Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 7, pp. 1335–1342
DOI : 10.47974/JSMS-1113

On estimating the variance of the sample median

Hadi Alizadeh Noughabi *
Department of Statistics
University of Birjand
Birjand 97135
Iran

Mohammad Shafaei Noughabi †
Department of Mathematics and Statistics
University of Gonabad
Gonabad 95715
Iran

Abstract

Recently, Ozen et al. [6] proposed the nonparametric bootstrap method for estimating the variance of sample median and showed that this method has better efficiency for symmetric or right skewed distributions when the sample size is small. The present article proposes a new estimator for the variance of sample median. The mean squared error, variance, and bias of the new estimator is calculated and compared with those of the other estimators. The results of our simulation study show the new estimator has better efficiency for different distributions in terms of MSE.

Subject Classification: *(2000) 62G05, 62G30.*

Keywords: *Variance, Sample median, Estimation, Simulation study, Mean squared error.*

1. Introduction

In statistics, the median is a measure of central tendency that represents the value that separates a set of data into two equal parts. It is the middle value in a sorted list of data, or the average of the two middle values if there is an even number of values.

* *E-mail:* **alizadehhadi@birjand.ac.ir (Corresponding Author)**
† *E-mail:* **mohamad.shafaee@gmail.com**

The median is a well-known statistic and useful for data with extreme values or outliers than other measures of central tendency like the mean. This makes it a robust measure of the typical value of a set of data, particularly if data are skewed.

The median has many applications in statistics, including:

1. **Descriptive statistics:** The median is a commonly used descriptive statistic that provides a quick summary of the typical value of a set of data.
2. **Inferential statistics:** The median is used in various inferential statistical approaches, such as confidence intervals, hypothesis testing, and regression analysis.
3. **Survival analysis:** The median is often used in survival analysis to estimate the time at which a certain proportion of a population is expected to experience an event of interest.
4. **Data visualization:** The median is often used in data visualization techniques, such as box plots and violin plots, to provide a visual representation of the central tendency of a set of data.

Therefore, the median is a useful statistic in statistics that provides a robust measure of the typical value of a set of data, and has a wide range of applications in various fields, see for example Maritz (1974).

The sample median and estimating the variance of it is an important problem in statistics, particularly in situations where the underlying distribution is unknown or non-normal. The sample median is a robust estimator for the population median, which makes it preferable to the sample mean in many cases, especially when dealing with skewed and heavy-tailed distributions.

The variance of the sample median can estimate using various methods, including asymptotic theory, resampling techniques, and bootstrap methods. Asymptotic theory provides an approximation of the variance based on the central limit theorem (CLT), assuming that the sample size is large enough. However, this method may not be reliable when size of the sample is small or when the underlying distribution is non-normal.

Resampling techniques, such as the jackknife and the permutation test, provide a non-parametric approach to estimating the variance of the median. The jackknife involves systematically leaving out one observation at a time and calculating the sample median each time, while the permutation test involves randomly permuting the observations and calculating the sample median for each permutation. These methods are computationally intensive but can be more accurate than asymptotic methods.

Bootstrap methods involve generating multiple samples from the original sample by randomly sampling with replacement, and then calculating the sample median under each sample obtained by bootstrap. The variance of the sample median can then be estimated from the distribution of the bootstrap medians. Bootstrap methods are flexible and can be used with any sample size and any distribution, but they can also be computationally intensive.

Suppose that $X_1, \ldots, X_n$ are a sample from density *f* with distribution function *F*. It is well-known that the sample median *m* is defined as $m = \inf\{x \mid F(x) \geq 0.5\}$. If $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denotes the ordered sample, then the median is defined for *n* odd, as $X_{((n+1)/2)}$ and for *n* even, as $(X_{(n/2)} + X_{(n/2+1)})\ /2$.

It is well known that the distribution of the median of a sample of size n, from a population having the density function $f(x)$, is asymptotically normal with mean *m* and variance $\{4n(f(m))^2\}^{-1}$, where *m* is the median of the population. (See, for example, Cramer [1]). However, it is clear that if the density function $f(x)$ is unspecified, the formula of asymptotic variance will not be useful. Moreover, for large sample sizes the asymptotic variance formula can be valid. Therefore, when the sample size is small, new methods were suggested. Comparison of this approximation with the true variance for some well-known distributions is performed by Rider [8]. He did some comparisons for small sample sizes based on relative errors. Some estimators of the variance of sample median are suggested by authors, but they are used for small sample sizes, either Maritz and Jarrett [5], Harrell and Davis [4], Sheather [9], Price and Bonett [7]. A resampling method called bootstrap is suggested by Efron [2]. By using this method, he investigated the asymptotic variance. Moreover, Ghosh et al. [3] proposed a formula for the variance of the median and discussed the validity if it is supported by some well-known theoretical rules. Also, Efron [2] stated the non-parametric bootstrap technique. Clearly it required more computations because it is using Monte Carlo approximation.

Recently, Ozen et al. [6] proposed the nonparametric bootstrap method for estimating the variance of sample median and showed that their method has better efficiency for symmetric or right skewed distributions when the sample size is small.

In this study, we consider the sample median and propose an estimator for variance of it and compare it with two existing methods under symmetric, left or right skewed populations. In Section 2, we state our method and our simulation studies. Our results of comparisons are given in Section 3. Some suggestions and conclusions are presented in Section 4.

2. The Proposed Method

In this section, we consider the sample median and present some estimators for its variance.

Ghosh et. al [3] proposed an estimator for the variance of the sample median as follows.

$$\hat{\sigma}_1^2 = \left\{ \sum_{i=1}^{n} p_n(i) \mathrm{X}_{(\mathrm{i})}^2 - \left(\sum_{j=1}^{n} p_n(j) \mathrm{X}_{(\mathrm{j})} \right)^2 \right\},$$

where

$$p_n(i) = \begin{cases} \sum_{j=0}^{a-1} \left[b_{j,n}\left(\frac{i-1}{n}\right) - b_{j,n}\left(\frac{i}{n}\right) \right] & \text{for } i \geq 2, \\ 1 - \sum_{j=0}^{a-1} b_{j,n}\left(\frac{i}{n}\right) & \text{for } i = 1, \end{cases}$$

and $b_{j,n}(u) = \binom{n}{j} u^j (1-u)^{n-j}$, $a = \left[\frac{n}{2}\right] + 1$.

Non-parametric bootstrap approach is other technique for estimation which is suggested by Efron [2]. If the probability density function f is unspecified, method of the non-parametric bootstrap can apply for the parameter estimation. This method is straightforward and also it has an explicit algorithm. Suppose that $X_1, \ldots, X_n$ are a sample with size of n. We generate B bootstrap samples from the original random sample with replacement sampling. Then the statistic of interest $\hat{\theta}$ is computed for each bootstrap sample. Hence, we can obtain the distribution of the statistic $\hat{\theta}$. In the present paper, we obtain the median based on 1000 bootstrap iterations and then variance of them is computed. We denote this estimator with $\hat{\sigma}_2^2$.

We propose the following estimator for variance of the sample median.

$$\hat{\sigma}^2_{new} = \frac{D_m}{n},$$

where $D_m = \frac{1}{n}\sum_{i=1}^{n}|X_i - \text{median}(X_i)|$ is the mean absolute deviation from the median.

Using a simulation study, we compare the above estimators for samples of sizes 20, 50, 100 and 150 based on 10,000 iterations. We compute the mean square errors (MSE), variances, biases, and means of the mentioned estimators. Also, efficiencies are computed to determine the best approach. We did the bias computations based on the asymptotic variance formula mentioned in the previous section. The value of *m* is the expected median value under the considered model. Efficiencies (Eff) of the proposed method are computed with respect to the other methods and shown as:

$$Eff_i = 100 \times \frac{MSE(\hat{\sigma}_i^2)}{MSE(\hat{\sigma}_{new}^2)}, \quad i = 1, 2.$$

We investigate the performance of the considered estimators under different distributions, e.g., symmetric, right and left skewed. We consider the generalized extreme value (GEV), chi-square distribution and the standard normal distribution. These distributions are considered by Ozen et al. [6].

3. Performance of the estimators

We perform a simulation study to analyze the behaviors of the considered estimators. Some comparisons are done among the considered estimators. First, we generate 100,000 samples of size *n*, and then the values of the considered estimators are obtained. Finally, we compute the mean, bias, variance and mean squared error (MSE) of the mentioned estimators. Moreover, efficiencies of the estimators are obtained. We use the GEV, chi-square, and normal distributions which are the same three distributions considered by Ozen et al. [6].

We generate data sets from the normal distribution with mean zero and standard deviation one and in Table 1 we display the results of our simulations. Efficiency results show that the new estimator gives lower variance and MSE values. From Table 1, it seen that the suggested technique works well and has a good performance when sample size is small.

We use the chi-square distribution with degree of freedom one and generate right-skewed data sets. The results of our simulations are given in Table 2. We can see that the performance of the new estimator is very good. Efficiencies are high and the values of bias (and also MSE) of the new method are lower than the other methods under different sample sizes. Clearly, the proposed approach has better efficiency than others when sample size is small.

We generate left-skewed data sets using the GEV distribution with shape parameter one. In Table 3, results of our simulations are displayed. Clearly, efficiency of the new estimator is high and give lower bias and MSE values. It is obvious that the new method is more efficient for different sample sizes than the other methods.

Table 1

Comparison of the considered estimators for the normal distribution

Estimator	Mean	Bias	Variance	MSE	Eff_1	Eff_2
	$n = 20$					
#1	0.092932	0.014393	0.003215	0.0034224		
#2	0.085978	0.007438	0.002681	0.0027362	206.554	165.139
#3	0.038379	-0.040160	0.000044	0.0016569		
	$n = 50$					
#1	0.035505	0.004088	0.000332	0.000349		
#2	0.035132	0.003716	0.0003275	0.000341	139.911	136.656
#3	0.015710	-0.015705	0.00000289	0.000249		
	$n = 100$					
#1	0.017276	0.001568	0.000058	0.000061		
#2	0.016832	0.001124	0.00005536	0.000056	99.682	92.733
#3	0.007921	-0.007787	0.00000036	0.000061		

Table 2

Comparison of the considered estimators for the chi-square distribution

Estimator	Mean	Bias	Variance	MSE	Eff_1	Eff_2
	$n = 20$					
#1	0.104811	0.048497	0.009627	0.011978		
#2	0.083806	0.027492	0.006536	0.007288	2952.637	179.654
#3	0.041537	-0.014776	0.000187	0.000406		
	$n = 50$					
#1	0.031022	0.008496	0.000461	0.000533		

Contd...

#2	0.028338	0.005813	0.000395	0.0004289	1225.902	986.554
#3	0.016936	-0.005590	0.000012	0.000043		
	$n = 100$					
#1	0.013666	0.002403	0.0000576	0.0000634		
#2	0.012962	0.001699	0.0000504	0.0000533	699.595	588.244
#3	0.008521	-0.002744	0.00000154	0.00000906		

Table 3

Comparison of the considered estimators for the GEV distribution

Estimator	Mean	Bias	Variance	MSE	Eff_1	Eff_2
	$n = 20$					
#1	0.059440	0.009440	0.002031	0.0021199		
#2	0.061225	0.011226	0.002139	0.0022639	608.782	650.139
#3	0.033419	-0.016581	0.000073	0.0003482		
	$n = 50$					
#1	0.022680	0.002680	0.000175	0.0001825		
#2	0.023325	0.003325	0.0001958	0.0002068	406.8225	460.897
#3	0.013670	-0.006329	0.0000048	0.00004487		
	$n = 100$					
#1	0.010989	0.000989	0.0000285	0.0000295		
#2	0.010851	0.000851	0.0000284	0.0000291	285.446	281.842
#3	0.006880	-0.003119	0.00000060	0.0000103		

Form Tables 1-3, it is evident that the new estimator has a good performance and works well under different distributions. When the underling distribution is normal, the new estimator has a better performance than the competitors in terms of variance. For the chi-square population, again the new estimator has the least variance and MSE than the competitors. Under generalized extreme value distribution, we can see from Table 3 that the new estimator has a better performance in compared to the competitor estimators.

Generally, based on our simulation results, we can conclude that the proposed estimator behaves better than the competing estimators.

5. Conclusions

We considered the sample median and its variance. We suggested an estimator for variance of the sample median. We presented the mean, bias, variance and MSE of the proposed estimator. We used a simulation study to compute MSE of the new estimator under various distributions, and compared the proposed estimator in terms of MSE with other estimators. Our simulation results indicated that the proposed estimator performs well under different distributions and also it is an efficient method.

References

[1] H. Cramer, *Mathematical Methods of Statistics*. Princeton U. Press, Princeton (1946).

[2] B. Efron, "Bootstrap methods: Another look at the Jackknife," *Annals of Statistics*, vol. 7, pp. 1-26 (1979).

[3] M. Ghosh, W.C. Parr, K. Singh, and G.J. Babu, "A note on bootstrapping the sample median," *The Annals of Statistics*, vol. 12, no. 3, pp. 1130-1135 (1984).

[4] F.E. Harrell and C.E. Davis, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, no. 3, pp. 635-640 (1982).

[5] J.S. Maritz and R.G. Jarrett, "A note on estimating the variance of the sample median," *Journal of the American Statistical Association*, vol. 73, no. 361, pp. 194-196 (1978).

[6] H. Ozen, E. Colak, C. Bal, F. Mutlu, and K. Ozdamar, "Is nonparametric bootstrap an appropriate technique for estimating variance of the sample median?" *Journal of Statistics and Management Systems*, vol. 24, no. 7, pp. 1523-1531 (2021).

[7] R.M. Price and D.G. Bonett, "Estimating the variance of the sample median," *Journal of Statistical Computation and Simulation*, vol. 68, no. 3, pp. 295-305 (2001).

[8] P.R. Rider, "Variance of the median of small samples from several special populations," *Journal of the American Statistical Association*, vol. 55, no. 289, pp. 148-150 (1960).

[9] S.J. Sheather, "A finite sample estimate of the variance of the sample median," *Statistics & Probability Letters*, vol. 4, no. 6, pp. 337-342 (1986).

Received October, 2022