Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/030301

Proposed Mean (Robust) in the Presence of Outlier

Md. Moyazzem Hossain[*]

Department of Statistics, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh.

Received: 18 Feb. 2016, Revised: 15 Mar. 2016, Accepted: 17 Mar. 2016.
Published online: 1 Sep. 2016.

Abstract: The term "arithmetic mean" is preferred in some contexts in mathematics and statistics because it helps distinguish it from other means, such as the geometric mean and the harmonic mean. In addition to mathematics and statistics, the arithmetic mean is used frequently in fields such as economics, sociology, and history, and it is used in almost every academic field to some extent. The first recorded time that the arithmetic mean was extended from 2 to n cases for the use of estimation was in the sixteenth century. While the arithmetic mean is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers. This paper attempts to propose a mean in the presence of outlier and compare the results to the well established and most frequently used arithmetic mean. The results reveal that the proposed mean is less affected by the outlier than the arithmetic mean.

Keywords: Mean, Robust, Outlier.

1 Introduction

The first recorded time that the arithmetic mean was extended from 2 to n cases for the use of estimation was in the sixteenth century. From the late sixteenth century onwards, it gradually became a common method to use for reducing errors of measurement in various areas (Plackett, 1958 [1]; Churchill Eisenhart, 1971 [2]). At the time, astronomers wanted to know a real value from noisy measurement, such as the position of a planet or the diameter of the moon. Using the mean of several measured values, scientists assumed that the errors add up to a relatively small number when compared to the total of all measured values. The method of taking the mean for reducing observation errors was indeed mainly developed in astronomy (Plackett, 1958 [1]; Bakker Arthur, 2003 [3]). A possible precursor to the arithmetic mean is the mid-range (the mean of the two extreme values), used for example in Arabian astronomy of the ninth to eleventh centuries, but also in metallurgy and navigation (Churchill Eisenhart, 1971 [3]).In the 16th century average meant a customs duty, or the like, and was used in the Mediterranean area. It came to mean the cost of damage sustained at sea. From that came an "average adjuster" who decided how to apportion a loss between the owners and insurers of a ship and cargo (https://en.wikipedia.org/wiki/Average). A second English usage, documented as early as 1674 and sometimes spelled "averish," is as the residue and second growth of field crops, which were considered suited to consumption by draught animals ("avers") (Ray, John, 1674 [4]). Jevons, W. S. (1835-1882) distinguishes between a ‘mean’ (the approximation of a definite existing quantity) and an ‘average’ or ‘fictitious mean’ (an arithmetical average) (Stanford Encyclopedia of Philosophy [5]).

The root is found in Arabic as *awar*, in Italian as *avaria*, in French as *avarie* and in Dutch as *averij*. It is unclear in which language the word first appeared (https://en.wikipedia.org/wiki/Average). The term "arithmetic mean" is preferred in some contexts in mathematics and statistics because it helps distinguish it from other means, such as the geometric mean and the harmonic mean. In addition to mathematics and statistics, the arithmetic mean is used frequently in fields such as economics, sociology, and history, and it is used in almost every academic field to some extent. While the arithmetic mean is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers (values that are very much larger or smaller than most of the values). Usually, the outlier is removed from the data set for further analysis which reduces the degrees of freedom. Notably, for skewed

[*]Corresponding author e-mail: mmhmm.justat@gmail.com

distributions, the arithmetic mean may not accord with one's notion of "middle", and robust statistics, such as the median, may be a better description of central tendency (https://en.wikipedia.org/wiki/Arithmetic mean). However, the main disadvantage of median is that it is not based on all observation in data set. This paper attempts to propose a mean in the presence of outlier and compare the results to the well established and most frequently used arithmetic mean. The results reveal that the proposed mean is less affected by the outlier than the arithmetic mean.

2 Methods

2.1 Arithmetic Mean

The arithmetic mean (or mean or average) is the most commonly used and readily understood measure of central tendency. In statistics, the term average refers to any of the measures of central tendency. The arithmetic mean is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations. Symbolically, if we have a data set containing the values $x_1, x_2, ..., x_n$. The arithmetic mean $\overline{x}$ is defined by the formula

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1).$$

2.2 Proposed Mean

This paper proposed a formula for mean which is robust that i.e., too much less affected by outlier. Suppose, we have a data set containing the values $x_1, x_2, ..., x_n$. Among them it is assumed that it contains at least one outlier namely x_n .

Thus, the proposed formula for mean is

$$\overline{x}_R = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (2)$$

where, $w_i = \frac{1}{|x_i - \overline{x}|}$ and $\overline{x}$ represent the arithmetic mean is defined by the formula $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

3 Results and Discussion

To compare the accuracy of the proposed mean to the famous formula for arithmetic mean we consider a data set contains 49 generated observations from different distributions and one observation (outlier) is added artificially. Here, we generate 49 observations from Binomial (49, 0.7), Poisson (10), Uniform (10, 100) and Normal (10, 10). The generated observations and artificially added an outlier are given in Table 1. The 50th observation considered as outlier for each distribution.

Table 1: Generated observations from different distributions

Sl. No.	Binomial	Poisson	Uniform	Normal	Sl. No.	Binomial	Poisson	Uniform	Normal
1	34	8	83.81939	-4.98133	26	37	9	74.88998	23.35891
2	31	8	91.11454	-7.55197	27	35	14	91.85339	9.522083
3	28	12	97.41264	-5.38087	28	36	13	14.72976	14.96091
4	34	9	17.14133	20.87392	29	34	8	37.35954	-5.85213
5	34	7	57.59697	4.805876	30	38	11	63.02988	0.628953
6	30	12	25.71093	13.7935	31	32	14	88.32118	-4.51608
7	36	10	82.64107	11.52554	32	32	6	41.60039	7.655823
8	29	15	13.94147	12.98312	33	31	9	31.1713	8.66767
9	32	11	91.08982	13.00151	34	37	10	28.66634	4.751511
10	37	12	29.52605	25.18169	35	34	7	19.19584	-8.23419
11	34	6	40.44404	-8.8893	36	34	8	48.59066	19.43892

Sl. No.	Binomial	Poisson	Uniform	Normal	Sl. No.	Binomial	Poisson	Uniform	Normal
12	29	9	16.37776	-3.30495	37	35	7	62.37342	5.764517
13	31	13	25.39232	24.12368	38	35	6	75.43107	-11.4921
14	30	10	74.29121	16.43008	39	35	5	97.12699	3.182178
15	37	12	85.57451	18.33193	40	28	12	32.20954	11.84601
16	40	12	85.41795	4.069049	41	34	10	83.13272	10.67025
17	34	16	36.95853	18.36985	42	32	6	26.7629	7.336124
18	36	8	37.63146	-2.31799	43	31	9	56.92129	11.17966
19	29	11	39.18882	17.98302	44	32	8	67.5399	9.803003
20	32	6	10.37629	15.34972	45	37	12	94.64675	25.45623
21	40	10	74.43129	3.494475	46	35	7	54.72671	18.21344
22	36	7	98.79696	15.71885	47	39	7	70.4706	-11.9792
23	33	8	53.02652	7.412147	48	32	13	49.76897	6.931194
24	37	12	12.68624	8.040255	49	34	9	33.81909	11.56037
25	35	10	98.65139	23.7914	**50**	**100000**	**100000**	**100000**	**100000**

Now, we compute the average for the 50 observations for four data sets using the formula given in (1) and (2). We also compute the mean for 49 observations i.e., avoiding the outlier for four data sets. The results are presented in Table 2.

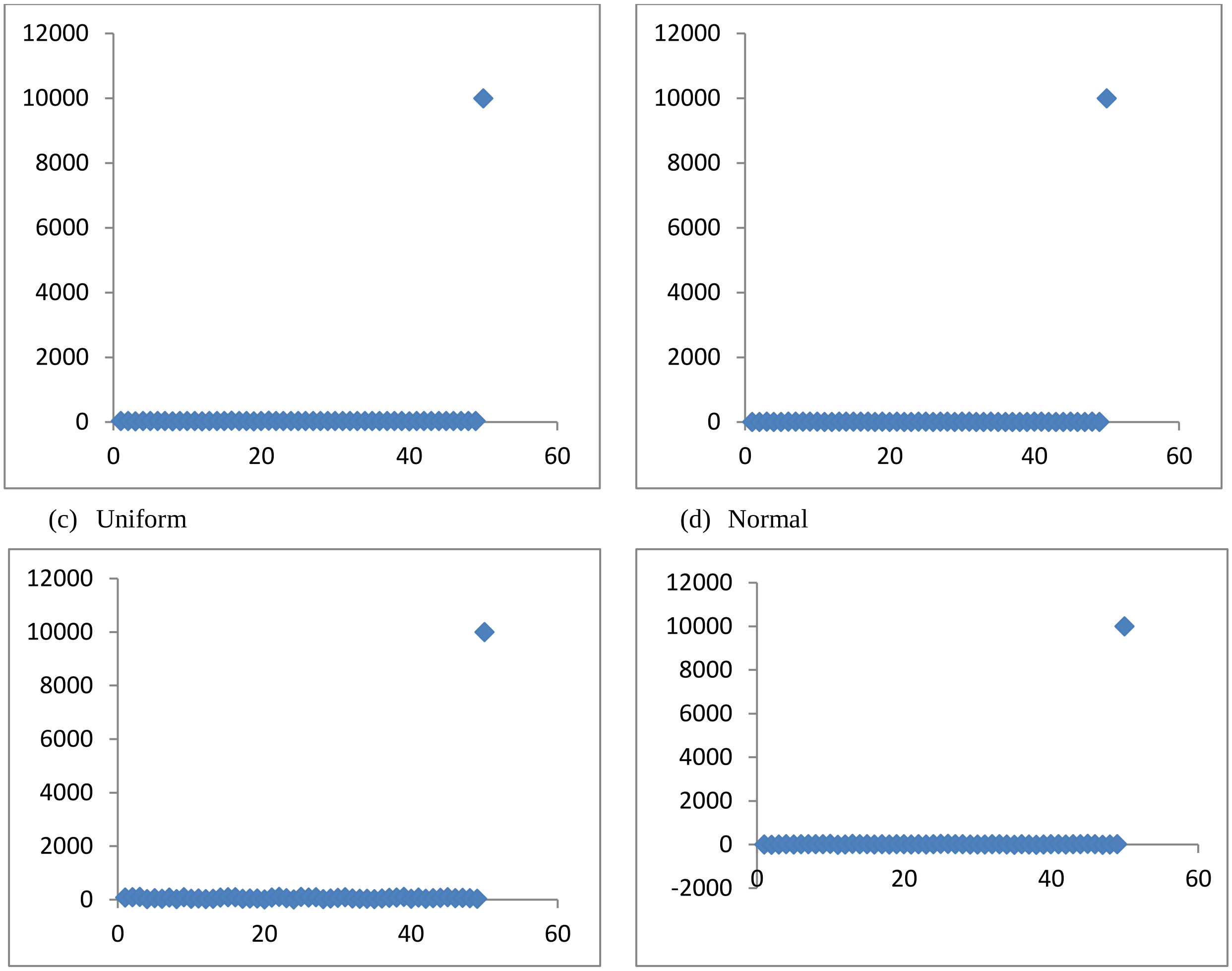
Table 2: Results of four data sets

Average	Binomial	Poisson	Uniform	Normal
Minimum (for 49 observations)	28	5	10.37629	-11.9792
Maximum (for 49 observations)	40	16	98.79696	25.45623
Outlier	10000	10000	10000	10000
$\overline{x}$ (for 49 observations)	33.81633	9.673469	55.58322	8.402189
$\overline{x}$ (presence of outlier)	233.14	209.48	254.472	208.234
$\overline{x}_R$ (presence of outlier)	38.00867	13.86706	63.67622	13.06334

From the generated data sets given in Table 1, it is observed that the minimum value is 28 with maximum value 40 in case of Binomial distribution. For Poisson distribution the minimum value is 5 and maximum value is 16 whereas 10.37629 is the minimum value and 98.79696 is the maximum value for Uniform distribution. In case of Normal distribution, the minimum value is -11.9792 and maximum value is 25.45623. Artificially an outlier 10,000 in each data set. In case of Binomial distribution, the mean for 49 observations is around 34 and for 50 observations i.e., in the presence of outlier it is about 233 which is very much affected by the outlier whereas the proposed mean gives the mean about 38. For the other data sets, the results are looking same. Thus, it is may conclude that the usual arithmetic mean is very much affected by the outlier whereas the proposed mean is less affected by the outlier [Table 2]. The graphical presentation of the data sets considered in this paper is shown in Figure 1.

(a) Binomial

(b) Poisson

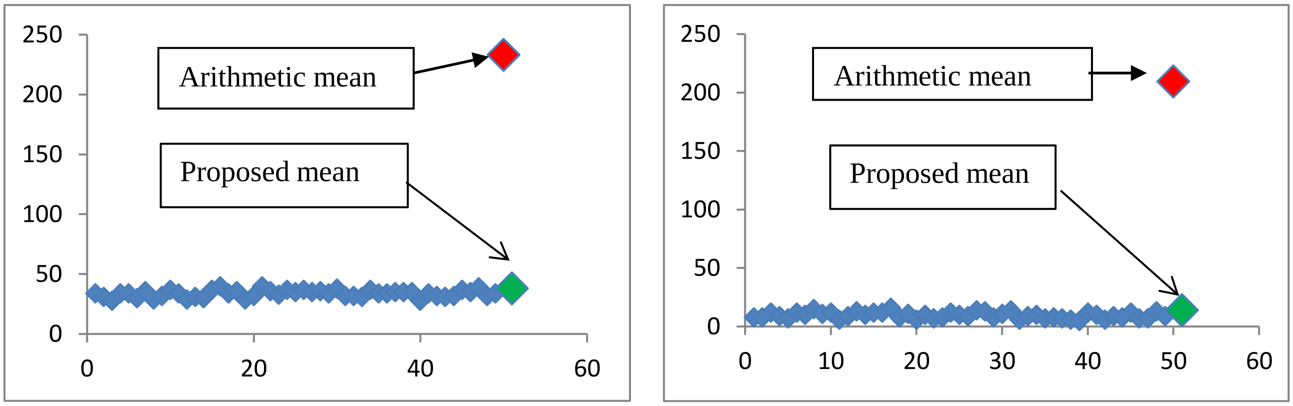


(c) Uniform (d) Normal

Figure 1: Graphical presentation of the data sets considered in this paper.

The graphical presentation of the data sets excepting outlier (10,000) with the arithmetic mean and proposed mean for the observations (in the presence of outlier) are given below (Figure 2). For all cases it is observed that the proposed mean are more and more close to the observations than the arithmetic mean in the presence of outlier. Thus, this paper strongly suggest to the proposed mean in the presence of outlier.

(a) Binomial (b) Poisson



(c) Uniform (d) Normal

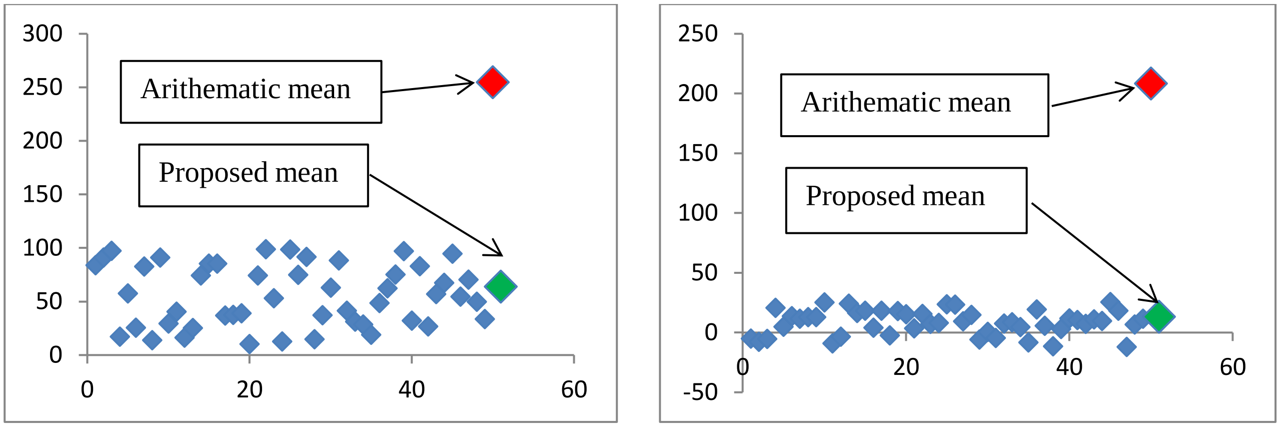


Figure 2: Graphical presentation of the data sets (except outlier) with arithmetic mean and proposed mean considered in this paper.

4 Conclusion

This paper compares the results of proposed mean to the well established and most frequently used arithmetic mean in the presence of outlier. The results of the four data sets considered in this study reveal that the proposed mean is less affected by the outlier than the frequently used arithmetic mean. Thus, this paper suggests to use the proposed mean in the presence of outlier in any field.

Acknowledgement

The authors are very much thankful to the editor-in-chief and the anonymous referee for critically examining the manuscript and giving valuable suggestions to improve it.

References

[1] Plackett, R. L., “Studies in the History of Probability and Statistics: VII. The Principle of the Arithmetic Mean", Biometrika **45(1/2)**, 130-135 (1958). doi:10.2307/2333051.

[2] Churchill Eisenhart, "The development of the concept of the best mean of a set of measurements from antiquity to the present day", Unpublished presidential address, American Statistical Association, 131st Annual Meeting, Fort Collins, Colorado, (1971).

[3] Bakker Arthur, "The early history of average values and implications for education", Journal of Statistics Education **11(1)**, 17-26 (2003).

[4] Ray, John, A Collection of English Words Not Generally Used, London: H. Bruges, (1674). Retrieved 18 May 2015.

[5] Stanford Encyclopedia of Philosophy, Available on (http://plato.stanford.edu/entries/william-jevons/). Retrieved 10 February 2016.