Original Research Article

A new modified confidence interval estimator of location parameter for skewed distribution

ABSTRACT

A new modified confidence interval has been proposed which incorporates a new point estimator of the location parameter mean for skewed data distribution. Traditionally, the trimmed mean confidence interval estimator (Trm-ci) is a robust method for dealing with the skewness of the underlying data distribution. However, the Trm-ci method trims a certain fraction of endpoint observations to address skewness, which may result in a loss data information. The Students' t confidence interval (t-ci) estimator, while the most efficient estimator at normal models, becomes impractical in a situations where observed data is subject to non-normality due to robustness. In between the two, the median confidence interval estimator (Med-ci) is expected to retain the robustness of Trm-ci and the efficiency of t-ci. The idea behind the proposed new modified confidence interval estimator (Mod-ci) is to consider both the sample mean and sample median simultaneously, while also using end-point information without trimming any observations. As such, the proposed Mod-ci is expected to be as good as or better than other underlying methods regarding robustness and efficiency when dealing with skewed data distributions. In this study, we examine the performance of the new method compared to most commonly used methods through examples and simulating data from skewed distribution with varying degree of skewness. The results of examples and simulations suggest that the proposed method is as good as, or better than other estimators relevant to this study, as measured by estimated coverage probability and the width of associated confidence interval. Therefore, this method is recommended for practical application while dealing with real-life data exhibiting skewness.

***Keywords**: coverage probability, confidence interval estimate of mean, modified t-ci, skewed distribution, simulation*

1. INTRODUCTION

One of the most important tasks in statistical analyses is to estimate the unknown location parameter around which most of the data values tend to cluster. For example, given a sample from a continuous distribution one may have interest in estimating the unknown mean or median of the distribution. The sample mean is the most efficient location estimator given the data distribution is normal (Casella and Berger, 2024; Hogg, McKean, and Craig, 2018). In the violation of normality, however, the estimator mean is not robust. The sample median, on the other hand, is the most robust location estimator in the presence of skewness or outlying observations in the data distribution (Hartwig, et al., 2020, Wilcox 2021). As an alternative to sample mean or median, the trimmed mean is more robust than the mean and more efficient than the median for data with normal models (Hampel, et al., 2011, Portnoy and He, 2000). Indeed, the trimmed mean has become extremely popular due to the fact that it is less sensitive to extreme deviations and heavy-tailed distributions than the ordinary sample mean for years. For example, one may refer to Tukey and McLaughlin (1963), Bickel (1965), and Huber (1972) for accounts of its history and properties. Fortunately, or unfortunately enough, the trimmed mean always trims a fixed fraction of data points at both ends of a data set, no matter whether these points are "good" or "bad". As such, the performance of trimmed mean may not be satisfactory when the underlying data are very "good" or contain "bad" observations only at one end. As such, researchers investigate many alternative estimators of the location parameter in dealing with data distribution with skewness or outlying observations.

In this article, we propose to estimate the unknown population mean μ by a modified estimator, which is a function of sample mean and median, to deal with data with skewness or outlying observations. The modified estimator uses end point data values, but does not trim any data values unlike the trimmed mean.

We study the property of the proposed estimator asymptotically. We assess the performance of the new modified estimator in constructing confidence interval estimator by comparing it with CI estimators involving median and trimmed means via examples and simulations from skewed distribution. It is expected that while keeping robustness of the trimmed mean or median, it retains the efficiency measured by the estimated coverage probability and width of confidence interval estimators.

The organization of the remaining paper is as follows. The literature review has been considered in section 2, along with subsections 2.1-2.5 to define popular confidence interval methods briefly to be relevant to this study. The proposed new CI estimation method is addressed in section 3. Two real life examples have been incorporated in section 4. A simulation study computing estimated coverage probability and width of various CI estimates, with data simulated from distributions with varying degree of skewness has been provided in section 5. We conclude on overall performance of various CI estimation methods by a few concluding remarks in section 6.

2. Literature review

Given a sample $X_1 X_2, \cdots, X_n$ from a distribution with an unknown mean μ and standard deviation σ, we wish to estimate the population mean μ via a confidence interval estimate to ensure the necessary safe guard against the sampling error and estimation certainty. Under the assumption that the sample comes from a normal distribution with a known standard deviation σ, a $100(1-\alpha)\%$ confidence interval (CI) estimator of μ is given by

$$\left[\bar{X} - z_{\alpha/2} \times \sigma/\sqrt{n}, \bar{X} + z_{\alpha/2} \times \sigma/\sqrt{n}\right] \quad (1)$$

where $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $z_{\alpha/2}$ is the upper $(\alpha/2)$th percentile of the standard normal distribution. In reality, however, σ is very unlikely to be known and it is estimated by the sample standard deviation to construct various confidence interval estimate of μ. Several versions of confidence interval estimates exist in literature where σ is estimated from the sample. For example, the Student's t-ci (Student, 1980) is the most efficient and useful CI estimate for μ at normal models. Many researchers, e.g., Johnson (1978), Kleijnen et al. (1986), Meeden (1999), Willink (2005), Kibria (2006), Shi and Kibria (2007), Islam and Shapla (2018), a few to mention. Johnson (1978) considers confidence interval estimator of μ by adjusting the t-ci with an unbiased estimator of third corrected moment. Islam and Shapla (2018) investigated several modifications to t-ci by incorporating trimmed-mean based methods, which trim a certain fraction of data values form both end. Kibria (2006) considers mean absolute deviation about median (Mad-ci) and median confidence interval estimator (Med-ci), which are computationally much easier than Johnson (1978). By comparing Johnson (1978) estimator, Kibria (2006) noted that the width of Student' t-ci and Johnson's methods are same. While many versions of t-ci exist in literature to deal with data with skewness, in real life applications, however, mean, median or trimmed-mean based confidence interval estimators are popular among the practitioners.

In this article, a new modified confidence interval estimator (Mod-ci) of μ has been proposed which incorporates a new point estimator $\hat{\mu}$ as a function of sample mean and sample median. The finite sample performance of Mod-ci has been compared with popular estimators as such as student's t confidence interval estimator (t-ci), mean absolute deviation about median (Mad-ci) and trimmed t confidence interval (Trm-ci) and median t confidence interval (Med-ci), using real-life data having both positive and negative skewness for practical relevancies. The proposed Mod-ci has also been compared with underlying confidence interval estimators by simulation from skewed distribution with varying degree of skewness and sample sizes, in terms of computed coverage probability (covp) and associated width of interval estimators.

2.1 Student's t-ci

Given that the confidence interval estimator in (1) is impractical in reality due to the fact that the population standard deviation σ is most likely to be unknown, Student (1980) proposed the classic t-ci estimate of μ. When the sample size n is small, the $100(1-\alpha)\%$ CI for μ is due to Student (1980) is given by

$$\left[\bar{X} - t_{\alpha/2,n-1}\frac{s_1}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{s_1}{\sqrt{n}}\right] \tag{2}$$

where $s_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}}$ and $t_{\alpha/2,n-1}$ is the upper $(\alpha/2)th$ percentile of Student's t distribution with $(n-1)$ degrees of freedom.

The Student's t-ci is the most popular confidence interval estimator in literature due to the fact that under the normal model it is the most efficient method, and therefore, it is omnipresent in statistical applications for making inference. However, if the data sample comes from the population with skewness, the Student's t-ci has poor coverage property. To overcome this problem, median and trimmed-mean based confidence interval estimators are popular alternatives to deal with non-normal or skewed population.

2.2 Mad-ci

Let the unknown population standard deviation σ be estimated by the mean absolute deviation about median as follows:

$$s_2 = \sqrt{\frac{\sum_{i=1}^{n}|x_i-\tilde{X}|}{n-1}} \tag{3}$$

where $\tilde{X}$ is the sample median defined by

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{for n is odd} \\ \frac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{for n is even} \end{cases} \tag{4}$$

As such, an ad hoc Mad-ci of μ can be constructed for skewed distribution as follows:

$$\left[\bar{X} - t_{\alpha/2} \times s_2/\sqrt{n}, \bar{X} + t_{\alpha/2} \times s_2/\sqrt{n}\right] \tag{5}$$

2.3 Med-ci

Let the unknown population standard deviation σ be estimated by the mean deviation about median as follows:

$$s_3 = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\tilde{X})^2}{n-1}} \tag{6}$$

As such, an ad hoc median t-ci (Med-ci) to deal with data with skewness, due to Kibria (2006), can be constructed as follows:

$$\left[\bar{X} - t_{\alpha/2} \times s_3/\sqrt{n}, \bar{X} + t_{\alpha/2} \times s_3/\sqrt{n}\right] \tag{7}$$

2.4 Trimmed based estimator

Let the point estimator of μ is given by the α-trimmed mean $\bar{X}_\alpha$ as follows

$$\bar{X}_\alpha = \frac{\sum_{i=[n\alpha]}^{n-[n\alpha]} X_{(i)}}{n-2[n\alpha]} \tag{8}$$

where $[n\alpha]$ is the greatest integer in $n\alpha$ for $0 < \alpha < 1$. Also, let an estimate of σ be given by the ad hoc estimator s_4 as follows:

$$s_4 = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{X}_\alpha)^2}{n-1}} \tag{9}$$

Following Islam and Shapla (2018), an ad hoc α-trimmed confidence interval estimator (Trm-ci) can be constructed for skewed distribution as follows:

$$\left[\bar{X} - t_{\alpha/2} \times s_4/\sqrt{n}, \bar{X} + t_{\alpha/2} \times s_4/\sqrt{n}\right] \tag{10}$$

3. Methodology

By incorporating the robustness of sample median for skewed or non-normal distribution, and efficiency of sample mean for normal model, a new modified confidence interval estimator (Mod-ci) has been proposed in this section. A new point estimator $\hat{\mu}$ of μ has been incorporated in Mod-ci, by taking into account the sample mean and median, simultaneously, while using the end-point data information without any trimming of observations. The proposed Mod-ci of μ is given by

$$\left[\hat{\mu} - t_{\alpha/2,n-1}\frac{s_5}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\frac{s_5}{\sqrt{n}}\right] \tag{11}$$

where

$$s_5 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \tag{12}$$

$$\hat{\mu} = \begin{cases} \bar{X}, & \text{if } \hat{\xi}_{n\alpha} < \bar{X} < \xi_{n(1-\alpha)} \\ \tilde{X}, & \text{other wise} \end{cases} \tag{13}$$

and $\hat{\xi}_{n\alpha}$ is an estimate of αth quantile ξ_α for the distribution of X given the sample $X_1, X_2, \cdots, X_n$, where $0 \le \alpha < 1$.
This method provides a guidance as to when to use sample mean $\bar{X}$ or sample median $\tilde{X}$ as an estimator $\hat{\mu}$.

An algorithm to choose $\hat{\mu}$ is as follows:
(i) Compute the sample mean $\bar{X}$ and the sample median $\tilde{X}$, along with sample αth and $(1-\alpha)$th quantiles given by

$$\hat{\xi}_{n\alpha} = X_{(n*\alpha)} \text{ and } \hat{\xi}_{n(1-\alpha)} = X_{(n*(1-\alpha))} \tag{14}$$

The observations at or below $\hat{\xi}_{n\alpha}$, or at or above $\hat{\xi}_{n(1-\alpha)}$ are trimmed by the trimmed mean $\bar{X}_\alpha$ in equation (8), in order to compute Trm-ci of equation (10).
(ii) If the sample mean $\bar{X}$ lies between $\hat{\xi}_{n\alpha}$ and $\hat{\xi}_{n(1-\alpha)}$, then use $\hat{\mu} = \bar{X}$, otherwise, as an estimator use $\hat{\mu} = \tilde{X}$, the sample median, unlike trimming any observations done by the trimmed mean $\bar{X}_\alpha$ or Trm-ci.

By the nature of the choice of the sample mean $\bar{X}$ or sample median $\tilde{X}$ for the point estimator $\hat{\mu}$, the information of end-point observations have been retained and utilized, without trimming of any observations and hence no information is being lost.

4. EXAMPLES

In this section, two examples have been considered, one with positive skewness and the other with negative skewness, for practical relevancies, to see how different confidence interval methods compare in the presence of skewness in the data.

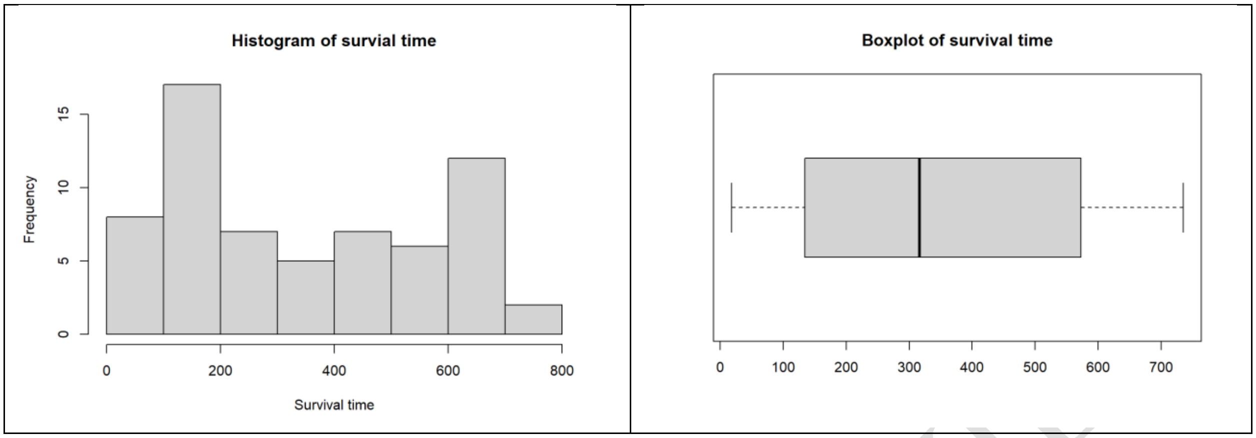
Example 1

Data below refers to survival times (in days) of a sample of 64 guinea pigs from a study by Doksum (1974) to be considered for comparing various confidence interval estimation methods.

36	18	91	89	87	86	52	50	149	120
119	118	115	114	114	108	102	189	178	173
167	167	166	165	160	216	212	209	292	279
278	273	341	382	380	367	355	446	432	421
421	474	463	455	546	545	505	590	576	569
641	638	637	634	621	608	607	603	688	685
663	650	735	725						

To determine the shape of survival time distribution, histogram and boxplot have been presented in Figure 1. Some other quantitative summary measures such as skewness, mean and median, as well as test of normality via lillie.test in R have been considered.

Figure 1. Histogram and boxplot of survival time of guinea pigs data in Example 1



From the histogram and boxplot in Figure 1 it is apparent that the survival time of guinea pigs is positively skewed. The skewness of survival time is 0.22, which supports the fact that the survival time of guinea pigs is positively skewed. The mean and median of survival time are 345.2 and 316.5, which also suggest that survival time is positively skewed as mean is higher than the median. The test of normality reveals a p-value of 0.00046, which suggests that the data is not normally distributed.

While testing the null hypothesis that the population mean or median is 345 days, the p-value of the t-test is found to be 0.993, and the p-value of Wilcoxon signed rank test is found to be 0.841. Therefore, based on the results of t-test or Wilcoxon test, it could be concluded that the population data has the mean or median of 345 days. From the 95% (chosen arbitrarily in this study) confidence interval estimates reported in Table 1, it is evident that all underlying confidence interval estimates contain the unknown mean of 345 days, set hypothetically by noting the sample mean of 345.2 days.

Table 1. 95% CIs of mean survival and width of corresponding CI using data in Example 1

Methods	CI estimate	Width
t-ci	[290, 401]	111
Mad-ci	[296, 395]	99
Med-ci	[289, 401]	112
Trm-ci	[283, 394]	111
Mod-ci	[290, 401]	111

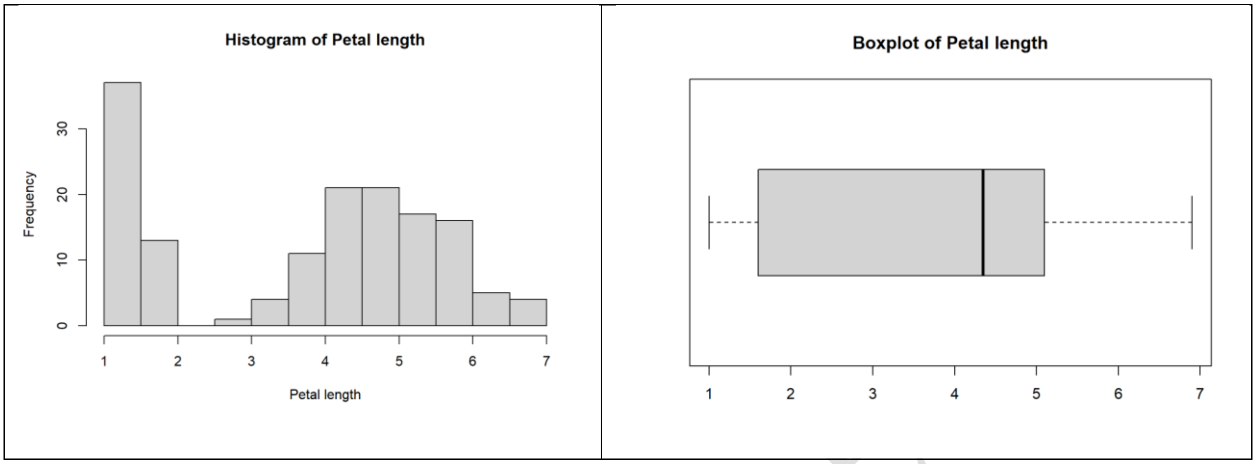
Lengthwise, Mad-ci has the shortest width (99), while t-ci, Trm-ci and Mod-ci all having jointly the second shortest length (111), with Med-ci having the highest length of 112 days. This may lead to the conclusion that the new Mod-ci may retain the efficiency of Student's t-ci and robustness of Med-ci in dealing with data with skewness. Therefore, the performance of Mod-ci is as good as Trm-ci or t-ci, and better than Med-ci, which may imply that Mod-ci might have retained the efficiency of t-ci and robustness of Trm-ci.

Example 2

In this example, an R iris dataset Petal length has been considered for constructing confidence interval using underlying method discussed in this study. Because the dataset remains in the public domain, one can get access and verify the results as needed. The Petal length refers to a sample of size 150, which is non-normal with a mean of 3.76. Assume that the Petal length in the population has a hypothetical mean of $\mu = 4$.

To determine the shape of Petal length distribution, the histogram and boxplot has been presented in Figure 2. The other quantitative summary such as skewness, mean and median has also been noted. The normality Petal length has been tested via lillie.test in R.

Figure 2. Histogram and boxplot for Petal length data in Example 2



From the histogram and boxplot in Figure 2, it is apparent that the Petal length is negatively skewed. The skewness of Petal length is -0.27, supporting that Petal length distribution is negatively skewed.

The R lillie.test of normality, with the p-value of 0.0031, provides the evidence to conclude that the Petal length distribution is not normal, at 5% level of significance. The result of t test, with a p-value of 0.09525 or the Wilcoxon signed rank test, with a p-value of 0.124 may lead to the conclusion that the sample comes from population with the mean or median Petal length of 4.

Now, as attention is being paid to the results of 95% confidence interval estimates reported in Table 2, it has been noted that all underlying methods include the hypothesized population mean of 4.

Table 2. 95% CIs and corresponding width for data in Example 2

Methods	CI estimate	Width
t-ci	[3.47, 4.04]	0.57
Mad-ci	[3.52, 4.00]	0.48
Med-ci	[3.46, 4.06]	0.60
Trm-ci	[3.48, 4.04]	0.56
Mod-ci	[3.47, 4.04]	0.57

As has been considered widthwise, Mad-ci has the shortest width (0.48), which also has marginally captured the hypothesized mean of 4 by the upper confidence limit. The Trm-ci has the second smallest width (0.56). The t-ci and Mod-ci are jointly with the same width of 0.57 is outperforming Med-ci which has length of 0.60.

As we look towards the performance of Example 2, it has clearly been revealed that Mod-ci is as good as t-ci, and better than Med-ci, or nearly close to Trm-ci, while outperforms Med-ci.

Considering the performance of the various confidence interval estimators in two examples, it can be concluded that Mod-ci might have retained the efficiency of t-ci and robustness of Trm-ci. As such, one should not have any hesitation in recommending the new method for practicing while dealing with data with skewness.

5. SIMULATION AND RESULT DISCUSSION

It is well understood that to justify the usefulness of any method and recommend it for practical usages, it is best to evaluate the method via simulation. In this section, a simulation study has been conducted by simulating data from skewed distribution with varying degree of skewness so as to compare the performance of the various confidence interval estimators.

Note that a gamma distribution is well known for its skewness in modeling data having skewness. As such, a gamma $G(\beta, \sigma)$ distribution, with shape parameter β and scale parameter σ, having the density function

$$f(x) = \frac{x^{\beta-1}\exp(-\frac{x}{\sigma})}{\sigma^{\beta}\Gamma(\beta)}; x > 0, \beta, \sigma > 0 \qquad (14)$$

has been considered. This distribution has skewness parameter $\gamma = 2/\sqrt{\beta}$, which allows simulation from varying skewness.

Since the mean of this distribution is $\mu = \beta\sigma$, in simulations the value of $\sigma = \frac{1}{\beta}$ is chosen arbitrarily to fix the mean at $\mu = 1$. The values of β has been set to 16, 4, 1, 1/4, 1/16, 1/36 to allow skewness values to 0.5, 1, 2, 4, 8, 12, respectively. In all simulations, the Monte Carlo size is 10,000, chosen arbitrarily, relatively large than is in common practice, as higher the Monte Carlo size more accuracy in the estimation could be reached. The simulation results of this study have been reported in Tables 3-9 for sample size varying between 10 and 100, arbitrarily.

The estimated or simulated coverage probability is the proportion of 10,000 CIs over all MC simulations containing the true mean $\mu = 1$. The width of confidence interval is estimated from average of all 10,000 confidence intervals for each given sample size.

As a computational tool, in all computation and simulation, the statistical software R (R Core Team, 2024) has been utilized in this article.

Table 3. Simulated coverage probability and width of 95% CIs of mean with skewness=0.50

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.94	0.87	0.95	0.94	0.94	0.35	0.26	0.35	0.35	0.35
15	0.95	0.87	0.95	0.94	0.95	0.27	0.21	0.28	0.27	0.27
20	0.95	0.87	0.95	0.94	0.95	0.23	0.18	0.23	0.23	0.23
25	0.95	0.88	0.95	0.94	0.95	0.20	0.16	0.21	0.20	0.20
30	0.95	0.88	0.95	0.94	0.95	0.18	0.14	0.19	0.19	0.18
35	0.95	0.87	0.95	0.94	0.95	0.17	0.13	0.17	0.17	0.17
40	0.95	0.88	0.95	0.93	0.95	0.16	0.12	0.16	0.16	0.16
45	0.95	0.87	0.95	0.93	0.95	0.15	0.12	0.15	0.15	0.15
50	0.95	0.88	0.95	0.93	0.95	0.14	0.11	0.14	0.14	0.14
100	0.95	0.88	0.95	0.92	0.95	0.10	0.08	0.10	0.10	0.10
Min	**0.94**	**0.87**	**0.95**	**0.92**	**0.94**	**0.10**	**0.08**	**0.10**	**0.10**	**0.10**
Max	**0.95**	**0.88**	**0.95**	**0.94**	**0.95**	**0.35**	**0.26**	**0.35**	**0.35**	**0.35**

Table 4. Simulated coverage probability and width of 95% CIs of mean with skewness =1

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.93	0.86	0.94	0.93	0.93	0.69	0.51	0.71	0.69	0.69
15	0.93	0.85	0.94	0.92	0.93	0.54	0.40	0.55	0.54	0.54
20	0.94	0.87	0.95	0.92	0.94	0.46	0.35	0.47	0.46	0.46
25	0.95	0.86	0.95	0.92	0.95	0.40	0.31	0.41	0.41	0.40
30	0.94	0.87	0.95	0.91	0.94	0.37	0.28	0.37	0.37	0.37
35	0.95	0.87	0.95	0.91	0.95	0.34	0.26	0.35	0.34	0.34
40	0.95	0.87	0.95	0.90	0.95	0.32	0.24	0.32	0.32	0.32
45	0.94	0.87	0.95	0.90	0.94	0.30	0.23	0.30	0.30	0.30
50	0.95	0.87	0.95	0.89	0.95	0.28	0.22	0.29	0.28	0.28
100	0.95	0.86	0.95	0.85	0.95	0.20	0.15	0.20	0.20	0.20
Min	**0.93**	**0.85**	**0.94**	**0.85**	**0.93**	**0.20**	**0.15**	**0.20**	**0.20**	**0.20**
Max	**0.95**	**0.87**	**0.95**	**0.93**	**0.95**	**0.69**	**0.51**	**0.71**	**0.69**	**0.69**

Table 5. Simulated coverage probability and width of 95% CIs of mean with skewness =2

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.90	0.82	0.91	0.87	0.90	1.32	0.92	1.40	1.34	1.32
15	0.91	0.81	0.92	0.87	0.91	1.05	0.73	1.11	1.06	1.05
20	0.91	0.82	0.92	0.85	0.91	0.89	0.62	0.93	0.90	0.89
25	0.93	0.83	0.93	0.86	0.93	0.80	0.55	0.84	0.80	0.80
30	0.93	0.83	0.94	0.83	0.93	0.73	0.51	0.76	0.74	0.73
35	0.93	0.83	0.94	0.84	0.93	0.67	0.47	0.70	0.68	0.67
40	0.93	0.82	0.94	0.81	0.93	0.63	0.44	0.66	0.63	0.63
45	0.93	0.82	0.94	0.80	0.93	0.59	0.41	0.62	0.60	0.59
50	0.94	0.83	0.95	0.78	0.94	0.56	0.39	0.59	0.57	0.56
100	0.94	0.83	0.95	0.61	0.94	0.39	0.27	0.41	0.40	0.39
Min	**0.90**	**0.81**	**0.91**	**0.61**	**0.90**	**0.39**	**0.27**	**0.41**	**0.40**	**0.39**
Max	**0.94**	**0.83**	**0.95**	**0.87**	**0.94**	**1.32**	**0.92**	**1.40**	**1.34**	**1.32**

Table 6. Simulated coverage probability and width of 95% CIs of mean with skewness=4

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.80	0.68	0.82	0.74	0.80	2.35	1.33	2.59	2.42	2.35
15	0.83	0.68	0.85	0.77	0.83	1.92	1.04	2.11	1.95	1.92
20	0.85	0.68	0.87	0.73	0.85	1.68	0.90	1.84	1.73	1.68
25	0.86	0.67	0.88	0.73	0.86	1.49	0.78	1.63	1.52	1.49
30	0.87	0.67	0.89	0.68	0.87	1.37	0.71	1.49	1.41	1.37
35	0.88	0.67	0.90	0.69	0.88	1.27	0.66	1.39	1.31	1.27
40	0.89	0.67	0.91	0.64	0.89	1.19	0.61	1.30	1.23	1.19
45	0.90	0.67	0.92	0.65	0.90	1.14	0.58	1.24	1.17	1.14
50	0.90	0.66	0.92	0.59	0.90	1.08	0.55	1.18	1.11	1.08
100	0.93	0.66	0.94	0.33	0.93	0.77	0.38	0.84	0.79	0.77
Min	**0.80**	**0.66**	**0.82**	**0.33**	**0.80**	**0.77**	**0.38**	**0.84**	**0.79**	**0.77**
Max	**0.93**	**0.68**	**0.94**	**0.77**	**0.93**	**2.35**	**1.33**	**2.59**	**2.42**	**2.35**

Table 7. Simulated coverage probability and width of 95% CIs of mean with skewness=8

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.60	0.43	0.61	0.54	0.60	3.48	1.44	3.79	3.63	3.48
15	0.66	0.40	0.67	0.58	0.66	3.00	1.11	3.22	3.08	3.03
20	0.69	0.39	0.70	0.57	0.69	2.74	0.94	2.91	2.84	2.77
25	0.72	0.39	0.73	0.59	0.72	2.48	0.82	2.63	2.56	2.51
30	0.75	0.39	0.76	0.57	0.75	2.37	0.75	2.50	2.45	2.39
35	0.77	0.38	0.78	0.58	0.77	2.27	0.70	2.38	2.34	2.29
40	0.78	0.39	0.79	0.55	0.78	2.08	0.63	2.18	2.15	2.10
45	0.80	0.38	0.81	0.55	0.80	2.03	0.60	2.12	2.09	2.04
50	0.80	0.37	0.81	0.52	0.80	1.95	0.57	2.03	2.01	1.96
100	0.86	0.38	0.87	0.33	0.86	1.46	0.40	1.52	1.50	1.46
Min	**0.60**	**0.37**	**0.61**	**0.33**	**0.60**	**1.46**	**0.40**	**1.52**	**1.50**	**1.46**
Max	**0.86**	**0.43**	**0.87**	**0.59**	**0.86**	**3.48**	**1.44**	**3.79**	**3.63**	**3.48**

Table 8. Simulated coverage probability and width of 95% CIs of mean with skewness=12

Sample sizes	Coverage probability of various CI methods					Width of various CI methods				
	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t-ci	Mad-ci	Med-ci	Trm-ci	Mod-ci
10	0.44	0.28	0.45	0.40	0.44	3.95	1.43	4.23	4.14	3.95
15	0.51	0.26	0.52	0.46	0.51	3.50	1.09	3.68	3.60	3.57
20	0.56	0.26	0.57	0.47	0.56	3.31	0.94	3.45	3.42	3.38
25	0.60	0.25	0.60	0.49	0.60	3.10	0.82	3.21	3.19	3.16
30	0.61	0.25	0.62	0.49	0.62	2.89	0.73	2.99	2.97	2.95
35	0.65	0.25	0.66	0.51	0.66	2.86	0.70	2.95	2.94	2.91
40	0.68	0.25	0.68	0.50	0.68	2.72	0.64	2.80	2.79	2.77
45	0.69	0.25	0.70	0.51	0.69	2.57	0.59	2.64	2.63	2.62
50	0.71	0.26	0.72	0.51	0.72	2.56	0.57	2.63	2.62	2.60
100	0.78	0.25	0.79	0.44	0.79	2.00	0.40	2.04	2.04	2.03
Min	**0.44**	**0.25**	**0.45**	**0.40**	**0.44**	**2.00**	**0.40**	**2.04**	**2.04**	**2.03**
Max	**0.78**	**0.28**	**0.79**	**0.51**	**0.79**	**3.95**	**1.43**	**4.23**	**4.14**	**3.95**

As we look at the simulated results carefully when the skewness is minimum (0.5), in Table 3, the Mod-ci is as good t-ci or Med-ci, in terms of estimated coverage probability(covp). The Mad-ci is worst in terms of covp, which never attains the desired or expected level of covp (i.e. 0.95). However, with average width consideration, Mad-ci has the smallest width. Among other methods, t-ci, Med-ci and Mod-ci are similar in width, outperforming Trm-ci. This simulation suggests that when skewness is lower, t-ci, Med-ci and Mod-ci are equally good. It is interesting to note that even with lower skewness the Mad-ci suffer in covp, which is very important aspect for an estimator. When the covp is comparable, only then average width should come to play in deciding which estimator is performing the best.

For a quick reference, one may refer to last two rows of Tables 3-8 for overall performance reported in terms of minimum (Min) and maximum (Max) coverage probability (covp) and width for a given skewness over varying sample sizes. The summary of Min and Max covp and width over all simulation for varying sample sizes and skewness has been reported in Table 9.

For example, as is noted in Table 3 when skewness is 0.5, three methods t-ci, Med-ci and Mod-ci attain the Max coverage probability of 0.95, which is expected from any estimator for a 95% confidence interval. The Mad-ci is behind the expectation, with a covp between 0.87 and 0.88, never attaining the expected level of 0.95. However, looking at the width criteria, Mad-ci is best, while underperforming significantly in coverage probability criteria. The researchers have to decide what is more important to them given a set of choices.

Of course, the consideration of width is important but not by compromising the covp. By taking all simulation cases and facts into considerations, the proposed Mod-ci provides expected coverage probability of 0.95, with only one exception where coverage probability is reported to be 0.94 (of course it is acceptable) when the sample size 10 and skewness is 0.5. It can be argued that the Mod-ci might have retained the efficiency of t-ci or robustness of Med-ci, which constantly have the coverage probability of 0.95.

Table 9. Min and Max covp and width, over all simulations, along with specified skewness (skew) of data

Methods of CI	Min covp (skew)	Max covp (skew)	Min length (skew)	Max width (skew)
t-ci	0.44 (12)	**0.95 (0.50)**	0.10 (0.50)	**3.95 (12)**
Mad-ci	**0.22 (12)**	0.88 (0.50)	**0.08 (0.50)**	**1.44 (08)**
Med-ci	0.45 (12)	**0.95 (0.50)**	0.10 (0.50)	4.23 (12)
Trm-ci	0.33 (12)	0.94 (0.50)	0.10 (0.50)	4.14 (12)
Mod-ci	0.44 (12)	**0.95 (0.50)**	0.10 (0.50)	**3.95 (12)**

Looking critically in all simulations, for sample sizes between 10 to 100, as skewness increases from 0.5 to 12, the Mad-ci fails to meet the expectation with attained coverage probability (covp) of min=0.22 (skew=12) and max=0.88 (skew=0.5), never reaching to the expectation of 0.95. It clearly suggests that Mad-ci has a severe underestimation in performance.

When skewness is 12 (**Table** 8), all methods suffer in reaching the desired covp of 0.95. The highest covp when skewness is 12 is attained by Med-ci and Mod-ci, both with covp=0.79, followed by t-ci with covp=0.78 Trm-ci with covp=0.51 and Mad-ci with covp=0.28. These results clearly suggest that when skewness is very high, estimation problem still exists and a further search for a better confidence interval approach is still required. However, despite the limitations highlighted in Table 8 regarding higher skewness, it is important to note that the newly proposed Mod-ci offers the advantage of simultaneously observing both the mean and the median during estimation. This dual consideration provides a higher degree of confidence compared to other estimation procedures. In particular, when the Mad-ci is underperforming, the Trm-ci, which trims a certain percentage of observations from both ends, may lead to the loss of valuable data information.

6. Conclusion

In this article, a new modified confidence interval estimator for μ, called Mod-ci, is proposed and evaluated against several widely used confidence interval estimators that are relevant to the proposed method. The other methods considered in this study include Student's t confidence interval (t-ci), the mean absolute deviation about the median (Mad-ci), the median t confidence interval (Med-ci), and the trimmed-mean confidence interval (Trm-ci), particularly in the presence of skewed data distributions. As always, the t-ci estimator can be relied upon when the underlying data distribution is normal. In the absence of normality, particularly with skewed data, the Mad-ci estimator may be preferred over others if the focus is on minimizing confidence interval length or width (Kibria, 2006). However, this method suffers from poor coverage probability. The simulation results in this study clearly demonstrate that as skewness increases from 0.5 to 12, the performance of Mad-ci deteriorates, showing poor coverage probability while its width remains the smallest across all simulations. The Med-ci estimator performs well when both sample size and skewness are small. The Trm-ci estimator also shows decreasing performance as skewness increases. In contrast, the Mod-ci estimator performs comparably to the Med-ci and t-ci, and outperforms the Trm-ci as skewness increases. Researchers must prioritize either coverage probability or width depending on their needs. Traditionally, to recommend a confidence interval estimator, a simulation is conducted, and performance is assessed based on coverage probability and width. If coverage probabilities are comparable across methods, average width can then be considered. Of course, coverage probability should never be compromised for the sake of width. Considering the simulation results and real-life examples, this study suggests that Mod-ci combines the efficiency of t-ci and the robustness of Med-ci or Trm-ci in both coverage probability and width. However, it is also noted that for high skewness values (≥4), estimation issues persist, and further research is needed to develop better methods for handling data with extreme skewness. In conclusion, Mod-ci provides the advantage of considering both the mean and median during estimation, offering a higher degree of confidence compared to other methods. Therefore, it is recommended for use in real-world scenarios involving skewed data.

Disclaimer (Artificial intelligence)

Authors hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Competing Interests

Authors have declared that no competing interests exist.

Acknowledgement

The authors are thankful to the reviewers and the editor for providing useful comments and suggestions which has improved the presentation of the article to a great extent.

REFERENCES

Bickel, P. J. (1965). On some robust estimates of location*. Ann. Math. Statist*., 36, 847–858.

Casella, G. & Berger, R. (Author). (2024). *Statistical Inference*, 2nd ed., Chapman and Hall/CRC.

Doksum, K. (1974). Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case. *Ann. Statist.* 2 (2): 267-277. DOI: 10.1214/aos/1176342662.

Hampel, F. R, Ronchetti, E. M., Peter J. Rousseeuw, P.J. & Stahel, W.A. (2011). Robust Statistics: The Approach Based on Influence Functions, Wiley Series in Probability and Statistics.

Hartwig, F.P., Smith, G.D., Schmidt, A. F., Sterne, J.A.C., Higgins, J.P.T. & Bowden J. (2020). The median and the mode as robust meta-analysis estimators in the presence of small-study effects and outliers. *Res Syn Meth.* 11: 397–412. https://doi.org/10.1002/jrsm.1402

Hayden, R. W. (2005). A Dataset that is 44% Outliers. Journal of Statistics Education, 13(1). https://doi.org/10.1080/10691898.2005.11910642

Johnson, N.J. (1978). Modified t Tests and Confidence Intervals for Asymmetrical Populations. *Journal of the American Statistical Association*, 73, pp. 536-544.

Hogg, R., McKean, J. & Craig, A. (2018). Introduction to Mathematical Statistics, 8th Edition. Pearson.

Islam, K. & Shapla, T. J. (2018). On Performance of Confidence Interval Estimate of Mean for Skewed Populations: Evidence from Examples and Simulations. *Mathematical Theory and Modeling*, 8(3), pp. 41-51.

Kibria, B.M.G. (2006). Modified Confidence Intervals for the Mean of the Asymmetric Distribution. *Pakistan Journal of Statistics*, 22(2), pp. 111-123.

Kleijnen J.P.C., Kloppenburg, G.L.J. & Meeuwsen, F.L. (1986). Testing the mean of asymmetric population: Johnson’s modified t test revisited. *Communications in Statistics- Simulation and Computation*, 15, 715-732.

Meeden, G. (1999). Interval Estimators for the Population Mean for Skewed Distributions with a Small Sample Size. *Journal of Applied Statistics*, 26(1), 81-96.
R version 3.3.2 (2016-10-31). *The R Foundation for Statistical Computing*.

Portnoy, S. & He, X. (2000). A robust journey in the new millennium. J. Amer. Statist. Assoc. 95(452), 1331–1335.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Shi, W. & Kibria, B.M.G. (2007). On some confidence intervals for estimating the mean of a skewed population. *Int. J. Math. Educ. Sci. Technol.* 38(3), pp. 412-421.
Student (1908). The probable error of a mean. *Biometrika* 6 (1): 1–25.

Shi, W. and Kibria, B.M.G. (2007). On some confidence intervals for estimating the mean of a skewed population. *Int. J. Math. Educ. Sci. Technol.* 38(3), pp. 412-421.

Student (1980). The Probable Error of a Mean. *Biometrika, 6(1), 1-25***.**

Tukey, J. W., and McLaughlin, D.H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample (Trimming/Winsorization 1). *Sankhya, Ser. A*, 25, 331–352.

Wilcox, R. (2021). Introduction to Robust Estimating and Hypothesis Testing, 5th Edition, Academic Press.

Willink, R. (2005). A Confidence Interval and Test for the Mean of an Asymmetric Distribution. *Communications in Statistics- Theory and Methods*, 34, 753-766.