MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Confidence Interval Estimating the Mean of Normal Distribution and Skewed Distribution

Siti Fairus Mokhtar[a,b]*, Zahayu Md Yusof[c], Hasimah Sapiri[a]

[a]School Quantitative Sciences, Universiti Utara Malaysia, Kedah, Malaysia; [b]School of Mathematical Sciences, College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Kedah Branch, Kedah, Malaysia; [c]Centre for Testing, Measurement & Appraisal (CeTMA), Universiti Utara Malaysia, Kedah, Malaysia

Abstract The confidence interval is an important statistical estimator of population location and dispersion parameters. The purpose of this paper is to comprehend CI utilising various techniques. This includes classical CI, percentile bootstrap method, bootstrap-*t* and proposed bootstrap-*t* decile mean method. Distributions that are skewed and normal are used to generate data. The efficiency of the proposed method is evaluated on the basis of an extensive simulation study. The simulation findings show that the performance of the Student-*t* and three bootstrap approaches varies dramatically depending on sample size and skewness type. The coverage probability and length of the proposed confidence interval are compared with certain existing and widely used confidence intervals. For illustrative purposes, two real-life data sets are analysed, which, to some extent, support the simulation study conclusions. This paper's findings will be useful to a variety of researchers with practical experience in the fields of science and social sciences.

Keywords: Confidence intervals, bootstrapping, bootstrap-*t*, parameter estimations.

Introduction

Many established statistical theories rely on the normality assumption as their foundation, such as Neyman's (1937) estimation theory for constructing confidence intervals (CIs). However, in practical situations, a significant portion of data doesn't adhere to this assumption of normality. The Student-*t* CI, as well as the classical normal CI are two of the most beneficial CIs. The normal CI necessitates a sample size of 30 or greater. However, practical experiments may require smaller sample sizes, leading researchers to opt for the Student-*t* CI as opposed to the normal one. These intervals offer more comprehensive insights into the population characteristic of interest compared to a point estimator.

The Student-*t* is related to two issues. First, the Student-*t* distribution is symmetric and predicted according to the normalcy assumption. Thus, the $(1-\alpha)100\%$ CI pertaining to the population mean (μ) is also centred on the normality assumption. But in practice, the normalcy assumption is not met. Numerous writers have addressed the weakness of the Student-*t* technique in certain circumstances, including Boos & Hughes-Oliver (2000), David (1998), Desharnais *et al*. (2015), and Wilcox (2021). Prior studies have demonstrated that when considering small sample sizes and asymmetric distributions, the Student-*t* performs well because coverage probability (CP) approaches the nominal confidence coefficient despite its length and variability being larger than other CIs. When considering an asymmetric distribution and small sample sizes, Student-*t* performs well because its CP approaches the nominal confidence coefficient (Boos & Hughes-Oliver, 2000; Shi & Golam Kibria, 2007; Wang, 2001; Zhou & Dinh, 2005). The purpose of this paper is to comprehend CI utilising various techniques. This includes classical CI, percentile bootstrap method, bootstrap-*t* and proposed bootstrap-*t* decile mean method.

This work's remaining sections are arranged as follows: Section 2 demonstrates the intended CI. A simulation analysis was performed in Section 3 to assess the CP and length performance of the underlying CIs using the proposed approaches. In Section 4, the real data applications are displayed. The conclusions round out in Section 5.

***For correspondence:**
fairusmokhtar706@gmail.com

Received: 05 Feb. 2024
Accepted: 04 Sept. 2024

Methods for Estimation of the CI for the Population Mean

This particular section presents the methods for estimating the CI of the population mean (μ) pertaining to a skewed distribution. Assume that $X_1, X_2, \ldots, X_n$ is a randomly selected sample from a positively skewed distribution with unknown μ and σ. It is dispersed in an identical and independent manner. The bootstrap technique and Student-*t* are discussed in this study. The following presents the (1 – α) 100% CI pertaining to the population mean (μ) using several methods.

Classical CI for the Population Mean

This interval is a more reliable method of testing hypotheses, especially when σ is unknown, or sample size are limited. Its foundation is the normality assumption (Student 1908). The (1–α)100% CI for μ can be from Eq. 1 when σ is known.

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{1}$$

Eq. 2 can be used to construct the (1–α)100% CI for μ, also referred to as the Student-*t*, for small sample sizes and unknown σ.

$$\bar{X} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}} \tag{2}$$

in which $t_{(\alpha/2, n-1)}$ denotes the upper *α/2* percentage point of the Student-*t* distribution with (n-1) degrees of freedom. Here, the Student-*t* approach method is not very resilient when there are extreme deviations from normality (Boos and Hughes-Oliver 2000). Furthermore, the Student-*t* may not be the ideal CI for asymmetric distributions since it depends on the normality assumption.

The classical CI is widely used in statistical literature and practice due to its effectiveness, particularly in normal models. However, Student-*t* CI has poor coverage when the sample population is biased. The bootstrap method is an alternate strategy for estimating parameters if one disregards the presumption of normality (Flowers-Cano *et al*., 2018).

Percentile Bootstrap CI

Percentile bootstrap CI is another name for the bootstrap method in general (Abu-Shawiesh *et al*., 2022; Pek *et al*., 2017). An algorithmic method for generating a (1−α)100% percentile bootstrap CI pertaining to the population mean is as follows (Pek *et al*., 2017):

(i) For this bootstrap sample, resample the observed sample using a replacement, then determine the sample mean.
(ii) Repeat Step 1 *M* times.
(iii) After sorting *M* bootstrapped sample means, the (1−α)100%percentile bootstrap CI for the population mean is obtained from the (α/2)100th and (1-α/2)100th percentiles given.

The development of this CI has been concerned with the resampling technique, which is a difficult process and has a strong performance in theoretical CP. Still, it tends to be inconsistent in real practice depending on the bootstrap distribution (Sinsomboonthong *et al*., 2020). Furthermore, this method is challenging to compute, making it difficult to use without statistical programming, whereas the Bootstrap-*t* methods suggested in this study are simple to implement.

Bootstrap-*t*

Bootstrap-*t* CI is sometimes referred to as studentized or percentile-*t* bootstrap (Berrar, 2019). The bootstrap-*t* approach outperforms the Student-*t* method for varied sample sizes (Zhao *et al*., 2021). Moreover, the bootstrap-*t* CI is not always symmetrically constructed, although the conventional CI is typically symmetric (Berrar, 2019). The CI constructions for the expected value determined by random variables with a normal distribution are similar to the bootstrap-*t* construction. In each of the scenarios examined, the bootstrap-*t* approach uses these data to create less biased CI that is more accurate than earlier bootstrapping techniques (Hoyle & Cameron, 2003).

With this technique, the distribution of statistics $Z=\frac{\hat{\theta}-\theta}{\widehat{se}}$ can be inferred directly from the data. Z can be computed in the following manner for every set of bootstrap samples.

$$Z^*(b)=\frac{\hat{\theta}^*(b)-\theta}{\widehat{se}^*(b)} \tag{3}$$

Eq. 3 where $\hat{\theta}^*(b)$ is the $\hat{\theta}$ estimate for the b^{th} bootstrap sample, standard errors are often used to assign approximate CI to a parameter θ , and $\widehat{se}^*(b)$ denotes the $\hat{\theta}^*$ estimated standard error for the b^{th} bootstrap sample. Next, the α^{th} percentile of $Z^*(b)$ is computed as the value $\hat{t}^{(\alpha)}$. The CI is then calculated using Eq. 4 (Barker 2005).

$$\left(\hat{\theta}-\hat{t}^{(1-\alpha)}\times\widehat{se},\hat{\theta}-\hat{t}^{(\alpha)}\times\widehat{se}\right) \tag{4}$$

The Proposed Bootstrap-*t* Decile Mean

The robust CI bootstrap-*t* decile mean is a suggested modification of the Student-*t* CI centred on the decile mean as well as the decile standard deviation. The bootstrap-*t* decile mean is then calculated by using Eq. 5, where *DM* is the decile mean, SD_{DM} is the decile standard deviation, as well as *n* is the sample size (Abu-Shawiesh, Sinsomboonthong, and Kibria 2022).

$$CI = DM \pm t_{(\frac{\alpha}{2},n-1)}\frac{SD_{DM}}{\sqrt{n}} \tag{5}$$

Simulation Study

A simulation study has been performed to compare the CI performance because it is challenging to theoretically evaluate various CIs (Shi & Golam Kibria, 2007). R programming languages are used to conduct all of the simulation outcomes.

Performance Evaluation

In this section, the performance of the CI for the three distributions' population means (Normal, Chi-Square, and Lognormal) was compared. The research's main objective is to evaluate the CI's performance in determining the population mean (μ) of three distributions. The performance of CI employing CP is based on the findings of simulation studies and length is based on real data applications (Islam & Shapla, 2018; Moslim *et al*., 2019; Omar & Abu, 2011; Waguespack *et al*., 2020).

(i) For a given parameter, the CP of the CI is close to a nominal value of 0.95 under normal distribution (Kyselý 2010). When a distribution is not normal, the CP of the CI for a parameter can be appreciably below 0.95 (Niwitpong and Kirdwichai 2008; Waguespack, Krishnamoorthy, and Lee 2020). The suitable performing CI in each case had a CP is greater than or equal to nominal confidence level of 0.95 (Chankham, Niwitpong, & Niwitpong 2022).

$\text{Coverage probability}=\frac{m}{s}$, in which *m* resembles the number of true values that are contained within the CI and *s* number of bootstrap replications.

(ii) Length (i.e. difference between the lower and upper limits). Moreover, the shortest length implies a more precise estimation and improved performance of CI. The shortest length gives a better CI (Islam and Shapla 2018; Omar and Abu 2011). In addition, length decreases with increasing sample size (Waguespack, Krishnamoorthy, and Lee 2020).

Probability Distributions for the Simulation Study

In order to investigate the impact of skewness and assess the effectiveness of the CI estimators pertaining to the population mean (μ) of the distribution, this study examines two scenarios for the simulated observations: i) normal distribution and ii) skewed distribution. The term "skewness" describes the probability distribution's departure from symmetry. According to Sharma *et al*. (2009), a distribution that is positively skewed has a longer tail on the right side, whereas one that is negatively skewed has a longer tail on the left. Pertaining to skewed distributions, employ Eq. 6 and Eq. 8. to duplicate data from three probability distributions with different skewness levels.

Case (a): Normal distribution

There is no skewness, and the distribution is symmetric. Given a normal distribution that consists mean μ and standard deviation σ, $N(\mu,\sigma^2)$, the probability density function (pdf) is given like below:

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};-\infty<x<\infty,-\infty<\mu<\infty,\sigma>0 \tag{6}$$

The population mean μ as well as the population standard deviation σ n this study's simulation algorithm are set to μ=0 and σ=1, respectively.

Case (b): Chi-square distribution

$$f(x;k)=\begin{cases}\frac{1}{\Gamma(k/2)2^{k-1}}x^{(k/2)-1}e^{-x/2} & x>0\\ 0 & otherwise\end{cases} \tag{7}$$

$\mu=k$ and $\sigma^2=2k$ represent the mean as well as the variance of the chi-square distribution, accordingly. The distribution's coefficient of skewness is $\sqrt{8/k}$. The parameter k for the chi-square distribution is specified as k = 5 and k = 50 in this study's simulation algorithm.

The particular distributions as well as skewness coefficients employed in this simulation study are displayed in Table 1.

Case (c): Lognormal distribution

The lognormal distribution is a tail-heavy distribution and is used when uncertainty estimates are expected to be positively skewed. These two non-normal distributions may not cover most of the skewed distributions. The Weibull and Gamma have approximately similar properties to log-normal but are less tail-heavy than log-normal (Tong, Saminathan, and Chang 2016). The probability density function of the lognormal distribution is defined as Eq. 8.

$$f(x)=\frac{e^{-((\ln(x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}};x>\theta;m,\sigma>0 \tag{8}$$

where σ is the shape parameter (and is the standard deviation of the log of the distribution), θ is the location parameter and m is the scale parameter (and is also the median of the distribution). The coefficient of skewness of the distribution is

$\sqrt{e^{\sigma^2}-1}(e^{\sigma^2}+2)$ which is always positive.

Table 1. The parameters of normal, chi-square, and lognormal distribution

Probability distribution	Parameters	Skewness
Normal	$\mu=0,\sigma=1$	0.0000
Chi-square	k=5 k=50	1.0690 0.4000
Lognormal distribution	$\mu=1,\sigma=0.5$ $\mu=1,\sigma=1$	1.75 6.18

The Simulation Technique

The following is the algorithm for the simulation study (Abu-Shawiesh et. al. 2018):

i) Decide on the number of simulation times (*M*), sample size (*n*), as well as significance level (α).
ii) Using the R software, create a random sample of size (*n*), $X_1, X_2, \ldots, X_n$, which is independently and identically distributed and originates from a normal, chi-square, as well as lognormal distribution with parameter with the selected population skewness.
iii) Using the formulas from Section 2, construct CI at a (1-α)100% confidence level.
iv) Determine whether each CI contains μ, and compute the expected length of the CI for those CIs that contain the mean.
v) Repeat steps (i) through (iv) *M* times. Then, as evaluation criteria, compute the CP.

According to the simulation's outcome, 10,000 random sample sizes of *n*=10, 20, 30, 50, 100, 200, as well as 400 were generated (Abu-Shawiesh, Sinsomboonthong, and Kibria 2022; Dey *et al*. 2019; Zhao *et al*. 2021). Select between 1000, 2000, and 5000 simulations. The most popular 95% CI (α = 0.05) is used for the confidence coefficient. It is commonly known that the CP will be exact or very near to (1−α) if the data come from a symmetric distribution or if *n* is large.

Simulation Results

According to the data gathered and presented in Table 2, the CI for CP Student-*t* CI is similar for bootstrap 1000, bootstrap 2000, and bootstrap 5000. The likelihood of coverage rose as the sample size grew. The bootstrap-*t* DM, on the other hand, appears to understate the likelihood of coverage for every sample size in data with a normal distribution. The traditional technique performs well in terms of coverage for normal data. Bootstrap-*t* DM constantly provides lower CP relative to the goal value, which leads to the least effective performance strategy. It is true that as sample sizes increase, so do CIs.

Table 2. CP of the 95% CIs for the N(0,1) distribution's population mean with skewness

Bootstrap	Sample size	Student-*t*	Percentile	Bootstrap-*t*	Bootstrap-*t* DM
1000	10	0.9485	0.8985	**0.95**	0.507
	20	**0.9502**	0.9247	0.956	0.116
	30	**0.9525**	0.9317	0.954	0.022
	50	0.9492	0.942	**0.95**	0
	100	**0.9505**	0.9403	0.951	0
	200	**0.9532**	0.9456	0.955	0
	400	**0.9537**	0.9476	0.958	0
2000	10	0.9485	0.9002	**0.951**	0
	20	**0.9502**	0.9262	0.954	0
	30	**0.9525**	0.9393	0.9505	0
	50	**0.9492**	0.9392	0.951	0
	100	**0.9505**	0.948	0.9485	0
	200	**0.9532**	0.9435	0.954	0
	400	**0.9537**	0.9499	0.958	0
5000	10	**0.9485**	0.9065	0.9526	0.0002
	20	0.9502	0.9251	**0.95**	0
	30	0.9525	0.9326	**0.9516**	0
	50	**0.9492**	0.9417	0.9518	0
	100	0.9505	**0.9456**	0.9514	0
	200	0.9532	0.9476	**0.9516**	0
	400	0.9537	**0.9503**	0.9506	0

Nonnormality is classified into three categories: values less than 1.0 for skewness and kurtosis, values for moderate nonnormality range from 1.0 to approximately 2.3, while those for severe nonnormality exceed 2.3 (Lei & Lomax, 2005). In a similar vein, Bulmer (1979) noted that distribution skewness is fairly symmetrical between 0 and 0.5, moderately skewed between 0.5 and 1, as well as highly skewed beyond 1. To examine how well the CI performs under skewed distributions, generate random samples from several skewed distributions with skewness ranging from 0.4 to 6.18 (Banik & Kibria, 2010). Table 3 shows the low skewness, Table 4 and 5 show the moderate skewness, while Table 6 shows the high skewness.

According to the data shown in Table 3, the CI for CP Student-*t* is similar to Bootstrap 1000, 2000, and 5000. In every interval, the CP drops as the sample size increases. 50% of Student-*t* and 50% of bootstrap-*t* perform well in terms of coverage. However, bootstrap-*t* DM constantly provides a CP that is less than the ideal value, leading to the least effective approach. This class of distributions has almost the same length for Student-*t*, percentile, and bootstrap-*t*. When increasing the sample size, CP will decrease. It shows that bootstrap-*t* DM is good for a small sample size. The result is similar for bootstrap-*t* DM (Abu-Shawiesh, Sinsomboonthong, & Kibria 2022).

Table 3. CP of the 95% CIs for the population mean of chi-square with degrees of freedom 50 chi-square distribution and skewness 0.4000

Bootstrap	Sample size	Student-*t*	Percentile	Bootstrap-*t*	Bootstrap-*t* DM
1000	10	0.9444	0.8954	**0.953**	0.861
	20	0.9447	0.9306	**0.949**	0.752
	30	**0.9504**	0.9312	0.951	0.657
	50	**0.9496**	0.9367	0.94	0.431
	100	**0.9485**	0.9459	0.948	0.093
	200	**0.948**	0.9474	0.957	0.001
	400	**0.9491**	0.946	0.955	0
2000	10	0.9444	0.9011	**0.949**	0.853
	20	0.9447	0.9208	**0.9495**	0.7355
	30	**0.9504**	0.9338	0.948	0.6235
	50	**0.9496**	0.9399	0.9475	0.403
	100	**0.9499**	0.9401	0.9505	0.0815
	200	**0.948**	0.9479	0.955	0.001
	400	0.9491	0.9461	**0.9495**	0
5000	10	0.9444	0.9014	**0.9478**	0.8462
	20	0.9447	0.9292	**0.9516**	0.7248
	30	**0.9504**	0.9345	0.952	0.6068
	50	**0.9496**	0.945	0.949	0.3732
	100	**0.9485**	0.9475	0.9518	0.072
	200	0.948	0.9472	**0.9516**	0.0006
	400	0.9491	0.9474	**0.95**	0

It is evident from the data collected and shown in Table 4 that the CI for CP Student-*t* is similar to bootstrap 1000, 2000 and 5000. As sample sizes increase, the CP decreases in bootstrap-*t* DM. It shows that bootstrap-*t* DM is good for a small sample size. The result is similar for bootstrap-*t* DM (Abu-Shawiesh, Sinsomboonthong, & Kibria 2022). The majority of bootstrap-*t* gives good results compared to other methods. Bootstrap-*t* DM provides a CP consistently below the target value, resulting in the least effective method.

Table 4. CP of the 95% CIs for the population mean of chi-square with degrees of freedom 5 chi-square distribution and skewness 1.0690

Bootstrap	Sample size	Student-*t*	Percentile	Bootstrap-*t*	Bootstrap-*t* DM
1000	10	0.9299	0.8824	**0.959**	0.912
	20	0.9394	0.9155	**0.945**	0.859
	30	0.9422	0.924	**0.952**	0.807
	50	0.9457	0.9355	**0.947**	0.695
	100	0.9469	0.9422	**0.948**	0.398
	200	0.9461	0.9456	**0.9456**	0.099
	400	0.9489	0.9486	**0.943**	0.001
2000	10	0.9299	0.8851	**0.9565**	0.911
	20	0.9394	0.913	**0.9475**	0.863
	30	0.9422	0.9292	**0.9505**	0.8025
	50	0.9457	0.931	**0.948**	0.702
	100	0.9469	**0.9473**	0.947	0.4205
	200	0.9461	**0.9495**	0.945	0.0895
	400	0.9489	**0.9495**	0.949	0.0005
5000	10	0.9299	0.8867	**0.9526**	0.9058
	20	0.9394	0.917	**0.9504**	0.8634
	30	0.9422	0.9294	**0.9518**	0.81
	50	0.9457	0.9421	**0.9512**	0.7122
	100	0.9469	0.9425	**0.9476**	0.4472
	200	0.9461	0.9461	**0.9484**	0.101
	400	**0.9489**	0.9534	0.9482	0.0016

The simulated results in Table 5 show the performance of lognormal distribution with skewness 1.75 (moderately skewed). The simulation results indicate that the CP of bootstrap-*t* DM were higher than or close to the nominal level in majority scenarios (Chankham, Niwitpong, & Niwitpong 2022). The result is similar for bootstrap-*t* DM (Abu-Shawiesh, Sinsomboonthong, & Kibria 2022).

Table 5. CP of the 95% CIs for the population mean of lognormal distribution with mean 1 and standard deviation with 0.5 with skewness 1.75

Bootstrap	Sample size	Student-*t*	Percentile	Bootstrap-*t*	Bootstrap-*t* DM
1000	10	0.9239	0.8727	0.955	**0.96**
	20	0.9318	0.9114	0.969	**0.963**
	30	0.9396	0.9216	0.962	**0.96**
	50	0.9414	0.9359	0.953	**0.966**
	100	0.9484	0.936	0.953	**0.952**
	200	0.9518	0.9454	**0.959**	0.903
	400	0.9534	0.9444	**0.951**	0.812
2000	10	0.9239	0.8802	0.959	**0.9585**
	20	0.9318	0.9167	0.965	**0.958**
	30	0.9396	0.9276	0.9555	**0.9605**
	50	0.9414	0.9317	0.953	**0.964**
	100	0.9484	0.9419	0.954	**0.9525**
	200	0.9518	0.9443	**0.9555**	0.9055
	400	0.9534	0.9477	**0.956**	0.828
5000	10	0.9239	0.8727	0.9538	**0.9584**
	20	0.9318	0.9114	0.956	**0.958**
	30	0.9396	0.9216	0.9505	**0.9616**
	50	0.9414	0.9359	0.9502	**0.9628**
	100	0.9484	0.936	0.9498	**0.9514**
	200	0.9518	0.9454	**0.951**	0.9082
	400	0.9534	0.9444	**0.9518**	0.8178

The simulated results in Table 6 show the performance of lognormal distribution with skewness 6.18 (highly skewed). Bootstrap-*t* showed that the CP was higher than or close to 0.95 (Chankham, Niwitpong, and Niwitpong 2022). In addition, the efficiency of the proposed method is similar to Bootstrap-*t*. Bootstrap-*t* DM closes the Bootstrap-*t* for small sample sizes. It is found that bootstrap-*t* and bootstrap-*t* DM are robust because the CP is higher than Student-*t* and Percentile for a small sample size (Abu-shawiesh, Saghir; 2019). In conclusion, the simulation study shows that for small sample sizes and moderate to highly skewed distributions, bootstrap-*t* DM performs a suitable method in the sense CP was higher than or close to 0.95.

Table 6: CP of the 95% CIs for the population mean of lognormal distribution with mean 1 and standard deviation with 1 with skewness 6.18

Bootstrap	Sample size	Student-*t*	Percentile	Bootstrap-*t*	Bootstrap-*t* DM
1000	10	0.8428	0.7998	**0.95**	0.947
	20	0.8665	0.8583	**0.956**	0.932
	30	0.8884	0.8749	**0.954**	0.914
	50	0.899	0.9072	**0.95**	0.889
	100	0.9185	0.914	**0.951**	0.805
	200	0.931	0.9321	**0.955**	0.649
	400	0.9421	0.9355	**0.958**	0.337
2000	10	0.8428	0.8138	0.951	**0.949**
	20	0.8665	0.8647	**0.954**	0.932
	30	0.8884	0.885	**0.9505**	0.919
	50	0.899	0.9005	**0.951**	0.8865
	100	0.9185	0.9151	**0.9485**	0.8195
	200	0.931	0.9327	**0.9535**	0.667
	400	0.9421	0.9393	**0.958**	0.3185
5000	10	0.8428	0.8131	0.9574	**0.9534**
	20	0.8665	0.8584	**0.952**	0.9362
	30	0.8884	0.8759	**0.955**	0.9276
	50	0.899	0.8972	**0.95**	0.8936
	100	0.9185	0.918	**0.9514**	0.827
	200	0.931	0.9359	**0.9516**	0.6716
	400	0.9421	0.9367	**0.9506**	0.3146

Real Data Applications

The various approaches to building a CI around the normally distributed and non-normally distributed data are demonstrated in this section using two real-world examples. This paper will show that there can be significant differences in the outcomes of the methods discussed here.

Example 1

Liquidity is the easier the asset or security can be converted into cash (Ahmad *et al.* 2019). The data describe standard liquidity for housing construction in Malaysia from 2000 to 2018:

0, 0.00136868, 0.00283931, 0.00443463, 0.00613946, 0.00794009, 0.00982426, 0.0117811, 0.0138008, 0.015875, 0.0179959, 0.020157, 0.0223523, 0.0245765, 0.0268252, 0.0290941, 0.0313798, 0.0336789, 0.0359888.

The histogram and box plot are displayed in Figure 1. The Shapiro-Wilk test is taken into consideration when examining the sample data's normal distribution. It was discovered that the Shapiro-Wilk test statistic possesses a p-value > 0.05 (W= 0.95194, p-value = 0.426). Thus, at significance level α = 5%, it can be inferred that standard liquidity is normally distributed.

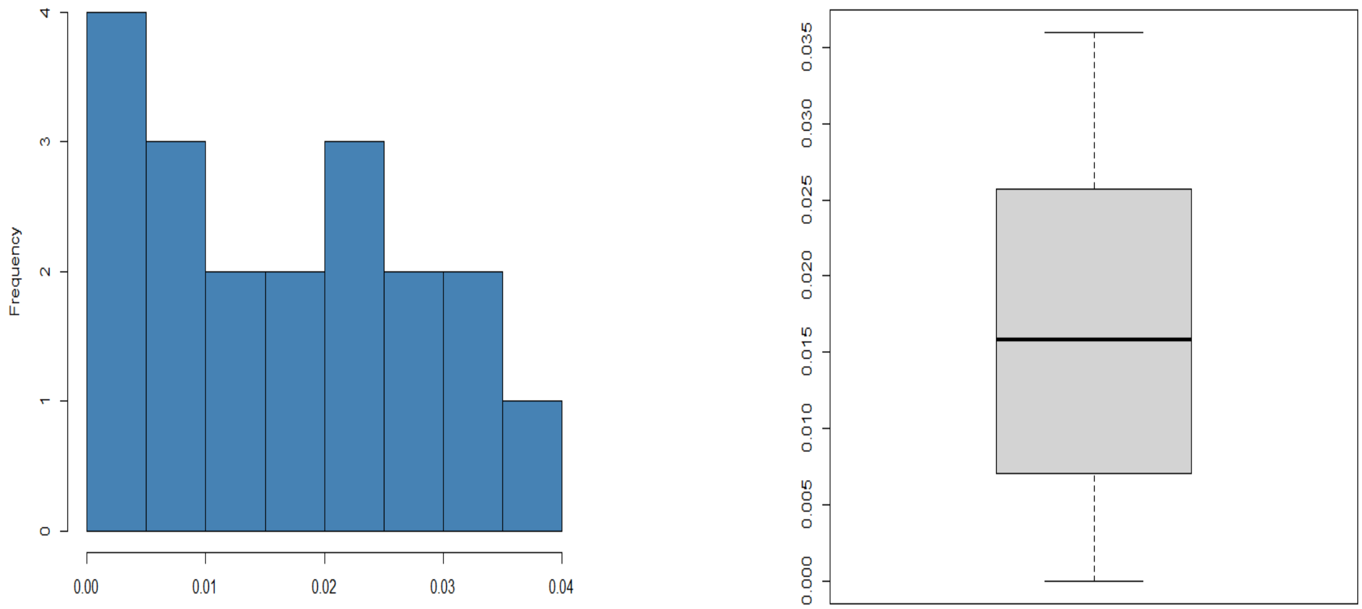

Figure 1. Boxplot and histogram of standard liquidity for housing construction data

Table 7. The 95% CIs pertaining to the average standard liquidity for housing construction in Malaysia

Methods	Estimated CI limit		Length
	Lower limit	Upper Limit	
Student-*t*	0.01110977	0.02215885	0.01104908
Percentile	0.000615906	0.034949345	0.034333439
Bootstrap-*t*	0.01110977	0.02215885	0.01104908
Bootstrap-*t* DM	0.01121523	0.02161599	**0.01040076**

The corresponding lengths and CIs for these are provided in Table 7. Out of all the intervals, Table 7 demonstrates that the bootstrap-*t* DM CI possesses the smallest width. Classical Student-*t* and bootstrap-*t* have the same value. Note that the data are extremely skewed, and the sample size is small. As a result, bootstrap-*t* DM CI outperforms the other CI in terms of shorter length. The outcomes of this example validated the findings of the simulation study.

Example 2

The number of psychotropic drug users was determined from a random sample of n=20 from various drug categories in order to examine the average use of psychotropic drugs among users of non-antipsychotic drugs. According to Johnson and McFarland (1993), the number of users is represented by the following data:

43.4, 24, 1.8, 0, 0.1, 170.1, 0.4, 150, 31.5, 5.2, 35.7, 27.3, 5, 64.3, 70, 94, 61.9, 9.1, 38.8 and 14.8.

Upon verification, the data exhibit a positive skewness of 1.57, a mean of 42.37, as well as a standard deviation of 48.43. Figure 2 displays a boxplot and histogram of the data values indicating a positive skew. Table 8 provides the suggested CI and corresponding lengths.

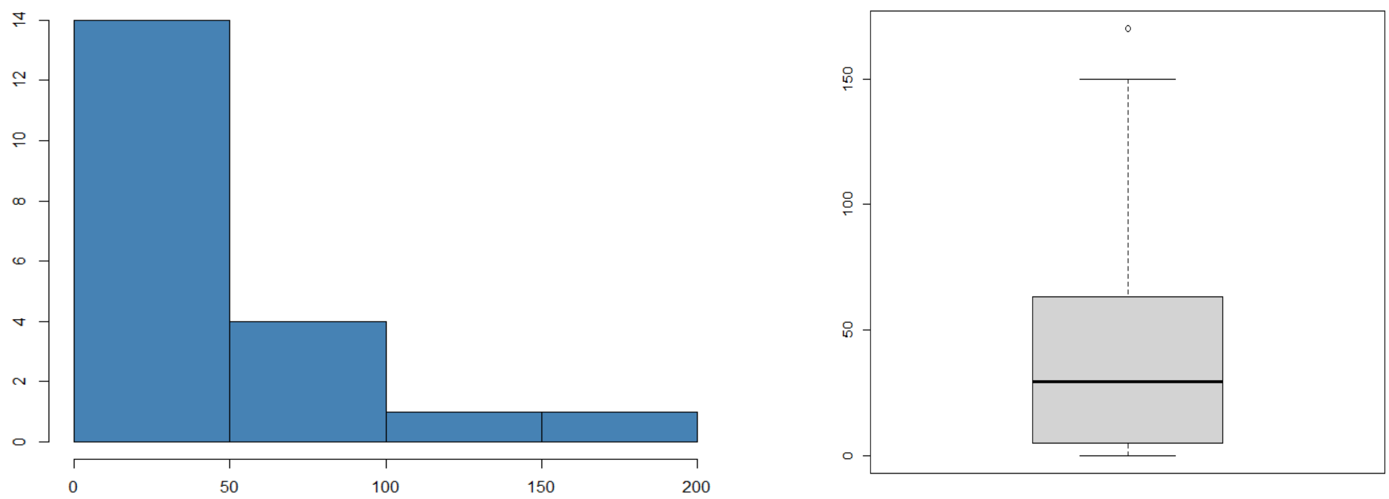

Figure 2. Boxplot and histogram of psychotropic drug exposure data

Table 8. The 95% CIs for the average use of psychotropic drugs methods

Methods	Estimated CI limit		Length
	Lower limit	Upper Limit	
Student-*t*	19.70406	65.03594	45.33188
Percentile	0.0475	160.5525	160.505
Bootstrap-*t*	19.70349	65.03651	45.33302
Bootstrap-*t* DM	19.34062	50.28383	**30.9432**

The corresponding lengths and CIs for these are provided in Table 8. The bootstrap-*t* DM CI possesses the smallest width, as shown in Table 8. It is then followed by classical Student-*t*, bootstrap-*t*, as well as percentile CIs. The widest CI is the classical Student-*t*. It is noteworthy that the sample size is small as well as data are moderate. As a result, bootstrap-*t* DM CI outperforms the other CI in terms of shorter length. The outcomes of this example validated the findings of the simulation study.

Conclusions

This research examines several parameter settings for CI estimation. To compare the CI performance, a simulation study was conducted as a theoretical comparison is not practical. Data gets generated using several distributions, including normal and skewed distributions. CP and length are regarded as good indicator criteria. The most commonly used approach for estimating the mean is Student-*t* (Akyüz & Abu-Shawiesh, 2020). The robust CI bootstrap-*t* DM is presented as a modification to the Student-*t* CI depending on the decile mean as well as the decile standard deviation (Abu-Shawiesh *et al*., 2022). The simulation findings show that the performance of the Student-*t* and three bootstrap approaches (percentile, bootstrap-*t*, and bootstrap-*t* DM) varies dramatically depending on sample size and skewness type. When the distribution is normal and low skewed, the simulation study reveals that Student-*t* outperforms other techniques. According to the simulation study's findings, the optimal CI determined by CP for situations with moderate to high skewness is bootstrap-*t*. However, bootstrap-*t* DM is preferable for small sample sizes and moderately to significantly skewed data. Consequently, researchers must determine whether CP or length is more significant when selecting a study CI because it is difficult to identify a CI with a close CP to 0.95 and a short length (Abu-Shawiesh *et al*., 2018; Abu-Shawiesh & Saghir, 2019). The study's discoveries are demonstrated by analyzing two actual datasets, validating the simulation outcomes. In conclusion, the proposed CI techniques outperformed the traditional Student-*t* CI in terms of population mean estimation. This paper's findings will be useful to a variety of researchers with practical experience in the fields of science and social sciences.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The authors are thankful to the reviewers and the editor for providing some useful comments and suggestions.

References

[1] Abu-Shawiesh, *M. O. A.*, *et al.* (2018). Confidence intervals based on absolute deviation for population mean of a positively skewed distribution. *International Journal of Computational and Theoretical Statistics, 5*(1), 1–13. https://doi.org/10.12785/ijcts/050101

[2] Abu-Shawiesh, *M. O. A.*, Sinsomboonthong, *J.*, & Kibria, *B. M. G.* (2022). A modified robust confidence interval for the population mean of distribution based on deciles. *Statistics in Transition New Series, 23*(1), 109–128. https://doi.org/10.21307/stattrans-2022-007

[3] Abu-Shawiesh, *M. O. A.*, Saghir, *A.* (2019). Robust confidence intervals for the population mean: Alternatives to the Student-t confidence interval. *Journal of Modern Applied Statistical Methods, 18*(1), 1–21. https://doi.org/10.22237/JMASM/1556669160

[4] Ahmad, *M. F.*, Sapiri, *H.*, Misiran Bakun, *M.*, Hashim, *Z.*, & Abdul Halim, *T. F.* (2019). A system dynamics model of Malaysia house pricing. *International Journal of Engineering and Advanced Technology, 8*(6 Special Issue 3), 57–64. https://doi.org/10.35940/ijeat.F1010.0986S319

[5] Barker, *N.* (2005). A practical introduction to the bootstrap using the SAS system. *Proceedings of the Pharmaceutical Users Software Exchange Conference*, Paper PK02. http://www.lexjansen.com/phuse/2005/pk/pk02.pdf

[6] Berrar, *D.* (2019). Introduction to the non-parametric bootstrap. *April*, 0–15. https://doi.org/10.1016/B978-0-12-809633-8

[7] Boos, *D. D.*, & Hughes-Oliver, *J. M.* (2000). How large does n have to be for Z and t intervals? *American Statistician, 54*(2), 121–128. https://doi.org/10.1080/00031305.2000.10474524

[8] Chankham, *W.*, Niwitpong, *S. A.*, & Niwitpong, *S.* (2022, December). Confidence intervals for ratio of coefficients of variation of inverse Gaussian distribution. In *Proceedings of the 2022 International Conference on Big Data, IoT, and Cloud Computing* (pp. 1–5).

[9] David, *H. A.* (1998). Early sample measures of variability. *Statistical Science, 13*(4), 368–377. https://doi.org/10.1214/ss/1028905831

[10] Desharnais, *B.*, Camirand-Lemyre, *F.*, Mireault, *P.*, & Skinner, *C. D.* (2015). Determination of confidence intervals in non-normal data: Application of the bootstrap to cocaine concentration in femoral blood. *Journal of Analytical Toxicology, 39*(2), 113–117. https://doi.org/10.1093/jat/bku127

[11] Efron, *B.* (1979). Bootstrap method: Another look at jackknife. In *Breakthroughs in Statistics* (Vol. 7, pp. 569–593).

[12] Flowers-Cano, *R. S.*, Ortiz-Gómez, *R.*, León-Jiménez, *J. E.*, Rivera, *R. L.*, & Cruz, *L. A. P.* (2018). Comparison of bootstrap confidence intervals using Monte Carlo simulations. *Water (Switzerland), 10*(2), 1–21. https://doi.org/10.3390/w10020166

[13] Hoyle, *S. D.*, & Cameron, *D. S.* (2003). Confidence intervals on catch estimates from a recreational fishing survey: A comparison of bootstrap methods. *Fisheries Management and Ecology, 10*(2), 97–108. https://doi.org/10.1046/j.1365-2400.2003.00321.x

[14] 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), 41–51.

[15] Kyselý, *J.* (2010). Coverage probability of bootstrap confidence intervals in heavy-tailed frequency models, with application to precipitation data. *Theoretical and Applied Climatology, 101*(3), 345–361. https://doi.org/10.1007/s00704-009-0190-1

[16] Lee, *K. M.*, Lee, *M. H.*, Lee, *J. S.*, & Lee, *J. Y.* (2020). Uncertainty analysis of greenhouse gas (GHG) emissions simulated by the parametric Monte Carlo simulation and nonparametric bootstrap method. *Energies, 13*(18). https://doi.org/10.3390/en13184965

[17] Moslim, *N. H.*, Zubairi, *Y. Z.*, Hussin, *A. G.*, Hassan, *S. F.*, & Mokhtar, *N. A.* (2019). A comparison of asymptotic and bootstrapping approaches in constructing confidence intervals of the concentration parameter in von Mises distribution. *Sains Malaysiana, 48*(5), 1151–1156.

[18] Niwitpong, *S.*, & Kirdwichai, *P.* (2008). Adjusted Bonett confidence interval for standard deviation of non-normal distributions. *Thailand Statistician, 6*(January), 1–16.

[19] Omar, *M.*, & Abu, *A.* (2011). A simulation study on some confidence intervals for the population standard deviation. *Sort, 35*(2), 83–102.

[20] Pek, *J.*, Wong, *A. C. M.*, & Wong, *O. C. Y.* (2017). Confidence intervals for the mean of non-normal distribution: Transform or not to transform. *Open Journal of Statistics, 7*(3), 405–421. https://doi.org/10.4236/ojs.2017.73029

[21] Shi, *W.*, & Golam Kibria, *B. M.* (2007). On some confidence intervals for estimating the mean of a skewed population. *International Journal of Mathematical Education in Science and Technology, 38*(3), 412–421. https://doi.org/10.1080/00207390601116086

[22] Sinsomboonthong, *J.*, Abu-Shawiesh, *M. O. A.*, & Kibria, *B. M. G.* (2020). Performance of robust confidence intervals for estimating population mean under both non-normality and in presence of outliers. *Advances in Science, Technology and Engineering Systems, 5*(3), 442–449. https://doi.org/10.25046/aj050355
[23] Student. (1908). The probable error of a mean (pp. 1–25).
[24] Waguespack, *D.*, Krishnamoorthy, *K.*, & Lee, *M.* (2020). Tests and confidence intervals for the mean of a zero-inflated Poisson distribution. *American Journal of Mathematical and Management Sciences, 39*(4), 383–390. https://doi.org/10.1080/01966324.2020.1777914
[25] Wang, *F.-K.* (2001). Non-normal data. *Quality and Reliability Engineering International, 267*(259), 257–267. https://doi.org/10.1002/qre400
[26] Wilcox, *R. R.* (2021). A note on computing a confidence interval for the mean. *Communications in Statistics: Simulation and Computation, 0*(0), 1–3. https://doi.org/10.1080/03610918.2021.2011926
[27] Zhao, *S.*, Yang, *Z.*, Musa, *S. S.*, Ran, *J.*, Chong, *M. K. C.*, Javanbakht, *M.*, He, *D.*, & Wang, *M. H.* (2021). Attach importance of the bootstrap t-test against Student's t-test in clinical epidemiology: A demonstrative comparison using COVID-19 as an example. *Epidemiology and Infection, 149*. https://doi.org/10.1017/S0950268821001047
[28] Zhou, *X. H.*, & Dinh, *P.* (2005). Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics, 6*(2), 187–200. https://doi.org/10.1093/biostatistics/kxi002