Malaysian Journal of Mathematical Sciences 9(1): 1-20 (2015)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Analyzing Length or Size Biased Data: A Study on the Lengths of Peas Plants

[1*]A.H.M. Rahmatullah Imon and [2]Keya Rani Das

[1]Department of Mathematical Sciences, Ball State University, Muncie,
IN 47306, USA

[2]Department of Statistics, Bangabandhu Sheikh Mujibur Rahman
Agricultural University, Salna, Gazipur 1706, Bangladesh

E-mail: imon_ru@yahoo.com

*Corresponding author

ABSTRACT

Conventional statistical data analysis techniques largely depend on assumptions like randomness, normality, independence and similarity of the data. But in reality we often observe that these assumptions do not hold. Among them the randomness is considered as the most important one because if the data are not random the entire inferential procedure breaks down. Faulty sampling technique is mostly responsible for nonrandom samples but in environmental studies often we observe data no matter how carefully we design the sampling technique the data become biased either in length or size. Normality is another very important issue in statistical inference because all conventional sampling distributions and test statistics heavily rely on normality of the data. If we knew the appropriate distribution of the data we can analyze those in different ways, but we often observe data which may not match with the well-known distributions and nonparametric statistics is the only alternative there. In this paper we develop a procedure of analyzing data sets which are length or size biased. For this type of data we have developed a biased correction technique first and then apply bootstrap method on corrected data for the inferential purpose. We present a very interesting example in this paper which clearly shows the merit of employing our proposed procedure in analyzing this type of data.

Keywords: Transect sampling, Outlier, Weighted distributions, Robust statistics, Bootstrap.

1. INTRODUCTION

Every simple step in statistical inference is guided by some kind of assumptions whose existence are essential for a valid inferential statement.

For example, all four major test statistics *z, t,* χ^2 and *F* are valid only when the sample observations come from a normal distribution. Now one question might come to our mind, what is wrong if the assumptions are violated? According to Tukey (1960) ‘A tacit hope in ignoring deviations from the ideal model was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope is often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.’

In classical setup when we collect and summarize data for any kind of inference, whether it is clearly stated or not, we assume that

(i) Observations are random.
(ii) They are independent and are identically distributed.
(iii) They have come from a normal distribution.
(iv) All observations are equally reliable i.e., there is no outlier in the data.

Randomness is a key assumption for statistical inference because it forms the basis of the entire inferential procedure. Observing random samples may be difficult in practice and often the practitioners use convenient sampling techniques for collecting data. We can only use summary statistics if the data are not random in nature. But in environmental statistics we often observe data which have bias either in length or size or both. If we sample fish in a pond by catching them in a net, there will be encounter bias (more usually called size bias). This is because the mesh size will have the effect of lowering the incidence of the smaller fish in the catch- some will slip through the net. No matter how carefully we design the sample, the outcome will be biased and any randomness test will reject the hypothesis of randomness for this data. We have already mentioned the importance of normality assumption in inference. The violation of the normality assumption may lead to the use of suboptimal estimators, invalid inferential statements and inaccurate predictions. When we definitely know that the population of the data is not normal we may try with other parent distributions. But we often see that the data may not fit any of the well-known distributions. We have to use non-parametric method such as bootstrap in such a situation. Existence of outliers may often cause non-normality of the data. Its existence also violates the similarity assumption because when outliers occur we can no longer claim that observations are identically distributed. We often use robust statistics when outliers are present in the data. At first we briefly review the assumptions of normality and randomness and describe some commonly used procedures for checking these assumptions. We also review detection techniques for outliers and introduce a popular nonparametric method, bootstrap. In the next section, we develop a new method for estimating variance and confidence interval when non randomness occurs due to length or size bias. Then we introduce an interesting data which is the length of peas plants. We apply all commonly used techniques for checking the major assumptions and observe that this data fail all the assumptions. Finally, we apply different alternative methods to analyze the data and observe that the bias corrected bootstrap method most adequately fits this data.

2. DIAGNOSTIC CHECKS AND NONPARAMETRIC AND ROBUST METHODS

In this section we examine basic three assumptions required for the application conventional statistical analysis which are normality, data screening and randomness. At first we check the normality assumption for the data. This is the most crucial diagnostic check as the entire classical statistics are based on the normality assumption of observations. At the time of the development of the classical statistics there was a general believe among the statisticians that the data set follow a normal distribution. It was observed that most of the classical data such as height, weight etc followed normal distribution. In the last hundred years, attitudes towards the assumption of a normal distribution in statistical models have varied from one extreme to another. To quote Pearson (1905), 'Even towards the end of the nineteenth century not all were convinced of the need for curves other than normal.' By the middle of this century Geary (1947) made this comment '*Normality is a myth; there never was and never will be a normal distribution*.' Now it is evident that nonnormal data are more prevalent in nature. A nice review of different tests for normality is available in Imon (2003).

The simplest graphical display for checking normality is the normal probability plot. This method is based on the fact that if the ordered observations are plotted against their cumulative probabilities on normal probability paper, the resulting points should lie approximately on a straight line. A test based on the correlation of the observations and the expectation of normalized order statistics is known as the Shapiro–Wilk test. A test based on empirical distribution function is known as the Anderson–Darling test. A test based on the coefficients of skewness and kurtosis is known as Bowman–Shenton test. This test is popularly known as the Jarque–Bera test. If we denote the sample size by n, the sample skewness by S and the sample kurtosis by K, then the Jarque–Bera test statistic is defined as

$$\text{JB} = [n / 6]\,[\,S^2 + (K-3)^2 / 4] \tag{1}$$

The standard theory tells us that a normal distribution has skewness 0 and the value of the kurtosis is 3. So a departure from these two values will indicate non-normality and that is how this test statistic was developed. The JB statistic follows a chi-square distribution with 2 degrees of freedom.

A well-known reason for normality may be the existence of outliers. Checking of outliers, which is also popularly known as data screening, has become an essential part of data analysis. According to Barnett and Lewis (1994), we shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. Hampel *et al*. (1986) claim that a routine data set typically contains about 1-10% outliers, and even the highest quality data set cannot be guaranteed free of outliers. One immediate consequence of the presence of outliers is that they may cause apparent non-normality and the entire classical inferential procedure might breakdown in the presence of outliers. An excellent review for the detection of outliers is available in Hadi, Warner and Imon (2009). A very simple graphical display that can be used in detecting outliers is the box plot where an observation is declared as an outlier if it falls outside the range

$$(Q_1 - 1.5\ \text{IQR},\ Q_3 + 1.5\ \text{IQR}) \tag{2}$$

where Q_1 and Q_3 are the first and third quartiles of the data and IQR = $Q_3 - Q_1$ is known as the interquartile range (IQR).

A very simple and popular technique for the detection of outliers is the so-called ‘three-sigma’ rule. If we assume a normal distribution, a single value may be considered as an outlier if it falls outside a certain range of the standard deviation. A traditional measure of the ‘outlyingness’ of an observation x_i with respect to a sample is the ratio between its distance to the sample mean and the sample standard deviation (SD):

$$t_i = \frac{x_i - \bar{x}}{s} \tag{3}$$

where $\bar{x}$ and s are respective mean and standard deviation of the data. Since the empirical rule of a normal distribution tells us that 99.7% of observations will fall in the interval $|t_i| \leq 3$, we declare an observation to be an outlier when it falls in the region $|t_i| > 3$. Although the three-sigma rule has an extensive use in data analysis as an outlier detection technique, it is now evident [see Imon (2005)] that this rule often fails to identify outliers. The reason is simple. In the three-sigma rule we use sample mean and sample standard deviation both of which get highly affected in the presence of outliers. Hampel *et al*. (1986) revised the three-sigma rule by using the robust plug-in technique to obtain a robust *t*-like statistic by replacing mean by median and standard deviation by the normalized median absolute deviation (MADN). Thus the modified statistic becomes

$$t_i' = \frac{x_i - \text{Median}(x)}{\text{MADN}(x)} \tag{4}$$

Here the median absolute deviation (MAD) is defined as MAD (*x*) = Med {|*x* – Med (*x*)|}. To make the MAD comparable to the SD in terms of efficiency, the normalized MAD defined as MADN (*x*) = MAD (*x*) / 0.6745 is considered. Observations with $|\ t_i'\ | > 3$ are identified as outliers. Chorminski and Tkacz (2010) made a comparative study on the effectiveness of a variety of outlier detection techniques and come up with the conclusion that Hample’s method performs best overall.

Another very important consideration of data analysis is the test for randomness. The run test (see Hogg and Tanis (2010)) is the most popular test for checking randomness of data. If a data set contains *n* observations replace each observation by L if it falls below the median and by U if it falls above the median. Then we count the number of runs denoted by *r*. If *n* is even, the number of observations of each group will be the same, i.e., $n_1 = n_2$. If *n* is odd, conventionally we put $n_2 = n_1 + 1$. The critical region is of the form $r \leq c_1$ or $r \geq c_2$. When n_1 and n_2 are large (say, each is at least equal to 10), *r* can be approximated by a normal random variable with

$$\mu = \frac{2n_1\ n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \tag{5}$$

The test statistic is

$$z = \frac{r - \mu}{\sigma} \tag{6}$$

The critical region for this test is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$. If the calculated value of z falls in the critical region we reject the hypothesis that the data is random, otherwise we accept the hypothesis that the data is random.

If the data is random, normal and free from outliers we are free to allow all kind of conventional analyses. But what we should do if the data does not pass the normality assumption? If the only reason for nononormality is the existence of outliers, we can use robust statistics. The term robustness signifies insensitivity to small deviations from the assumption. That means a robust procedure is nearly as efficient as the classical procedure when classical assumptions hold strictly but is considerably more efficient over all when there is a small departure from them. One objective of robust techniques is to cope with outliers by trying to keep small the effects of their presence. Consequently, we should require resistant estimators. To quote Ryan (1997), 'A resistant estimator is one that is relatively unaffected by large changes in a small part of the data or small change in a much larger part of the data.' An excellent review of different robust methods used in statistics is available in Maronna *et al*. (2006). Median, trimmed mean, Huber's M are very popular robust measure of central tendency. Robust estimates of dispersion are normalized median absolute deviation, *S* estimator etc. We have robust estimates of skewness, kurtosis, correlation coefficient, regression coefficient to name a few. It is now generally believed that corresponding to every classical statistic there exists a robust alternative.

If the data is not normal, but follows some other known distribution such as exponential, lognormal, gamma we can still find an alternative way of analyzing this data. But what happens if the data do not match any known distribution? We cannot apply any conventional method to analyze data and for this reason we apply nonparametric methods in this situation. The essential difference between classical and nonparametric statistics is this: in classical statistics we assume a known form (distribution) of data and then collect sample from there. But in nonparametric method we do not assume any distribution of the data, rather we explore the data to find out its most appropriate pattern, if at all. Otherwise, we calculate all measures in an empirical way. Among a variety of nonparametric methods, bootstrap has become very popular with the statisticians. Bootstrapping is a modern, computer intensive, general purpose approach to statistical inference, falling with a broader class of resampling methods.

Bootstrap was first introduced by Efron (1979) for assigning the measure of accuracy of the estimates using the idea of resampling from a sample. The key idea is to resample from original data to create replicate data sets from which variability of the quantities of interest can be assessed without long winded analytical calculation. In the real world, the unknown probability distribution F gives the data $S=(x_1, x_2,...x_n)$ by random sampling; from S we calculate the statistic of interest $\hat{T}$ = $t(S)$. In the bootstrap world, $\hat{F}$ generates S_b^*= { x_{b1}^*, x_{b2}^*, …, x_{bn}^*} by random sampling giving $\hat{T}_b^*$ = $t($ S_b^*). There is only one observed value of $\hat{T}$, but we can generate as many bootstrap replications $\hat{T}_b^*$ as affordable. Next, we compute the statistic T for each of the bootstrap samples; that is $\hat{T}_b^*$ = $t($ S_b^*). Then the distribution of $\hat{T}_b^*$ around the original estimate $\hat{T}$ is analogous to the sampling distribution of the estimator T around the population parameter θ . For example the average of the bootstrapped statistics,

$$\bar{\hat{T}}^* = \hat{E}^*\left(\hat{T}^*\right)= \frac{\sum_{b=1}^{B} \hat{T}_b^*}{B} \qquad (7)$$

is the estimate of the expectation of the bootstrap statistics; then $Bia\hat{s}^* = \bar{\hat{T}}^*$ – T is an estimate of the bias of T. Similarly, the estimated bootstrap variance of T is

$$\hat{V}^*\left(\hat{T}^*\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{T}_b^* - \bar{\hat{T}}^*\right)^2 \qquad (8)$$

that estimates the sampling variance of T. The random selection of bootstrap samples is not an essential aspect of the nonparametric bootstrap. At least in principle, we could enumerate all bootstrap samples of size n. Then we could calculate $E^*\left(\hat{T}^*\right)$ and $V^*\left(\hat{T}^*\right)$ exactly, rather than having to estimate them. The bootstrap confidence interval (BCI) is often used instead of the classical confidence interval especially when the distribution is not symmetric or the parent distribution is unknown. The percentile confidence interval is given as

$$(k_{\alpha/2}, k_{1-\alpha/2}) \qquad (9)$$

3. LENGTH OR SIZE BASED SAMPLING AND WEIGHTED DISTRIBUTIONS

In statistics we often observe data which have bias either in length or size or both. If we sample fish in a pond by catching them in a net, there will be encounter bias (more usually called size bias). This is because the mesh size will have the effect of lowering the incidence of the smaller fish in the catch- some will slip through the net. If we were to sample harmful industrial fibers (in monitoring adverse health effects) by examining fibers on a plane sticky surface by line-intercept methods, the similar problem may arise. In this case our data would consist of the lengths of fibers crossed by the intercept line as shown below.

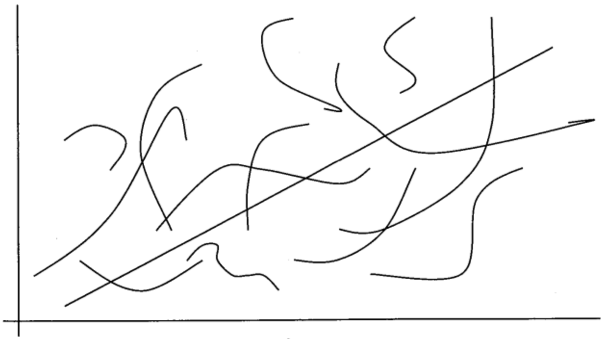


Figure 1: An example of transect sampling

Our interest will be in the distribution of sizes, but the sampling methods just described are clearly likely to produce seriously biased results. Here we are bound to obtain what are known as length-biased or size-biased samples, and statistical inference drawn from such samples will be seriously flawed because they relate to distribution of measured sizes, not to the population at large (as shown in Figure 2), which will our real interest. Thus we will typically overestimate the mean both in the fish and in the fiber examples, possibly to a serious extent.

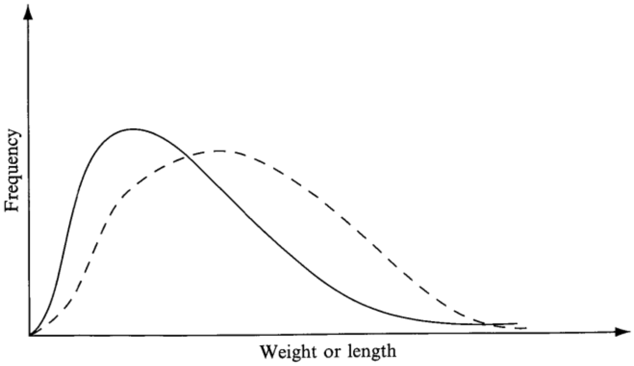


Figure 2: An example of original and length or size biased distribution

Here we introduce one kind of weighted distribution to remove or reduce the bias in the data. Suppose *X* is nonnegative random variable with mean μ and variance σ^2, but what we actually sample is a random variable X^*. A special but popular case of the size-biased distribution (see Barnett (2004)) has the p.d.f.

$$f^*(x) = xf(x)/\mu \tag{10}$$

The variable actually sampled has expected value

$$E(X^*) = \int \left[x^2 f(x)/\mu\right] dx = \mu\left(1 + \frac{\sigma^2}{\mu^2}\right) \tag{11}$$

So if we take a random sample of size *n*, then the sample mean of the observed data $\bar{x}^*$ is biased upward by a factor $\left(1 + \frac{\sigma^2}{\mu^2}\right)$. Here the problem is that we do not know the true values of μ and σ^2. However, the statistic

$$\bar{x} = \frac{\bar{x}^* \sum_{i=1}^{n} 1/x_i^*}{n} \tag{12}$$

provides an intuitively appealing estimate (see Barnett (2004)) of the bias factor $\left(1+\frac{\sigma^2}{\mu^2}\right)$. But the estimator of the mean is not good enough to provide the basic properties of the data. Now we would propose an estimate of σ^2 for this distribution. From (11) and (12) we obtain

$$V\left(X^*\right)=E\left(X^{*2}\right)-\left[E\left(X^*\right)\right]^2=\int\left[x^3 f\left(x\right)/\mu\right]dx-\left[\mu\left(1+\frac{\sigma^2}{\mu^2}\right)\right]^2 \tag{13}$$

Now

$$E\left(X^{*2}\right)=\int\left[x^3 f\left(x\right)/\mu\right]dx=\frac{1}{\mu}\left[\mu_3+3\mu\left\{E\left(X^2\right)\right\}-3\mu^2\mu+\mu^3\right]$$

$$=\frac{1}{\mu}\left[\mu_3+3\mu\left(\mu^2+\sigma^2\right)-2\mu^3\right]$$

$$=\frac{1}{\mu}\left[\mu_3+3\mu\,\sigma^2+\mu^3\right]$$

$$=3\sigma^2+\mu^2+\frac{\mu_3}{\mu} \tag{14}$$

where $\mu_3=E\left(X-\mu\right)^3$. Hence using (13) and (14) we obtain

$$V\left(X^*\right)=3\sigma^2+\mu^2+\frac{\mu_3}{\mu}\;-\left[\mu\left(1+\frac{\sigma^2}{\mu^2}\right)\right]^2$$

$$=3\sigma^2+\mu^2+\frac{\mu_3}{\mu}\;-\mu^2-2\sigma^2-\frac{\sigma^4}{\mu^2}$$

$$=\sigma^2+\frac{\mu_3}{\mu}\;-\frac{\sigma^4}{\mu^2} \tag{15}$$

Since the expression (15) contains both σ^2 and σ^4 both it may not be possible to get a plausible estimator of σ^2 like (12), but we can rewrite (15) as

$$3\sigma^2 = V(X^*) - \frac{\mu_3}{\mu} + \mu^2 \left[\left(1 + \frac{\sigma^2}{\mu^2} \right)^2 - 1 \right] \tag{16}$$

Thus using liner approximations in (16) a reasonable estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{3} \left[S^2(X^*) - \frac{m_3}{\overline{x}} + \overline{x}^2 \left\{ \left(bias\ \ factor \right)^2 - 1 \right\} \right] \tag{17}$$

where $S^2(X^*)$ is the estimated sample variance from the biased data m_3 is the estimated third central moment of the data.

4. MODELING PEAS PLANT DATA ANALYSIS, RESULTS AND DISCUSSIONS

In this section we first introduce a data that we use in our study. This primary data set which contains 1000 peas plants (pisum satiuvum) is collected from Bangabandhu Sheikh Mujibur Rahman Agricultural University using the transect sampling method as described in the previous section. This data is presented in Table A1 in the Appendix. We examine basic three assumptions required for the application conventional statistical analysis which are normality, data screening and randomness.

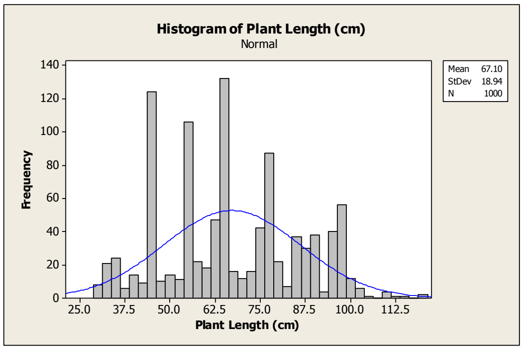


Figure 3: Histogram of the lengths of peas plants

Figure 3 presents a histogram of the length of peas plants and apparently this data does not look like normal. Here we have employed three different normality tests here. The first one is the normal probability plot, the second one is the Anderson–Darling test and the third one is the Jarque–Bera (Bowman–Shenton) test.

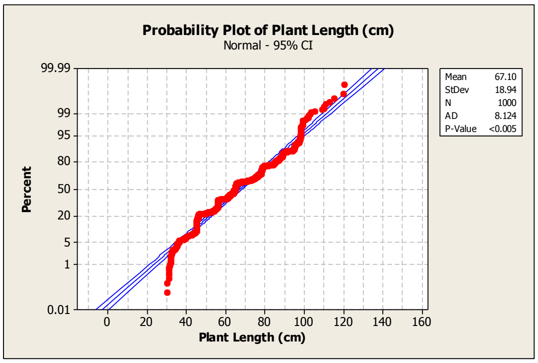


Figure 4: Normal probability plot of the lengths of peas plants

The normal probability plot of the data as shown in Figure 4 does not clearly show a normal pattern. We also observe from this table that the p-value of the Anderson-Darling statistic for this data is less than 0.005 which clearly rejects normality. We also employ the Jarque-Bera test for this data. The sample skewness and kurtosis for this data are 0.155719 and 2.19055 respectively that yield the value of the statistic as 22.24 which is much higher than the cut-off value of this test which is 5.99 at the 5% level of significance. Thus we can conclude that there is enough evidence to believe that this data do not follow a normal distribution and conventional statistical analyses should not be appropriate for this data.

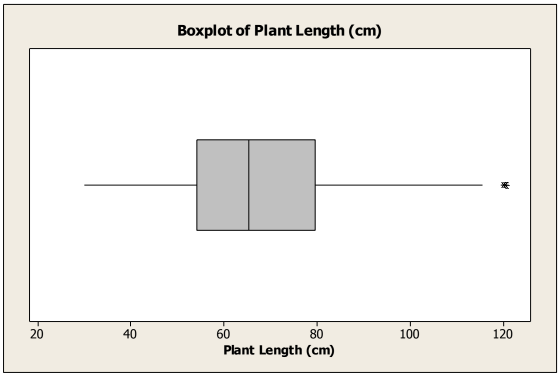


Figure 5: Box plot of the lengths of peas plants

Now we do an outlier analysis to the length of peas plant data. Most of the popular detection methods including Hampel's test do not identify any observations as outlier, however the rule based on the inter-quartile range identify two observations (cases 39 and 74) as outliers. We observe exactly the same kind of picture from the box plot of the data as shown in Figure 5.

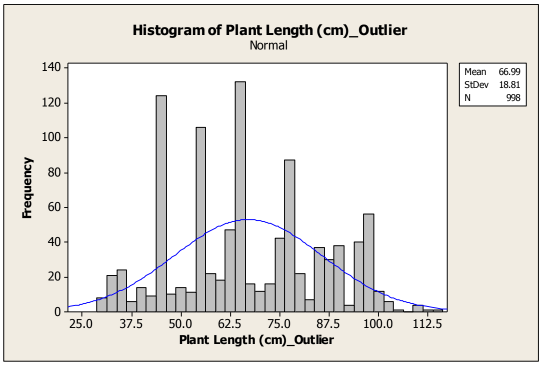


Figure 6: Histogram of the lengths of peas plants without outliers

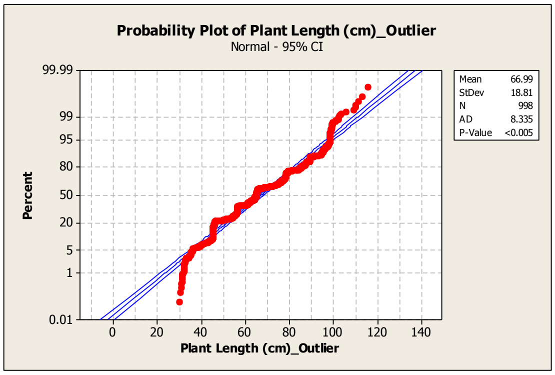
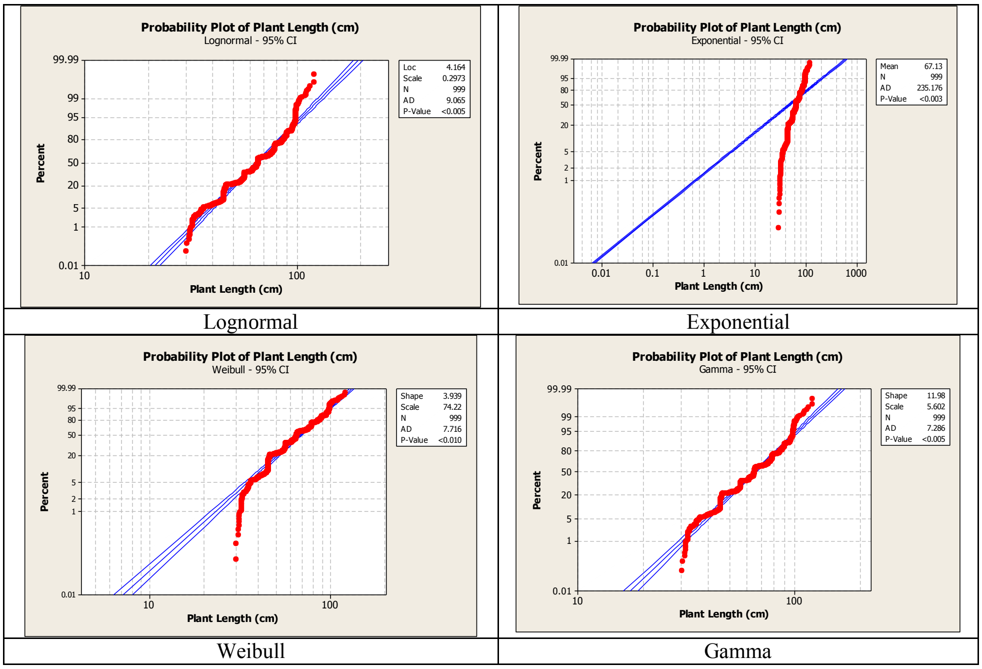


Figure 7: Normal probability plot of the lengths of peas plants without outliers

To understand the effect of outliers we remove observations 39 and 74 and repeat all steps that we did before. Figure 6 presents the histogram of the length of peas plants without outliers. The plot looks very similar to Figure 1 and their normal probability plot as shown in Figure 7 shows that there is not much improvement in the results when outliers are omitted. The p-value of the Anderson-Darling statistic for this data is less than 0.005 which clearly rejects normality. For the full data set the value of the Jarque-Bera statistic was 22.24 and now after the omission of outliers is 23.71. This means that the omission of outliers did not improve the normality pattern of the data. If we look at the summary statistics we observe that the mean of the full data is 67.100 cm with a standard deviation of 18.943 cm.

After the omission of the outlier the values of the mean and standard deviation are 66.694 cm and 18.811 cm respectively. Since there exists outliers in the data we employ the robust estimation technique to estimate the mean and standard deviation of the lengths of peas plants and the resulting values are 65.803 and 18.559 respectively. These values are very close to one another and that tell us that we do not have unusually big outlier in the data.



Lognormal

Exponential

Weibull

Gamma

Figure 8: Probability plot of the lengths of peas plants

We are convinced at this point that the lengths of peas plants data is not normal and the omission of the outliers also cannot help. Hence we try to fit this data with some other well-known distributions. The box plot as given in Figure 5 shows that the data is a bit skewed to the right, we consider four different alternatives: lognormal, exponential, Weibull, and gamma. The normal probability plots and the Anderson-Darling tests as shown in Figure 8 clearly indicate that neither of the distributions can adequately fit the data.

Next we move to checking another important assumption, that is the assumption of the randomness. We employ the run test as mentioned in (5) and (6). Here the observed number of runs is 347 and the expected number of runs is 490.632.

Hence the p-value for the run test is 0.000 and thus the peas plant data clearly rejects the null hypothesis of randomness. One question immediately comes to our mind, what is wrong with the data? Was there anything wrong with the sampling design? We carefully monitored the entire procedure and observed that there was no flaws, but the data was collected by transect sampling method, which is susceptible to length bias and consequently the test for randomness may fail. Here we employ the bias correction techniques as they were described in (10) – (17). Using (12) we obtain the bias correction factor as 1.09210. Using this bias factor the corrected mean length of peas plant becomes 61.069 cm which is about 6.03 cm less than the original mean. This difference is highly significant at any level of significance. The corrected standard deviation of the lengths is 16.98 cm which is about 1.94 cm less than the original standard deviation, which is significantly different from the original value. The above results make much sense. Since the data was collected by the transect sampling method it is highly likely that taller plants were selected more than smaller ones and hence the corrected mean is significantly smaller than the original one.

Finally we employ the bootstrap technique to estimate the mean, standard deviation and confidence interval of mean for the lengths of peas plants. Although we have a relatively large sample size of 1000 we fail to find its appropriate parent distribution. So for the computation of mean, standard deviation and especially for finding the confidence interval of the parameters it is better to use the bootstrap technique which does not require any assumption regarding the parent distribution of data. Here we work with the bias corrected data. We use the statistical package R for bootstrapping and the results are based on 10000 replications. It is intersting to note that the bootstrap mean is exacly equal to the bias corrected mean. The standard deviation is marginally smaller than the bias corrected one. But the most intersting feature of this method is the confidence interval of the mean. Here the 95% bootstrap confidence interval is (60.915, 61.224) with the confidence length 0.309 cm. This result clearly indicates that how precisely bootstrap estimates the mean length of peas plants.

Table 1 offers a comparison of different estimation methods used to analyze the lengths of peas plants. Here we compute the mean, the standard deviation, 95% confidence interval of the mean and the confidence length. We compare six different sets of methods, the classical method, classical method without outliers, robust method based on Huber’s weight function, our newly proposed bias corrected method applied on classical, robust and bootstrap techniques.

TABLE 1: Summary statistics for lengths of peas plants using different estimation methods

Estimation Method	Mean	Standard deviation	95% Confidence interval	95% Confidence length
Classical	67.100	18.943	(65.926, 68.274)	2.348
Classical without outliers	66.694	18.811	(65.524 ,67.864)	2.341
Robust	65.803	18.559	(64.653, 66.953)	2.300
Bias corrected classical	61.069	16.980	(60.018, 62.120)	2.105
Bias corrected robust	60.152	16.960	(59.101, 61.203)	2.102
Bias corrected bootstrap	61.069	16.960	(60.915, 61.224)	0.309

When we compare the means we observe that the bias corrected means are about 5-6 cm smaller than the uncorrected means. The bias corrected standard deviations are about 2 cm smaller than uncorrected standard deviations. These differences are huge in a sample of size 1000 and it clearly reemphasizes our concern that when the data is length biased no matter how shophsticated method we use, unless we correct the bias we will not get the correct results. Among the six sets of results the standard deviations are the least for the bias corrected robust and the bias corrected bootstrap methods. But we get an astonishing result when we look at the 95% confidence lengths for the means. This is only 0.039 cm for the bootstrap method whereas they are more than 2 cm for all other five methods. Although these results look surprising, it makes much practical sense. We must not forget the fact the first five methods use normal assumption while constructing the confidence interval. But here there is clear evidence that the data do not follow a normal distribution. Bootstrap confidence interval is computed based on only the empirica values and does not require any assumption regarding the parent distribution of the data and hence it produces the shortest interval. Thus the bias corrected bootstrap method produces the best set of results for the lengths of peas plants data.

5. CONCLUSION

In this paper our main objective was to find an appropriate method for analyzing data when the data may appear as nonrandom because of length or size bias and at the same time may not follow a normal distribution. We develop a method for correcting bias in mean and standard deviation. Finally we applied these biased corrected methods to analyze the lengths of peas plants. Since there is enough evidence that the data do not follow a normal distribution, the bias corrected bootstrap method is proven as the most appropriate method for analyzing the data.

ACKNOWLEDGEMENTS

We gratefully acknowledge all valuable comments and suggestions made by the reviewer. The first would like to thank Ball State University for funding a part of this research through President's travel grant.

REFERENCES

Barnett, V. (2004). *Environmental Statistics: Theory and Methods*. New York: Wiley.

Barnett, V. and Lewis, T. B. (1994). *Outliers in Statistical Data*, 2nd ed. New York: Wiley.

Chorminski, K. and Tkacz, M. (2010). Comparison of outlier detection methods in biomedicaI data. *Journal of Medical Informatics and Technologies*. **16**: 89-94.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *Annals of Statistics*. **7**: 1-26.

Geary, R.C. (1947). Testing for normality. *Biometrika*. **34**: 209-242.

Hadi, A.S., Imon, A. H. M. R. and Werner, M. (2009). Detection of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*. **1**: 57–70.

Hampel, F. R., Ronchetti, E. M., Rousseeuw P. J. and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: Wiley.

Hogg, R.V. and Tanis, E. A. (2010). *Probability and Statistical Inference*, 8th ed. New York: Prentice Hall

Imon, A. H. M. R. (2003). Regression residuals, moments, and their use in tests for normality. *Communications in Statistics-Theory and Methods*. **32**: 1021–1034.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*. **32**: 929–946.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York: Wiley.

Pearson, K. (1905). On the general theory of skew correlation and non-linear regression. *Biometrika*. **4**: 171-212.

Ryan, T. P. (1997). *Modern Regression Methods*. New York: Wiley.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.), CA: Stanford University Press, 448–485.

APPENDIX

Table A1: Lenghts of peas plants (in cm)

40.10	45.60	75.30	98.30	89.00	71.20	60.50	44.50	64.20	91.20
45.20	88.10	85.10	98.30	77.60	75.20	60.20	85.20	67.10	64.20
40.10	102.10	65.20	98.20	79.80	76.30	64.30	87.20	83.20	65.30
88.20	54.20	45.30	78.40	78.90	78.10	62.30	65.30	60.20	67.20
77.30	60.10	65.20	56.20	99.80	79.40	61.30	110.20	94.50	55.60
74.10	77.60	45.20	78.20	94.60	98.10	64.30	45.20	113.20	46.20
71.10	84.50	45.10	89.20	94.50	95.20	64.30	56.20	64.50	88.20
74.50	46.10	32.10	45.20	95.10	75.10	65.10	61.30	63.10	64.30
31.40	31.20	65.40	56.20	95.30	74.20	62.30	71.20	73.20	61.30
35.00	55.10	98.50	56.20	78.60	65.00	61.30	75.30	72.00	98.20
35.20	46.50	78.50	78.50	95.10	54.20	61.50	79.20	94.50	79.10
35.40	78.30	78.40	45.40	45.60	78.50	61.50	80.20	87.30	61.20
66.20	89.10	32.10	65.20	98.30	45.20	77.30	40.20	64.20	46.30
65.20	45.30	65.40	76.20	65.20	65.20	79.30	41.00	31.20	87.20
54.20	56.20	98.60	86.20	45.30	98.20	65.40	63.20	64.30	62.20
45.20	34.10	87.50	84.20	56.80	65.30	54.30	52.80	65.20	76.20
45.20	36.50	98.30	95.20	89.50	45.20	65.20	89.60	93.20	76.20
56.30	35.20	65.50	75.10	78.10	65.20	98.20	79.50	56.20	56.20
98.00	30.20	78.20	78.30	56.20	78.10	54.20	64.10	45.10	46.20
99.20	99.30	45.10	98.60	36.20	89.20	65.30	61.30	65.30	33.20
98.20	98.60	56.20	98.60	34.60	56.20	65.40	56.50	61.30	34.20
78.30	78.40	56.20	75.60	98.30	32.20	65.30	55.20	62.30	60.20
62.20	45.10	45.30	85.30	97.20	65.20	89.50	49.30	82.30	84.60
87.40	56.20	89.50	95.30	76.50	66.50	78.40	51.20	46.20	95.60
98.10	78.20	78.40	74.30	73.10	98.20	56.10	76.50	42.30	76.20
101.10	88.00	78.50	65.20	74.50	78.00	52.10	89.40	73.60	79.50
102.10	83.10	65.20	63.20	54.20	65.50	56.20	51.30	56.10	46.20
101.00	82.10	45.10	88.50	56.10	87.40	32.60	61.30	46.20	46.20
45.50	45.60	56.20	78.50	78.20	44.60	36.00	36.20	95.30	41.20
78.60	95.00	56.20	98.00	89.10	98.50	84.60	52.20	62.30	32.20
91.30	77.00	45.20	65.00	45.00	78.60	87.20	71.23	70.20	55.20
91.30	77.50	78.50	65.50	56.40	95.10	65.20	50.20	64.30	98.30
94.30	74.20	89.50	45.30	89.20	65.40	32.20	63.20	68.30	64.30
97.30	71.30	45.10	98.20	91.20	63.20	65.20	61.20	53.20	64.20
98.60	34.50	45.00	75.20	94.30	56.10	45.30	45.20	36.10	56.10
77.20	36.50	36.20	65.20	96.10	45.20	65.20	65.30	39.40	76.10
77.10	98.20	45.30	54.30	85.20	98.50	98.30	56.20	60.50	46.20
102.20	76.20	56.10	53.20	75.00	78.50	65.20	45.20	46.30	84.50
120.40	75.30	46.20	56.50	86.20	45.90	45.20	62.30	59.10	95.60
55.10	45.10	78.10	45.60	56.20	65.20	56.20	61.30	76.20	64.30
54.20	75.20	88.20	98.30	45.30	87.20	78.20	95.30	46.20	61.30
57.20	65.00	45.10	87.30	56.30	79.20	89.20	65.20	63.20	64.30
78.10	64.20	66.20	65.30	96.20	89.20	56.20	45.20	83.20	65.20
64.20	99.20	78.20	79.20	85.30	56.20	45.20	65.20	46.20	67.20
87.70	98.60	45.00	54.60	96.30	45.20	65.30	87.20	69.20	77.50
77.30	97.40	66.20	66.50	78.20	56.20	65.30	98.20	86.20	60.20
55.20	78.50	55.00	85.30	56.30	32.20	65.20	81.20	46.30	43.10
79.10	46.10	99.10	98.30	45.10	39.00	45.20	72.20	76.10	42.10
74.10	46.20	95.20	65.30	63.10	38.50	65.30	64.20	31.20	61.30
75.30	32.10	85.30	65.20	89.20	69.20	98.30	45.10	64.30	62.30

34.30	45.30	35.20	45.20	45.20	76.30	78.20	56.20	67.30	46.20
36.10	65.20	54.20	45.30	56.20	74.50	89.20	75.10	65.30	94.00
88.20	56.30	51.20	56.30	78.00	89.50	49.20	84.20	35.20	62.20
87.70	78.10	56.40	55.20	56.20	75.60	50.20	40.20	64.30	73.50
94.60	77.20	44.60	77.30	45.20	86.40	45.20	51.20	95.30	94.30
35.10	89.30	77.30	66.30	56.20	86.10	31.20	65.20	75.10	96.10
78.30	85.10	98.50	78.50	63.20	98.50	64.30	54.20	65.00	86.10
75.30	45.30	65.20	98.50	56.20	75.40	61.20	53.20	76.20	45.60
97.60	45.10	45.20	45.50	45.20	76.20	56.20	36.20	43.10	88.20
75.20	65.20	56.20	65.20	56.20	56.40	89.20	37.50	46.00	81.20
77.40	64.30	56.10	55.20	56.20	32.50	78.20	55.60	76.20	73.10
65.20	66.30	45.10	78.50	56.30	65.00	56.20	59.40	50.20	64.20
64.30	69.30	31.50	45.30	85.20	45.60	89.20	76.50	64.30	103.20
98.20	77.20	65.40	66.20	85.90	56.80	56.20	84.60	61.20	84.50
78.30	45.30	87.20	55.20	98.50	56.00	45.20	95.60	64.20	67.10
55.10	45.10	99.20	45.30	78.20	81.20	31.00	84.30	95.30	92.40
44.30	56.30	55.40	65.40	36.20	82.10	32.50	54.20	64.30	94.50
56.20	58.30	65.00	56.90	45.30	75.60	65.30	65.10	61.30	57.50
77.60	69.30	64.20	73.50	56.20	45.10	64.20	78.20	62.30	56.10
84.30	87.30	55.20	87.30	89.20	56.20	62.30	45.20	64.30	50.00
95.20	45.30	45.20	84.50	78.20	97.50	61.30	63.20	79.50	71.50
97.30	45.10	78.50	79.20	56.20	84.50	64.30	56.20	46.10	64.30
33.20	56.20	45.50	56.40	45.20	62.30	96.50	46.20	56.10	68.50
120.00	50.10	66.20	78.30	35.20	61.30	97.50	56.20	30.00	39.50
31.00	45.60	45.20	56.10	65.00	65.20	48.50	89.50	61.30	58.60
35.10	78.60	33.20	45.20	64.10	45.20	64.20	78.40	46.30	47.60
65.40	89.10	35.20	32.60	111.20	65.20	68.30	65.20	53.20	43.90
65.20	45.10	94.70	61.20	110.20	89.50	67.50	79.50	89.10	72.50
87.30	56.20	85.60	65.50	56.30	78.50	79.50	46.50	56.10	84.60
95.20	77.00	86.10	64.30	78.50	78.40	45.20	41.20	54.30	99.50
97.30	56.10	84.20	87.30	98.30	65.30	56.20	56.20	64.30	74.60
95.50	45.00	89.20	91.20	65.30	45.90	78.50	45.10	62.30	53.40
45.60	78.60	55.20	97.50	45.20	97.60	56.20	56.20	61.00	90.80
45.20	89.10	56.00	78.50	56.20	87.60	46.20	78.20	33.20	109.40
39.20	78.00	69.50	96.20	63.20	45.60	46.20	56.10	55.60	78.60
38.50	45.60	56.20	78.50	56.20	35.60	32.20	43.50	64.30	42.90
49.30	78.70	32.20	78.20	45.30	39.50	89.50	65.10	69.20	61.90
50.40	89.90	87.50	105.60	56.20	40.90	78.10	46.20	85.20	83.50
56.20	89.20	45.20	44.30	85.30	98.30	60.20	46.30	87.30	43.90
45.60	56.10	65.30	68.20	96.20	65.40	64.20	56.20	65.30	68.50
97.20	54.20	65.00	67.50	78.00	78.00	61.30	85.60	64.30	66.00
54.20	44.30	79.20	99.50	56.20	95.20	95.20	73.20	79.90	43.70
74.60	62.30	75.10	98.70	89.50	45.60	71.20	56.20	89.20	73.10
85.30	61.30	45.20	66.20	78.50	56.00	80.40	46.20	65.30	94.50
75.20	77.60	56.30	46.50	65.20	78.60	65.20	42.10	65.20	38.40
73.50	88.30	85.30	55.20	115.50	98.50	62.30	60.20	61.30	52.90
95.60	99.20	65.20	56.20	32.20	45.60	63.10	73.50	65.20	72.60
98.10	99.40	45.30	45.20	65.20	56.20	64.20	91.20	46.20	84.90
97.60	102.50	65.20	56.00	64.20	32.60	61.20	54.30	49.20	73.80
95.60	103.40	65.30	89.20	61.30	55.60	43.10	52.10	50.20	39.50