

Alternative Measures of Standard Deviation Coefficient of Variation and Standard Error

Md. Siraj-Ud-Doulah

Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

Abstract The Paper examines the performances of some popular and frequently used measures of dispersion such as standard deviation, coefficient of variation and standard error and survey that they may not perform as good as anticipating in the presence of non-normality or outliers. The usefulness of the proposed measures is scrutinized with the frequently used measures of dispersion by bootstrap and jackknife computer based techniques as well as Monte Carlo simulation approach. In this paper, I propose new alternative measures of dispersion, namely DM-standard deviation, coefficient of deviation and DM-standard error. These measures should be fairly robust. The results demonstrate that DM-standard deviation, coefficient of deviation and DM-standard error outperforms than the standard deviation, coefficient of variation and standard error in a broad range of frequently occurring situations.

Keywords Bootstrap, Coefficient of Variation, Coefficient of Deviation, DM-standard Deviation, DM-standard Error, Jackknife, Monte Carlo Simulation

1. Introduction

Various statistical methods, when we are interested in describing the entire distribution of some observations and characteristics of individuals, there are two measures: central tendency and dispersion. Former measures locate the center of a distribution but they do not reveal how the observations are spread out on each side of a center. On the other hand, the latter measures are numbers that attempt to answer questions like: how do the observations spread out around the typical value? Both are used to describe the distribution of observations or characteristics of the individual under study for any accurate descriptive summary of the data. But there is evidence that the existing measures of dispersion may perform poorly in the presence of non-normality or when outliers arise in data. There are frequently used different measures of dispersion. The theoretical and practical application of these measures have been documented in several books [1-3, 5, 7, 8, 10-15, 18, 20, 21, 23, 24] and journal articles. Yet, most researchers continue to search and develop advance measures of dispersion for making correct inferences. I briefly review some of the more fundamental problems with conventional measures. Range depends solely on the highest and lowest

values, it would be misleading if any of the two extreme values has very high or low magnitude. Quantile deviation (QD) is a better measure than range, but both fail to provide measurement of scatter of the observations, relative to the typical value. In addition, it does not enter into any of the higher mathematical relationships that are basic to inferential statistics. Mean deviation (MD) is highly sensitive to unusual observations in a series. The most popular and frequently used scale measure is standard deviation (SD), but it is very much sensitive to outlying observations. Therefore, I propose an alternative measure of scale is DM-standard deviation () namely, which I discuss in section 2. But when the means of data sets vary considerably, we do not get an accurate picture of the relative variability in two sets just by comparing the SD. In that situation, we use another measures of dispersion tends to the overcome this difficulty. The coefficient of range is rarely used on a measure of dispersion because of its inherent difficulties in interpretation. Coefficient of QD fails to identify discrepancy of the observations. Coefficient of MD is sensitive to outliers. The most popular and commonly used measure is coefficient of variation (CV), which based on mean and standard deviation (SD), so, it's sensitive to contamination of data. Consequently, I develop a sound measure of CV is coefficient of deviation (CD) explicitly, which discuss in section 2. The concept of standard error (SE) is the one and only best understood with reference to a sampling distribution. Since, SE is calculated by sample variance, which is however known to be extremely sensitive to outliers. Hence, to overcome the sensitivity, I propose a new measure of SE is DM-standard

* Corresponding author:

sdoulah_brur@yahoo.com (Md. Siraj-Ud-Doulah)

Published online at <http://journal.sapub.org/statistics>

Copyright © 2018 The Author(s). Published by Scientific & Academic Publishing

This work is licensed under the Creative Commons Attribution International

License (CC BY). <http://creativecommons.org/licenses/by/4.0/>

error () namely, which I discuss in section 2. The performance comparisons of classical popular and frequently used measures as well as my newly proposed measures are discussed by utilizing real life data sets in section 3. Bootstrap and jackknife properties of classical and proposed estimators are illustrated in section 4 and 5 respectively. The performances of the classical and proposed estimators are investigated in section 6 through a Monte Carlo simulation studies in section 4.

2. Propose SD_{DM} , CD & SE_{DM} Estimators

The sample average and sample standard deviation are the classical estimators of the location and scale parameter of a statistical distribution. They are however unreliable in presence of outliers. For that reason, I propose sound measure of scale, so that it can give right outcome any typical data situation. I now define some well-known and simple robust estimator of location. The most popular robust estimator of location is decile mean (DM). A survey on DM is given by [19]. Here, I formulate DM as-

$$DM = \frac{D_1 + D_2 + \dots + D_9}{9}$$

Where D_1, D_2, \dots, D_9 are 9 deciles from grouped or ungrouped data.

Now I define an alternative scale estimator is

$$SD_{DM} = \sqrt{DM (X_i - DM)^2}$$

$$\Rightarrow SD_{DM} = \sqrt{\frac{\text{sum of 9 deciles of } (X_i - DM)^2}{9}}$$

Next, I define a coefficient of deviation (CD) is computed as a ratio of SD_{DM} of the distribution to the DM of the same distribution.

$$\text{Symbolically, } CD = \frac{SD_{DM}}{DM}$$

The CD is usually expressed in percentage, in which case

$$CD = \frac{SD_{DM}}{DM} \times 100$$

And then I define, the term, 'standard error' applies to DM, denoted by SE_{DM} , is computed as

$$SE_{DM} = \frac{SD_{DM}}{\sqrt{n}}$$

3. Real Data Example

The score data has taken from Mardia, Kent and Bibby (1979), which contains 88 students' examination marks, each gave five tests in two different categories: one is closed book examination, which included the subjects, in mechanics and vectors and another is open book

examination, which included the subjects, in algebra, analysis and statistics. Now I find out which type of examination is sound for students. At first, I check the outliers independently for closed and open book tests by the robust Med-MAD method [4, 22], it detects 4 outliers (case 1, 81, 85 and 87) for closed book test as well as 5 outliers (case 1, 2, 3, 87 and 88) for open book test. Actual data set and deleting these outliers I revisit the data set which result has shown below-

Table 1. Comparison of Classical and Propose Estimators

Estimators	Closed Test		Open Test	
	With Outliers	Without Outliers	With Outliers	Without Outliers
SD	27.0749955	23.6340035	46.8391038	32.3858560
CV	0.3023604	0.2597483	0.3115762	0.2333964
SE	2.8862042	2.5786812	4.9930652	3.5548095
SD_{DM}	22.8995229	21.1144091	33.4190103	30.26586
CD	0.2543763	0.2312922	0.2295087	0.21774
SE_{DM}	2.4410973	2.3037709	3.5624785	3.32211

From Table 1 I demonstrated that when the data set holds outliers the classical estimators have given opposite conclusion that is the closed test is much better than open test book performance for student. Alternatively, my newly proposed estimators have given correct outcome when the data set contain outliers or free from outliers. It is to be worth-mentioned that Mardia, Kent and Bibby (1979) analyzed this data set by principal component analysis (PCA) method and recommended that the open book test is much better than closed book test for students.

4. Bootstrap

The bootstrap is a recently developed technique for making certain kinds of statistical inferences. It is only recently developed because it requires modern computer power to simplify the often intricate calculations of traditional statistical theory [6]. Here, I consider a real data set for getting the right finding that is the score test data set has taken from [16]. This data set contains 88 observations of student's test. Now, I perform a bootstrap study in order to evaluate the performance of the classical measures that are standard deviation (SD), coefficient of variation (CV), standard error (SE) and our newly proposed estimators of dispersion that is DM-standard deviation (SD_{DM}), coefficient of deviation (CD), DM-standard error (SE_{DM}). I investigate the standard error, bias and percentile length of those six estimators of dispersion when the data set holds outlying observations or free from outlying observations. I utilize those formulae for getting output of the bootstrap program, for some convenient value of 10,000 replications. The results of bootstrap are shown in Table 2 and Table 3.

From Table 2 & 3 I demonstrated that the standard error (SE), bias and percentile length of my proposed estimators

are very low than classical estimator of dispersion when the data set holds outliers or free from outliers. Notice that our classical tools declare that closed test is much better than open test of book performance when data set contains outliers. It is to be worth-mentioned that bootstrap is very

sensitive to outliers. Yet, when the data set contains outliers or free from outliers, my newly proposed measures declares that open test is very much helpful achieving high score than closed test book performance.

Table 2. Bootstrap Results Comparison of Six Different Measures of Dispersion for Score Data set With Outliers and Without Outliers

Estimators	Standard Deviation (SD)		Coefficient of Variation (CV)		Standard Error (SE)		DM-standard deviation (SD_{DM})		Coefficient of Deviation (CD)		DM-standard error (SE_{DM})	
Test Type	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test
Data Situation	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers
Actual Value	22.9	21.11	33.42	30.15	0.2295	0.2313	0.2544	0.2129	2.441	2.304	3.562	3.62
Bootstrap Value	23.2861	21.2128	33.9293	29.8371	0.232297	0.23396	0.2596	0.211084	2.48062	2.31657	3.60785	3.58992
S.E	2.395	1.68	5.021	2.413	0.02306	0.02014	0.21615	0.01647	0.2242	0.1833	0.4025	0.245
Bias	0.3861	0.1028	0.5093	-0.3129	0.002797	0.00266	0.00520	-0.001816	0.03962	0.01257	0.04585	-0.03008
									-0.2299	1.0121	1.584	4.37
											2.185	2.185
									-0.004265	0.02177	0.01893	0.02075
									-0.002141	0.01893	0.02075	0.01535
									-0.00210	0.02075	0.01535	0.01535
									-0.001932	0.01535	0.01535	0.01535
									-0.02822	0.02241	0.1663	0.3699
									-0.02007	0.1663	0.3699	0.2276
									-0.05963	0.3699	0.2276	0.2276
									-0.03114	0.2276	0.2276	0.2276

Table 3. Comparison of Percentile Length of Six Different Estimators of Dispersion for Score Data with Outliers and without Outliers by Bootstrap

Estimators	Standard Deviation (SD)		Coefficient of Variation (CV)		Standard Error (SE)		DM-standard deviation (SD_{DM})		Coefficient of Deviation (CD)		DM-standard error (SE_{DM})	
Test Type	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test	Closed Test	Open Test
Data Situation	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers
Percentile	0.05	0.95	0.05	0.95	0.05	0.95	0.05	0.95	0.05	0.95	0.05	0.95
Length	6.93	5.49	10.91	7.88	0.0658	0.0659	0.0859	0.0535	0.723	0.596	1.137	0.872
									4.95	4.19	5.73	4.23
											0.0586	0.0502
											0.0445	0.0369
											0.023	0.55
											0.893	0.602

and another is free from outliers. Now, I compute a jackknife report in order to examine the performance of the classical measures i.e., standard deviation (SD), coefficient of variation (CV), standard error (SE) and our newly proposed estimators of dispersion i.e., DM-standard deviation (), coefficient of deviation (CD), DM-standard error (). I investigate those six estimators to find out appropriate inference tools in compare to standard error, bias and percentile length of jackknife. The jackknife focuses on the samples that leave out one observation at a time. Now the outcome of jackknife are given above-

As it can be seen, in the presence or absence of outlying observations, the standard error (SE), bias and percentile length of the classical estimators of dispersion have given high values than my newly proposed estimators of dispersion. Mention that popular and commonly used classical estimators certify that closed test is much better than open test book performance when outliers occur in the data set. Although, it is well known that jackknife is very much sensitive to outliers, in the meantime, my proposed measures announce that open test is very much effective than closed test book performance for student when the data set holds outliers or free from outliers, my newly proposed estimators which provide the best results in this cases; such results are given in Table 4 & 5.

6. Monte Carlo Simulation Study Report

In the previous section, I consider real life data to see how the outliers affect the right conclusion. I also analyze this data set by bootstrap and jackknife computer based method to find out the robust measure for right inferences. But these are not enough to get a final answer to a problem. In this section, I study a Monte Carlo simulation that is planned to compare the power performance of the classical estimators for appropriate solution in any situation of data. I simulate artificial two sets data, A and B, namely from uniform distribution. Firstly I make sure that the data set A is much better than the data set B. In general, I also simulate to compare the performance of my newly proposed estimators with classical popular and frequently used estimators of dispersion and compute the simulated mean, bias and standard error (SE). I now consider five different sample sizes, $n=50, 100, 200, 500$ and 1000 . Then I generate sample from uniform distribution based on A and B and in general. Each experiment is run 10,000 times and the results are shown below-

Table 6. Power comparison of different estimators of dispersion

Estimators	Power (in percentage)									
	CV	CD	CV	CD	CV	CD	CV	CD	CV	CD
n	50		100		200		500		1000	
A	11.46	97.19	14.08	99.89	21.69	100	33.79	100	42.94	100
B	88.54	2.81	85.92	0.11	78.31	0	66.21	0	57.06	0

Table 7. Simulation Performance Comparison of Different Estimators

Estimators	Actual Value	Simulated Value	S.E	Bias
	n=50			
SD	61.3563795	57.6044521	0.5760445	3.7519275
CV	0.579070329	0.580670787	0.000612220	0.00160045
SE	8.018350915	8.149444051	0.005361576	0.13109313
SD_{DM}	55.1299815	54.8725493	0.5487255	-0.2574322
CD	0.581342608	0.553399374	0.000580793	-0.0279432
SE_{DM}	9.215495569	7.764348133	0.005259535	-1.4511474
	n=100			
SD	58.1547975	57.6970078	0.5769701	0.4577896
CV	0.510562201	0.557513743	0.000409313	0.04695154
SE	5.643008691	5.763477345	0.002698305	0.12046865
SD_{DM}	57.6860589	55.5502858	0.5555029	-2.1357731
CD	0.616175953	0.578912512	0.000400295	-0.0372634
SE_{DM}	5.817906585	5.553043202	0.002526841	-0.2648633
	n=200			
SD	50.1342795	55.5343112	0.5773431	5.4000318
CV	0.555285558	0.578187696	0.000299138	0.02290213
SE	5.555055387	6.047684318	0.002645403	0.49262893
SD_{DM}	55.7001041	57.7381886	0.5553819	-2.0380845
CD	0.597877348	0.559805594	0.000292907	-0.0380717

SE_{DM}	4.125784490	4.079644093	0.001302364	-0.0461403
n=500				
SD	58.3567133	57.7234621	0.5772346	0.6332512
CV	0.550218171	0.577658248	0.000188169	0.02744007
SE	2.482114481	2.509731322	0.000587134	0.02761684
SD_{DM}	56.6400787	56.1177185	0.5611772	-0.5223602
CD	0.597368232	0.561565345	0.000181706	-0.0358028
SE_{DM}	2.627204253	2.581672749	0.000515339	-0.0455315
n=1000				
SD	58.0521474	57.7392475	0.5773925	0.3129000
CV	0.549221234	0.562272413	0.000138702	0.013051180
SE	1.825768973	1.826015845	0.000257914	0.00024687
SD_{DM}	56.2986188	56.1928434	0.5619284	-0.1057754
CD	0.597843425	0.577512468	0.000129218	-0.0203309
SE_{DM}	1.816411606	1.776954847	0.000254864	-0.0394567

From Table 6 I observed that in different sample sizes the accepting power of coefficient of deviation (CD) is very high than coefficient of variation (CV) when the data set A is true my simulation. The performance of coefficient of variation (CV) is not satisfactory as its power is less than 50% for this simulation in different samples.

From Table 7 I showed that the standard error and bias of my newly proposed estimators are very low than classical estimators in different sample sizes, which offer the sound outcome in this simulation.

7. Conclusions

In this paper, I have considered six estimators of dispersion to find out the right decision. In section 3 I have taken real data set for correct judgment. I demonstrated that the classical estimators of dispersion have given opposite results when the data set contains outliers. Alternatively, my proposed estimators have provided appropriate choice in both cases. In section 4 and 5 I observed that the standard error (SE), bias and percentile length of my proposed estimators presented very low than the classical estimators of dispersion by bootstrap and jackknife computer based methods. In section 6, Monte Carlo simulation confirmed that the newly proposed estimators perform better than any other classical estimators of dispersion for different sample sizes. Therefore, from the survey of aforesaid discussion I observe that my newly proposed three estimators perform very sound in every situation. So, I recommend utilizing these.

ACKNOWLEDGEMENTS

The author would like to thank the obscure reviewers for their helpful comments to enrich the quality of this paper.

REFERENCES

- [1] D.R. Anderson, J. Dennis, and A. Thomas, Introduction to Statistics, West Publishing Company, New York, 1981.
- [2] S.S. Blank, Descriptive Statistics, Appleton Century Crofts Publication, Michigan, 2010.
- [3] K. Black, Business Statistics: Contemporary Decision Making, John Wiley & Sons, 2009.
- [4] V. Barnett, and T. Lewis, Outlier in Statistical Data, 3rd Edition, Wiley, New York, 1994.
- [5] J.L. Devore, and K.N. Berk, Modern Mathematical Statistics with Applications, 2nd Ed., Springer, 2011.
- [6] B. Efron, and R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, 1993.
- [7] E.A. Freedman, Statistics, John Wiley & Sons, 1991.
- [8] W.L. Hays, and R.L. Winkler, Statistics Probability Inference and Decision, Holt, Rinehart and Winston Publication, Michigan, 2010.
- [9] P.J. Huber, and E.M. Ronchetti, Robust Statistics, A John Wiley & Sons Inc. Publication, 2009.
- [10] W.L. Hays, Statistics, Harcourt Brace College Publishers, California, 2008.
- [11] R.V. Hogg, and J. Ledolter, Applied Statistics for Engineer and Social Scientists, 2nd Edition, Macmillan Publishing Company, New York, 1992.
- [12] Z.A. Karian, and E.A. Tanis, Probability and Statistics: Explorations with MAPLE, 2nd Edition, Prentice Hall, NJ, 1999.
- [13] R.N. Lussier, Research Methods and Statistics for Business, Waveland Publication Inc, 2011.
- [14] R.N. Lussier, Basic Descriptive Statistics for Decision Making, Addison Wesley Publication, 2011.
- [15] J. Levin, and J. A. Fox, Elementary Statistics in Social Research, Longman, 1997.

- [16] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.
- [17] R.A. Maronna, R.D. Martin, and V.J. Yohai, *Robust Statistics*, Wiley, New York, 2006.
- [18] A. Neil, Hassett, M.J. Weiss, *Introductory Statistics*, 3rd Ed., Addison Wesley Publication, 1991.
- [19] Rana, S., Doulah, M.S.U., Midi, H., and Imon, A.H.M.R., 2012, Decile Mean: A New Robust Measure of Central Tendency., *Chiang Mai Journal of Sciences*, 39(3), 478-485.
- [20] O.W. Winkler, *Descriptive Statistics*, Springer, New York, 2009.
- [21] N.A. Weiss, *Introductory Statistics*, 7th Ed., Addison Wesley Publication, 2004.
- [22] Wilcox, R.R., and Keselman, H.J., 2003, Modern Robust Data Analysis Methods: Measure of Central Tendency., *Psychological Methods*, 8(3), 254-274.
- [23] N. Weiss, *Descriptive Statistics*, McGraw-Hill, 2002.
- [24] N.A. Wiss, *Elementary Statistics*, 4th Ed., Addison Wesley Publication, 1998.