Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/040203

Variance in the Presence of Outlier: Weighted Variance

Md. Moyazzem Hossain[*]

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

Received: 24 Dec. 2016, Revised: 10 Mar. 2017, Accepted: 13 Mar. 2017
Published online: 1 May 2017

Abstract: In Statistics, an outlier is an observation point that is distant from other observations. Usually, the outlier is removed from the data set for further analysis as a consequence the degrees of freedom is lost. The variance has a central role in Statistics. It is extremely affected by the outlier. This paper propose the new formula for computing the variance called weighted variance that are not unduly affected by outliers. Thus, this paper suggests to use the proposed variance (weighted variance) in the presence of outlier in any field of study.

Keywords: Outlier, Variance

1 Introduction

The term variance was first introduced by Ronald Fisher [2] in his paper entitled *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. Kagan and Shepp [1] stated that the second moment of a random variable attains the minimum value when taken around the mean of the random variable. The variance has a central role in Statistics. It is used in descriptive statistics, statistical inference, hypothesis testing, goodness of fit, Monte Carlo sampling, amongst many others. This makes it a central quantity in numerous fields such as physics, biology, chemistry, economics, and finance. The variance of a data set is calculated by taking the arithmetic mean of the squared differences between each value and the mean value. Squaring the difference has some advantages like (i) squaring makes each term positive so that values above the mean do not cancel values below the mean, (ii) squaring adds more weighting to the larger differences, and in many cases this extra weighting is appropriate since points further from the mean may be more significant and so on.

An outlying observation, or ”outlier” is one that appears to deviate markedly from other members of the sample in which it occurs (see Grubbs [3]). The arithmetic mean is extremely affected by outlier as a consequence variance also affected by outlier. Usually, the outlier is removed from the data set for further analysis which reduces the degrees of freedom. Hossain [4] proposed a formula for computing mean in the presence of outlier. He shows that the proposed mean is less affected by the outlier. Here, his proposed mean is used to define the weighted variance. The main aim of this paper is to propose the new formula for computing the variance called weighted variance that are not unduly affected by outliers.

2 Methods

2.1 Variance

The variance or standard deviation (positive square root of variance) is the most commonly used and readily understood measure of dispersion. It is used to measure the variability of a data set. Variance is defined as the squares of deviations of given observations from their arithmetic mean.Symbolically, if we have a data set containing the values $x_1, x_2, x_3, \ldots, x_n$.

The variance is defined by the formula

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \tag{1}$$

where, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

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

2.2 Proposed Weighted Variance

This paper propose a formula for variance which is called weighted variance and is too much less affected by the outlier compare to the well-established variance. It is supposed that a data set having an outlier and containing the values $x_1, x_2, x_3, \ldots, x_n$. Among them it is assumed that it contains at least one outlier say, x_n. The proposed formula for variance is given by,

$$S_w^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_w)^2}{n} \tag{2}$$

where, $\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$, $w_i = \frac{1}{|x_i - \bar{x}|}$, $\bar{x}$ represent the arithmetic mean is defined by the formula $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$. Here, $\bar{x}_w$ is proposed by Hossain [4].

3 Results and Discussion

To compare the result of the well-established variance with the proposed variance we consider three data sets each of which contains 49 observations that are generated from different distributions and one observation (outlier) is added artificially. Here, we generate 49 observations from Binomial (49, 0.75), Uniform (0,100) and Normal (50, 100) respectively. The generated observations and an added outlier are given in Table 1. The 50th observation considered as outlier for each distribution.

Table 1: Generated observations from different distributions with an outlier

Serial No.	Binomial	Uniform	Normal	Serial No.	Binomial	Uniform	Normal
1	36	36.055	65.416	26	38	59.468	46.103
2	42	29.984	59.737	27	34	85.418	44.038
3	36	67.675	50.307	28	34	54.772	44.567
4	36	37.919	45.066	29	35	75.176	58.908
5	37	1.9990	38.710	30	37	68.358	38.024
6	34	61.043	47.740	31	43	56.240	46.753
7	37	52.858	55.693	32	31	48.454	46.483
8	39	52.171	57.713	33	35	79.766	60.575
9	36	35.176	61.770	34	37	85.357	43.432
10	43	77.667	50.191	35	37	49.852	38.755
11	33	22.694	63.309	36	34	23.072	53.67
12	36	79.055	39.600	37	33	21.796	36.458
13	41	58.843	55.999	38	37	64.937	38.388
14	39	53.291	55.654	39	35	84.121	48.935
15	38	77.233	43.285	40	37	31.114	75.84
16	41	7.3150	54.314	41	33	75.481	54.516
17	32	90.002	42.977	42	41	39.402	50.239
18	37	37.150	60.200	43	37	47.349	53.661
19	37	18.049	40.246	44	33	2.0510	33.834
20	36	34.449	67.435	45	43	19.254	55.487
21	35	79.604	46.795	46	40	22.889	53.505
22	29	96.182	36.075	47	41	98.737	26.671
23	39	35.002	51.375	48	35	3.4700	47.086
24	36	68.834	49.785	49	35	64.858	46.578
25	37	18.549	54.949	50	**1000**	**1000**	**1000**

Now, mean and variance is computed for the 50 observations for the data sets using the formula (1) and (2). Avoiding the outlier for the three data sets considered in this paper i.e., on the basis of 49 observations, mean and variance also computed here. Finally, all the results are presented in Table 2.

The generated data sets given in Table 1, having the minimum value 29, 2 and 26.67 for the data set comes from Binomial, Uniform and Normal distribution respectively. In addition, the maximum value of these data sets (49 observations) are 43, 98.74 and 75.84 respectively. Artificially an outlier **1000** has been added in each data set. Now, in case of Binomial distribution, the mean and variance for 49 observations are about 37 and 10 respectively whereas for 50 observations i.e., in the presence of outlier they are approximately 56 and 18569. Here, it is observed that the variance is extremely affected by the outlier whereas the proposed variance gives the result about 20 which is closer to variance of the data set excluding the outlier. For the other data sets, the results are looking same. Thus, it is may conclude that the usual variance is very much affected by the outlier whereas the proposed variance is less affected by the outlier [Table 2].

Table 2: Results of three data sets

Statistic	Binomial	Uniform	Normal
Minimum (for 49 observations)	29	2.00	26.67
Maximum (for 49 observations)	43	98.74	75.84
Outlier	1000	1000	1000
$\bar{x}$ (for 49 observations)	36.67	50.82	49.73
$\bar{x}$ (presence of outlier)	55.94	69.80	68.74
$\bar{x}_w$ (presence of outlier)	37.60	65.47	57.24
S_x^2(for 49 observations)	9.81	699.43	92.22
S_x^2 (presence of outlier)	18569.57	18704.19	18150.54
S_w^2(presence of outlier)	20.16	42.36	26.95

4 Conclusion

This paper compares the results of proposed variance to the well-established and most frequently used variance in the presence of outlier. The results of the three data sets considered in this study reveal that the proposed variance is less affected by the outlier than the usual variance. Thus, this paper suggests to use the proposed variance (weighted variance) in the presence of outlier in any field of study.

Acknowledgement

The author is very much thankful to the editor in chief and the anonymous referee for carefully examining the manuscript and for helpful comments that improved this paper.

References

[1] A. Kagan and L. A. Shepp, Why the variance?, *Statistics & Probability Letters*, **38**, 329-333 (1998).

[2] R. Fisher, The correlation between relatives on the supposition of Mendelian Inheritance, *Philosophical Transactions*, **52**, 399433 (1918).

[3] F. E. Grubbs, Procedures for detecting outlying observations in samples, *Technometrics*, **11(1)**, 121 (1969). doi:10.1080/00401706.1969.10490657.

[4] M. M. Hossain, Proposed Mean (Robust) in the Presence of Outlier, *Journal of Statistics Applications & Probability Letters*, **3(3)**, 103-107 (2016).

Md. Moyazzem Hossain is working as an Assistant Professor of Statistics at Jahangirnagar University, Bangladesh. His research interests are on regression analysis, econometrics, time series forecasting, multivariate analysis and environmental statistics. He has published research articles in reputed international journals. Currently, he also served as an Editorial board member of American Journal of Statistics and Probability.