Dhaka Univ. J. Sci. 60(1): 109-113, 2012 (January)

Globally Robust Confidence Intervals for Location

M. Ershadul Haque and Jafar A. Khan

Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1000, Bangladesh

Received on 25. 05. 2011. Accepted for Publication on 20. 08. 2011

Abstract

Classical inference considers sampling variability to be the only source of uncertainty, and does not address the issue of bias caused by contamination. Naive robust intervals replace the classical estimates by their robust counterparts without considering the possible bias of the robust point estimates. Consequently, the asymptotic coverage proportion of these intervals of any nominal level will invariably tend to zero for any proportion of contamination.

In this study, we attempt to achieve reasonable coverage percentages by constructing globally robust confidence intervals that adjust for the bias of the robust point estimates. We improve these globally robust intervals by considering the direction of the bias of the robust estimates used. We compare the proposed intervals with the existing ones through an extensive simulation study. The proposed methods have reasonable coverage percentage while the existing method show very poor coverage as sample size increases.

Keywords: Coverage percentage (CP), Average length (AL), The pseudo-standard deviation (S_{PS}), The median absolute deviation from the sample median (MAD), Student t confidence interval.

I. Introduction

Data collected in a broad range of applications frequently contain one or more atypical observations called outliers; that is, observations that are well separated from the majority of the data, or in some way deviate from the general pattern of data. These outliers in some sense also contaminate the data.

An outlier may have a serious adverse influence of confidence interval or its coverage probability. Classical inference considers sampling variability to be the only source of uncertainty, and does not address the issue of bias caused by contamination. These outliers may have a strong influence on the classical (Student t) confidence interval in the sense that they pull the width of the confidence interval to much in their direction and alter the coverage probability.

The literature showed that the sample median and inter-quartile range or the sample median and median absolute deviations are indeed more resistant to departures from normality and presence of outliers. Some study incorporates this observation into constructing some interval estimators for the mean of the normal distribution with contaminated data. The sample median (MD) is used to estimate the parameter μ, whereas the population standard deviation σ is estimated by its robust counterparts such as: the Inter-Quartile Range, the Median Absolute Deviation from the sample median (MAD), Gini's mean difference (G). The confidence interval for μ, constructed by these estimators is called naive confidence interval.

Park and Cho (2003)[10] proposed robust design to develop improvement in industrial production. They showed that the sample mean and variance are useful estimates under normality without contamination and the sample median and MAD or the sample median and the IQR are more useful under a contaminated normal. Confidence intervals constructed by this method are resistance to outliers only when the mean of the clean and contaminated data are same, without any restriction of their variances. But, in the contaminated data, if (1-α) proportion of data comes from N(0,1) and α proportion comes from N(τ,1), the method did not work as we shown in our study. Therefore, if outliers are far from the clean data the coverage percentage invariably tend to zero for any nominal level of contamination.

Naive robust intervals replace the classical estimates by their robust counterparts without considering the possible bias of the robust point estimates. Consequently, the asymptotic coverage proportion of these intervals of any nominal level will invariably tend to zero for any proportion of contamination.

A confidence interval called globally robust if it is stable (in the sense of keeping coverage at or above the nominal level) and informative (in the sense of keeping a reasonable average length) not only at the central method, but also over the entire contamination neighborhood.

Adrover et al. (2004)[2] defined globally robust confidence intervals for the location among other things which takes into consideration a large scale of contaminated distributions. They considered intervals that are stable in the sense of achieving coverage near the nominal level and informative in the sense of having short lengths by taking into account the potential bias of the estimates but they did not consider the direction of the bias of the robust estimates used.

In this study, we attempt to obtain globally robust intervals by considering the direction of bias of the robust estimator used. Therefore it is certain that the average length of the proposed method is always less than the existing globally robust confidence intervals method (results not shown in the table). Our result showed that the proposed confidence intervals S_{PS} t* and MAD t* satisfy the two conditions of globally robust confidence intervals under normal and contaminated normal distributions.

The objective of this study is to investigate the coverage percentages of the proposed confidence intervals with the existing ones for different proportion of contamination and for different sample sizes. Such investigations are carried by a simulation procedure to determine the coverage percentage (CP) and the average length (AL) of each confidence interval under the normal assumption with and without contaminated data and then select confidence interval which is more resistance against the presence of outliers or maintain a CP close to the desired nominal confidence interval 100(1-α)% and more informative (smaller AL).

II. Some Robust Estimators and Naive Intervals

We introduce several robust estimators against outliers that are used for constructing naive intervals for mean μ where σ is unknown.

The sample median (MD): The sample median for a random sample of n observations $X_1, X_2, \ldots, X_n$ is defined as follows:

$$MD = \begin{cases} X_{(\frac{n+1}{2})}, & If\ n\ is\ odd \\ \dfrac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & If\ n\ is\ even \end{cases}$$

The sample median is best known for being insensitive to outlier. Under the normal distribution, the efficiency of the sample median drops off rapidly towards its asymptotic value of 0.64 as sample size increases. The sample median has a maximum 50% breakdown point[11]. Also, the sample median is difficult to handle in mathematical equations, does not use all available values and can be misleading in distributions with a long tail because it discards so much information[4]. Even that the sample median has emerged as a good estimator and is generally considered as an alternative average to the sample mean especially when outliers are present in the data. For a normal distribution with mean μ and standard deviation σ, the standard error for the sample median is given by $\sigma_{MD} = 1.253\frac{\sigma}{\sqrt{n}}$.

The pseudo-standard deviation (S_{ps}) and S_{ps} t confidence interval: The pseudo-standard deviation S_{ps} based on the IQR can be written as:

$$S_{PS} = \frac{IQR}{1.349}$$

Under the normal distribution with mean μ and standard deviation σ, the scale estimate is unbiased estimator of σ. It has a breakdown point of 25%, but an efficiency of only 0.37[13].

The S_{ps} t confidence interval[1] is a modification of the Student t confidence interval based on the MD, as an estimate of μ and the pseudo-standard deviation, S_{ps} as an estimate of σ. Therefore the S_{ps} t confidence interval for μ as:

$$MD \pm 1.253 t_{\alpha/2, n-1} \frac{S_{PS}}{\sqrt{n}}$$

The median absolute deviation from the sample median (MAD) and MAD t confidence interval:

For a random sample $X_1, X_2, \ldots, X_n$ with a sample median (MD), the median absolute deviation from the sample median is defined as follows:

$$MAD = 1.4826\, Median\,\{|X_i - MD|\};\quad i = 1,2,\ldots,n$$

The MAD is a more robust scale estimator than the sample standard deviation, measures the deviation of the data from the median. It was proposed first by Hampel (1974)[7], who attributed it to Gauss. It is often used as an initial value for the computation of more efficient robust estimators. The statistic b_nMAD will be an approximately unbiased estimator of σ where, b_n is an correction factor needed to make b_nMAD unbiased when $X_1, X_2, \ldots, X_n$ are normally distributed[11]. This correction factor is given for $n \leq 9$ by:

n:	2	3	4	5	6	7	8	9
b_n:	1.196	1.495	1.363	1.206	1.200	1.140	1.129	1.107

and when $n > 9$ then:

$$b_n = \frac{n}{n - 0.8}$$

Abu-Shawiesh *et al.,* (2009) also consider the MAD t confidence interval for μ, is given as:

$$MD \pm 1.253 t_{\alpha/2, n-1} b_n \frac{MAD}{\sqrt{n}}$$

This confidence interval is based on the sample median, MD, as an estimate for μ and the median absolute deviation from the sample median, MAD, as an estimate for σ.

The Downton estimator (σ*) and Downton t confidence interval: Downton (1966)[6] introduced a family of estimators based on ordered sample values. Among this family of estimators, Downton proposed σ* as an estimator for the standard deviation σ of a normal population. Let $X_1, X_2, \ldots, X_n$ bea random sample from a normal distribution with mean μ and standard deviation σ. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$denotes the corresponding order statistics. The Downton's estimator (σ*) is given by:

$$\sigma^* = \frac{2\sqrt{\pi}}{n(n-1)} \sum_{i=1}^{n} [i - \frac{1}{2}(n+1)] X_{(i)}$$

Downton estimator has been also studied by David (1968)[5], where he showed that this estimator is equivalent to Gini's mean difference which is a robust estimator of the standard deviation σ[8]. Therefore, the Downton estimator can be written using the Gini mean difference, G, as:

$$\sigma^* = \frac{1}{2}\sqrt{\pi}G$$

$$where, G = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}|X_i - X_j|$$

Nair (1936)[9] found that for a normal distribution $\frac{1}{2}\sqrt{\pi}G$ may be used as an unbiased estimator for σ. The Downton estimator has been recommended as a robust scale estimator by Iglewicz (1983)[14] but this gives a little extra protection against outliers[12]. Barnett et al. (1967)[3] studied Downton's estimator and obtained its first four moments in a closed form.

The Downton t confidence interval for μ, is given as:

$$MD \pm 1.253t_{\alpha/2,n-1}\frac{\sigma^*}{\sqrt{n}}$$

III. The Proposed Confidence Intervals

In practice outlier can arise at any one side of the centre of data. If the outliers (may or may not be drastic) occurred at the right side of the centre of data then for the entire population, mean > median, consequently, the naive procedure constructed a confidence interval for μ whose centre is greater than the length calculated as the distance of maximum bias of median and μ. Conversely, if for the entire data mean < median, the naive procedure constructed a confidence interval for μ whose centre is smaller than the length calculated as the distance of maximum bias of median and μ.

The maximum bias of median (MB) can be shown to be $F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right)$, where F(.) is a cumulative distribution function of a N(0,1) variate.

Assumption

In this process, it is assumed that our data contains maximum 25% outliers. Therefore the maximum bias of median used for this study is

$$MB = F^{-1}\left(\frac{1}{2(1-0.25)}\right) = 0.43$$

Therefore the confidence interval suggested in this study is given as:

$$MD - t_{\alpha/2,n-1}\widehat{\sigma_{MD}} - 0.43 \times I(\bar{X} > MD) < \mu$$
$$< MD + t_{\alpha/2,n-1}\widehat{\sigma_{MD}} + 0.43 \times I(\bar{X} < MD)$$

It has a breakdown point of 25%.

Where: I(A) is an indicator function such that $I(x) = \begin{cases} 1; & x \in A \\ 0; & Otherwise \end{cases}$

The S_{ps} t* confidence interval

$$MD - 1.253t_{\alpha/2,n-1}\frac{S_{PS}}{\sqrt{n}} - 0.43 \times I(\bar{X} > MD)$$
$$< \mu < MD + 1.253t_{\alpha/2,n-1}\frac{S_{PS}}{\sqrt{n}} + 0.43 \times$$
$$I(\bar{X} < MD)$$

The MAD t* confidence interval

$$MD - 1.253t_{\alpha/2,n-1}b_n\frac{MAD}{\sqrt{n}} - 0.43$$
$$\times I(\bar{X} > MD) < \mu < MD +$$
$$1.253t_{\alpha/2,n-1}b_n\frac{MAD}{\sqrt{n}} + 0.43 \times I(\bar{X} < MD)$$

IV. Results

In this study, we are interested in comparing and studying the behavior of the proposed confidence intervals with the others under the normal distribution at different proportion of contamination and how the presence of outliers affects them by using a simulation study. The programs by the programming language R for window version 2.9.2 are used to run the simulation and to make the necessary tables. We generated 1,000 normal samples of different sizes, contains various proportion of contamination by considering the following two situations:

- Uncontaminated distribution (Clean data) where all samples are generated from the standard normal distribution i.e., N(0,1)
- Contaminated distribution where outliers are introduced in the data in four different combinations as follows:

❖ **C05N30:** A situation where 95% observation come from N(0,1) and 5% from N(30,1).

❖ **C10N30:** A situation where 90% observation come from N(0,1) and 10% from N(30,1).

❖ **C15N30:** A situation where 85% observation come from N(0,1) and 15% from N(30,1).

❖ **C20N30:** A situation where 80% observation come from N(0,1) and 20% from N(30,1).

The simulated results for coverage percentage (CP) and average length (AL) of the confidence intervals different levels of contamination are given in the Table 1-5.

Table. 1. Coverage percentage and average length for the standard normal distribution

	95% confidence intervals for the existing methods						95% confidence intervals for the proposed methods			
	S_{PS} t		MAD t		Downton t		S_{PS} t*		MAD t*	
n	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
20	93.8	1.10	94.2	1.17	95.5	1.17	95.2	1.53	95.3	1.60
50	94.8	0.69	95.1	0.71	96.1	0.71	95.2	1.12	95.4	1.14
100	94.1	0.49	94.1	0.50	95.3	0.50	94.5	0.92	94.4	0.93
200	95.8	0.35	95.4	0.35	95.6	0.35	95.9	0.78	95.5	0.78
500	94.5	0.22	94.7	0.22	95.0	0.22	94.9	0.65	95.2	0.65

Table. 2. Coverage percentage and average length for the 5% contaminated normal distribution

	95% confidence intervals for the existing methods						95% confidence intervals for the proposed methods			
	S_{PS} t		MAD t		Downton t		S_{PS} t*		MAD t*	
n	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
20	94.0	1.17	95.4	1.26	100	4.17	98.1	1.60	98.5	1.69
50	93.5	0.72	94.7	0.75	100	2.14	98.6	1.15	98.8	1.18
100	92.1	0.52	92.7	0.53	100	1.72	99.8	0.95	99.8	0.96
200	91.4	0.37	91.4	0.37	100	1.20	100	0.80	100	0.80
500	81.1	0.23	81.3	0.23	100	0.76	100	0.66	100	0.66

Table. 3. Coverage percentage and average length for the 10% contaminated normal distribution

	95% confidence intervals for the existing methods						95% confidence intervals for the proposed methods			
	S_{PS} t		MAD t		Downtont		S_{PS} t*		MAD t*	
n	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
20	92.1	1.26	93.9	1.35	100	6.86	98.8	1.69	99.0	1.78
50	89.2	0.79	91.0	0.81	100	4.06	99.8	1.22	99.8	1.24
100	85.1	0.57	86.3	0.57	100	2.81	100	1.00	100	1.00
200	73.1	0.40	73.6	0.40	100	1.97	100	0.88	100	0.83
500	40.4	0.25	40.2	0.25	100	1.24	100	0.68	100	0.68

Table. 4. Coverage percentage and average length for the 15% contaminated normal distribution

	95% confidence intervals for the existing methods						95% confidence intervals for the proposed methods			
	S_{PS} t		MAD t		Downton t		S_{PS} t*		MAD t*	
n	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
20	91.1	1.39	94.3	1.48	100	9.23	99.3	1.82	99.5	1.91
50	83.6	0.90	84.7	0.91	100	5.71	99.8	1.33	99.8	1.34
100	74.8	0.63	74.2	0.62	100	3.77	99.9	1.06	99.9	1.05
200	50.3	0.45	49.5	0.44	100	2.62	100	0.88	100	0.87
500	8.6	0.28	7.7	0.27	100	1.66	100	0.71	100	0.70

Table. 5. Coverage percentage and average length for the 20% contaminated normal distribution

	95% confidence intervals for the existing methods						95% confidence intervals for the proposed methods			
	S_{PS} t		MAD t		Downton t		S_{PS} t*		MAD t*	
n	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
20	91.8	1.61	94.6	1.65	100	11.29	99.7	2.04	99.9	2.08
50	81.2	1.03	80.5	0.99	100	6.66	99.8	1.46	99.9	1.42
100	64.3	0.74	58.6	0.69	100	4.61	99.9	1.16	99.9	1.12
200	28.6	0.52	22.2	0.49	100	3.22	100	0.95	100	0.92
500	1.3	0.33	0.7	0.30	100	2.03	100	0.76	100	0.73

V. Discussion

The performance of the proposed methods for the normal distribution when there are no outliers is examined first. Also, the estimated coverage percentage and the average length for all confidence interval methods are displayed in Table 1. The results in table 1 suggest that the proposed methods have coverage percentage closed to the nominal confidence coefficient when sampling from a normal distribution which is as expected. However, it is less informative than the Student t followed by other methods.

Table 2, 3, 4 and 5 give the estimated coverage percentages and the average lengths for all confidence interval methods under a normal distribution with 5, 10, 15 and 20% contamination respectively. The results in tables 2-5 demonstrated that, although the Student t followed by other method gives smaller average lengths, their asymptotic coverage percentages are close to zero except the Downton t. From these tables it is clear that our proposed methods have the coverage percentage at or above the nominal level, which is also satisfied by Downton t method. But the Downton method has the AL much more than our proposed methods. Consequently, the Downton t method is less informative than our methods. Inspection of these tables also suggest that MAD t* followed by S_{PS} t* confidence intervals are more resistant to drastic outliers. It is evident also, that for the all sample sizes and contaminated normal distribution, MAD t* and S_{PS} t* intervals are resistant to contaminated data and had good coverage percentages with average interval lengths, but MAD t* is better for large sample sizes ($n \geq 100$). At any rate of contamination, we suggest that the MAD t* followed by S_{PS} t* confidence intervals should be used when the population is normal with outliers. The MAD t* followed by S_{PS} t* are more resistant to outliers (may or may not be drastic) than other method.

…………

1. Abu-Shawiesh, M.O., F.M. Al-Athari and H.F. Kittani, 2009. Confidence interval for the mean of a contaminated normal distribution. JAS **9(15)**: 2835-2840.
2. Adrover, J., M. Salibian-Barrera and R. Zamar, 2004. Globally robust inference for the location and simple linear regression models. J. Statis. Plann. Inform., **119:353-375**.
3. Barnett, F., K. Mullen and J.G. Saw, 1967. Linear estimates of a population scaleparameter. Biometrika, **54**: 551-554.
4. Betteley, G., N. Mettrick, E. Sweeney and D. Wilson, 1994. Using Statistics in Industry: Quality Improvement Through Total Process Control. 1st Edn., Prentice Hall International Ltd., London.
5. David, H.A., 1968 Gini's mean difference rediscovered. Biometrika, **55**: 573-575.
6. Downton, F., 1966. Linear estimates with polynomial coefficients. Biometrika, **53**: 129-141.
7. Hampel, F.R., 1974. The influence curve and its role in robust estimation. J. Am. Statist. Assoc., **69**: 383-393.
8. Kendall, M. and A. Stuart, 1958. The Advanced Theory of Statistics, Distribution Theory 3rd Edn., Charles Griffin and Co. Ltd., London.
9. Nair, U.S., 1936. The standard error of Gini's mean difference. Biometrika, **28**: 428-436.
10. Park, C. and B.R. Cho, 2003. Development of robust design under contaminated and non-normal Data. Qual. Eng., 15: 463-469.
11. Rousseeuw, P.J. and C. Croux, 1993. Alternatives to the median absolute deviation. J. Am. Statist. Assoc., **80**: 1273-1283.
12. Sarhan, A.E. and B.G. Greenberg, 1962. Contributions to Order Statistics. 1st Edn., John Wiley and Sons, New York, ISBN: 978-0471754206.
13. Staudte, R.G. and S.J. Sheather, 1990. Robust Estimation and Testing. 2nd Edn., John Wiley and Sons, New York, ISBN: 978-0471-85547-7.
14. Iglewicz, B., 1983. Robust Scale Estimators and Confidence Intervals for Location. In: Understanding Robust and Exploratory Data Analysis, Hoaglin, D.C., F. Mosteller and J.W. Tukey (Eds.). John Wiley and Sons, New York, ISBN: 0-471-38491-7, 405-431.