A New Modified Confidence Interval Estimate of Mean for Skewed Distribution: Applications and Simulation

ABSTRACT

A new modified point estimate of mean has been proposed for skewed data distribution. The proposed estimate has been utilized in the construction of a new modified confidence interval estimate of the unknown population mean. The usefulness of the new method of confidence interval estimates has been justified by real-life examples where the data distribution is subject to skewness. The performance of the new confidence interval (ci) method has been compared with traditional Students' t-ci, mean absolute deviation (mad) t-ci, median t-ci and trimmed t-ci by real-life examples and simulations from skewed distributions. While doing simulation, we consider varying degree of skewness in the population distribution to study the sensitivity of underlying methods with respect to skewness. Results of examples and simulation suggest that the proposed method is as good as or better than other estimator relevant to this study, and as such we would recommend this method for practicing while dealing with real-life data with 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.

2. METHODS

Given a sample $X_1X_2, \cdots, X_n$ 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) estimate 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, 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 (2018), investigated several modifications to t-CI to deal with skewness in data distribution. While many versions of t-CI estimates exist in literature to deal with data with skewness or outlying observation, in real life applications, however, mean, median or trimmed-mean based confidence interval estimates are popular among the practitioners.
In this article, we propose a new CI estimate motivated by a new modified estimate of mean and its asymptotic normality property. The finite sample of performance of this CI estimate has been justified and compared with popular CIs such as such as student's t, mean absolute deviation about median (mad-med) t and trimmed t CI, using real-life data having both positive and negative skewness for practical relevancies. The modified CI has also been compared with underlying CI estimation methods by simulation from skewed distribution with varying degree of skewness and sample sizes, in terms of computed coverage probability and width of CI estimates.

The organization of the remaining paper is as follows. In section 2, we define Student's t and other CI estimates 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.

2.1 Student's t-CI

Given that the CI estimate 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

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

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 CI in literature due to the fact that under the normal model it is the most efficient CI estimation 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 CI estimates are popular alternatives to deal with non-normal or skewed population.

2.2 Mad Med t-CI

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

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

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

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

As such, an ad hoc mad-med t-CI can be constructed for skewed distribution as follows:

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

2.3 Med t-CI

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

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

As such, an ad hoc med t-CI for skewed distribution can be constructed as follows:

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

Kibria (2006) proposes to use this CI estimate of μ, called median t-CI (med t), in dealing with data with skewness.

2.4 Trimmed t-CI

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

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

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

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

As such, an ad hoc α-trimmed t-CI can be constructed for skewed distribution as follows:

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

Islam (2018) proposes to use this CI estimate of μ, called median t-CI (med t), in dealing with data with skewness.

3. NEW PROPOSED t-CI

By incorporating the robustness of sample median for skewed or non-normal distribution, and efficiency of sample mean for normal model, we propose a blended or mixed version of the new estimator for the population mean μ. The basic idea is to construct of a t-CI estimate on the basis of a new estimate of the population mean which incorporates mean and median functionally.

3.1 The point estimate of μ

Under the new proposed method, a point estimate of μ is given by

$$\hat{\mu} = \begin{cases} \bar{X}, & \text{if } \hat{\xi}_{n\alpha} < \bar{X} < \xi_{n(1-\alpha)} \\ \tilde{X}, & \text{other wise} \end{cases} \quad (11)$$

An equivalent expression of $\hat{\mu}$ is given by

$$\hat{\mu} = \bar{X} I\left(\hat{\xi}_{n\alpha} < \bar{X} < \xi_{n(1-\alpha)}\right) + \{1 - I(\cdot)\}\tilde{X} \quad (12)$$

where

$$I\left(\hat{\xi}_{n\alpha} < \bar{X} < \xi_{n(1-\alpha)}\right) = \begin{cases} 1, & \text{if } \hat{\xi}_{n\alpha} < \bar{X} < \xi_{n(1-\alpha)} \\ 0, & \text{other wise} \end{cases} \quad (13)$$

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 $0 \leq \alpha < 1$.

This method provides a guidance as to when we use mean or when to use median for estimating μ by noting the boundary of a desired quantiles of the sample.

3.2 Proposed new t-CI estimate of μ

Unlike trimming observations from both ends by trimmed t-CI, we expect that the proposed new test retains the robustness of median and efficiency of mean in the presence of skewness and non-normality of data. As such we propose to construct of a t-CI estimate based on the estimator $\hat{\mu}$ as is given by

$$\left[\hat{\mu} - t_{\alpha/2,n-1}\,\frac{s_5}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n-1}\,\frac{s^*}{\sqrt{n}}\right] \tag{14}$$

where

$$s_5 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \tag{15}$$

This t-CI estimate is termed as modified t-CI (mod t-CI).

4. EXAMPLES

In this section, we consider two examples, 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.

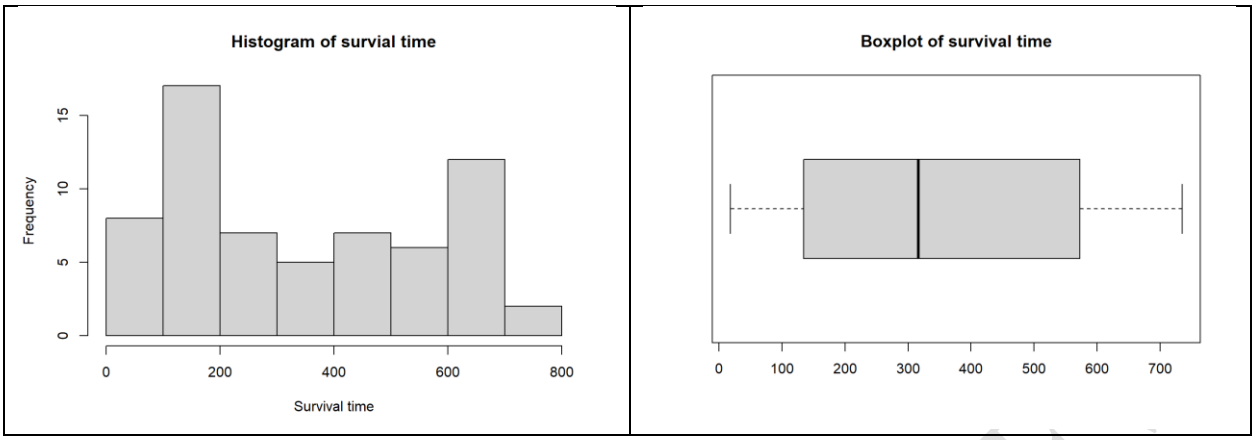
Example 1

Data below refers to survival times (in days) of a sample of 64 guinea pigs from a study by Doksum (1974). We wish to find t-CI estimates by underlying 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 distribution of the data let us have a look at the graphs in Figure 1 for the survival time of guinea pigs, along with some quantitative summary measures such as skewness, mean and median, as well as test of normality via lillie.test in R:

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, we could conclude that the population data has the mean or median of 345 days. Now, let us have a look at the results of 95% (chosen arbitrarily in this study) confidence interval estimates reported in Table 1 so as to see if the underlying confidence interval estimates contain the true 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	Length
t-ci	[290, 401]	111
Mad-ci	[296, 395]	99
Med-ci	[289, 401]	112
Trim-ci	[283, 394]	111
Mod-ci	[290, 401]	111

As we see from the 95% CIs reported in Table 1, all methods have captured the hypothesized mean $\mu = 345$. Lengthwise, Mad *t* CI has the shortest width (99). The student's t, trimmed t and modified trimmed t have jointly the second shortest length (111), while the median trimmed t has the highest length of 112 days. This may lead to the conclusion that the new modified *t* CI may retain the efficiency of Student's *t* and robustness of median *t* CIs in dealing with data with skewness.

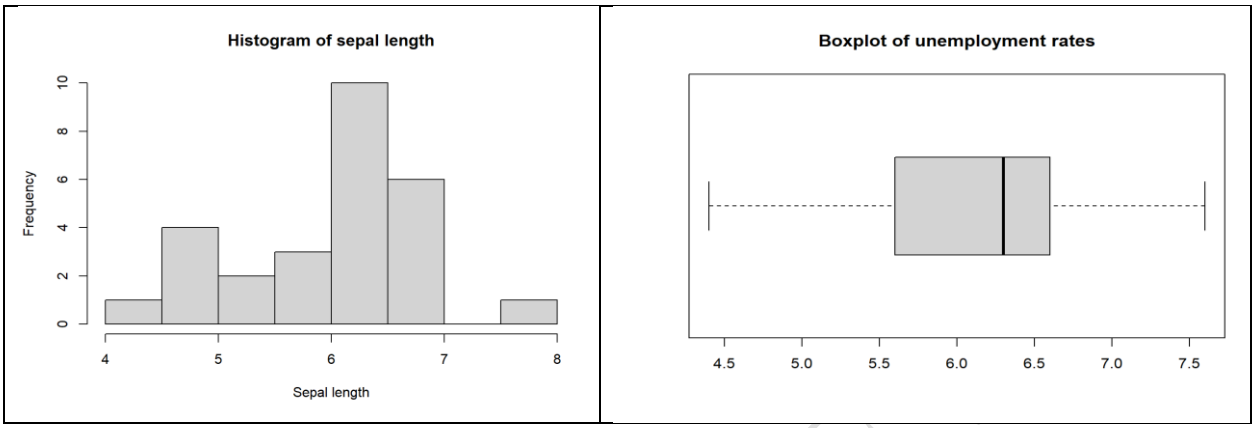
Example 2

Data below refers to a sample of size 27, of Sepal length, taken from R dataset iris, which is non-normal with a mean of 5.84. Assume that iris data is a population with a mean of $\mu = 5.84$ and we wish to see if the CI estimates of based on the given sample captures the mean and also wish to compare width of CI estimates of underlying estimates.

7	6.7	6.9	6.9	6.3	6.3	7.6	6.5	5.2
4.4	6.4	5	6.3	5.7	6.4	5.7	5	4.6
6.1	6.7	4.8	6.7	6.3	6.4	5.8	5.5	6.3

To determine the shape of sepal length distribution on the basis of the sample, let us have a look at the following graphs, and other quantitative summary such as skewness, mean and median, as well as test of normality via lillie.test in R:

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



From the histogram and boxplot in Figure 2 it is apparent that the sepal length is negatively skewed. The skewness of the sepal length is -0.45, which also provides evidence in the support that the sepal length distribution is negatively skewed. The mean and median of sepal length are 6.1 and 6.3, mean<median, which may also suggest that sepal length is negatively skewed.

The R lillie.test of normality has been carried out to test the null hypothesis that the sepal length distribution is normal. The p-value of 0.0031 provides the evidence to conclude that the sepal length distribution is not normal, at 5% level of significance. We have noted that in the entire iris dataset in R, the mean sepal length is 5.84. As such we wish to test if the population mean is 5.58 by setting it to the null hypothesis. The result of t test reveals a p-value of 0.1756, while the Wilcoxon signed rank test reveals the p-value of 0.1594. Given the results of these two tests, we could conclude that the sample comes from population with the mean or median 5.84.

Now, for an obvious reason, we wish to draw our attention to the results of 95% confidence interval estimates reported in Table 2 for various underlying methods to see if they capture the true hypothesized value of 5.84.

As we see, from the results of the 95% CIs reported in Table 2, all five confidence interval estimates capture the hypothesized mean of 5.84. Therefore, the use of the new proposed confidence interval estimates is justifiable for a situation where data may contain skewness or outlying observations.

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

Methods	CI estimate	Length
t-ci	[5.74, 6.37]	0.63
Mad-ci	[5.82, 6.30]	0.48
Med-ci	[5.72, 6.39]	0.67
Trim-ci	[5.76, 6.40]	0.64
Mod-ci	[5.74, 6.37]	0.63

While considering the width of confidence intervals, Mad t CI has the shortest width (0.48), while Med CI estimate has the highest width (0.67). The Student's t and modified CI estimates have the second lowest width, beating the trimmed mean or median CI methods. This example may also lead to the conclusion that in the presence of skewness, the new modified test may retain the efficiency of mean or robustness of median in width of CI estimates consideration.

Therefore, following the success of the new proposed method in constructing CI estimate of mean for data distribution with skewness, one should not have any hesitation in recommending the new proposed confidence interval estimation method for practicing.

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, carry out a simulation by generating samples from skewed distribution. Note that a gamma distribution is well known for its skewness in modeling data having skewness. As such, to compare the finite sample performance of underlying CI estimates, we consider the population distribution to be gamma $G(\beta,\sigma)$ with density function specified by

$$f(x)=\frac{x^{\beta-1}\exp(-\frac{x}{\sigma})}{\sigma^{\beta}\Gamma(\beta)};x>0,\beta,\sigma>0 \quad (16)$$

where β is a shape and σ is a scale parameter. For the specified gamma distribution, the measure of skewness parameter $\gamma=2/\sqrt{\beta}$. Since the mean of this distribution is $\mu=\beta\sigma$, and for simulation we choose $\sigma=\frac{1}{\beta}$ arbitrarily to fix the mean at $\mu=1$ for all simulation allowing varying skewness by changing values of the parameter β.

Indeed, we arbitrarily choose β values at 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 coverage probability is estimated as the proportion of 10,000 CIs over all MC simulations containing the true mean. The width of confidence interval is estimated from average of all 10,000 confidence intervals for each given sample size.

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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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				
	tci	Mad-ci	Med-ci	Trm-ci	Mod-ci	t	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 this underlying study. We see that all four methods perform almost equally well except the Mad-ci when the skewness of 0.5 (Table 9) in coverage probability criteria. The Mad-ci has coverage probability of 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 proposed modified CI (Mod-ci) provides coverage probability of 0.95 as is expected, with only one exception with coverage probability of 0.94 (of course acceptable) when the sample size 10. Lengthwise, all four methods except Mad-t ci are comparable. It can be argued that the Mod-ci may retain the efficiency of t-ci or robustness of Med-ci, which constantly have the coverage probability of 0.95. Min coverage probability of Trm-ci is 0.92 and the max is 0.93, never attaining the expected confidence coverage of 0.95. In all consideration, the new proposed Mod-ci the meeting the expectation in performance.

Table 9. Min and Max coverage probability (covp) and width, along with 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)

With the increase in skewness from 0.5 to 12, the Mad-ci fails to meet expectation with attained coverage probability (covp) of min=0.22 (skew=12) and max=0.88 (skew=0.5), never meeting the expectation of 0.95, and thereby shows severe underestimated performance. The min-max covp and min-max width are reported in Table 7, which clearly demonstrate that the new Mod-ci meets expectation in all simulated cases with covp and length criteria, by retaining efficiency of mean and robustness of median. The most importantly, however, it should be remembered that the new proposed estimate gives the leverage of observing both mean and median at the same time while doing the estimation, and as such it provides some degree of confidence over other estimation procedures. Particularly, Mad-ci is underperforming, and so is the Trm-ci which trims a certain percentage of observations form both ends, may be a reason to lose some information.

6. conclusion :

As is always, the best recommendation is to use the t-CI interval if there is sufficient evidence to support that the data in an underlying study comes from a normal distribution. However, due to simplicity of implementation, the traditional and popular methods of confidence interval estimation methods in the presence of skewness or outlying observation is median based CI or trimmed-mean based CI. While other recommendations are available with relatively different approaches and scenarios, none of them is believed to be uniformly good in all forms of skewness situation. In this study, we propose a modified version of a point estimate of the unknown mean, which is a function of sample mean and median. The idea is the incorporate the median in the computation when the sample mean is outside of the two desired end point quantiles, unlike trimming any observations from both end done in the trimmed mean approach. We recommend to use this estimate in the computation of sample variance and thereby in the construction of the confidence interval estimate when there is skewness or outlying observations in the data distribution. By real-life examples, we justified the usefulness of the new method in the context of the other relevant measures. The result of the simulation study is also supportive of the proposed new method. As such, we recommend the new approach for practicing while dealing with data with skewness and, or outlying observations.

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.

U.S. Bureau of Labor Statistics (2024). Unemployment Rates for States. https://www.bls.gov/web/laus/laumstrk.htm

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.

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.