Estimating the Population Mean Under Contaminated Normal Distribution Based on Deciles

Moustafa Omar Ahmed Abu-Shawiesh (mabushawiesh@hu.edu.jo)
The Hashemite University (HU)
Juthaphorn Sinsomboonthong
Kasetsart University (KU)
Nadia Saeed
University of the Punjab (PU)
Golam Kibria
Florida International University

Research Article

Keywords: Robust measure, confidence interval, decile mean, decile mean standard deviation, decile mean standard error, average width, coverage probability, contamination

Posted Date: February 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1314340/v1

Estimating the Population Mean under Contaminated Normal Distribution Based on Deciles

Moustafa Omar Ahmed Abu-Shawiesh*
Department of Mathematics, Faculty of Science
The Hashemite University (HU), P.O. Box 330127, Zarqa 13133, Jordan

Juthaphorn Sinsomboonthong
Department of Statistics, Faculty of Science
Kasetsart University (KU), Bangkok 10900, Thailand

Nadia Saeed
College of Statistical and Actuarial Sciences
University of the Punjab (PU)
Lahore 54000, Pakistan

Golam Kibria
Department of Mathematics and Statistics, Florida International University
University Park Campus, Miami, FL 33199, USA

Abstract

Confidence intervals are important statistical methods used to estimate the location and dispersion parameters of the population. A new robust interval estimator, which is an adjustment of the Student-t confidence interval for estimating population mean based on the decile mean and standard deviation is consider in this research. The efficiency of this proposed interval estimator is evaluated using an extensive Monte-Carlo simulation study. The coverage probabilities and average widths of the proposed interval estimator are compared with some existing widely used interval estimators under normal and contaminated normal distributions. The simulation results show that the proposed interval estimator performs very well in terms of attaining high coverage probability and shorter average width. For illustration purposes, real-life data sets are analyzed which supported the findings obtained from the simulation study to some extent. In summary, our results confirmed that the type of estimator used to construct the confidence interval affects the performance of the interval estimator, and the proposed version of the interval estimator performs better than the other estimators evaluated herein. Consequently, we recommend the new robust confidence interval for the practitioners to be used for estimating of the population mean when the contamination in the data of the distribution is present. The proposed confidence interval of the population mean can be easily calculated by using R program which is providing in this appendix.

Keywords: Robust measure; confidence interval; decile mean; decile mean standard deviation; decile mean standard error; average width; coverage probability; contamination.

Mathematics Subject Classification 62F10; 62F35.

*** Correspondence Author:** mabushawiesh@hu.edu.jo

1. Introduction

Most of the statistical estimation theories are developed on the assumption of normality. One of these theories to construct the confidence interval (CI) is through estimation developed by Neyman (1937). The confidence interval is an important statistical method used to estimate the population location and dispersion parameters and is defined as a range of values determines how precise the estimates of a parameter are (Abu-Shawiesh et al. 2019). However, in real-life the data set do not always follow normality assumption and are not mound shaped, instead of they are skewed or contaminated. Therefore, the violation of this assumption and presence of outliers can badly affect the performance of the said method. For example, right tailed data are common in numerous fields of modeling such as health science (Ghosh and Polansky 2016, Banik and Kibria 2010, Baklizi and Kibria 2009 and Zhou et al. 2001), environmental science (Mudelsee and Alkio, 2007), biological science (Andersson 2004 and Gregoire and Schabenberger 1999), engineering science and others. Furthermore purity of the water and customer waiting times, chemical process like impurity levels in semiconductor and beta particle emissions in nuclear reaction from many real-world distributions follow abnormal and contaminated distributions (Hurdey and Hurdey 2004, Miller and Miler 1995, Bissell 1994 and James 1989). In these circumstances, it is not suitable to construct the confidence intervals depend upon the assumption of normality and therefore, there is a necessity to adjust and modify the confidence intervals for contaminated distributions.

Researchers are habitually fascinated in making inference about the location and dispersion parameters of the populations. This inference can be made by finding a suitable confidence interval (CI). A CI is an interval estimate that will capture the true parameter value when the samples are collected repeatedly. A convenient way to perform the test of hypothesis on behalf of CI is to reject the null hypothesis if the assumed parameter lies beyond the limits calculated from 100(1-α)% CI. In this paper, our interest is to consider the problem of estimation of the location parameter with certain pre-specified confidence level. It is a usual practice to use normal theory to construct CI for making inferences about the mean of the population. However, in practice the normality assumption may violates in many situations. For example, health related data are skewed (see, for example, Li et al. 2011, Baklizi and Kibria 2009, Zhou et al. 2001) and data from Chemistry and Bioanalysis (Desharnais et al. 2015) are not normal.

Confidence interval based on ordinal-t or z-statistic suffers when samples come from skewed or contaminated populations. Several methods are readily available to overcome this issue.

Some of them are constructed for higher order terms by correcting the studentized t-statistic (Hall 1992 and Johnson 1978) while some are using bootstrap technique. However, in many applications, it seems more appropriate to use an estimator that will be resistant against outliers (Rostron et al 2020).

The purpose of this research is to evaluate and compare the performance of a proposed confidence interval based on robust estimators for the population mean with various existing interval estimators under normal and contaminated normal distributions. Since the aforementioned researchers considered several confidence intervals under different simulation conditions, their performance are not comparable as a whole. This justifies the uniqueness of our study. Moreover in this paper, we reviewed some widely used interval estimators and proposed a new robust interval estimator for the population mean based on the decile mean and the decile mean standard deviation. Since comparison is not possible theoretically, the performance evaluation is made on the basis of simulation study in terms of attaining the nominal values of confidence probabilities and widths of the intervals.

The rest of the paper is organized as follows: upcoming section describes the operational definitions of decile, decile mean and decile standard deviation. The proposed and existing interval estimators are discussed in Section 3. A simulation study is described in Section 4. The Results are discussed in section 5. Real-life data under normal and contaminated normal distributions are analyzed in Section 6. Finally, section 7 ends with some concluding remarks.

2. The Decile Mean and the Decile Mean Standard Deviation

The sample mean ($\bar{X}$) and sample standard deviation (S) of a statistical distribution are the classical estimators of the location and scale parameters respectively. They are the most popular and frequently used measures, however they are unreliable in the presence of non-normality or when outliers arise in data. Due to this reason, in this paper, we will replace them by well-known and simple robust estimators of location and scale. The most popular robust estimators of location and scale parameters are decile mean (DM) and decile mean standard deviation (SD_{DM}) respectively. Furthermore, the standard error of the decile mean standard deviation (SD_{DM}) will be defined.

2.1 The Deciles

The central tendency is a measure of location for a typical data set. The mean, median and trimmed mean of a sample data can be used as measures on central tendency which can adequately locate the center of data under well behaved normal distribution but under abnormal distributions with presence of outliers, the traditional location measure may behave poorly. A new measure of central tendency proposed by Rana et al. (2012) called decile mean (DM) since it is based on deciles. This measure is more informative than median as it automatically discards extreme observation from the both tails. The decile mean outperforms traditional measures i.e. mean, median and trimmed mean and hence it is called a robust measure of central tendency.

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a given population with mean (μ) and standard deviation (σ), the deciles, which are a measure of position, are the values (nine in numbers) of the variable that divide any ordered data set $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ into ten equal parts so that each part represents $\frac{1}{10}$ of the sample or population and are denoted by $D_1, D_2, \dots, D_9$, where the first decile (D_1) is the value of order statistics that exceed $\frac{1}{10}$ of the observations and less than the remaining $\frac{9}{10}$ and the ninth decile (D_9) is the value in order statistic that exceeds $\frac{9}{10}$ of the observations and is less than $\frac{1}{10}$ remaining observations. The fifth decile (D_5) is equal to the sample median (MD). The deciles determine the values for 10%, 20% … and 90% of the data set. The deciles can be calculated as follows:

(1) Order the observations $X_1, X_2, \dots, X_n$ according to the magnitudes of the values to get the ordered data set $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.

(2) To find the value of the m^{th} decile where $m = 1, 2, \dots, 9$, the following simple formula can be used:

$$D_m = \text{The value of the } \left(\frac{m\,(n+1)}{10}\right)^{th} \text{ observation} \tag{1}$$

where n is the total number (sample size) of observations.

2.2 The Decile Mean

The decile mean denoted by DM for the random sample $X_1, X_2, \dots, X_n$ can be calculated by summing of all deciles $D_1, D_2, \dots, D_9$ and divided the sum by the number of deciles. Hence, the formula to find the decile mean (DM) from nine deciles is given by:

$$DM = \frac{\sum_{i=1}^{9} D_i}{9} = \frac{D_1 + D_2 + \dots + D_9}{9} \tag{2}$$

The one of the main advantages of the decile mean (DM) is robust for outliers as compared to existing measures as it is based on 80% of a data set. It is referred as a robust estimator in this regard. Moreover, the bootstrap method to investigate the sampling distribution of the newly proposed decile mean (DM) with three other popular and commonly used measures of location, i.e., the sample mean, median and trimmed mean was first used by Rana et al. (2012). They found that the newly proposed decile mean (DM) has the following properties:

i. In general median, trimmed mean and decile mean do not affect by the outliers but the distribution of decile mean is quite normal even under the presence of outliers while the distributions of median and trimmed mean are far from normality.

ii. The proposed decile mean estimator has fewer amounts of bias and standard error as compared to other estimators considered for study. Hence decile mean is considered best for the purpose.

iii. Using the empirical approach based on Monte-Carlo simulation study, it is verified that the proposed decile mean estimator has small amount of bias relative to other estimators of central tendency i.e. sample mean, median and trimmed mean. The proposed estimator has also showed higher efficiency in terms of standard error and a drastic improvement is achieved in terms of substantially less standard error values as compared to other three estimators. The trend is same for different sample sizes.

iv. The decile mean holds remarkable properties as the bootstrap distribution is nearly normal. Although other estimators such as median and trimmed mean also hold some desirable asymptotic properties but decile mean pertain its asymptotically normal behavior for small to moderate samples.

The bootstrap simulation study demonstrates that the decile mean (DM) is more precise measure of central tendency in terms of possessing smaller bias and standard errors in a variety of situations and hence can be recommended to use as an effective measure of central tendency or location.

2.3 The Decile Mean Standard Deviation

The decile mean standard deviation (SD_{DM}) is a robust measure of dispersion proposed by Doullah (2018) as an alternative to the sample standard deviation (S). Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a given population with mean (μ) and standard deviation (σ), the decile mean standard deviation (SD_{DM}) can be calculated by using the following formula:

$$SD_{DM} = \sqrt{\frac{\sum_{i=1}^{9}(X_i - DM)^2}{9}} \tag{3}$$

Since, the standard error (SE) is calculated by sample standard deviation, which is considered to be extremely sensitive for the detection of outliers. Due to this reason, Doullah (2018) defined the standard error of the decile mean standard deviation (SD_{DM}), denoted by SE_{DM} , as follows:

$$SE_{DM} = \frac{SD_{DM}}{\sqrt{n}} \tag{4}$$

According to Doullah (2018), by using a Monte Carlo simulation study that is planned to compare the performance of the newly proposed robust estimators with the classical popular and frequently used estimators of dispersion based on computing the simulated mean, bias and standard error (SE), the results showed that the proposed estimators are more efficient than classical ones under different sample sizes, which offer the sound outcome in this simulation. The results also demonstrated that classical estimators of dispersion have shown opposite results in presence of outliers. Alternatively, proposed robust estimators have provided appropriate choice in both cases.

3. Methods for Computing the Confidence Interval for Mean

Let $X_1, X_2, \dots, X_n$ be identically independently distributed (IID) random sample of size n from a population with mean μ and standard deviation σ. The aim is to construct such an interval estimate for the population mean μ using specified confidence level. Several methods to construct such a confidence interval are addressed in literature such as parametric and non-parametric, modified t and bootstrap approaches. In this study, we will concentrate parametric and modified t approaches only. The 100(1 – α)% confidence interval (CI) for the population mean (μ) by different approaches are presented in the sections below.

3.1 The Parametric t-Approach

The parametric method to construct the 100(1 – α)% confidence interval is widely used approach to estimate population mean (μ). Under this approach, we will consider the following two confidence intervals under known and unknown population standard deviation. Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a normal distribution with mean (μ) and variance (σ^2), that is, $X_1, X_2, \dots, X_n \sim N(\mu,\ \sigma^2)$. Then, the 100(1 – α)% confidence interval for the population mean μ is constructed as:

$$C.I. = \bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{5}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean, σ is true population standard deviation and $Z_{1-\frac{\alpha}{2}}$ is $(1 - \alpha/2)^{th}$ quantile of the standard normal distribution $N(0, 1)$. In real life, however, it is impracticable to have a known population standard deviation (σ) and hence an estimate of σ is used by sample data. Also we can use the sample standard deviation if sample size is large ($n > 30$) and can apply the normal distribution to construct the 100(1 – α)% confidence interval for the population mean (μ) as:

$$C.I. = \bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \tag{6}$$

where $S = \sqrt{(n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$. On the other hand, for a small sample size n ($n \leq 30$) and unknown population standard deviation (σ), the student's t distribution can be used to construct 100(1 – α)% confidence interval for the population mean μ (Student 1908) as:

$$C.I. = \bar{X} \pm t_{(\frac{\alpha}{2},\ n-1)} \frac{S}{\sqrt{n}} \tag{7}$$

where $t_{(\alpha/2,\ n-1)}$ is the upper $\alpha/2$ percentage point of the student-t distribution with $(n - 1)$ degrees of freedom. As the Student-t CI is based on the assumption of normality, hence it is not considered best in all situations specifically when the outliers are present in the data. According to the Boos and Hughes-Oliver (2000), the CI based on Student-t statistic is not to be considered very sensitive if extreme deviations from normality is observed.

3.2 The Modified Parametric t-Approach

If the random sample $X_1, X_2, \dots, X_n$ follows a non-normal distribution, the distribution of the t-statistic cannot be same as students t distribution due to the reason that the skewness factor has a large impact on the validity of the t-distribution, see for example, Yanagihara and Yuan (2005). As a remedy to this effect of skewness, a large number of methods for constructing the 100(1 – α)% CI for the population mean (μ) have been proposed to eliminate the effect of non-normality by modifying the t-statistic. Here a brief review of most important of these methods is provided.

3.2.1 The Chebyshev Theorem Based t-Approach

Hogg and Craig (1978) proposed the following approach based on the Chebyshev inequality, which is defined as:

$$C.I. = \bar{X} \pm \sqrt{((\alpha/2)^{-1} - 1)}\ \hat{\sigma}_{\bar{X}} \tag{8}$$

where $\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$ is considered as an estimated standard error of the sample mean ($\bar{X}$).

3.2.2 The Johnson t-Approach

Johnson (1978) proposed the following modification of CI for the population mean (μ) by incorporating the first term of inverse Cornish–Fisher expansion:

$$C.I. = \left[\bar{X} + \frac{\hat{\mu}_3}{6nS^2}\right] \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{S}{\sqrt{n}} \tag{9}$$

where $\hat{\mu}_3 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}$ is the estimator of the third central moment of the population (μ_3). It is important to note that the width of student-t and Johnson-t confidence interval are considered same (see Kibria 2006).

3.2.3 The Chen t-Approach

Chen (1995) modified the central limit theorem approach using the Edgeworth expansion and proposed the following CI for the population mean (μ):

$$C.I. = \bar{X} \pm \left[t_{\left(\frac{\alpha}{2},\ n-1\right)} + \frac{\hat{\gamma}\left(1 + 2t^2_{\left(\frac{\alpha}{2},\ n-1\right)}\right)}{6\sqrt{n}} + \frac{\hat{\gamma}^2\left(t_{\left(\frac{\alpha}{2},\ n-1\right)} + 2t^2_{\left(\frac{\alpha}{2},\ n-1\right)}\right)}{9n} \right] \frac{S}{\sqrt{n}} \tag{10}$$

where $\hat{\gamma} = \frac{\hat{\mu}_3}{S^3}$ is the estimate of the coefficient of skewness.

3.2.4 The Yanagihara and Yuan t-Approach

To reduce the effect skewness and bias in mean, Yanagihara and Yuan (2005) proposed the following CI for the population mean (μ):

$$C.I. = \left[\bar{X} + \frac{(S\,\hat{k}_3)}{\left((4n)\left(2 + \frac{15}{n}\right)\right)}\right] \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{S}{\sqrt{n}} \tag{11}$$

where $\hat{k}_3 = \frac{\left(\sum_{i=1}^{n}(X_i - \bar{X})^3/n\right)}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2/n\right)^{3/2}}$.

3.2.5 The Shi and Kibria Median t-Approach

It is evident that sample median is appropriate measure of central tendency for abnormal distributions instead of sample mean as later is preferable when distribution is symmetric and free from outliers. Due to this remarkable property, it is reasonable to construct a confidence interval based on sample standard deviation in terms of the deviations taken from sample median. Shi and Kibria (2007) proposed the following CI for the population mean μ as:

$$C.I. = \bar{X} \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{\tilde{S}_1}{\sqrt{n}} \tag{12}$$

where $\tilde{S}_1 = \sqrt{(n-1)^{-1}\sum_{i=1}^{n}(X_i - MD)^2}$ and $MD = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & if \quad n\ is\ odd \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & if \quad n\ is\ even \end{cases}$.

3.2.6 The Shi and Kibria MAD t-Approach

In terms of the absolute deviation from mean rather than the sample mean for defining the sample standard deviation, Shi and Kibria (2007) proposed another CI for the population mean (μ) given as follows:

$$C.I. = \bar{X} \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{\tilde{S}_2}{\sqrt{n}} \tag{13}$$

where $\tilde{S}_2 = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right|$ is the estimated Mean Absolute Deviation (MAD).

3.2.7 The Abu-Shawiesh, Banik and Kibria AADM t-Approach

Abu-Shawiesh et al. (2018) proposed a modified form of CI for the population mean (μ) based on student's-t statistic for skewed distribution called AADM-t CI and given as follows:

$$C.I. = \bar{X} \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{AADM}{\sqrt{n}} \tag{14}$$

where $AADM = \frac{\sqrt{\pi/2}}{n}\sum_{i=1}^{n}|X_i - MD|$ given by Gastwirth (1982) is the average absolute deviation from the sample median. As stated by Gastwirth (1982), the distribution of AADM estimator is asymptotically normal distributed and it is also a consistent estimator of the population standard deviation σ.

3.3. The Proposed Robust $DMSD_{DM}$ t-Approach

In this subsection, a modified Student-t confidence interval for population mean under non-normal and contaminated populations is proposed. It is a simple adjustment of the Student-t CI based on the decile mean (DM) and the decile mean standard deviation (SD_{DM}) as given in Equations (2) and (3). This will refer to as $DMSD_{DM}$-t CI for mean. Thus, the proposed 100(1 – α)% $DMSD_{DM}$-t CI for the population mean (μ) is given as:

$$C.I. = DM \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{SD_{DM}}{\sqrt{n}} \tag{15}$$

4. Simulation Study

In this section, the performance of nine confidence interval estimators for the population mean is compared on the basis of simulation study conducted for the purpose. A set of possible useful CIs are considered and compared them with the proposed robust method and tried to confirm that it will be a good and useful one to estimate the population mean (μ) of a distribution. The simulation method is easy in implementation and provides results that are close to the exact case if large number of runs is used. In order to decide better confidence intervals on the basis of their performance, the popular performance measures such as the coverage probability (CP) and average width (AW) of confidence interval are considered. A smaller width indicates a better confidence interval when the coverage probabilities are the same. Further, the higher coverage probability indicates a better confidence interval when the widths are the same. The simulation study is conducted by using SAS version 9.4 programming. In this research, the most common confidence interval i.e. 95% confidence interval is used to check whether the confidence interval methods are sensitive to sample size. The sample size n = 10, 20, 30, 40, 50 and 100 were randomly generated 100,000 times. For each set of samples, at general level of confidence i.e. 95% confidence intervals are constructed for the considered methods. The coverage probability (CP) and the average width (AW) of the confidence interval are obtained by using the following two formulas:

$$CP = \frac{P(L \leq \theta \leq U)}{100{,}000} \quad \text{and} \quad AW = \frac{\sum_{i=1}^{100{,}000} (U_i - L_i)}{100{,}000} \tag{16}$$

In order to study the effect of non-normality and the presence of outliers in the performance of the classical confidence interval, the modified confidence intervals and the proposed robust confidence interval for the population mean (μ), the two cases for the simulated observations; normal and contaminated normal ones; are considered in this study. The details of these two cases are as follows:

Case (a): Normal Distribution

A symmetric distribution has no skewness. Therefore, the coefficient of skewness for a normal distribution is zero since it is a symmetric one. The probability density function (pdf) of a normal distribution with mean μ and standard deviation σ, $N(\mu, \sigma^2)$, is given as follows:

$$f(x;\ \mu, \sigma) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & , \quad -\infty < x < \infty\ ;\ -\infty < \mu < \infty\ ,\ \sigma > 0 \\ 0 & , \quad otherwise \end{cases} \tag{17}$$

In the simulation algorithm of this study, the population mean μ and the population standard deviation σ are set as $\mu = 20$ and $\sigma = 5, 10, 20$.

Case (b): Contaminated Normal Distribution

In this case, to assess the sensitivity of confidence intervals under the presence of outliers, the set of observations is generated from the mixture distributions by artificially incorporated outlier values that are called contaminated normal distributions. Let $X \sim N(\mu, \sigma^2)$ denote the normal distribution with mean μ and standard deviation σ and let (1 - δ) and δ be the mixing probabilities, then the contaminated normal probability density function is given as follows:

$$f(x;\ \mu, \sigma) = (1 - \delta)\ N(\mu, \sigma^2) +\ \delta\ N(\mu, \lambda\sigma^2) \tag{18}$$

where λ > 1 is a parameter that determines the standard deviation of the wider component. The idea is that the "main" distribution $X \sim N(\mu, \sigma^2)$ is slightly "contaminated" by a wider distribution. The different levels of mixing probabilities δ = 0.1, 0.2 and 0.3 which represents 10%, 20% and 30% "contamination" respectively, are used for simulation study. Moreover we assign λ = $10.5^2, 15.5^2$ as the scale multipliers. In this section, the mean and standard deviation of a normal distribution are defined by μ = 1 and σ = 1 respectively, so that the uncontaminated component is normally distributed with $N(1, 1)$. The following six cases are constructed for the probability density function of the contaminated normal distribution as the linear combination of $N(1, 1)$ and $N(1, 10.5^2)$ densities shown in equations (19) to (21), and the probability density function (pdf) of a contaminated normal distribution as the linear combination of $N(1, 1)$ and $N(0, 15.5^2)$ densities shown in equations (22) to (24):

- **Case 1**: The contamination of 10% values in the normal distribution with mean 1 and $\sigma^2 = 10.5^2$ as $N(1, 10.5^2)$, while 90% values are generated by $N(1, 1)$. Hence it will retain 10% artificial outliers.

  $$CN(1, 10.5^2)\_10 = 0.9\ N(1,1) + 0.1\ N(1, 10.5^2) \tag{19}$$

  where CN stands for contamination.

- **Case 2**: The contamination of 20% values in the normal distribution with mean 1 and $\sigma^2 = 10.5^2$ as $N(1, 10.5^2)$, while 80% values are generated by $N(1, 1)$. Hence it will retain 20% artificial outliers.

  $$CN(1, 10.5^2)\_20 = 0.8\ N(1,1) + 0.2\ N(1, 10.5^2) \tag{20}$$

- **Case 3**: The contamination of 30% values in the normal distribution with mean 1 and $\sigma^2 = 10.5^2$ as $N(1, 10.5^2)$, while 70% values are generated by $N(1, 1)$. Hence it will retain 30% artificial outliers.

  $$CN(1, 10.5^2)\_30 = 0.7\ N(1,1) + 0.3\ N(1, 10.5^2) \tag{21}$$

- **Case 4**: The contamination of 10% values in the normal distribution with mean 1 and $\sigma^2 = 15.5^2$ as $N(1, 15.5^2)$, while 90% values are generated by $N(1, 1)$. Hence it will retain 10% artificial outliers.

  $$CN(1, 15.5^2)\_10 = 0.9\ N(1,1) + 0.1\ N(1, 15.5^2) \tag{22}$$

- **Case 5**: The contamination of 20% values in the normal distribution with mean 1 and $\sigma^2 = 15.5^2$ as $N(1, 15.5^2)$, while 80% values are generated by $N(1, 1)$. Hence it will retain 20% artificial outliers.

$$CN(1, 15.5^2)\_20 = 0.8\, N(1,1) + 0.2\, N(1, 15.5^2) \tag{23}$$

- **Case 6**: The contamination of 30% values in the normal distribution with mean 1 and $\sigma^2 = 15.5^2$ as $N(1, 15.5^2)$, while 70% values are generated by $N(1, 1)$. Hence it will retain 30% artificial outliers.

$$CN(1, 15.5^2)\_30 = 0.7\, N(1,1) + 0.3\, N(1, 15.5^2) \tag{24}$$

5. Results and Discussion

The simulation results for all cases under study are shown in Tables 1-5 and Figures 1-3. The performances of 95% confidence intervals of the population mean using nine methods are evaluated in case of normally distributed data as shown in Tables 2-3 and Figure 1. It is observed that the coverage probabilities of six confidence intervals i.e. Student-t, Johnson-t, Chen-t , YY-t, Median-t and AADM-t confidence intervals, are close to the specified confidence coefficient level for all sample sizes and Cheby-t confidence interval tends to have the coverage probability beyond 0.95, whereas the Mad-t and DMSD_{DM}-t confidence intervals under 0.95. When considering the performance of nine confidence intervals in term of the average width, it is found that almost all the studied confidence intervals tend to have no difference of the average width, except Cheby-t confidence interval that has the highest average width for all sample sizes. However, the average width of Cheby-t confidence interval tends to decrease when the sample size enlarges. Moreover it is also seen that proposed DMSD_{DM}-t confidence interval maintained reasonable and consistent coverage probability at lower average width amongst all other intervals at each sample size.

In the second case, the performance of nine confidence intervals for contaminated normal distributions is evaluated and the results are shown in Tables 4-5 and Figures 2-3. It is found that the proposed DMSD_{DM}-t confidence interval is most efficient in terms of coverage probability and average width for these contaminated distributions. That is, it tends to provide the smallest average width and coverage probability around the specified level whatever the sample size will be, especially in the situations of 10% and 20% outliers as shown in Figures 2-3. Although the coverage probability of Cheby-t confidence interval is the highest among all distributions under study, but it provides largest average width for all sample sizes and hence cannot be considered most precise confidence interval.

Table 1: Coverage probabilities and average widths of the 95% CIs for mean with $N(20, 5)$

n	Performance Measures	Confidence Interval Methods								
		Student-t	Cheby-t	Johnson-t	Chen-t	YY-t	Median-t	Mad-t	AADM-t	$DMSD_{DM}$-t
5	CP	0.9506	0.9999	0.9506	0.9477	0.9505	0.9539	0.8855	0.9438	0.9260
	AW	7.0	19.2	7.0	7.1	7.0	7.1	5.4	6.8	6.2
10	CP	0.9496	1.0000	0.9495	0.9482	0.9495	0.9520	0.8826	0.9459	0.9174
	AW	4.6	13.8	4.6	4.6	4.6	4.7	3.6	4.6	4.1
20	CP	0.9494	1.0000	0.9495	0.9489	0.9495	0.9511	0.8819	0.9464	0.9142
	AW	3.7	11.3	3.7	3.7	3.7	3.7	2.9	3.7	3.3
30	CP	0.9499	1.0000	0.9499	0.9493	0.9499	0.9513	0.8820	0.9481	0.9138
	AW	3.2	9.8	3.2	3.2	3.2	3.2	2.5	3.2	2.8
50	CP	0.9501	1.0000	0.9502	0.9499	0.9501	0.9513	0.8818	0.9483	0.9129
	AW	2.8	8.8	2.8	2.8	2.8	2.8	2.2	2.8	2.5
100	CP	0.9501	1.0000	0.9501	0.9501	0.9501	0.9510	0.8818	0.9490	0.9119
	AW	2.0	6.2	2.0	2.0	2.0	2.0	1.6	2.0	1.8

Table 2: Coverage probabilities and average widths of the 95% CIs for mean with $N(20, 10)$

n	Performance Measures	Confidence Interval Methods								
		Student-t	Cheby-t	Johnson-t	Chen-t	YY-t	Median-t	Mad-t	AADM-t	$DMSD_{DM}$-t
5	CP	0.9506	0.9999	0.9506	0.9477	0.9505	0.9539	0.8855	0.9437	0.9260
	AW	13.9	38.4	13.9	14.1	13.9	14.2	10.8	13.6	12.5
10	CP	0.9496	1.0000	0.9495	0.9482	0.9495	0.9519	0.8826	0.9458	0.9173
	AW	9.2	27.6	9.2	9.3	9.2	9.3	7.3	9.1	8.2
20	CP	0.9494	1.0000	0.9494	0.9489	0.9495	0.9511	0.8820	0.9464	0.9142
	AW	7.4	22.6	7.4	7.4	7.4	7.5	5.9	7.3	6.6
30	CP	0.9499	1.0000	0.9500	0.9493	0.9499	0.9513	0.8821	0.9481	0.9138
	AW	6.4	19.6	6.4	6.4	6.4	6.4	5.0	6.3	5.6
50	CP	0.9501	1.0000	0.9502	0.9499	0.9501	0.9512	0.8819	0.9483	0.9129
	AW	5.7	17.6	5.7	5.7	5.7	5.7	4.5	5.6	5.0
100	CP	0.9501	1.0000	0.9501	0.9500	0.9501	0.9509	0.8818	0.9490	0.9119
	AW	4.0	12.5	4.0	4.0	4.0	4.0	3.1	3.9	3.5

Table 3: Coverage probabilities and average widths of the 95% CIs for mean with $N(20, 20)$

n	Performance Measures	Confidence Interval Methods								
		Student-t	Cheby-t	Johnson-t	Chen-t	YY-t	Median-t	Mad-t	AADM-t	$DMSD_{DM}$-t
5	CP	0.9506	0.9999	0.9506	0.9477	0.9506	0.9539	0.8856	0.9437	0.9260
	AW	27.8	76.8	27.8	28.3	27.8	28.4	21.7	27.1	25.0
10	CP	0.9496	1.0000	0.9495	0.9482	0.9495	0.9520	0.8826	0.9458	0.9173
	AW	18.5	55.1	18.5	18.6	18.5	18.7	14.6	18.2	16.4
20	CP	0.9494	1.0000	0.9494	0.9489	0.9495	0.9511	0.8819	0.9464	0.9142
	AW	14.8	45.2	14.8	14.9	14.8	14.9	11.7	14.7	13.1
30	CP	0.9499	1.0000	0.9500	0.9493	0.9499	0.9513	0.8821	0.9481	0.9138
	AW	12.7	39.2	12.7	12.7	12.7	12.8	10.1	12.6	11.3
50	CP	0.9501	1.0000	0.9502	0.9499	0.9501	0.9512	0.8818	0.9483	0.9129
	AW	11.3	35.1	11.3	11.3	11.3	11.4	9.0	11.2	10.0
100	CP	0.9501	1.0000	0.9502	0.9500	0.9501	0.9510	0.8818	0.9490	0.9119
	AW	7.9	24.9	7.9	7.9	7.9	7.9	6.3	7.9	7.0

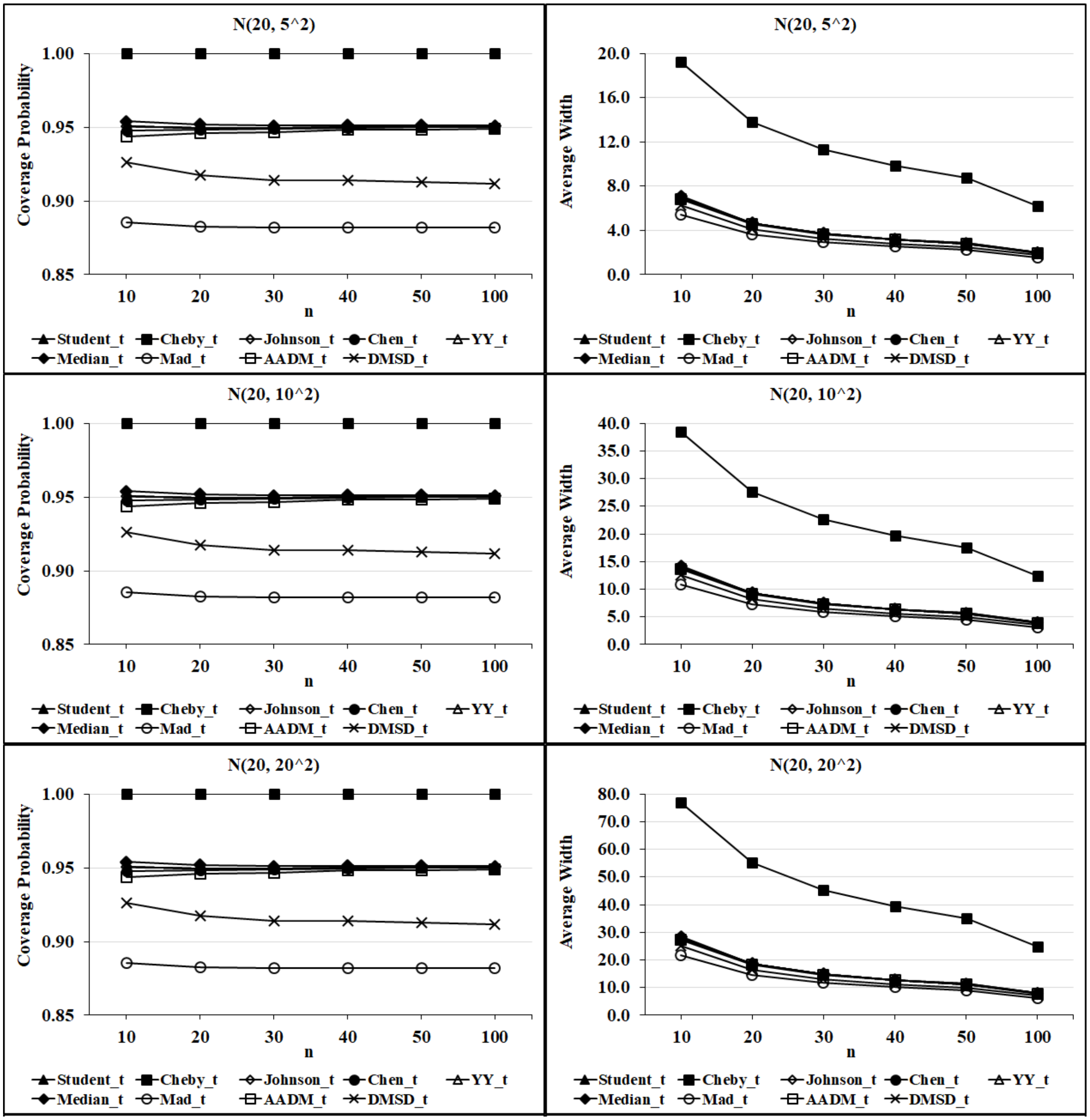

Figure 1: Coverage probabilities and average widths of the 95% CIs for mean of Normal distributed data

Table 4: Coverage probabilities and average widths of the 95% CIs for mean of contaminated normal distribution as the linear combination of $N(1, 1)$ and $N(1, 10.5^2)$

n	Performance Measures	Confidence Interval Methods								
		Student-t	Cheby-t	Johnson-t	Chen-t	YY-t	Median-t	Mad-t	AADM-t	DMSD$_{DM}$-t
		Contaminated Normal Distribution $CN(1, 10.5^2)$_10								
5	CP	0.9834	1.0000	0.9821	0.9735	0.9829	0.9848	0.8816	0.9652	0.9734
	AW	4.2	11.5	4.2	5.0	4.2	4.3	2.5	3.2	3.2
10	CP	0.9859	1.0000	0.9826	0.9563	0.9841	0.9879	0.6865	0.8748	0.9492
	AW	2.9	8.8	2.9	3.5	2.9	3.0	1.6	2.0	1.5
20	CP	0.9802	1.0000	0.9729	0.9398	0.9761	0.9839	0.6497	0.7974	0.9360
	AW	2.4	7.4	2.4	2.7	2.4	2.4	1.2	1.5	1.0
30	CP	0.9738	1.0000	0.9643	0.9373	0.9678	0.9788	0.6431	0.7768	0.9318
	AW	2.1	6.5	2.1	2.3	2.1	2.1	1.0	1.3	0.8
50	CP	0.9674	1.0000	0.9577	0.9356	0.9608	0.9724	0.6357	0.7647	0.9302
	AW	1.9	5.9	1.9	2.0	1.9	1.9	0.9	1.1	0.7
100	CP	0.9553	1.0000	0.9482	0.9443	0.9503	0.9587	0.6289	0.7457	0.9247
	AW	1.3	4.2	1.3	1.4	1.3	1.3	0.6	0.8	0.5
		Contaminated Normal Distribution $CN(1, 10.5^2)$_20								
5	CP	0.9921	1.0000	0.9906	0.9747	0.9914	0.9936	0.8922	0.9790	0.9865
	AW	6.1	16.9	6.1	7.0	6.1	6.4	3.8	4.7	4.9
10	CP	0.9785	1.0000	0.9710	0.9380	0.9742	0.9877	0.7321	0.8866	0.9608
	AW	4.2	12.6	4.2	4.6	4.2	4.3	2.3	2.9	2.7
20	CP	0.9653	1.0000	0.9561	0.9396	0.9598	0.9754	0.7010	0.8389	0.9457
	AW	3.4	10.5	3.4	3.7	3.4	3.5	1.8	2.3	2.1
30	CP	0.9582	1.0000	0.9507	0.9424	0.9533	0.9667	0.6912	0.8195	0.9343
	AW	3.0	9.2	3.0	3.1	3.0	3.0	1.6	1.9	1.7
50	CP	0.9563	1.0000	0.9499	0.9460	0.9521	0.9632	0.6827	0.8071	0.9292
	AW	2.7	8.2	2.7	2.7	2.7	2.7	1.4	1.7	1.5
100	CP	0.9510	1.0000	0.9469	0.9492	0.9480	0.9543	0.6674	0.7834	0.9306
	AW	1.9	5.9	1.9	1.9	1.9	1.9	0.9	1.2	1.0
		Contaminated Normal Distribution $CN(1, 10.5^2)$_30								
5	CP	0.9921	1.0000	0.9899	0.9600	0.9912	0.9954	0.8825	0.9843	0.9865
	AW	7.7	21.2	7.7	8.5	7.7	8.0	4.9	6.2	6.3
10	CP	0.9640	1.0000	0.9562	0.9418	0.9597	0.9796	0.7805	0.9097	0.9492
	AW	5.2	15.6	5.2	5.5	5.2	5.3	3.1	3.9	3.8
20	CP	0.9569	1.0000	0.9507	0.9462	0.9530	0.9679	0.7528	0.8782	0.9328
	AW	4.2	12.9	4.2	4.4	4.2	4.3	2.4	3.0	3.0
30	CP	0.9540	1.0000	0.9487	0.9466	0.9505	0.9625	0.7352	0.8569	0.9204
	AW	3.6	11.2	3.6	3.7	3.6	3.7	2.1	2.6	2.5
50	CP	0.9537	1.0000	0.9496	0.9505	0.9508	0.9600	0.7331	0.8516	0.9135
	AW	3.2	10.1	3.2	3.3	3.2	3.3	1.8	2.3	2.2
100	CP	0.9509	1.0000	0.9485	0.9502	0.9492	0.9542	0.7166	0.8284	0.8777
	AW	2.3	7.2	2.3	2.3	2.3	2.3	1.2	1.6	1.5

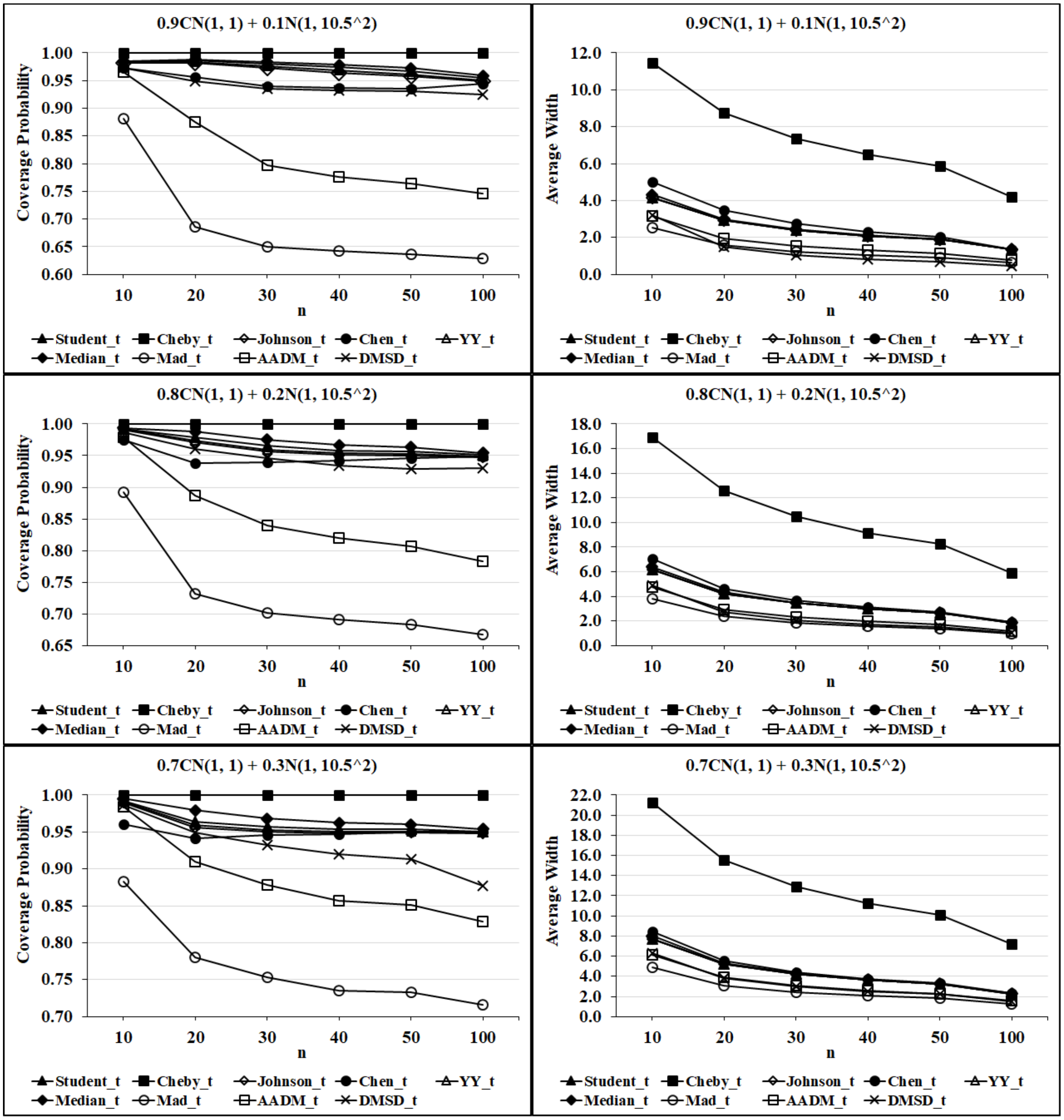

Figure 2: Coverage probabilities and average widths of the 95% CIs for mean of contaminated normal distributed data as the linear combination of $N(1, 1)$ and $N(1, 10.5^2)$

Table 5: Coverage probabilities and average widths of the 95% CIs for mean of contaminated normal distribution as the linear combination of $N(1, 1)$ and $N(1, 15.5^2)$

n	Performance Measures	Confidence Interval Methods								
		Student-t	Cheby-t	Johnson-t	Chen-t	YY-t	Median-t	Mad-t	AADM-t	DMSD$_{DM}$-t
		Contaminated Normal Distribution $CN(1, 15.5^2)$_10								
5	CP	0.9834	1.0000	0.9821	0.9735	0.9829	0.9848	0.8816	0.9652	0.9734
	AW	4.2	11.5	4.2	5.0	4.2	4.3	2.5	3.2	3.2
10	CP	0.9859	1.0000	0.9826	0.9563	0.9841	0.9879	0.6865	0.8748	0.9492
	AW	2.9	8.8	2.9	3.5	2.9	3.0	1.6	2.0	1.5
20	CP	0.9802	1.0000	0.9729	0.9398	0.9761	0.9839	0.6497	0.7974	0.9360
	AW	2.4	7.4	2.4	2.7	2.4	2.4	1.2	1.5	1.0
30	CP	0.9738	1.0000	0.9643	0.9373	0.9678	0.9788	0.6431	0.7768	0.9318
	AW	2.1	6.5	2.1	2.3	2.1	2.1	1.0	1.3	0.8
50	CP	0.9674	1.0000	0.9577	0.9356	0.9608	0.9724	0.6357	0.7647	0.9302
	AW	1.9	5.9	1.9	2.0	1.9	1.9	0.9	1.1	0.7
100	CP	0.9553	1.0000	0.9482	0.9443	0.9503	0.9587	0.6289	0.7457	0.9247
	AW	1.3	4.2	1.3	1.4	1.3	1.3	0.6	0.8	0.5
		Contaminated Normal Distribution $CN(1, 15.5^2)$_20								
5	CP	0.9921	1.0000	0.9906	0.9747	0.9914	0.9936	0.8922	0.9790	0.9865
	AW	6.1	16.9	6.1	7.0	6.1	6.4	3.8	4.7	4.9
10	CP	0.9785	1.0000	0.9710	0.9380	0.9742	0.9877	0.7321	0.8866	0.9608
	AW	4.2	12.6	4.2	4.6	4.2	4.3	2.3	2.9	2.7
20	CP	0.9653	1.0000	0.9561	0.9396	0.9598	0.9754	0.7010	0.8389	0.9457
	AW	3.4	10.5	3.4	3.7	3.4	3.5	1.8	2.3	2.1
30	CP	0.9582	1.0000	0.9507	0.9424	0.9533	0.9667	0.6912	0.8195	0.9343
	AW	3.0	9.2	3.0	3.1	3.0	3.0	1.6	1.9	1.7
50	CP	0.9563	1.0000	0.9499	0.9460	0.9521	0.9632	0.6827	0.8071	0.9292
	AW	2.7	8.2	2.7	2.7	2.7	2.7	1.4	1.7	1.5
100	CP	0.9510	1.0000	0.9469	0.9492	0.9480	0.9543	0.6674	0.7834	0.9306
	AW	1.9	5.9	1.9	1.9	1.9	1.9	0.9	1.2	1.0
		Contaminated Normal Distribution $CN(1, 15^2)$_30								
5	CP	0.9921	1.0000	0.9899	0.9600	0.9912	0.9954	0.8825	0.9843	0.9865
	AW	7.7	21.2	7.7	8.5	7.7	8.0	4.9	6.2	6.3
10	CP	0.9640	1.0000	0.9562	0.9418	0.9597	0.9796	0.7805	0.9097	0.9492
	AW	5.2	15.6	5.2	5.5	5.2	5.3	3.1	3.9	3.8
20	CP	0.9569	1.0000	0.9507	0.9462	0.9530	0.9679	0.7528	0.8782	0.9328
	AW	4.2	12.9	4.2	4.4	4.2	4.3	2.4	3.0	3.0
30	CP	0.9540	1.0000	0.9487	0.9466	0.9505	0.9625	0.7352	0.8569	0.9204
	AW	3.6	11.2	3.6	3.7	3.6	3.7	2.1	2.6	2.5
50	CP	0.9537	1.0000	0.9496	0.9505	0.9508	0.9600	0.7331	0.8516	0.9135
	AW	3.2	10.1	3.2	3.3	3.2	3.3	1.8	2.3	2.2
100	CP	0.9509	1.0000	0.9485	0.9502	0.9492	0.9542	0.7166	0.8284	0.8777
	AW	2.3	7.2	2.3	2.3	2.3	2.3	1.2	1.6	1.5

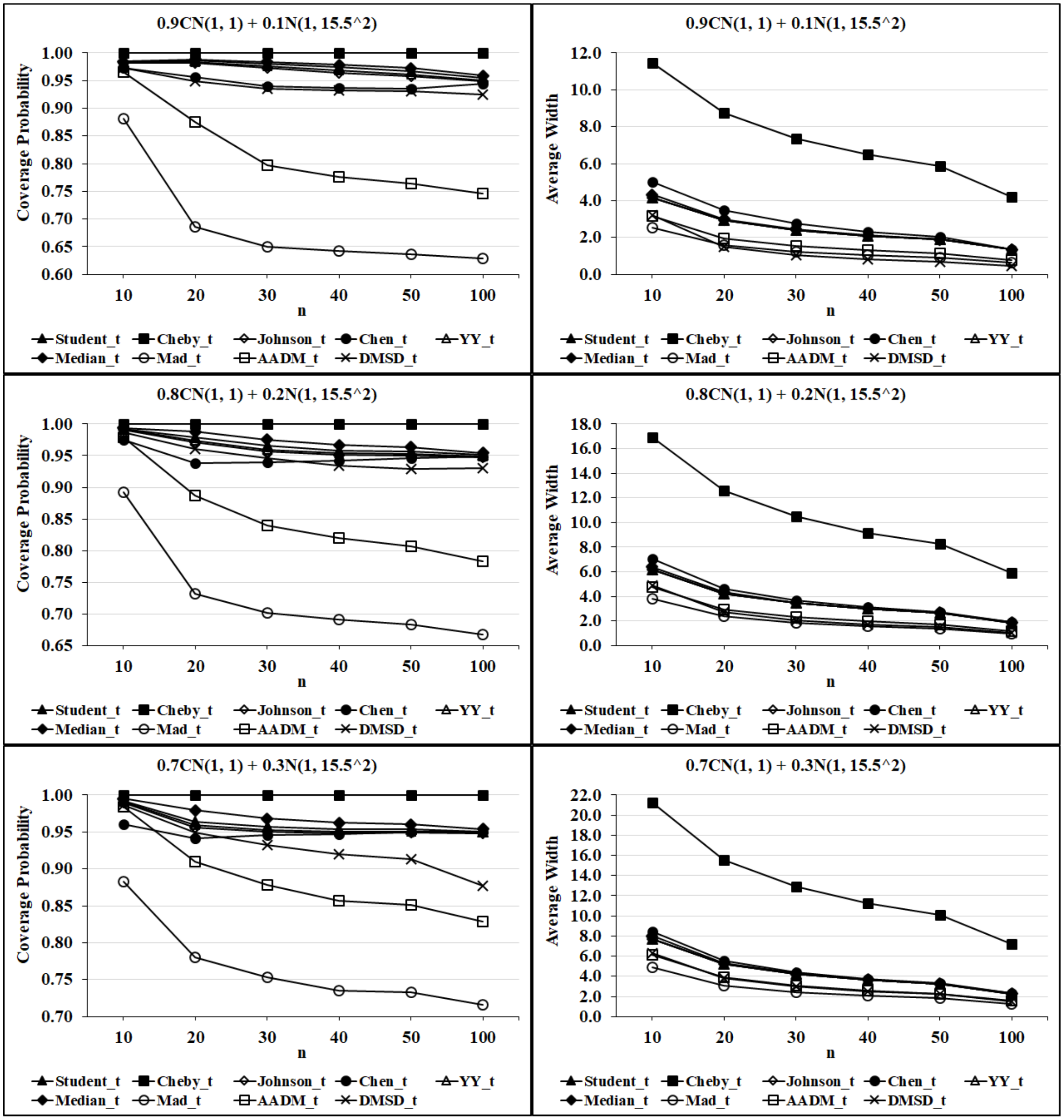

Figure 3: Coverage probabilities and average widths of the 95% CIs for mean of contaminated normal distributed data as the linear combination of $N(1, 1)$ and $N(1, 15.5^2)$

6. Applications Using Real Data

In this section, three real-life examples from normal and non-normal contaminated distributions are analyzed to illustrate the applications of the proposed $DMSD_{DM}$-t robust confidence interval.

6.1 Example 1: Data of Beeswax Melting Points

The first data set represent about melting points (°C) of beeswax are obtained from 59 sources which is considered from White et al. (1960). These data are illustrated in Table 6.

Table 6: The 59 sources of beeswax melting points (°C)

No.	X	No.	X	No.	X	No.	X
1	63.78	16	63.92	31	64.42	46	64.12
2	63.83	17	63.86	32	63.50	47	63.03
3	63.88	18	63.13	33	63.84	48	63.66
4	63.78	19	63.08	34	64.21	49	63.34
5	63.50	20	63.30	35	64.40	50	63.34
6	63.41	21	63.51	36	62.85	51	63.56
7	63.45	22	63.56	37	63.27	52	63.92
8	63.63	23	63.93	38	63.36	53	63.68
9	63.36	24	63.69	39	64.27	54	63.60
10	63.92	25	63.40	40	64.24	55	63.50
11	63.30	26	63.83	41	63.61	56	63.92
12	63.60	27	63.51	42	63.31	57	63.39
13	63.58	28	63.43	43	63.10	58	63.53
14	63.27	29	63.43	44	63.86	59	63.13
15	63.36	30	63.05	45	63.50		

According to Panichkitkosolkul (2015), it is known that the population mean of beeswax melting points (μ) is about 63.580 °C. To investigate the normal distribution of these sample data, Shapiro-Wilk test statistics is considered. It is found that the Shapiro-Wilk test statistic has a p-value > 0.05. Hence, it can be concluded that the beeswax melting points are normally distributed at significance level $\alpha = 5\%$ and these data are not contaminated with outliers by considering from boxplot in Figure 4.

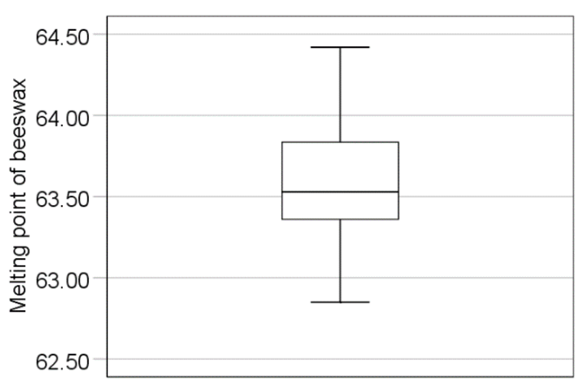

Figure 4: Boxplot of beeswax melting points data

After the investigation of beeswax melting points that conforms to normal distributed data assumption, the nine estimation methods of the 95% confidence intervals for the population mean of beeswax melting points are constructed and results are shown in Table 7.

Table 7: The 95% confidence intervals for the population mean of beeswax melting points data

Methods	Confidence Interval Limits		Widths
	Lower Limit	Upper Limit	
Student-t	63.4983	63.6793	0.1810
Cheby-t	63.3065	63.8711	0.5646
Johnson-t	63.4987	63.6797	0.1810
Chen-t	63.4947	63.6830	0.1883
YY-t	63.4986	63.6796	0.1810
Median-t	63.4970	63.6806	0.1836
Mad-t	63.5173	63.6603	0.1430
AADM-t	63.4992	63.6784	0.1792
$\mathbf{DMSD_{DM}}$-t	63.4885	63.6462	0.1577

From Table 7, the 95% confidence interval for the population mean of beeswax melting points data which is constructed by using Cheby-t method give the largest width, whereas the 95% of Mad-t and $DMSD_{DM}$-t confidence intervals give almost similar and smallest widths amongst all other intervals. As width of Mad-t interval is smaller than $DMSD_{DM}$-t but on the basis of simulation results discussed in section 5, that Mad-t interval can never approach reasonable coverage probability near to nominal value, we recommend $DMSD_{DM}$-t confidence interval. Hence, the results from this real-life example shown on Table 7 support the simulation study.

6.2 Example 2: Urinary Tract Infection (UTI) Data

The second sample data represent about the duration of male patient urinary tract infections (UTIs) in days and these show in Table 8. It was considered by many researchers, e.g., Santiago et al. (2013), Aslam et al. (2014), Azam et al. (2017) and Sinsomboonthong et al. (2020).

Table 8: Urinary tract infection data

No.	X	No.	X	No.	X
1	0.57014	19	0.12014	37	0.27083
2	0.07431	20	0.11458	38	0.04514
3	0.15278	21	0.00347	39	0.13542
4	0.14583	22	0.12014	40	0.08681
5	0.13889	23	0.04861	41	0.40347
6	0.14931	24	0.02778	42	0.12639
7	0.03333	25	0.32639	43	0.18403
8	0.08681	26	0.64931	44	0.70833
9	0.33681	27	0.14931	45	0.15625
10	0.03819	28	0.01389	46	0.24653
11	0.24653	29	0.03819	47	0.04514
12	0.29514	30	0.46806	48	0.01736
13	0.11944	31	0.22222	49	1.08889
14	0.05208	32	0.29514	50	0.05208
15	0.12500	33	0.53472	51	0.02778
16	0.25000	34	0.15139	52	0.03472
17	0.40069	35	0.52569	53	0.23611
18	0.02500	36	0.07986	54	0.35972

The normality test of urinary tract infection data using Shapiro-Wilk test statistics is considered, it is found this test statistic has a p-value < 0.05. Hence, it can be concluded that the urinary tract infection data are non-normally distributed at significance level $\alpha = 5\%$ and these data are contaminated with outliers by considering from boxplot in Figure 5. These results support the study of Santiago and Smith (2013) that the data are well fitted to an exponential distribution with a mean time of $\mu = 0.2100$ days.

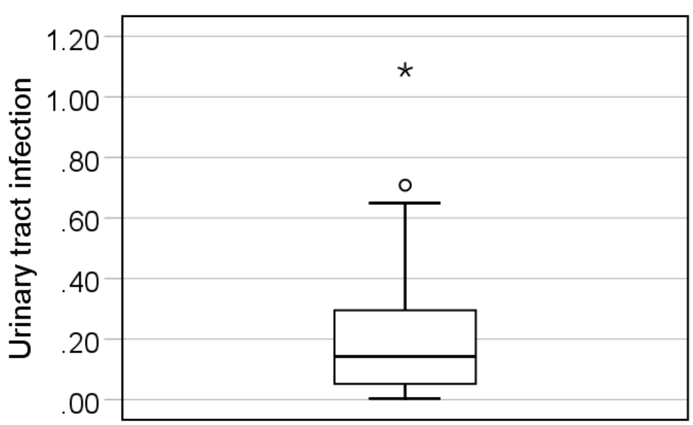

Figure 5: Boxplot of urinary tract infection data

After the investigation of urinary tract infection data that conforms to non-normal distributed data assumption, the nine estimation methods of the 95% confidence intervals for the population mean of urinary tract infection data are constructed and results are shown in Table 9.

Table 9: The 95% confidence intervals for the population mean of urinary tract infection data

Methods	**Confidence Interval Limits**		**Widths**
	Lower Limit	**Upper Limit**	
Student-t	0.1524	0.2681	0.1157
Cheby-t	0.0302	0.3904	0.3602
Johnson-t	0.1536	0.2693	0.1157
Chen-t	0.1396	0.2809	0.1413
YY-t	0.1532	0.2689	0.1157
Median-t	0.1495	0.2711	0.1216
Mad-t	0.1675	0.2530	0.0855
AADM-t	0.1567	0.2638	0.1071
$\mathbf{DMSD_{DM}}$-t	0.1423	0.2261	0.0838

From Table 9, the 95% confidence interval for the population mean of duration of male patient urinary tract infections in days which is constructed by using Cheby-t method give the largest width, whereas Mad-t and $DMSD_{DM}$-t confidence intervals give the smallest width amongst all. As a result of outliers in the data set, the simulation study was found that coverage probability of $DMSD_{DM}$-t method close to the nominal level than this of Mad-t confidence interval. Hence, the $DMSD_{DM}$-t method is suggested for population mean estimation under contaminated distribution. Further, the results support the findings of simulation study for the case of contaminated distribution.

6.3 Example 3: Psychotropic Drug Exposure Data

To study the average use of psychotropic drugs from non-antipsychotic drug users, the number of users of psychotropic drugs was reported for a random sample of twenty (n = 20) different categories of drugs. The data represent the number of users (Johnson and McFarland, 1993) was shown in Table 10

:

Table 10: Psychotropic Drug Exposure data

No.	X	No.	X
1	43.4	11	35.7
2	24	12	27.3
3	1.8	13	5
4	0	14	64.3
5	0.1	15	70
6	170.1	16	94
7	0.4	17	61.9
8	150	18	9.1
9	31.5	19	38.8
10	5.2	20	14.8

By using the Shapiro-Wilk test statistics to examine the normal distribution of these data, it is found this test statistic has a p-value < 0.05. Hence, it can be concluded that the psychotropic drugs are non-normally distributed at significance level $\alpha = 5\%$ and these data are contaminated with outliers by considering from boxplot in Figure 6.

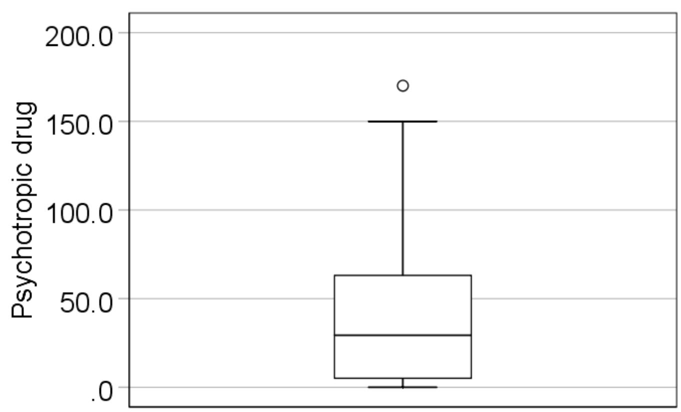

Figure 6: Boxplot of psychotropic drug exposure data

Table 11 represents the 95% confidence intervals for the average use of psychotropic drugs. It is shown that Cheby-t method gives the largest width and differs from other methods, whereas $DMSD_{DM}$-t confidence interval give the shortest width followed by Mad-t and AADM-t confidence intervals, therefore we conclude that the results from Table 11 support the simulation results of this study in the case of contaminated distribution.

Table 11: The 95% confidence intervals for the average use of psychotropic drug

Methods	Confidence Interval Limits		Widths
	Lower Limit	Upper Limit	
Student-t	19.8445	64.8955	45.0509
Cheby-t	-25.2605	110.0005	135.2611
Johnson-t	20.3850	65.4359	45.0509
Chen-t	13.4694	71.2706	57.8013
YY-t	20.1629	65.2139	45.0509
Median-t	19.0097	65.7303	46.7205
Mad-t	25.7607	58.9793	33.2185
AADM-t	22.7839	61.9561	39.1722
$DMSD_{DM}$-t	19.3406	50.2838	30.9432

7. Concluding Remarks

The present research addresses and reviews several interval estimators for estimating the mean of contaminated distributions. The robust confidence interval $DMSD_{DM}$-t which is an adjustment of the student-t confidence interval is proposed based on the decile mean and the decile standard deviation. A simulation study has been conducted to compare the performance of the estimators in term of coverage probability and average width. Simulation results evident that the proposed confidence interval performed better than the existing confidence intervals in terms of attaining nominal coverage probability and shorter average width especially when observations are sampled from a contaminated normal distribution. Even though, the Mad-t confidence interval tends to provide the smallest average width in case of observations are sampled from normal distributions, but the coverage probability of this interval is much lower than the nominal value when it compares with the proposed confidence interval. Hence, the performance of $DMSD_{DM}$-t method is better than Mad-t method in terms of balancing both the high coverage probability near to nominal value and low average width. The findings of this research will hopefully be fruitful for the researchers in various fileds of social, medical and physical sciences.

Data Availability Statement

The data that supports the findings of this study are provided in the application section of the manuscript.

References

Abu-Shawiesh, M. O. A., Banik, B. and Kibria, B. M. G. (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.

Abu-Shawiesh, M. O. A., Akyüz, H. E., & Kibria, B. M. G. (2019). Performance of some confidence intervals for estimating the population coefficient of variation under both symmetric and skewed distributions. Statistics, Optimization and Information Computing: An International Journal, 7(2), 277-290.

Andersson, P. G. (2004) Alternative confidence intervals for the total of a skewed biological population, *Ecology* 85, 3166–3171.

Aslam, M., N. Khan, M. Azam and C. H. Jun. (2014). Designing of a new monitoring t-chart using repetitive sampling. *Inf. Sci*, 269, 210-216. https://doi.org/10.1016/j.ins.2014.01.022

Azam, M., M. Aslam and C. H. Jun. (2017). An EWMA control chart for the exponential distribution using repetitive sampling plan. *Operations Research and Decisions*, 27(2), 5-19. doi: 10.5277/ord170201

Baklizi, A and Kibria, B. M. G. (2009). One and two sample confidence intervals for estimating the mean of skewed populations: An empirical comparative study, *J. Appl. Stat*. 36 (2009), 601–609.

Banik, B. and Kibria, B. M. G. (2010). Comparison of some parametric and nonparametric type one sample confidence intervals for estimating the mean of a positively skewed distribution, *Commun. Stat. Simul. Comput*. 39, 361–389.

Bissell, D. (1994). Statistical Methods for SPC and TQM. 1st Edition. Chapman & Hall: New York.

Chen, L. (1995). Testing the mean of skewed distributions, *J. Am. Statist. Assoc*. 90, 767–772.

Desharnais, B., Camirand-Lemyre, F, Mireault, P and Skinner, C. (2015). Determination of Confidence Intervals in Non-normal Data: Application of the Bootstrap to Cocaine Concentration in Femoral Blood, *Journal of Analytical Toxicology*, 39: 113 –117.

Doulah, M. S. U. (2018). Alternative measures of standard deviation coefficient of variation and standard error, International Journal of Statistics and Applications, 8(6): 309–315.

Gastwirth, J. L. (1982). Statistical Properties of a Measure of Tax Assessment Uniformity. Journal of the Statistical Inference and Planning, 6, 1-12.

Ghosh, S. and Polansky, A. M. (2016). New bootstrap confidence interval for means of a positively skewed distributions. *Communication sin Statistics-Thoery and Methods*. 45(23), 6915-6927.

Gregoire, T. G. and O. Schabenberger (1999). Sampling skewed biological populations: Behavior of confidence intervals for the population total, *Ecology* 80, 1056–1065.

Hall, P. (1992). On the removal of skewness by transformation, *J. R. Stat. Soc*. B 54, 221–228.

Hogg, R. V., Craig, A. T. (1978). *Introduction to Mathematical Statistics*. New York: Macmillan Publishing Company.

Hrudey S. E. and Hrudey E. J. (2004). Safe Drinking Water – Lessons from Recent Outbreaks in Affluent Nations. IWA Publishing, London, UK.

International Olympic Committee. (2019). London 2012 long jump men – Olympic athletics. Retrieved from https://www.olympic.org/london2012/athletics/long-jump-men.

James, P. C. (1989). CPK equivalencies, *Quality,* 28:75.

Johnson, N. J. (1978). Modified t-tests and confidence intervals for asymmetrical population, J. Am. Stat. Assoc. 73, 536–544.

Johnson, R. E., and McFarland, B. H. (1993). Antipsychotic drug exposure in a health maintenance organization. Medical Care, 31(5), 432-444.

Li, H.Q, Tang, M. L, Poon, W.Y. and Nian-Sheng Tang (2011): Confidence Intervals for Difference Between Two Poisson Rates, Communications in Statistics - Simulation and Computation, 40:9, 1478-1493

Miller, I. and Miller, M. (1995). Statistical Methods for Quality with Applications to Engineering and Management. Prentice-Hall, Inc.: Englewood Cliffs, New Jersey.

Montgomery, D. C. and Runger, G. C. (2003). Applied Statistics and Probability for Engineers. 3rd Edition, John Wiley & Son, Inc., Hoboken.

Mudelsee, M. and Alkio, M. (2007). Quantifying effects in two-sample environmental experiments using bootstrap confidence intervals, *Environ. Model. Softw*. 22, 84–96.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society Series A, 236, 333-380.

Panichkitkosolkul, W. (2015). Confidence interval for the coefficient of variation in a normal distribution with a known population mean after a preliminary t test. KMITL Sci. Tech. J., 15(1), 34-46.

Rana, S., Doulah, M. S. U., Midi, H., and Imon, A. H. M. R. (2012). Decil Mean: A New Robust Measure of Central Tendency, *Chiang Mai Journal of Sciences,* 39(3): 478-485.

Rostron, P.D., Fearn, T. & Ramsey, M.H. Confidence intervals for robust estimates of measurement uncertainty. *Accreditation and Quality Assurance*, **25,** 107–119 (2020). https://doi.org/10.1007/s00769-019-01417-4

Santiago, E. and J. Smith. (2013). Control charts based on the exponential distribution: adapting runs rules for the t chart. *Qual. Eng.*, 25(2), 85-96. https://doi.org/10.1080/08982112.2012.740646

Sinsomboonthong, J, Abu-Shawiesh, M. O. A. and 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 Journal*, 5: (3), 442-449.

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1-25.

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

White, J., M. Riethof and I. Kushnir. (1960). Estimation of microcrystalline wax in beeswax. *J. Assoc. Off. Anal. Chem.*, 43(4), 781-790. https://doi.org/10.1093/jaoac/43.4.781

Yanagihara, H. and Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British Journal of Mathematical and Statistical Psychology*, 58 (2), 209−237.

Zhou, H. S., C. Li, S. Gao, and W.M. Tierney. (2001). Methods for testing equality of means of health care costs in a paired design study, *Stat. Med*. 20, 1703–1720.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixd1.docx