



Decile Mean: A New Robust Measure of Central Tendency

Sohel Rana*[a, b], Md.Siraj-Ud-Doulah [c], Habshah Midi [a, b] and A.H.M. R. Imon [d]

[a] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia.

[b] Laboratory of Computational Statistics and Operations Research, Institute for Mathematical
Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

[c] Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh.

[d] Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, U.S.A.

*Author for correspondence; e-mail: sohel@science.upm.edu.my

Received: 19 August 2011

Accepted: 9 February 2012

ABSTRACT

In statistics, central tendency of a data set is a measure of the middle or location or typical or expected value of the data set. There are many different descriptive statistics that can be chosen as a measurement of the central tendency and under a well-behaved normal distribution few of them possess some nice and desirable properties. But there is evidence that they may perform poorly in the presence of non-normality or when outliers occur in data. We investigate the performances of some popular and commonly used measures of central tendency such as the mean, the median and the trimmed mean and observe that they may not perform as good as we expect in the presence of non-normality or outliers. In this paper, we proposed a new measure of central tendency which we call Decile Mean (DM) since it is based on deciles. This measure should be fairly robust as it automatically discard extreme observations or outliers from both tails but at the same time is more informative than the median or interquartile mean. The usefulness of the proposed measure is investigated by bootstrap and simulation approach. The results show that decile mean outperforms the mean, the median and the trimmed mean in every respect.

Keywords: mean, median, trimmed mean, decile, bootstrap, robustness, Monte Carlo simulation

1. INTRODUCTION

The overall goal of descriptive statistics is to provide a concise, easily understood summary of the characteristics of a data set. A data set can be summarized in several ways: measures of central tendency, measures of dispersion, measures of shape or relative position etc. In this paper, we focus on finding

the best measures of central tendency. A number of measures are available in the literature for central tendency [1-5]. Among them the arithmetic mean (more popularly known as the mean) is the most popular and commonly used measure. Although mean is based on all the observations, it is very much

affected by extreme values of the data as well as when data come from extremely asymmetrical distribution. Median is another popular measure and it is free from sensitive to extreme values but it depends on either the higher or lower half of a sample or a population or a probability distribution. In a sample of data, or a finite population, there may be no member of a sample whose value is identical to the median (in case of an even size) and, if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Geometric and harmonic mean are based on all the observations, but if any observations is zero, geometric mean becomes zero as well as harmonic mean becomes undefined and if any observations is negative, geometric mean becomes undefined. Truncated or trimmed mean or Windsorized mean is less sensitive to extreme values but it depends on half of the observations or depends on value of α , a proportion of the sample size. Assumed mean is something like the median but some distinct (firstly, they take a plausible initial guess which is same as median and this value is then subtracted from all the sample values, after completing this process, this value is subtracted from plausible initial guess). Some other means are available in the literature such as Frechet mean, power mean, f-mean or quasi-arithmetic mean [6-8]. All of these means are based on Euclidean distances but if any of the observations is zero, these means do not exist. For this reason, we define a new and simple measure of central tendency. We call it the Decile Mean (DM) which is introduced in section 2. The properties of this new measure are illustrated in section 3 with a real life data in the context of bootstrap. The performance of the proposed mean is investigated in section 4 through a Monte Carlo simulation experiment.

2. DECILE MEAN (DM)

In this section, we introduce some commonly used measure of central tendency and define a new one. Let us start with a probability-based statistical model. For a set of observations x_1, x_2, \dots, x_n , we assume that each observation x_i depends on the 'true value' m of the unknown parameter and also on some random error process. The simplest assumption is that the error acts additively, i.e.,

$$x_i = \mu + \varepsilon_i \quad (1)$$

where the errors are random variables. The model given in (1) is called the location model. If the observations are independent replications of same experiment under equal conditions, it may be assumed that (i) Errors are independent and (ii) Errors have same distribution function F_0 . It follows that x_1, x_2, \dots, x_n are independent with common distribution function

$$F(x) = F_0(x - \mu) \quad (2)$$

and we say that x 's are *i.i.d.* (independently and identically distributed) random variables. Assume that the distribution function has a density $f_0 = F_0'$. Then the joint density of the observations (the likelihood function) is

$$L(x_1, \dots, x_n; \mu) = \prod_{i=1}^n f_0(x_i - \mu) \quad (3)$$

The maximum likelihood estimate of m is the value $\hat{\mu}$, that maximizes $L(x_1, \dots, x_n; \mu)$

$$\hat{\mu} = \arg \max_{\mu} L(x_1, \dots, x_n; \mu) \quad (4)$$

where 'arg max' stands for 'the value maximizing'. If F_0 is everywhere positive (4) can be written as

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^n \rho(x_i - \mu) \quad (5)$$

where $\rho = -\log F_0$. If $F_0 = N(0, 1)$, then $f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and hence (5) is equivalent to

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^n (x_i - \mu)^2 \quad (6)$$

Differentiating (6) w.r.t. μ yields

$$\sum_{i=1}^n (x_i - \hat{\mu}) = 0 \quad (7)$$

which has $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ as solution. Hence, under normality the sample mean provides the best measure of location, but this might not be the case always. For example, if F_0 is the Laplace (double exponential) distribution $f_0(x) = \frac{1}{2} e^{-|x|}$ and hence (5) is equivalent to

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^n |x_i - \mu| \quad (8)$$

The standard theory tells us that in such a situation $\#(x_i > \hat{\mu}) = \#(x_i < \hat{\mu})$, which implies that the solution is the sample median.

The sample mean and sample median are approximately equal if the sample is symmetrically distributed about its center, but not necessarily otherwise. Statisticians were aware of the weakness of sample mean as an estimator of location parameter for over two hundred years, especially when extreme observations or outliers are present in the data, but it retains its popularity mainly because of the fact of reliance on the breakthrough result of Fisher [9]. In his work Fisher showed that both sample mean and sample median are unbiased, consistent and sufficient estimator of location parameter. But under normality sample mean is more efficient than sample median. When observations come from $N(\mu, \sigma^2)$, then

$$\text{Var}(\bar{x}) = \frac{(1-\epsilon) + \epsilon \sigma^2}{n} \text{ and } \text{Var}[\text{Med}(x)] = \pi \sigma^2 / 2n$$

This is true for a well-behaved normal data, but Tukey [5] showed that for a perturbed model

$$F = (1 - \epsilon) N(0, 1) + \epsilon N(0, \sigma^2)$$

we obtain

$$\text{Var}(\bar{x}) = \frac{(1-\epsilon) + \epsilon \sigma^2}{n} \text{ and } \text{Var}[\text{Med}(x)] \approx \frac{\pi}{2n(1-\epsilon + \epsilon/\sigma)^2}$$

Thus, the efficiency of sample mean suffers a huge setback in the presence of even a single outlier but this is not the case with the sample median. Despite this advantage median has several shortcomings for which we need to think about alternatives.

One popular choice is the trimmed mean which is obtained after discarding a proportion of the largest and smallest values. More precisely, let $\alpha \in [0, 1/2)$ and $m = [(n-1)\alpha]$ where $[\cdot]$ stands for the integer part. We define the α -trimmed mean as

$$\bar{X}_{\alpha} = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)} \quad (9)$$

The limit cases $\alpha = 0$ and $\alpha \rightarrow 0.5$ correspond to the sample mean and sample median, respectively.

Here we define the decile mean. In descriptive statistics, a decile is any of the nine values that divide the sorted data into ten equal parts, so that each part represents 1/10 of the sample or population for raw data as well as decile from frequency distribution. We have 9 deciles from ungrouped or grouped data denoted as D_1, D_2, \dots, D_9 . Sum of all deciles divided by the number of deciles is called the decile mean (DM). Hence, the formula to find the DM from 9 deciles is given by

$$DM = \frac{D_1 + D_2 + \dots + D_9}{9} \quad (11)$$

where D_1, D_2, \dots, D_9 are the deciles. The main advantage of the DM is less sensitive to extreme values than any other existing measures as well as it depends on the eighty percent of a sample, a population, or a probability distribution. It is referred as a robust estimator in this regard.

3. PROPERTIES OF THE DECILE MEAN USING BOOTSTRAP

It is not easy to find the distribution of measures of location. The distribution of the sample mean is normal only when the observations come from normal. The distribution of the sample median is asymptotic normal. The distribution of the trimmed mean is not known. We anticipate that it may be very difficult to find the theoretical distribution of proposed decile mean. But by virtue of bootstrap we can investigate different properties of this mean. Efron and Tibshirani's [10] bootstrap is a computer-based method of statistical inferences that can answer many questions about the properties of a statistic or estimator. The main advantage of bootstrap is that it

can give variance, bias, coverage and other probabilistic phenomena of any statistic. It can automatically produce accurate estimates in almost any situation. Here, we consider a real life data. Bangladesh consumption of petrochemical data is taken from Bhuyan [11]. This data consists of 65 observations. The data we have used can be categorized broadly into two groups: one is free from extreme observations or outliers and the other group contains outliers. It is worth mentioning that the original data do not contain any potential outliers. So to form the second group we have deliberately putted few outliers in the data. We use bootstrap to investigate the sampling distribution of four location estimator, the sample mean, the median, the trimmed mean and the decile mean and each of this result is based on 10,000 bootstrap replications.

At first we present some graphical displays for the data that do not contain any outlier. For each of the four measures of location we present histogram and normal probability plot and the graphs are

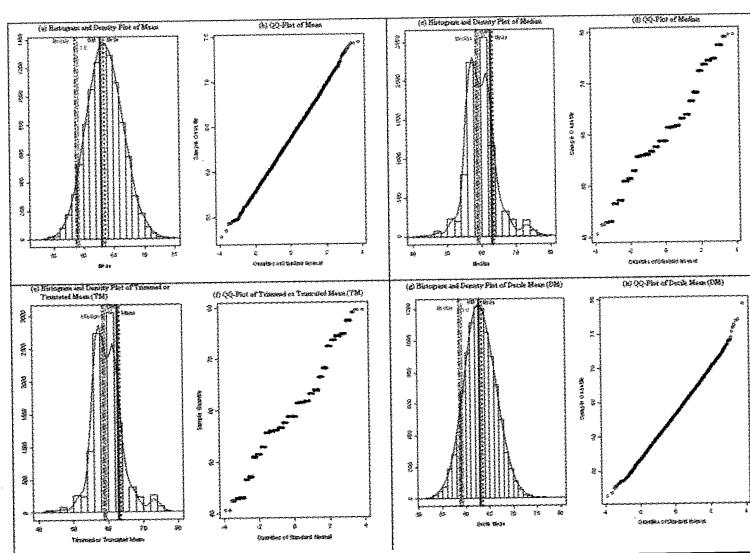


Figure 1. A graphical comparison of four measures of location when data are free from outliers.

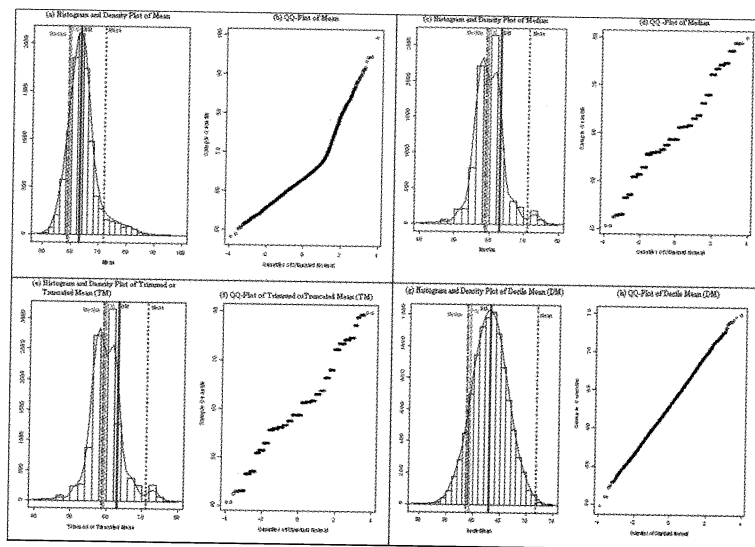


Figure 2. A graphical comparison of four measures of location when data contain outliers.

given in Figure 1. Figure 2 gives the same for the four measures, the essential difference, however, is here the data contain outliers.

When the data set are free from outliers we observe from Figure 1 that the histogram and density plot as well as QQ-plot of sample mean are reasonably normal in shape. On the contrary, from Figure 2 we observe that the histogram and density plot as well as QQ-

plot of mean are not normal-shaped when the data set contains outliers. Irrespective of the presence of outliers or not, we see that the histogram and density plots of median and TM look similar and QQ-plots are nearly the same, which confirm the standard theory that outliers do not affect median and trimmed mean to a great extent. But it is interesting to note that the distributions of both median and trimmed

Table 1. Bootstrap distribution of four measures of locations results for Bangladesh consumption of petrochemical data with and without outliers.

	Mean		Median		TM		DM	
	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers
Actual Value	63.54	70.99	58.7	58.7	59.22	59.22	62.97	62.97
Bootstrap Value	63.60	73.91	59.92	59.81	60.37	60.294	63.02	63.06
S.E	0.925	6.346	4.205	4.266	4.133	4.252	0.298	0.577
Bias	0.0590	2.919	1.22	1.111	1.147	1.074	0.0491	0.0855

mean are notably non-normal. The above figures show that the distribution of decile is quite normal in shape irrespective of the presence of outliers and both of them almost coincide.

Table 1 presents various bootstrap distributional results for four different measures of location, i.e., the mean, median, TM and the newly proposed DM under two different situations: with and without outlier. In this table, we compare bootstrap values with the actual values and also present standard error and bias of all four estimators considered here. We observe from this table that bias and standard error of mean are low when the data set is free from outliers. On the other hand, when the data set contains outliers both bias and standard error of mean are the most. Median and trimmed mean perform

almost equally and they are not affected by outliers. The newly proposed decile mean is also unaffected by outliers. But it is really interesting to note that both bias and standard error of DM are very small and among the four this estimator appears as the best in every respect.

4. MONTE CARLO SIMULATION

In this section, we report a Monte Carlo simulation study which is designed to compare the performance of the newly proposed decile mean with three other popular and commonly used measures of location, i.e., the mean, median and trimmed mean. We generate samples from uniform distribution for four different sample sizes, $n = 65, 100, 200$ and 500 . We run each experiment 10,000 times and the results based

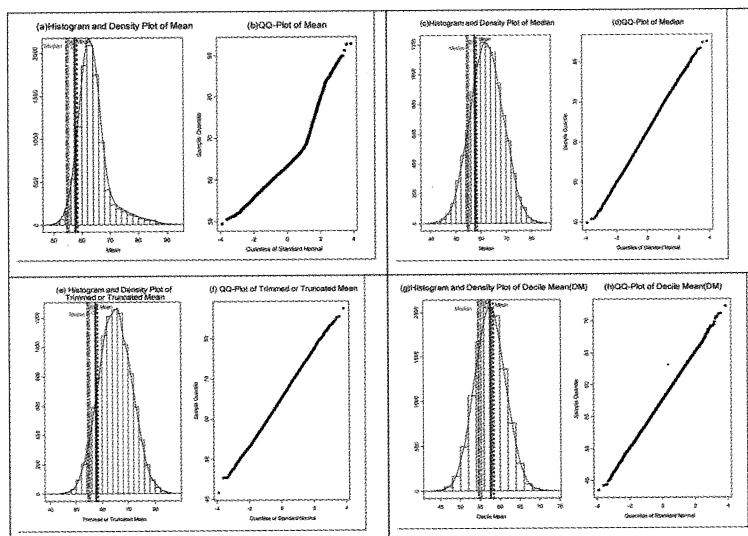


Figure 3. A graphical comparison of four measures of location.

on these 10000 replications. First we show the results graphically. Figure 3 presents the histogram and density plot as well as QQ-plot for four different measures considered in this study.

We observe from the top left panel of Figure 3 that the histogram and density

plot as well as QQ-plot of sample mean are notably non-normal in shape when data are generated from an uniform distribution. Similar graph of sample median and trimmed mean look reasonably normal in shape but these graphs indicate that the actual value of both median and trimmed mean are not close enough to the actual values. The decile

mean performs best overall in this situation.

Table 2 offers a comparison between four measures of central tendency based on a Monte Carlo simulation. As stated earlier, we generate samples from uniform distribution for four different sample sizes,

$n = 65, 100, 200$ and 500 . We run each experiment 10,000 times and compute the simulated mean, bias and standard error for mean, median, trimmed mean and decile mean.

Table 2. Simulation results of different measures of central tendency.

Measures	$n = 65$			
	Actual Value	Simulated value	S.E	Bias
Mean	58.39	65.0037	4.9187	6.6137
Median	54.59	64.9726	11.7411	10.3826
TM	55.21	62.498	8.4345	7.2879
DM	57.61	54.5975	0.0759	-3.0125
$n = 100$				
Mean	57.48518	63.24642	7.733963	5.761237
Median	53.18977	59.56381	8.683069	6.374042
TM	54.15946	58.57385	5.684472	4.414395
DM	57.53415	57.50687	1.032938	-0.02727909
$n = 200$				
Mean	62.68469	67.8731	6.853663	5.188408
Median	53.58317	57.55386	4.330939	3.970686
TM	52.8702	57.48928	5.947263	4.61908
DM	60.98788	57.47116	0.0929239	-3.516724
$n = 500$				
Mean	63.5289	72.49636	9.964239	8.967465
Median	58.03687	65.47162	10.338157	7.434745
TM	59.82883	67.83853	11.338925	8.009702
DM	60.22685	54.76411	0.00224319	-5.462739

Results presented in Table 2 show that for all four different sample sizes the decile mean performs best. All four estimators are biased, but this bias is the least for the decile mean. But when we look at the standard errors of different estimators we can observe the dramatic improvement achieved by the proposed decile mean. Its

standard error values are substantially less than that of the mean, median and trimmed mean for different sample sizes.

CONCLUSIONS

In this paper, our main objective was to propose a new measure of central tendency or location which represents the

data better than any existing measures. Although the sample mean, the sample median, the trimmed mean and other measures have some desirable properties, these properties hold asymptotically. But small to moderate samples are prevalent in nature and bootstrap distribution of these measures show that they are nowhere near the normal distribution in the presence of model violation. But the decile mean perform superbly here. Not only that, both the bootstrap and simulation study demonstrate that the Decile Mean (DM) is more accurate measure in terms of possessing smaller bias and lower standard errors in a variety of situations and hence can be recommended to use an effective measure of location.

REFERENCES

- [1] Bernstein S., *Schaum's Outline of Elements of Statistics I: Descriptive Statistics and Probability (Schaum's Outlines)*, New York: McGraw-Hill, 1999.
- [2] Hiai F. and Kosaki H., Comparison of various means for operators, *J. Funct. Analys*, 1999; **163**: 300-323.
- [3] Montresor A., Arithmetic or geometric means of eggs per gram are not appropriate indicators to estimate the impact of control measures in helminthes infections. *Trans. Royal Soc. Trop. Med. Hyg.*, 2007; **101**:773-776.
- [4] von Hippel P.T., Mean median and skew: Correcting a textbook rule, *J. Stat. Edu.*, 2005: 13.
- [5] Tukey J.W., A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.) 448-485, Stanford University Press, CA., 1960.
- [6] Bullen P.S., *The Power Means. Handbook of Means and Their Inequalities*, Netherlands: Kluwer, 2003.
- [7] Fiori S., Computation of the Frechet mean, variance and interpolation for a pool of neural networks over the manifold of special orthogonal matrices, *Int. J. Comp. Int. Studies*, 2009; **1**: 50-71.
- [8] Ume J.S. and Kim Y.H., Some mean values Related to the Quasi-arithmetic mean, *J. Math. Analys. Appl.*, 2000; **252**: 167-176.
- [9] Fisher R.A., On a distribution yielding the error functions of several well known statistics, *Proce. Int. Cong. Math.*, 1924; **2**: 805-813.
- [10] Efron B. and Tibshirani J.R. *An Introduction to the Bootstrap*, London: Chapman and Hall, 1993.
- [11] Bhuyan K.C., *Methods of Statistics*, Dhaka: Sahitya Prokashani, 2009.