www.istatistikciler.org
İstatistikçiler Dergisi 3 (2010) 8-16

İstatistikçiler Dergisi

Comparing measures of location when the underlying distribution has heavier tails than normal

Abdullah Fırat Özdemir

Dokuz Eylul University Department of Statistics
firat.ozdemir@deu.edu.tr

Abstract

In this study, two conventional (mean and median) and three robust (20% trimmed mean, one-step M-estimator and modified one-step M-estimator) measures of location are compared in terms of their asymptotic relative efficiencies and mean squared error when the underlying distribution is contaminated normal. When n=20, one-step M-estimator was best in five, modified one-step M-estimator was best in three and 20% trimmed mean was best in one situation, when n=40, one-step M-estimator and modified one-step M-estimator were best in four, 20% trimmed mean was best in one sampling situation covered.

Keywords: Contaminated normal distribution; Trimmed mean; One-step M-estimator; Modified one-step M-estimator; Asymptotic relative efficiency; Mean squared error.

Özet

Normal dağılımdan daha ağır kuyruklara sahip dağılımlarda konum ölçülerinin karşılaştırılması

Bu çalışmada iki geleneksel (aritmetik ortalama ve ortanca) ve üç dayanıklı (%20'lik budanmış ortalama, tek-adım M-tahmincisi, düzeltilmiş tek-adım M-tahmincisi) konum ölçüsü, bozulmuş normal dağılım'dan türetilen veriler kullanılarak asimptotik göreli etkinlik ve hata kareler ortalaması bakımından karşılaştırılmıştır. n=20 iken tek-adım M-tahmincisi beş defa, düzeltilmiş tek-adım M-tahmincisi üç defa ve %20'lik budanmış ortalama bir defa, n=40 iken tek-adım M-tahmincisi ve düzeltilmiş tek-adım M-tahmincisi dört defa, %20'lik budanmış ortalama bir defa en iyi konum ölçüsü olarak gözlenmiştir.

Anahtar sözcükler: *Bozulmuş normal dağılım; Budanmış ortalama; Tek-adım M-tahmincisi; Düzeltilmiş tek-adım M-tahmincisi; Asimptotik göreli etkinlik; Hata kareler ortalaması.*

1. Introduction

Standard methods for analyzing data and comparing groups are based on the assumption that observations are randomly sampled from a normally distributed population. When the normality assumption is violated and the null hypothesis H_0 is true, the most commonly used methods in statistics based on sample means seem to perform quite well in most situations. But if the alternative hypothesis H_1 is likely to be true, there are situations where violating the normality assumptions can still be ignored, but there are also situations where violating the normality assumption can be very serious because of loss of power.

The normal distribution is the most important distribution in all of statistics. But it can fail in terms of approximating the distribution of any continuous distribution. An aphorism given by Cramer has stated that "Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact" [2]. The normal distribution implies that only two numbers are required to tell us everything about the probabilities associated with a random variable: the population mean μ and population variance σ^2. Moreover, an assumption of normality implies that distributions must be symmetric.

Normal distribution may provide good approximation to most distributions that arise in practice. But unfortunately, empirical investigations indicate that departures from normality, that have practical importance, are rather common in applied work [4, 6, 9, 12]. In particular, distributions can be highly skewed, they can have heavy tails and random samples often have outliers. Outliers and heavy tailed distributions are serious practical problems because they inflate the standard error of the sample mean and the power can be relatively low when comparing groups.

Tukey argued that heavy tailed distributions are likely to occur in practice and understanding the implications of heavy tailed distributions has great practical utility [11] and, investigations into the characteristics of actual distributions support Tukey's view [1, 3, 6, 10, 12, 14].

There is only one simulation study containing modified one-step M-estimator in the literature. Wilcox (2005) compared these 5 estimators in terms of their sampling variances when n=10 and the data were generated from Normal, Lognormal, One-Wild and Slash distributions. Variance of the mean is minimum under normality but when data come from lognormal distribution variance of the median and when the data come from One-Wild and Slash distribution variance of the modified one-step m-estimator is minimum in Wilcox's study.

The rest of this article is organized as follows. Section 2 provides a description of these three robust measures of location and contaminated normal distribution. Section 3 reports the results of simulation study between two conventional (mean and median) and three robust (trimmed mean, one-step M-estimator and modified one-step M-estimator) measures of location. Section 4 gives some concluding remarks.

2. Three robust measures of location

One basic problem with the mean is that the tails of a distribution can dominate its value. If a measure of location is intended to reflect the typical subject of the population, the mean might fail because its value can be inordinately influenced by a very small proportion of the subjects who fall in the tails of a distribution. One strategy for reducing the effects of the tails of a distribution is simply to remove them, and this is the strategy employed by the trimmed mean. Other strategy is first empirically determine the amount of trimming then remove, this is the strategy employed by one-step M-estimator and modified one-step M-estimator.

2.1 The trimmed mean

The γ-trimmed mean is

$$\mu_t = \frac{1}{1-2\gamma}\int_{x_\gamma}^{x_{1-\gamma}} x dF(x). \qquad (1)$$

where x_γ and $x_{1-\gamma}$ are the gamma and 1-gamma quintiles $(0 \le \gamma < 0.5)$.

Estimator of population trimmed mean μ_t is computed as follows. Let $X_1,...,X_n$ be a random sample and let $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$ be the observations written in ascending order. Let $g = [\gamma n]$, where $[\gamma n]$ is the value of γn rounded down to the nearest integer. The sample trimmed mean is [15].

$$\overline{X}_t = \frac{X_{(g+1)} + ... + X_{(n-g)}}{n-2g} \tag{2}$$

2.2 One-step M-estimator

Let $\xi(X-\mu_m)$ be some function that measures the distance between X and some unknown constant μ_m, and let ψ be its derivative with respect to μ_m. Attention is restricted to those functions for which $E[\xi(X-\mu_m)]$, viewed as a function of μ_m, has a derivative.

A general approach to defining a measure of location is to take μ_m to be the value that minimizes $\xi(X-\mu_m)$. So in general, μ_m satisfies

$$E[\psi(X-\mu_m)] = 0 \tag{3}$$

When $\xi(X-\mu_m) = (X-\mu_m)^2$, $E[\psi(X-\mu_m)] = 0$ and $\mu_m = \mu$, the population mean. Various choices for ξ and ψ have been proposed. Here, the focus is on Huber's well known choice for ψ [5]

$$\psi(x) = \max\{-k, \min(k,x)\} \tag{4}$$

Huber's M-measure of location is usually estimated with an iterative estimation procedure such as the Newton-Raphson method. Even with only a single iteration, the resulting estimator has good asymptotic properties [7]. One iteration of this procedure yields one-step M-estimator of location. As an example, one-step M-estimator of location of the series 3, 4, 8, 16, 24, 53 can be computed as follows:

1. Let i_1 be the number of observations X_i such that $\dfrac{|X_i - M|}{MAD/0.6745} < -1.28$ where M is median of the given series and $MAD = MED\{|X_1 - M|, |X_2 - M|, ..., |X_n - M|\}$ is median absolute deviation. In this inequality, 1.28 is a constant given by Huber and 0.6745 is also a constant used for to make the denominator an unbiased estimator of population standard deviation σ for normal distribution[15]

2. Let i_2 be the number of observations X_i such that $\dfrac{|X_i - M|}{MAD/0.6745} > 1.28$

3. The one-step M-estimator of location in [13] is

$$\hat{\mu}_m = \frac{1.28(MAD/0.6745)(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2} \tag{5}$$
$$= 14.2$$

A fundamental difference between the one-step M-estimator and trimmed mean is that an M-estimator empirically determines the amount of trimming whereas a trimmed mean is based on a predetermined amount of trimming. A natural appeal of the one-step M-estimator is that if sampling is from a light tailed distribution, it might be reasonable to trim very few observations or none at all. If a distribution is skewed to the right, a natural reaction is to trim more observations from the right versus the left tail of the distribution.

2.3 Modified one-step M-estimator

When testing hypothesis, a slight variation has practical value. The term $1.28(MAD/0.675)(i_2 - i_1)$ in Eq. (5) arises for technical reasons [8]. Ignoring it yields the modified one-step M estimator (MOM). This modified one-step M-estimator simply averages values not declared outliers; but to get reasonably good efficiency under normality, the outlier detection rule used by the one-step M-estimator is modified

$$\hat{\mu}_{mom} = \frac{\sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2} \tag{6}$$

where i_1 is the number of observations for which $(X_i - M)/MADN < -2.24$, and i_2 is the number for which $(X_i - M)/MADN > 2.24$ (Hampel identifier) [15].

2.4 Contaminated normal distribution

Contaminated normal distributions are mixtures of two normal distributions. Generating an observation from a contaminated normal distribution means that an observation is randomly sampled from a standard normal distribution with probability $1-\varepsilon$, where ε can be any number between 0 and 1 otherwise, an observation is randomly sampled from a normal distribution with mean 0 and standard deviation K. Let $\Phi(x) = P(X \le x)$ be the standard normal distribution. The contaminated normal distribution is

$$H(x) = (1-\varepsilon)\Phi(x) + \varepsilon\,\Phi\left(x/K\right) \tag{7}$$

which has mean 0 and variance $1-\varepsilon+\varepsilon K^2$ [15]. A common choice for ε and K is 0.1 and 10, respectively, in which case

$$H(x) = 0.9\Phi(x) + 0.1\Phi\left(x/10\right) \tag{8}$$

which has mean 0 and variance 10.9. Figure 1 shows the standard normal and the contaminated normal probability density function corresponding to Eq. (8).

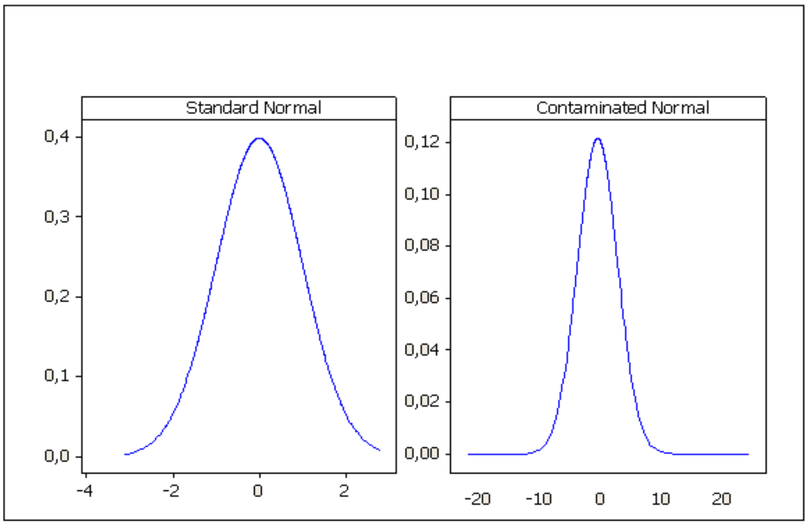

Figure 1. Probability density functions of standard normal and contaminated normal distribution corresponding to Eq. (8).

It can easily be observed that the tails of the contaminated normal are heavier than the standard normal and there is a large difference between variances of these two distributions. The reason for this large difference is that σ^2 is very sensitive to the tails of a distribution. In other words, a small proportion of the population of subjects can have an extremely large effect on its value [15].

3. Simulation study

In this simulation study, mean, median, 20% trimmed mean, one-step M-estimator and modified one-step M-estimator are compared in terms of their asymptotic relative efficiency values and mean squared error of $\hat{\theta}$ $\left(MSE\left(\hat{\theta}\right)\right)$. The asymptotic relative efficiency of estimator $\hat{\theta}_2$ with respect to $\hat{\theta}_1$ is defined as

$$EFF\left(\hat{\theta}_2,\hat{\theta}_1\right)=\frac{Var\left(\hat{\theta}_1\right)}{Var\left(\hat{\theta}_2\right)} \tag{9}$$

and if this ratio is less than 1, it can be said that $\hat{\theta}_1$ is asymptotically more efficient than $\hat{\theta}_2$. Asymptotic relative efficiency is a useful concept in that it allows us to make comparisons of competing estimators which are generally valid for sample sizes as low as 20 [8]. Mean squared error of an estimator is defined as

$$MSE\left(\hat{\theta}\right)=Var\left(\hat{\theta}\right)+\left[Bias\left(\hat{\theta}\right)\right]^2 \tag{10}$$

where $Bias\left(\hat{\theta}\right)=E\left(\hat{\theta}\right)-\theta$

To generate data from the contaminated normal distribution $CN\left(\varepsilon,K\right)$, three different contamination probabilities $\varepsilon=\left(0.05\right),\left(0.1\right),\left(0.2\right)$ and three different standard deviations $\sigma=5,10,20$ were used with sample sizes $n=20,40$. 10000 executions are made for each case by using the statistical software MINITAB 14. The MSE of all measures of location were reported in Table 1 and 2 for $n=20,40$ respectively.

Table 1. Mean squared error of mean, median, 20% trimmed mean, one-step M-estimator and modified one-step M-estimator when sample size n=20.

Distribution	MSE mean	MSE median	MSE 20%tmean	MSE osmest	MSE mosmest
N(0,1)	**0.04921**	0.07444	0.05721	0.05296	0.06235
CN(0.05,5)	0.15721	0.15504	0.12216	**0.11667**	0.12993
CN(0.05,10)	0.35165	0.15510	0.12066	**0.11736**	0.13027
CN(0.05,20)	1.16489	0.15379	0.12311	**0.1182**	0.13028
CN(0.1,5)	0.22618	0.16127	0.12658	**0.12354**	0.13307
CN(0.1,10)	0.61629	0.16986	0.13354	**0.13134**	0.13187
CN(0.1,20)	2.27206	0.16976	0.13696	0.13565	**0.13218**
CN(0.2,5)	0.36353	0.17781	**0.15334**	0.15645	0.15337
CN(0.2,10)	1.25135	0.19355	0.19458	0.18401	**0.14886**
CN(0.2,20)	4.69487	0.19599	0.28074	0.21287	**0.14027**

Table 2. Mean squared error of mean, median, 20% trimmed mean, one-step M-estimator and modified one-step M-estimator when sample size n=40.

Distribution	MSE mean	MSE median	MSE 20%tmean	MSE osmest	MSE mosmest
N(0,1)	**0.02446**	0.02920	0.02652	0.03799	0.03082
CN(0.05,5)	0.13175	0.11606	0.09195	**0.08776**	0.09925
CN(0.05,10)	0.32835	0.11875	0.09237	**0.08902**	0.09543
CN(0.05,20)	1.09971	0.11720	0.09311	**0.08817**	0.09655
CN(0.1,5)	0.20362	0.12351	0.09863	**0.09679**	0.10193
CN(0.1,10)	0.58318	0.12458	0.10522	0.10378	**0.10098**
CN(0.1,20)	2.13737	0.12919	0.10684	0.10680	**0.09631**
CN(0.2,5)	0.32275	0.14203	**0.12160**	0.12794	0.12177
CN(0.2,10)	1.17398	0.15117	0.16048	0.15127	**0.11844**
CN(0.2,20)	4.41400	0.15921	0.24414	0.17591	**0.10789**

Let Eff1, Eff2, Eff3 and Eff4 denote the asymptotic relative efficiencies of median(med), trimmed mean (tmean), one-step M-estimator(osmest) and modified one-step M-estimator (mosmest) with respect to mean; similarly Eff 5, Eff 6 and Eff7 denote the asymptotic relative efficiencies of trimmed mean, one-step M-estimator, modified one-step M-estimator with respect to median; Eff 8 and Eff 9 denote the asymptotic relative efficiency of one-step M-estimator and modified one-step M-estimator with respect to trimmed mean; and Eff 10 denote the asymptotic relative efficiency of modified one-step M-estimator with respect to one-step M-estimator. These values can be found in Table 3 and Table 4, for $n = 20, 40$ respectively.

Table 3. Asymptotic relative efficiencies of mean, median, trimmed mean, one-step M-estimator and modified one-step M-estimator for different CN distributions for $n = 20$ (wrt: with respect to).

ε	**σ**	**Eff1** **med wrt mean**	**Eff2** **tmean wrt mean**	**Eff3** **osmest wrt mean**	**Eff4** **mosmest wrt mean**	**Eff5** **tmean wrt med**	**Eff6** **osmest wrt med**	**Eff7** **mosmest wrt med**	**Eff8** **osmest wrt tmean**	**Eff9** **mosmest wrt tmean**	**Eff10** **mosmest wrt osmest**	**Choice**
N(0,1)		0.661	0.860	0.929	0.789	1.301	1.405	1.193	1.080	0.917	0.849	1.mean 2.osmest 3.tmean
0.05	**5**	1.014	1.286	1.347	1.209	1.268	1.328	1.192	1.047	0.940	0.897	1.osmest. 2.tmean 3.mosmest
	10	2.270	2.913	2.995	2.698	1.283	1.319	1.188	1.028	0.926	0.900	1.osmest. 2.tmean 3.mosmest
	20	7.574	9.461	9.854	8.942	1.249	1.301	1.180	1.041	0.945	0.907	1.osmest. 2.tmean 3.mosmest
0.1	**5**	1.402	1.786	1.830	1.699	1.273	1.305	1.211	1.024	0.951	0.928	1.osmest. 2.tmean 3.mosmest
	10	3.628	4.614	4.691	4.672	1.271	1.293	1.288	1.016	1.012	0.995	1.osmest. 2.mosmest 3.tmean
	20	13.379	16.583	16.743	17.190	1.239	1.251	1.284	1.009	1.036	1.026	1.mosmest. 2.osmest 3.tmean
0.2	**5**	2.045	2.370	2.323	2.370	1.159	1.136	1.159	0.980	0.999	1.020	1.tmean 2.mosmest 3.osmest
	10	6.464	6.430	6.800	8.407	0.994	1.051	1.300	1.057	1.307	1.236	1.mosmest. 2.osmest 3.med
	20	23.961	16.729	22.054	33.467	0.698	0.920	1.396	1.318	2.000	1.517	1.mosmest. 2.med 3.osmest

Table 4. Asymptotic relative efficiencies of mean, median, trimmed mean, one-step M-estimator and modified one-step M-estimator for different CN distributions for $n = 40$ (wrt: with respect to).

ε	σ	**Eff1** **med** **wrt** **mean**	**Eff2** **tmean** **wrt** **mean**	**Eff3** **osmest** **wrt** **mean**	**Eff4** **mosmest** **wrt** **mean**	**Eff5** **tmean** **wrt** **med**	**Eff6** **osmest** **wrt** **med**	**Eff7** **mosmest** **wrt** **med**	**Eff8** **osmest** **wrt** **tmean**	**Eff9** **mosmest** **wrt** **tmean**	**Eff10** **mosmest** **wrt** **osmest**	**Choice**
N(0,1)		0.644	0.837	0.992	0.793	1.300	1.432	1.232	1.101	0.947	0.860	1.mean. 2.osmest 3.tmean
0.05	**5**	1.135	1.432	1.501	1.327	1.261	1.322	1.168	1.047	0.926	0.884	1.osmest. 2.tmean 3.mosmest
	10	2.764	3.555	3.688	3.440	1.285	1.334	1.244	1.037	0.967	0.932	1.osmest. 2.tmean 3.mosmest
	20	9.378	11.807	12.468	11.383	1.258	1.329	1.213	1.056	0.964	0.913	1.osmest. 2.tmean 3.mosmest
0.1	**5**	1.648	2.064	2.103	1.998	1.252	1.275	1.211	1.018	0.967	0.949	1.osmest. 2.tmean 3.mosmest
	10	4.683	5.543	5.619	5.775	1.183	1.199	1.233	1.013	1.041	1.027	1.mosmest 2.osmest 3.tmean
	20	16.543	20.003	20.014	22.191	1.209	1.209	1.341	1.000	1.109	1.108	1.mosmest 2.osmest 2.tmean
0.2	**5**	2.272	2.653	2.522	2.650	1.167	1.110	1.166	0.950	0.998	1.050	1.tmean. 2.mosmest 3.osmest
	10	7.766	7.316	7.761	9.912	0.942	0.999	1.276	1.060	1.354	1.277	1.mosmest 2.med 3.osmest
	20	27.721	18.079	25.090	40.914	0.652	0.905	1.475	1.387	2.263	1.630	1.mosmest 2.med 3.osmest

4. Conclusion

In Table 3, when the samples are taken from standard normal distribution, all asymptotic relative efficiencies with respect to mean is less than 1. This means that sample mean is more efficient than all other measures of location. This is the reason that the first choice is the mean in this row. When the asymptotic relative efficiencies with respect to median is controlled, it can be seen that all are greater than 1 and which means median should not be the second choice. Remaining part of standard normal row of Table 3 tells us that the second most efficient measure of location should be one-step M-estimator and the third one should be 20% trimmed mean with respect to their asymptotic relative efficiencies values. All rows of both Table 3 and Table 4 are commented in the same manner and the orders of choices are determined.

When n=20, one-step M-estimator is preferred 5 times, modified one-step M-estimator is preferred 3 times and 20% trimmed mean is preferred 1 time out of 9 situations covered in terms of asymptotic relative efficiency

When n=40, one-step M-estimator is preferred 4 times, modified one-step M-estimator is preferred 4 times and %20 trimmed mean is preferred 1 time out of 9 situations covered in terms of asymptotic relative efficiency.

When we look at the order of choices in the Table 3 and 4, it is observed that the first choices have always the smallest MSE value that's why it can be said that two criterions have confirmed each other in all sampling situations.

When the sample is taken from N (0, 1), mean is clearly the best choice with the smallest MSE value. But when the underlying distribution has some heavier tails than normal, mean is the worst choice in all of the sampling situations covered in terms of both asymptotic relative efficiency and MSE criterions.

Heavy tailed distributions like contaminated normal are very common in applied work. It is known that these heavy tails are likely to create outlying observations which substantially inflate the sample mean and variance and reduce the power of the tests that utilize them.

The use of one-step M-estimator modified one-step M-estimator and trimmed mean and the use of statistical tests that utilize these robust location estimators should be considered as an alternative approach when the underlying distribution has heavier tails than normal. Some bootstrap techniques such as percentile bootstrap and bootstrap-t make it possible to use these three robust estimators for inferential purposes.

References

[1] D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey, (1972), *Robust Estimates of Location: Survey and Advaces*, Princeton University Press.

[2] H. Cramer, (1946), *Mathematical Methods of Statistics*, Princeton University Pres.

[3] F.R. Hampel, (1973), Robust Estimation: A condensed partial survey. Z. *Wahrscheinlichkeitstheorie Verw*, Gebiete 27, 87-104.

[4] M. Hill, and W.J. Dixon, (1982), Robustness in real life: A study of clinical laboratory data. *Biometrics* 38, 377-396.

[5] P.J. Huber, (1981), *Robust Statistics.* Newyork: Wiley.

[6] T. Micceri, (1989), The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin* 105, 156-166.

[7] R.J. Serfling, (1980), *Approximation Theorems of Mathematical Statistics*. Newyork: Wiley.

[8] R.G. Staudte., and S.J. Sheater, (1990), *Robust Estimation and Testing*. Newyork: Wiley.

[9] S.M. Stigler, (1973), Simon Newcomb, Percy Daniel and the history of robust estimation 1885-1920. *Journal of the American Statistical Association* 68, 871-879.

[10] S.M. Stigler, (1977), Do robust estimators work with real data. *Annals of Statistics* 5, 1055-1098.

[11] J.W. Tukey, (1960), A survey of sampling from contaminated normal distributions. In I.Olkin et al. (Eds), *Contributions to Probability and Statistics*. Stanford, CA: Stanford University

[12] R.R. Wilcox, (1990), Comparing the means of two independent groups. *Biometrical Journal* 32, 771-780.

[13] R.R. Wilcox, (2001), *Fundamentals of Modern Statistical Methods*. Springer-Verlag.

[14] R.R. Wilcox, (2003), Modern Robust Data Analysis Methods: Measures of Central Tendency. *Psychological Methods* 8, 254-274.

[15] R.R. Wilcox, (2005), *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Pres, Second Edition.