which gives us an unbiased estimator of $E^2(X_{t+1:2t+1})$ by substituting an unbiased estimators of $E(X_{j,j+1:j+1+m})$ and $E(X_{s,s+1:s+1+m})$

In particular, the distribution-free variance of $\hat{\mu}^1$ is

$$\hat{V}(\hat{\mu}^1) = \frac{36}{n}\left[\sum_i\sum_{<j}\frac{2(i-1)(n-j)}{n^{(4)}}\left[\frac{6((i-2)+(n-j-1))}{(n-4)} - 4 - \frac{9(i-2)(n-j-1)}{(n-4)(n-5)}\right]x_{i:n}x_{j:n} + \sum_{i=1}^{n}\frac{(i-1)^{(2)}(n-i)^{(2)}}{n^{(5)}}x_{i:n}^2\right]$$

which are an unbiased estimator to $V(\hat{\mu}^1)$, very easy to compute nd does not need any modelling assumption.

In large sample the asymptotic normality of $\hat{\mu}^1$ follows directly from the results of Hosking (1990) and Stigler (1974) where we have

$$\hat{\mu}^1 - \mu^1 \sim N(0, v_1)$$

where v_1 is the variance of $\hat{\mu}^1$.

Quantile normal plots of sampling distribution of $\hat{\mu}^1$ and $\hat{\mu}^2$ are given in Figure 2 from normal and double exponential distributions for sample size $n = 25$ which show that the normal approximation is quite good for $\hat{\mu}^1$ and $\hat{\mu}^2$.

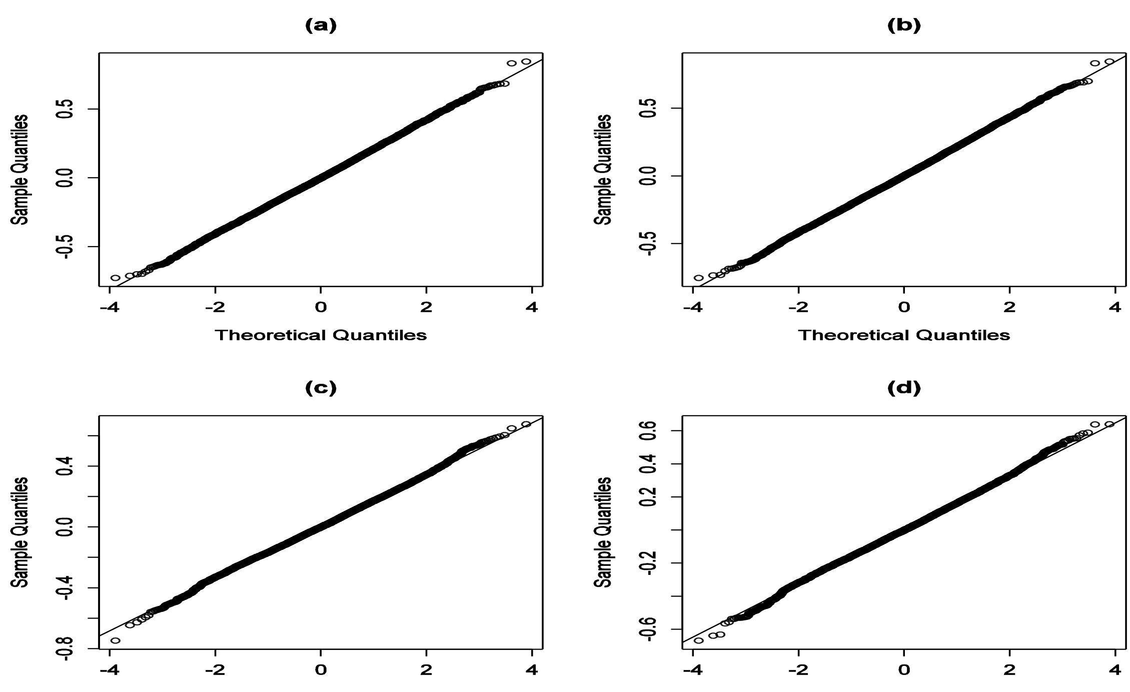

Figure 2 Quantile normal plots of sampling distribution of $\hat{\mu}^1$ and $\hat{\mu}^2$ from normal and double exponential distributions for sample size $n = 25$

5 Finite sample performance

In this section we perform a Monte Carlo study to assess the performance of the proposed estimator in finite sample cases. We use the breakdown point, stylized sensitivity curve and simulation to compare sample EM with sample mean, Huber's estimator, Hodges-Lehmann estimator and median from small sample sizes.

5.1 Robustness Criteria

We use the breakdown point and stylized sensitivity curve as a robust criteria to compare sample CM with other measures. The breakdown point is the minimum proportion of contaminated points in a sample that make the estimator unbounded; see, for example, Hoaglin et al. (1983) and Hampel et al. (1986). Following Donho and Huber (1983), the finite sample breakdown point of estimator T_n as

$$\varepsilon_n(T_n) = \frac{1}{n}\min(m; \max sup|T_n(z_1,\dots,z_n)| = \infty)$$

Where $x_1,\dots,x_n$ is a sample and $z_1,\dots,z_n$ is $x_1,\dots,x_n$ with the m values replaced by the contamination values $y_1,\dots,y_m$.

Then the infinite sample breakdown point is $\varepsilon_\infty(T_n) = lim_{n\to\infty}\varepsilon_n(T_n)$. For example if T_n is the sample mean, then replacing x_1 with $m = 1$ contamination points $y_1 = -\infty$ and $y_1 = \infty$ would imply that $|T_n(z_1,\dots,z_n)| = |T_n(y_1,x_2,\dots,x_n)| = \infty$. Thus, finite sample breakdown point of the mean is $1/n$ and its infinite sample breakdown point is zero. Another example, If T_n is the median of a sample of size n, then at least $m = \lfloor(n+1)/2\rfloor$ contamination points to bring the median higher or lower than any value, therefore, $\varepsilon_n(T_n) = \lfloor(n+1)/2\rfloor$ and $\varepsilon_\infty(T_n) = 1/2$.

For our estimator the breakdown point is $\varepsilon_n(T_n) = (t+1)/n$ which is high in small samples; for example, when $n = 10$, the breakdown point will be 0.20 and 0.30 for $t = 1$ and 2, respectively, but in large samples it will be small although the sample EM is still protect against few outliers. However, we could obtain 0.5 breakdown point by increasing t to $[\frac{n}{2}]$ but in this case it is not easy to find a distribution- free variance.

While the breakdown point is a measure of how much contamination an estimator can tolerate before becoming meaningless, the sensitivity curve describes the effect of a single contamination point on the estimator. Given the sample $x_1,\dots,x_n$, the sensitivity curve of estimator T_n is defined as

$$SC_n(x) = n[T_n(x_1,\dots,x_{n-1};x) - T_{n-1}(x_1,\dots,x_{n-1})]$$

where x is the value of a arbitrary which shows the effect on an estimate of adding or deleting an observation and when x_i is the expected values of order statistics or $x_i = F^{-1}((i-0.5))/n$ the $SC_n(x)$ is called stylized sensitivity curve; see, for example Andrew et al. (1972), and we expect the robust estimator to be bounded.

Figures 2 and 3 give stylized sensitivity curves of mean, sample conceptual mean and median from normal distribution with sample sizes 25 and 100and show that the stylized sensitivity curve of $\hat{\mu}^1$ and $\hat{\mu}^2$ is bounded like median but it is not bounded for mean.

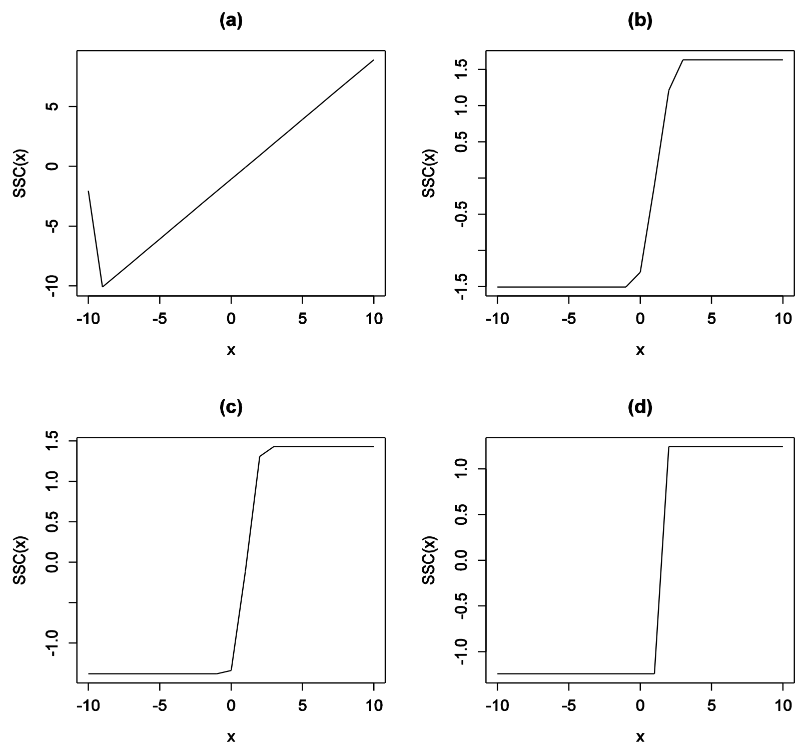

Figure 3 Stylized sensitivity curves from normal distribution with mean 0 and variance 1 of (a) $\hat{\mu}^0$, the sample mean (b) $\hat{\mu}^1$, (c) $\hat{\mu}^2$ and (d) the median using $n = 25$.

5.2 Simulation study

We perform a Monte Carlo study to assess the performance of the conceptual mean in finite sample cases. In order to investigate how it competes with some of the best location estimators sample mean $\bar{x}$, Huber's estimator, Hodges-Lehmann and median estimators as competitors, see Huber (1981). The Hodges-Lehmann estimator is the median of $n(n-1)/2$ pairwise averages $0.5(X_i + X_j)$ $i < j$.

Random samples are drawn from four distributions: normal and t-distribution with 5 degrees of freedom with sample sizes $n = 10, 20, 35, 50$ and 100. For each underlying distribution and value of n, 10000 random samples were generated.

The mean, extended mean with $t = 1$ and 2, Huber's estimator and the Hodges-Lehmann estimator were computed and the estimated variances of these estimators were then obtained based on these random samples. The simulation results are summarized in Table 1.

When the underlying distribution is normal, the extended mean work as good as other robust estimators, for example, $\hat{\mu}^1$ has asymptotic efficiency about 0.95 to sample mean, and as good as Huber's and Hodges-Lehmann estimators even when n is small and much better than median. In t_5 distribution EM is competitive to Huber's estimator and Hodges-Lehmann estimator and better than mean and median.

Table 1 Estimated variances and efficincies of mean, EM, median (med.), Hodges-Lehmann (H.L.) and Huber (Hub) from standard normal distribution (Norm.) and t-distribution with 5 degrees of freedom for sample sizes $n = 10, 20, 35, 50$ and 100 and number of replications 10000.

	n				
Measure	10	20	35	50	100
Mean	0.10119	0.04986	0.02823	0.01993	0.00986
$\hat{\mu}^1$	0.11059	0.05329	0.03017	0.02106	0.01045
Eff.	0.91	0.93	0.93	0.94	0.94
$\hat{\mu}^2$	0.12009	0.05662	0.03194	0.02224	0.01102
Eff	0.84	0.88	0.88	0.89	0.89
Med.	0.14032	0.07349	0.04430	0.03046	0.01565
Eff.	0.72	0.68	0.64	0.065	0.63
H.L.	0.10923	0.05282	0.02992	0.02087	0.01035
Eff.	0.93	0.94	0.94	0.95	0.95
Hub.	0.10788	0.05196	0.02953	0.02065	0.01022
Eff.	0.94	0.96	0.95	0.96	0.96
Mean	0.1685	0.08378	0.04758	0.03377	0.01657
Eff	0.85	0.84	0.81	0.81	0.81
$\hat{\mu}^1$	0.14149	0.06921	0.03820	0.02726	0.01328
Eff	1	1.02	1.01	1.01	1.01
$\hat{\mu}^2$	0.14494	0.06986	0.03829	0.02728	0.01326
Eff.	0.96	1	1.01	1.01	1.01
Med	0.16278	0.08521	0.04980	0.03438	0.01725
Eff	0.88	0.83	0.78	0.80	0.78
H.L.	0.14486	0.07042	0.03844	0.02743	0.01329
Eff.	0.99	1	1	1	1
Hub	0.14354	0.07055	0.03857	0.02758	0.01342

5.3 Example

We consider Cushny and Peebles data which is given in Staudte and Sheather (1990). The data show the differing effects of optical isomers of hyoscyamine hydrobromide in producing sleep. We compare sample conceptual mean $(t = 1)$ with mean and Huber's estimators which are given by Staudte and Sheather (1990) under normality assumption as

$$\bar{x} = 1.58 \quad \text{and} \quad se(\bar{x}) = 0.39$$

and

$$\hat{\omega} = 1.36 \quad \text{and} \quad se(\hat{\omega}) = 0.22$$

Where $\hat{\omega}$ is Huber’s estimate.

The conceptual sample mean which does not require any distribution assumption is given by

$$\hat{\mu}^1 = 1.363 \quad \text{and} \quad se(\hat{\mu}^1) = 0.18$$

If we compare with mean and Huber's estimator we find that $\hat{\mu}^1$ has the the same value as Huber's estimator and has standard error less than the sample mean and Huber’s estimator.

6 Conclusion

We defined a measure of location from a class of symmetric distributions in terms of order statistics in which the expectation of lower order statistics is replaced by the expectation of larger order statistics. The main attraction of sample EM is that the exact and distribution-free variances are obtained in closed form and easy to compute. The simulation study showed that the sample EM is very competitive to other robust estimators in both small and large sample sizes from the distributions we have studied.

References

Andrews D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey. Robust Estimates of Location. Princeton University Press, Princeton, NJ, 1st edition, 1972.

Cherno H., J. L. Gastwith, and M. V. Johns. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. Annals of Mathematical Statistics, 38:52_72, 1967.

David H.A. Order Statistics. Wiley: New York, 2nd edition, 1981.

Donho D.L. and P.J. Huber. The notion of a breakdown point. In P.J.Bickel, K.A. Doksum, and J.L. Hodges, editors, Fetschrift for Erich L.Lehmann, pages 131-138. Wadsworth, Belmont, CA, 1983.

Downton F. Linear estimates with polynomial coe_cients. Biometrika,53:129_141, 1966.

Elamir E.A.H. and A.H. Seheult. Exact variance structure of sample L-moments. Journal of Statistical Planning and Inference, 2004.

Hampel F.R., E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York, NY, 1st edition, 1986.

Hoaglin D.C., F. Mosteller, and J. W. Tukey. Understanding Robust and Exploratory Data Analysis. John Wiley and Sons, New York, NY, 1st edition, 1983.

Hosking J.R. L-moments: Analysis and estimation of distributions using linear combinations of statistics. Journal of the Royal Statistical Society B, 52:105-124, 1990.

Huber P.J. Robust Statistics. New York: Wiley, 1st edition, 1981.

Sillito G.P.. Interrelations between certain linear systematic statistics of samples from any continuous distribution. Biometrika, 38:377-382, 1951.

Staudte R.G. and S.J. Sheather. Robust Estimation and Testing. New York: Wiley, 1st edition, 1990.

Stigler S.M. Simon newcombe, perrcy daniell and the history of robust estimation 1885-1920. Journal of the American Statistical Association, 68:872-879, 1973.

Stigler S.M. Linear functions of order statistics with smooth weight function. Annals of Statistics, 2:676_693, 1974.

Tukey J.W. A survey of sampling from contaminated distributions. In Olkin et al. (Ed.), Contributions to probability and statistics, Stanford University Press, 1:448-485, 1960.