INTERVAL ESTIMATION IN THE PRESENCE OF AN OUTLIER

WONG YOKE CHEN
School of Business
The University of Nottingham Malaysia Campus
Email: YokeChen.Wong@nottingham.edu.my

POOI AH HIN
School of Business
Sunway University
Email: ahhinp@sunway.edu.my

ABSTRACT

Outliers are often ubiquitous in surveys that involve linear measurements. Knowing that the presence of such extreme points can grossly distort statistical analyses, most researchers are often tempted to conveniently eliminate them from the data set without much careful consideration. In this study, we investigate the performance of confidence intervals for the population mean under the various probabilities of outlier being caused by uncorrectable human errors. The sample under study is randomly generated and subscribed to a normal distribution, and it contains only one outlier at one of the two extreme ends. For the generated sample, we compute three types of nominally $100(1-\alpha)\%$ confidence interval for the population mean, namely, I_E (when the single outlier is expunged from the sample), I_R (the outlier is replaced) and I_U (a union of I_E and I_R). It is found that when the sample size is smaller, I_U has a satisfactory level of coverage probability for all values of *p*. However, for larger sample sizes, I_R and I_E would instead be the better ones as they have shorter expected lengths, in addition to having reasonable levels of coverage probabilities for a wide range of *p*.

Key words: Outlier, Confidence interval, Coverage probability

INTRODUCTION

In surveys involving linear measurements, one or more data points may be found to be far from the rest of the observed values in the set. These points are conveniently classified as outliers, and in most cases are simply removed from the data set without careful consideration. An outlier has been described with various phrases. Grubbs (1969) defines an outlier as an observation that “appears to deviate markedly from other members of the sample in which it occurs”. Moore and McCabe (1999) describe an outlier as an observation that lies outside the overall pattern of a distribution.

Most statistics books identify outliers as those observed values that are at least 1.5 times greater than the upper quartile or 1.5 less than the lower quartile of the inter-quartile range. Graphically, the commonly used techniques for detecting outliers are the normality plot, histogram, scatterplot and the box plot. The normality test uses the three-sigma rule to identify outliers. Another test is the Grubbs' Test (1969) which employs an analytic procedure for detecting an outlier, also under the assumption of normal distribution.

The presence of outliers is never to be underestimated. It can grossly distort statistical analyses. For instance, calculations of the mean and standard deviation can be massively distorted by a single extremely small or large data point. Outliers generally serve to increase error variance and cause a decrease in accuracy of the estimators. Failure to deal with outliers appropriately may run the risk of bias in estimating models.

As a result, many researchers would simply eliminate any outliers detected. A simple act of elimination of an extreme data point may well result in an accidental deletion of some interesting and unforeseen change of norm. The problem can become even more complex when there is more than one outlier or one variable in the analysis. After having taken steps to identify outliers, an experienced statistician would carefully review each outlier and consider cautiously its appropriateness for inclusion or exclusion in the data analysis.

Outliers can arise from several different mechanisms or causes. Human carelessness is one of the biggest contributors to the existence of outliers. Errors may occur in data collection, recording, or entry. Such errors can often be corrected by re-checking. However, if such human errors cannot be corrected at all, it would be best to just eliminate them from the data set.

Outliers may also be caused by an error in sampling whereby several members of a sample were inadvertently drawn from a different population instead of the target population. There is also the likelihood that the outliers are due to an intentional misreporting by the survey participant due to his unwillingness to reveal some truth. In both cases, when we are sure of these being the causes of the outliers, removal of the outliers would also be the most natural thing to do.

Both the prevailing physical conditions under which the research was carried out and the poor quality of the measuring equipment can contribute to a deficient measurement process. This source of exceptionally large measurement errors accounts for another common cause of outliers. Outliers could also be attributed to natural deviations from the population. Based on the 3-sigma rule, there is a 0.26% random chance that an outlier legitimately occurs in a normally distributed population. This means, the bigger the sample size, the higher the probability of an outlier occurring naturally.

In summary, checking for outliers should be a routine procedure of any data analysis. If the extreme data point is in error, it should be corrected, if possible; and removed, if we believe that the outlier is due to careless mistakes and a correction of the data point is impossible.

When we have no good reasons to believe that the extreme data point is due to careless mistakes, the classical way to estimate the population mean and standard deviation is by using respectively the median and the median absolute difference, or by a process called winsorisation (Tukey, 1960 and Huber, 1964). However there is not much work which has been done on the construction of a suitable confidence interval for the population mean.

In this study, we investigate the construction of confidence interval for population mean in the presence of only **one** outlier at one of the two extreme ends. The sample under study is randomly generated, subscribing to a normal distribution.

We assign the probability of p to the occurrence whereby the outlier is due to human errors, with no corrections possible. For instance, when $p = 0.4$, in the generation of N samples, each of size n, about 40% of these generated samples would contain a single outlier caused by human errors, and about 60% of them contain an outlier due to exceptionally large measurement errors.

For a generated sample the outlier is first expunged from the sample, and a nominally $100(1-\alpha)\%$ confidence interval for the population mean is constructed based on the resulting sample of size n - 1. For the same generated sample, the outlier is next replaced (removed and substituted) by a randomly generated value which is larger than the second largest value (or smaller than the second smallest value) in the original sample that contains the outlier, and a nominally $100(1-\alpha)\%$ confidence interval for the population mean is constructed using the resulting sample of size n. By taking a union of these two confidence intervals, we form a third confidence interval.

To determine the performance of a given confidence interval, we estimate its coverage probability and expected length. The coverage probability may be estimated by the ratio of the number of confidence intervals that contain the population mean to N, while the expected length by the average length of the N confidence intervals based on the generated samples.

A nominally $100(1-\alpha)\%$ confidence interval is said to perform adequately well if the estimated coverage probability is close to the stipulated target value of $1-\alpha$. Between two types of confidence intervals with approximately the same estimated coverage probabilities, the one with a shorter average length is deemed to be a better confidence interval.

The above three types of confidence intervals are compared using their estimated coverage probabilities and average lengths. The main findings are that when the sample size is about 10, the confidence interval formed by the union operator has a satisfactory level of coverage probability for all values of p. As for sample sizes of about 20 and 30, the confidence interval in the case when the outlier is replaced and the confidence interval in the case when the outlier is eliminated from the sample would be a better one, respectively, as they have shorter average lengths apart from having reasonable level of coverage probabilities for a wide range of p.

Other works on construction of confidence intervals in the presence of only one outlier can be found in Goh (2011) and Low (2011). When there are two independent normal random samples with common variance σ^2, means μ_1 and μ_2, and sizes n_1 and n_2, respectively, Goh (2011) assumes that there is an outlier from the first sample, and uses a similar method based on the union operation to construct a confidence interval for the difference of the means of the two samples. Low (2011) assumes that there is an outlier in the set of data generated from a simple linear regression model with normal random errors, and also uses a method based on the union operation to construct a confidence interval for the slope parameter. The present article differs from Goh (2011) and Low (2011) in the method of generating new observation to replace the removed outlier.

CONSTRUCTION OF CONFIDENCE INTERVAL IN THE PRESENCE OF AN OUTLIER

Suppose ($y_1, y_2, y_3, ...y_n$) is a normal random sample from the normal population with mean μ and variance σ^2. The sample mean and standard deviation are given, respectively, by $\bar{y} = \left(\sum_{i=1}^{n} y_i\right) \Big/ n$ and $s^2 = \left[\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) / (n-1)\right]^2$. The usual normal-theory (1- α) 100% confidence interval for μ is I = [L, U], where $L = \bar{y} - t\, s \big/ \sqrt{n}$ and $U = \bar{y} + t\, s \big/ \sqrt{n}$; t being the 100(1 - $\alpha/2$) percentile of the t distribution with $n-1$ degrees of freedom.

The coverage probability of confidence interval is given by $P_C = P(L \le \mu \le U)$ and the expected length of the confidence interval is given by $E_L = E(U - L)$.

Suppose there is an outlier in the sample and we do not know whether the outlier is due to human errors or exceptionally large measurement errors. In what follows, we study three types of nominally 100(1-α) % confidence intervals for the population mean in the presence of a single outlier.

(a) Outlier is Eliminated

We delete the outlier and rename the sample as $y'_1, y'_2, y'_3, ...y'_{n-1}$. Let the corresponding sample mean and sample variance be denoted respectively by $\bar{y}'$ and s'^2. A (1-α)100% confidence interval for μ is $I_E = [L_E, U_E]$, where $L_E = \bar{y}' - t's' \big/ \sqrt{n-1}$ and $U_E = \bar{y}' + t's' \big/ \sqrt{n-1}$, t' being the 100(1 - $\alpha/2$) percentile of the t distribution with n - 2 degrees of freedom. The coverage probability and expected length of the confidence interval I_E are given by $P_{CE} = P[L_E \le \mu \le U_E]$ and $E_{LE} = E(U_E - L_E)$, respectively.

(b) Outlier is Replaced

We next assume that the outlier in the sample is due to exceptionally large measurement errors and it would then be replaced. The confidence interval is constructed using the following procedure:

1. Sort the values in the sample in an ascending order: $y_{(1)}, y_{(2)}, y_{(3)}, ..., y_{(n)}$. Then remove the outlier, either $y_{(1)}$ or $y_{(n)}$ (depending on whether the outlier is at the lowest or the highest end) and find the median $\hat{M}$ of $y_{(2)}, y_{(3)}, ..., y_{(n-1)}$.

2. Calculate $\hat{\sigma}_M^2 = \dfrac{1}{n-3}\sum_{j=2}^{n-1}\left(y_{(j)} - \hat{M}\right)^2$.

3. Compute the factor $\hat{f}$ from the values of $\hat{M}$ and $\hat{\sigma}_M^2$ using the formula $\hat{f} = c_o + c_1\left(\hat{M}/\hat{\sigma}\right) + c_2\left(\hat{M}/\hat{\sigma}\right)^2$, where c_o, c_1 and c_2 are constants found in the next section.

4. Keep generating $y_{(n)}^*$ (in the case when $y_{(n)}$ is the outlier) using the distribution $y_{(n)}^* \sim N\left(\hat{M},\ \left(\hat{f}\hat{\sigma}_M\right)^2\right)$ until the generated $y_{(n)}^*$ is larger than $y_{(n-1)}$. On the other hand, when $y_{(1)}$ is the outlier, we generate $y_{(1)}^*$ using the distribution $y_{(1)}^* \sim$ $N\left(\hat{M},\ \left(\hat{f}\hat{\sigma}_M\right)^2\right)$ until the generated value is smaller than $y_{(2)}$.

5. Replace the deleted outlier with the generated $y_{(n)}^*$ (or $y_{(1)}^*$) and rename the sample as $y'_{(1)}, y'_{(2)}, y'_{(3)}, ..., y'_{(n)}$. Let the mean and variance of the resulting sample be $\bar{y}'$ and s'^2 , respectively.

6. Compute a nominally (1-α)100% confidence interval $I_r = \left[L_r, U_r\right]$ for μ , where $L_r = \bar{y}' - ts'/\sqrt{n}$
and $U_r = \bar{y}' + ts'/\sqrt{n}$.

7. Repeat Steps (4) – (6) above N_g times, and obtain the confidence interval $I_R = \left[L_R, U_R\right]$ of which L_R
and U_R are, respectively, the averages computed from the N_g values of L_r and U_r found in Step (6).

We obtain another confidence interval for μ by using the union operator: $I_U = I_E \cup I_R = \left[L_U, U_U\right]$, and estimate the following coverage probabilities and expected lengths of the confidence intervals I_R and I_U :

$$P_{CR} = P\left[L_R \le \mu \le U_R\right],\ E_{LR} = E\left(U_R - L_R\right)$$
$$P_{CU} = P\left[L_U \le \mu \le U_U\right],\ E_{LU} = E\left(U_U - L_U\right)$$

FORMULA FOR COMPUTING THE FACTOR f^*

Starting with a given value of (μ , σ), we generate N values of the vector of observations **y.** We next sort the components in each generated vector in an ascending order to $y_{(1)}, y_{(2)}, y_{(3)}, ..., y_{(n)}$, and find the median $\hat{M}$ of $y_{(2)}, y_{(3)}, ..., y_{(n-1)}$ and the value

$\hat{\sigma}_M^2 = \left[\left(\sum_{j=2}^{n-1} \left(y_{(j)} - \hat{M} \right)^2 \right) \Big/ (n-3) \right]^2$ which serves as an estimate of the population variance. The average R^* of the N values of $\hat{M}/\hat{\sigma}_M$ is then computed.

For a given trial fixed value $\tilde{f}$, we keep generating $y^*_{(n)}$ - in the case when $y_{(n)}$ is the outlier – using the distribution $y^*_{(n)} \sim N\left(\hat{M},\ \left(\tilde{f}\hat{\sigma}_M\right)^2\right)$ until the generated $y^*_{(n)}$ is larger than $y_{(n-1)}$. In the case when $y_{(1)}$ is the outlier, we generate $y^*_{(1)}$ using the distribution $y^*_{(1)} \sim N\left(\hat{M},\ \left(\tilde{f}\hat{\sigma}_M\right)^2\right)$until the generated value is smaller than $y_{(2)}$. We next apply Steps (5) – (7) to each generated vector of observations to find a nominally (1-α)100% confidence interval $I_R = [L_R, U_R]$ for μ and use the proportion of confidence intervals (out of the N confidence intervals) which covers μ to estimate the coverage probability of the confidence intervals when $\tilde{f}$ is used. We then find the value f^* of $\tilde{f}$ such that the coverage probability of the corresponding confidence intervals is approximately equal to the target value $1-\alpha$.

A number of other starting values of (μ , σ) are then chosen. For each chosen value of (μ , σ), the corresponding values of R^* and f^* are obtained. Figures 1, 2 and 3 depict the scatterplots of f^* against R^*, for the case when n = 10, n = 20 and n = 30, respectively. For each value of n, we use a regression procedure to obtain the fitted function of R^*: $f^* = c_0 + c_1 R^* + c_2 R^{*2}$. The values of c_o , c_1 and c_2 are given in the figures.

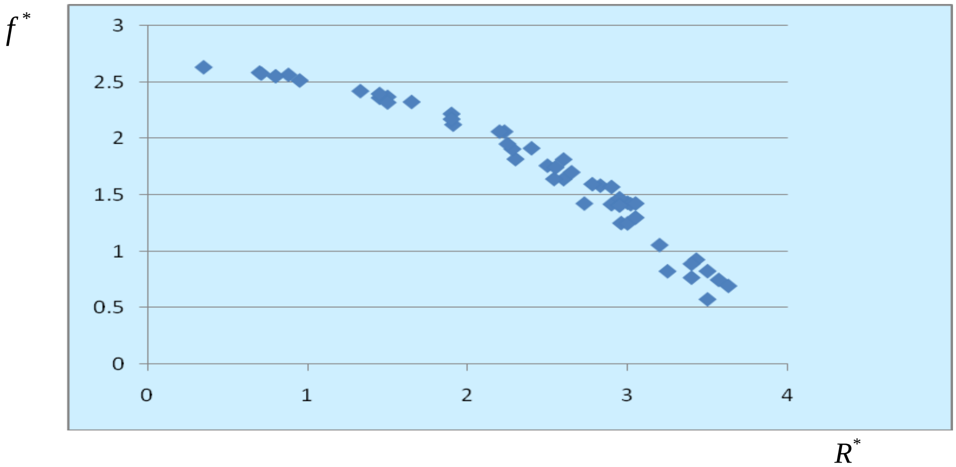

Figure 1. Scatterplot of $\left(f^*_{,}\ R^*\right)$; n = 10, fitted function is

$$f^* = 3.2217 + 0.7655 R^* - 0.6549 R^{*2}$$

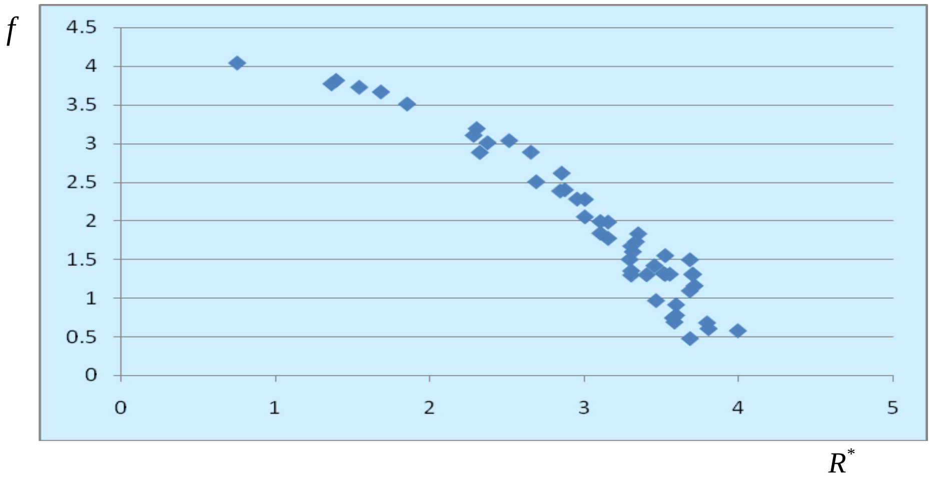

Figure 2. Scatterplot of $\left(f^*_{,}\ R^*\right)$; n = 20, fitted function is

$$f^* = 3.6962 + \ 0.1104R^* - 0.1881R^{*2}$$

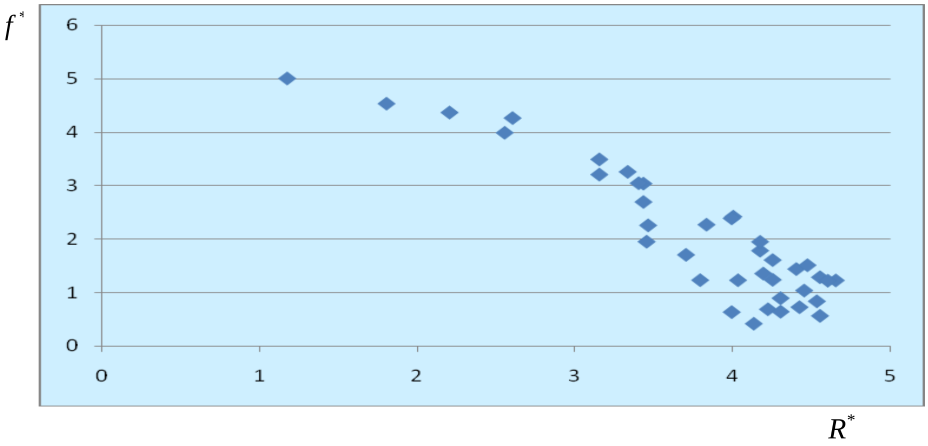

Figure 3. Scatterplot of $\left(f^*_{,}\ R^*\right)$; n = 30, fitted function is

$$f^* = 4.2848 + \ 0.1646R^* - 0.1509R^{*2}$$

NUMERICAL RESULTS

For each generated vector of observations $\mathbf{y}$, we find the confidence intervals I, I_E, I_R and I_U by using the procedures described above. For each type of confidence interval, we compute the corresponding estimated coverage probability and average length and record the results in Tables 1, 2 and 3.

Table 1 shows that when $n = 10$, the coverage probability of confidence interval I_E is very much less than the target value of 0.95 if p is small. Conversely, when p is sufficiently large, the coverage probability of confidence interval I_R is clearly less than 0.95. The coverage probability of confidence interval I_U is always slightly larger than 0.95, irrespective of the values of p, μ, σ and n.

Table 1. Coverage Probabilities and Expected Length of Confidence Intervals for N =10000, σ = 0.5, n = 10

μ	p	P_C	P_{CE}	P_{CR}	P_{CU}	E_L	E_{LE}	E_{LR}	E_{LU}
0.0	0.0001	0.9517	0.8990	0.9550	0.9663	0.6934	0.6472	0.8171	0.8630
	0.1	0.9513	0.8920	0.9580	0.9653	0.6974	0.6576	0.8363	0.8784
	0.3	0.9523	0.9053	0.9613	0.9723	0.6971	0.6781	0.8676	0.9055
	0.5	0.9553	0.9263	0.9700	0.9783	0.6942	0.6961	0.8975	0.9300
	0.7	0.9463	0.9347	0.9687	0.9783	0.6943	0.7181	0.9319	0.9585
	0.9	0.9517	0.9510	0.9777	0.9810	0.6961	0.7360	0.9659	0.9844
	0.9999	0.9543	0.9510	0.9143	0.9683	0.6939	0.7437	0.6939	0.8127
0.5	0.0001	0.9487	0.9053	0.9523	0.9703	0.6944	0.6487	0.8227	0.8689
	0.1	0.9477	0.9077	0.9583	0.9723	0.6984	0.6603	0.8468	0.8874
	0.3	0.9470	0.9110	0.9590	0.9723	0.6888	0.6706	0.8616	0.9001
	0.5	0.9560	0.9277	0.9653	0.9750	0.6987	0.7006	0.9108	0.9413
	0.7	0.9493	0.9303	0.9693	0.9767	0.6970	0.7145	0.9378	0.9635
	0.9	0.9490	0.9437	0.9660	0.9747	0.6956	0.7363	0.9708	0.9909
	0.99999	0.9473	0.9430	0.9100	0.9667	0.6908	0.7387	0.6908	0.8080
1.0	0.0001	0.9460	0.8993	0.9523	0.9633	0.6958	0.6503	0.8094	0.8601
	0.1	0.9577	0.9113	0.9587	0.9737	0.6990	0.6601	0.8254	0.8733
	0.3	0.9553	0.9167	0.9600	0.9693	0.6908	0.6717	0.8513	0.8926
	0.5	0.9513	0.9207	0.9630	0.9717	0.6973	0.6981	0.8918	0.9268
	0.7	0.9490	0.9303	0.9663	0.9747	0.6973	0.7173	0.9254	0.9524
	0.9	0.9387	0.9337	0.9667	0.9730	0.6941	0.7340	0.9516	0.9753
	0.99999	0.9580	0.9550	0.9117	0.9657	0.6936	0.7437	0.6936	0.8122
1.5	0.0001	0.9487	0.9067	0.9550	0.9690	0.6967	0.6487	0.7986	0.8485
	0.1	0.9450	0.9037	0.9507	0.9653	0.6996	0.6633	0.8188	0.8680
	0.3	0.9490	0.9170	0.9537	0.9667	0.6930	0.6743	0.8394	0.8826
	0.5	0.9467	0.9197	0.9610	0.9693	0.6923	0.6958	0.8751	0.9116
	0.7	0.9487	0.9380	0.9647	0.9733	0.6950	0.7145	0.9084	0.9382
	0.9	0.9603	0.9543	0.9713	0.9793	0.6954	0.7361	0.9404	0.9651
	0.99999	0.9540	0.9570	0.9213	0.9683	0.6951	0.7426	0.6951	0.8126
2.0	0.0001	0.9450	0.9027	0.9570	0.9667	0.6979	0.6498	0.8038	0.8527
	0.1	0.9493	0.9080	0.9610	0.9683	0.6948	0.6590	0.8203	0.8644
	0.3	0.9487	0.9127	0.9610	0.9703	0.6983	0.6792	0.8508	0.8904
	0.5	0.9500	0.9283	0.9667	0.9740	0.6993	0.6991	0.8816	0.9161
	0.7	0.9537	0.9397	0.9653	0.9743	0.6942	0.7145	0.9052	0.9345
	0.9	0.9363	0.9333	0.9643	0.9703	0.6926	0.7317	0.9330	0.9570
	0.9999	0.9520	0.9517	0.9117	0.9653	0.6966	0.7462	0.6966	0.8153
2.5	0.0001	0.9543	0.9080	0.9680	0.9750	0.7019	0.6558	0.8230	0.8691
	0.1	0.9440	0.9047	0.9590	0.9693	0.6956	0.6589	0.8288	0.8725
	0.3	0.9533	0.9200	0.9703	0.9770	0.6950	0.6768	0.8550	0.8925
	0.5	0.9507	0.9210	0.9680	0.9733	0.6961	0.6960	0.8867	0.9201

	0.7	0.9510	0.9343	0.9700	0.9753	0.6985	0.7174	0.9142	0.9428
	0.9	0.9480	0.9500	0.9773	0.9817	0.6930	0.7344	0.9398	0.9618
	0.9999	0.9427	0.9413	0.9003	0.9547	0.6944	0.7441	0.6944	0.8127
3.0	0.0001	0.9503	0.9017	0.9590	0.9677	0.6912	0.6445	0.8293	0.8674
	0.1	0.9567	0.9007	0.9680	0.9750	0.6974	0.6607	0.8482	0.8874
	0.3	0.9517	0.9203	0.9730	0.9777	0.6924	0.6740	0.8692	0.9023
	0.5	0.9457	0.9210	0.9697	0.9743	0.6941	0.6948	0.9010	0.9283
	0.7	0.9470	0.9293	0.9760	0.9807	0.6972	0.7187	0.9332	0.9554
	0.9	0.9513	0.9497	0.9807	0.9827	0.6936	0.7325	0.9529	0.9713
	0.9999	0.9487	0.9503	0.9113	0.9637	0.6966	0.7460	0.6966	0.8149

Table 2 reveals a slightly different observation. Although the coverage probability of confidence interval I_E is still less than target value 0.95 for the case when the value of *p* is small, the coverage probability of confidence interval I_R is not very much less than 0.95 when *p* is sufficiently large. As in the case when $n=10$, the coverage probability of confidence interval I_U when *n* = 20 is likewise slightly larger than 0.95, irrespective of the values of *p*, μ, σ and n. This means that when n = 20, both the confidence intervals I_R and I_U have satisfactory coverage probabilities. The performance of I_R and I_U can be deduced further by comparing their expected lengths. We observe that I_R would be a better confidence interval as it has a shorter expected length.

Table 2. Coverage Probabilities and Expected Length of Confidence Intervals for *N* =10000, σ = 0.5, n = 20

μ	p	P_C	P_{CE}	P_{CR}	P_{CU}	E_L	E_{LE}	E_{LR}	E_{LU}
0.0	0.0001	0.9477	0.9093	0.9460	0.9657	0.4630	0.4356	0.5477	0.5781
	0.1	0.9443	0.9127	0.9523	0.9703	0.4618	0.4392	0.5531	0.5792
	0.3	0.9543	0.9247	0.9627	0.9720	0.4635	0.4489	0.5688	0.5893
	0.5	0.9513	0.9287	0.9630	0.9723	0.4612	0.4548	0.5785	0.5950
	0.7	0.9467	0.9350	0.9680	0.9757	0.4613	0.4640	0.5924	0.6036
	0.9	0.9593	0.9523	0.9763	0.9817	0.4639	0.4729	0.6065	0.6127
	0.9999	0.9520	0.9540	0.9277	0.9660	0.4642	0.4781	0.4642	0.5228
0.5	0.0001	0.9437	0.9067	0.9453	0.9653	0.4610	0.4351	0.5456	0.5773
	0.1	0.9430	0.9153	0.9470	0.9647	0.4630	0.4395	0.5532	0.5803
	0.3	0.9480	0.9263	0.9533	0.9690	0.4594	0.4446	0.5627	0.5852
	0.5	0.9503	0.9363	0.9610	0.9747	0.4617	0.4548	0.5777	0.5950
	0.7	0.9530	0.9477	0.9733	0.9803	0.4625	0.4640	0.5917	0.6022
	0.9	0.9467	0.9420	0.9710	0.9743	0.4623	0.4728	0.6052	0.6113
	0.99999	0.9460	0.9460	0.9243	0.9597	0.4626	0.4756	0.4626	0.5206
1.0	0.0001	0.9550	0.9163	0.9483	0.9720	0.4622	0.4354	0.5415	0.5726
	0.1	0.9453	0.9127	0.9440	0.9637	0.4620	0.4388	0.5467	0.5758
	0.3	0.9480	0.9220	0.9583	0.9733	0.4607	0.4457	0.5589	0.5825
	0.5	0.9540	0.9373	0.9600	0.9750	0.4633	0.4572	0.5744	0.5929
	0.7	0.9567	0.9430	0.9673	0.9747	0.4617	0.4631	0.5856	0.5984
	0.9	0.9463	0.9433	0.9643	0.9673	0.4645	0.4744	0.6024	0.6097
	0.99999	0.9563	0.9560	0.9233	0.9647	0.4628	0.4763	0.4628	0.5211

1.5	0.0001	0.9503	0.9147	0.9437	0.9697	0.4613	0.4348	0.5352	0.5687
	0.1	0.9510	0.9153	0.9503	0.9700	0.4593	0.4369	0.5387	0.5694
	0.3	0.9487	0.9207	0.9507	0.9680	0.4611	0.4469	0.5545	0.5786
	0.5	0.9437	0.9270	0.9547	0.9693	0.4622	0.4561	0.5672	0.5870
	0.7	0.9467	0.9337	0.9660	0.9713	0.4624	0.4642	0.5816	0.5945
	0.9	0.9483	0.9393	0.9683	0.9727	0.4612	0.4708	0.5922	0.6001
	0.99999	0.9480	0.9450	0.9213	0.9570	0.4634	0.4773	0.4634	0.5213
2.0	0.0001	0.9413	0.9027	0.9403	0.9640	0.4608	0.4332	0.5307	0.5642
	0.1	0.9463	0.9157	0.9413	0.9670	0.4610	0.4384	0.5382	0.5692
	0.3	0.9457	0.9170	0.9527	0.9687	0.4600	0.4450	0.5488	0.5733
	0.5	0.9463	0.9277	0.9567	0.9717	0.4617	0.4544	0.5634	0.5831
	0.7	0.9533	0.9410	0.9683	0.9760	0.4614	0.4625	0.5767	0.5903
	0.9	0.9523	0.9490	0.9757	0.9807	0.4618	0.4713	0.5889	0.5981
	0.9999	0.9513	0.9520	0.9270	0.9643	0.4621	0.4755	0.4621	0.5200
2.5	0.0001	0.9513	0.9173	0.9463	0.9690	0.4629	0.4362	0.5346	0.5681
	0.1	0.9500	0.9103	0.9503	0.9677	0.4610	0.4383	0.5387	0.5687
	0.3	0.9440	0.9113	0.9500	0.9633	0.4628	0.4485	0.5533	0.5773
	0.5	0.9527	0.9357	0.9640	0.9743	0.4638	0.4579	0.5669	0.5867
	0.7	0.9477	0.9320	0.9597	0.9693	0.4613	0.4631	0.5764	0.5893
	0.9	0.9557	0.9520	0.9740	0.9767	0.4612	0.4703	0.5868	0.5959
	0.9999	0.9513	0.9513	0.9190	0.9627	0.4638	0.4770	0.4638	0.5219
3.0	0.0001	0.9523	0.9203	0.9480	0.9673	0.4602	0.4340	0.5350	0.5676
	0.1	0.9480	0.9220	0.9493	0.9700	0.4639	0.4412	0.5452	0.5761
	0.3	0.9507	0.9207	0.9557	0.9697	0.4625	0.4481	0.5553	0.5811
	0.5	0.9597	0.9323	0.9653	0.9780	0.4623	0.4555	0.5672	0.5861
	0.7	0.9507	0.9427	0.9670	0.9750	0.4626	0.4642	0.5799	0.5937
	0.9	0.9523	0.9510	0.9740	0.9780	0.4633	0.4729	0.5921	0.6002
	0.9999	0.9540	0.9537	0.9430	0.9707	0.4601	0.4732	0.5150	0.5513

When n = 30, all three confidence intervals I_E, I_R and I_U have satisfactory coverage probabilities (refer to Table 3). The interval I_E would now be the best confidence interval as it has the shortest expected length.

By using linear extrapolation of the results in Tables 1 – 3, we may further conclude that the confidence interval I_U would be the preferred one when the sample size n is less than or equal to 10, and the confidence interval I_E would instead be the most satisfactory one when n is bigger or equal to 30. For a given value of n between 11 and 29 but not close to 10, 20 or 30, the best confidence interval may be determined by using linear interpolation of the results in Tables 1 – 3.

Table 3. Coverage Probabilities and Expected Length of Confidence Intervals for N =10000, σ = 0.5, n = 30

μ	p	P_C	P_{CE}	P_{CR}	P_{CU}	E_L	E_{LE}	E_{LR}	E_{LU}
0	0.0001	0.9550	0.9317	0.9610	0.9777	0.3790	0.3617	0.4581	0.4810
	0.1	0.9557	0.9380	0.9607	0.9780	0.3797	0.3651	0.4625	0.4823
	0.3	0.9543	0.9303	0.9593	0.9713	0.3790	0.3686	0.4692	0.4850
	0.5	0.9543	0.9440	0.9703	0.9793	0.3789	0.3743	0.4776	0.4897
	0.7	0.9567	0.9503	0.9760	0.9830	0.3797	0.3797	0.4865	0.4941
	0.9	0.9563	0.9487	0.9813	0.9833	0.3785	0.3843	0.4929	0.4961
	0.9999	0.9593	0.9587	0.9383	0.9707	0.3795	0.3873	0.3795	0.4199
0.5	0.0001	0.9510	0.9310	0.9510	0.9743	0.3782	0.3610	0.4554	0.4787
	0.1	0.9600	0.9340	0.9577	0.9770	0.3781	0.3633	0.4595	0.4796
	0.3	0.9557	0.9360	0.9670	0.9767	0.3794	0.3694	0.4689	0.4856
	0.5	0.9560	0.9327	0.9647	0.9757	0.3792	0.3736	0.4757	0.4884
	0.7	0.9503	0.9430	0.9700	0.9790	0.3794	0.3790	0.4847	0.4924
	0.9	0.9573	0.9547	0.9803	0.9820	0.3785	0.3839	0.4922	0.4952
	0.9999	0.9563	0.9550	0.9383	0.9670	0.3795	0.3878	0.3795	0.4198
1.0	0.0001	0.9533	0.9297	0.9540	0.9727	0.3794	0.3614	0.4520	0.4746
	0.1	0.9537	0.9350	0.9540	0.9777	0.3804	0.3654	0.4582	0.4798
	0.3	0.9540	0.9377	0.9643	0.9780	0.3780	0.3681	0.4628	0.4799
	0.5	0.9557	0.9380	0.9707	0.9773	0.3782	0.3736	0.4714	0.4836
	0.7	0.9570	0.9463	0.9760	0.9813	0.3800	0.3803	0.4818	0.4897
	0.9	0.9613	0.9593	0.9837	0.9860	0.3783	0.3835	0.4878	0.4912
	0.9999	0.9560	0.9540	0.9327	0.9663	0.3785	0.3861	0.3785	0.4184
1.5	0.0001	0.9553	0.9360	0.9523	0.9773	0.3800	0.3624	0.4494	0.4741
	0.1	0.9567	0.9280	0.9543	0.9690	0.3771	0.3624	0.4500	0.4718
	0.3	0.9553	0.9380	0.9607	0.9743	0.3799	0.3698	0.4613	0.4794
	0.5	0.9603	0.9423	0.9660	0.9783	0.3804	0.3758	0.4701	0.4840
	0.7	0.9500	0.9433	0.9663	0.9743	0.3790	0.3791	0.4764	0.4842
	0.9	0.9553	0.9527	0.9780	0.9807	0.3788	0.3840	0.4839	0.4875
	0.99999	0.9587	0.9580	0.9373	0.9687	0.3794	0.3875	0.3794	0.4197
2.0	0.0001	0.9567	0.9320	0.9493	0.9747	0.3801	0.3625	0.4465	0.4725
	0.1	0.9563	0.9303	0.9527	0.9743	0.3796	0.3642	0.4491	0.4727
	0.3	0.9557	0.9350	0.9640	0.9787	0.3782	0.3683	0.4559	0.4745
	0.5	0.9590	0.9413	0.9650	0.9790	0.3779	0.3735	0.4640	0.4781
	0.7	0.9570	0.9463	0.9720	0.9813	0.3800	0.3802	0.4743	0.4838
	0.9	0.9560	0.9553	0.9800	0.9850	0.3783	0.3839	0.4802	0.4842
	0.9999	0.9583	0.9523	0.9330	0.9643	0.3778	0.3857	0.3778	0.4177
2.5	0.0001	0.9623	0.9350	0.9543	0.9770	0.3801	0.3625	0.4447	0.4707
	0.1	0.9483	0.9250	0.9440	0.9657	0.3783	0.3632	0.4464	0.4700
	0.3	0.9527	0.9373	0.9573	0.9733	0.3796	0.3697	0.4566	0.4756
	0.5	0.9520	0.9337	0.9620	0.9740	0.3778	0.3725	0.4612	0.4748
	0.7	0.9530	0.9453	0.9687	0.9743	0.3795	0.3801	0.4719	0.4805
	0.9	0.9577	0.9563	0.9780	0.9803	0.3788	0.3839	0.4792	0.4833
	0.9999	0.9527	0.9540	0.9330	0.9690	0.3784	0.3862	0.3784	0.4185
3.0	0.0001	0.9530	0.9250	0.9483	0.9650	0.3768	0.3595	0.4416	0.4678
	0.1	0.9587	0.9327	0.9567	0.9767	0.3794	0.3643	0.4481	0.4715
	0.3	0.9527	0.9380	0.9563	0.9720	0.3782	0.3679	0.4540	0.4721
	0.5	0.9620	0.9467	0.9697	0.9793	0.3805	0.3759	0.4650	0.4788
	0.7	0.9607	0.9497	0.9760	0.9837	0.3786	0.3786	0.4700	0.4789
	0.9	0.9523	0.9523	0.9757	0.9797	0.3796	0.3848	0.4793	0.4832
	0.9999	0.9590	0.9587	0.9483	0.9697	0.3806	0.3884	0.4046	0.4361

CONCLUDING REMARKS

In the present article it is assumed that the random errors are normally distributed and there is only one outlier in the given set of data. The numerical results in this study show that the choice of a suitable confidence interval for the population mean would depend on the sample size. Future research may be carried out to determine confidence intervals for the population mean and other parameters under a more general assumption of the distribution of the random errors in the presence of more than one outlier.

REFERENCES

Goh, S. T. (2011). *Confidence interval in the presence of an outlier* (M. Sc. (Statistics) Project Report, Institute of Mathematical Sciences, University of Malaya).

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1–21.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*(1), 73-101.

Low, C.Y. (2011). *Interval estimation of a slope parameter in the presence of an outlier* (M. Sc. (Statistics) Project Report, Institute of Mathematical Sciences, University of Malaya).

Moore, D. S. & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (Vol. 2), pp. 448-485). Stanford, CA: Stanford University Press.