Analyzing Length or Size Biased Environmental Data

A. H. M. Rahmatullah Imon
Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA
rimon@bsu.edu

Abstract

Environmental data are often quite different from conventional statistical data in nature and warrant different types of treatments. Randomness, independence and normality are key assumptions for any conventional statistical analysis. Among them the randomness is considered as the most important one because if data are not random the entire inferential procedure breaks down although lack of randomness is more prevalent in environmental data. Faulty sampling technique is mostly responsible for nonrandomness of samples but in environmental studies, no matter how carefully we design the sampling technique, we often observe data which are biased either in length or in 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 in environmental studies we often have to deal with data which may not match with well-known distributions and nonparametric statistics is the only alternative there. In this paper we develop a procedure of analyzing environmental data which are length or size biased. For this type of data we have developed a biased correction technique first and then apply bootstrap method to 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 three basic assumptions: randomness, independence, and normality, whose existence is essential for a valid inferential statement. In addition to these three we further assume that the data are free from outliers. 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 in 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 [see Hogg *et al*. (2015)] 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. 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. Tests based on the coefficients of skewness and kurtosis such as the Jarque–Bera test or the rescaled moment test [see Imon (2003)] have become popular. If outliers cause the breakdown of normality assumption we can employ outlier detection methods for the identification of outliers. A large body of literature is now available [see Barnett & Lewis (1994), Hadi *et al*. (2009)] for the identification of outliers. Among them tests based on robust statistics [see Hampel *et al*. (1986), Maronna *et al*. (2006)] are considered to be very effective. 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 can use non-parametric method such as bootstrap [see Efron (1979)] in such a situation.

2. Bias Correction for Length or Size Biased Data

In this paper we are mostly concerned about environmental data which have bias either in length or in size or both. We have already talked about encounter data which often cases size bias. Catching fishes with a net is an excellent example of this type of bias. 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 in Figure 1. The line-intercept sampling method is popularly known as the transect sampling method in environmental studies. Transect sampling is a very popular method in agriculture especially when we randomly select plants for assessing their growth. Plants crossed by the intercept line or stick are selected as random samples for a certain study. 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 1), which will be our real interest. Thus we will typically overestimate the mean both in fish, in fiber, and in the plant possibly to a serious extent.

Figure 1. An example of transect sampling

Here we assume that we have environmental data which are either length or size biased. 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.

$$f^*(x) = xf(x)/\mu \tag{1}$$

The variable actually sampled has expected value

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

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

$$\bar{x} = \frac{\bar{x}^* \sum_{i=1}^{n} 1/x_i^*}{n} \tag{3}$$

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 (1) and (2) we obtain

$$V(X^*) = E(X^{*2}) - [E(X^*)]^2 = \int [x^3 f(x)/\mu] dx - \left[\mu\left(1+\frac{\sigma^2}{\mu^2}\right)\right]^2 \tag{4}$$

Now

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

$$= \frac{1}{\mu}\left[\mu_3 + 3\mu(\mu^2+\sigma^2) - 2\mu^3\right] = \frac{1}{\mu}\left[\mu_3 + 3\mu\,\sigma^2 + \mu^3\right] = 3\sigma^2 + \mu^2 + \frac{\mu_3}{\mu} \tag{5}$$

where $\mu_3 = E(X-\mu)^3$. Hence using (4) and (5) we obtain

$$V(X^*) = 3\sigma^2 + \mu^2 + \frac{\mu_3}{\mu} - \left[\mu\left(1+\frac{\sigma^2}{\mu^2}\right)\right]^2 = 3\sigma^2 + \mu^2 + \frac{\mu_3}{\mu} - \mu^2 - 2\sigma^2 - \frac{\sigma^4}{\mu^2}$$

$$= \sigma^2 + \frac{\mu_3}{\mu} - \frac{\sigma^4}{\mu^2} \tag{6}$$

It is not easy to analytically solve the above equation and we can employ bootstrap technique to estimate the variance as given in (6).

3. 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, Gazipur, Bangladesh using the transect sampling method.

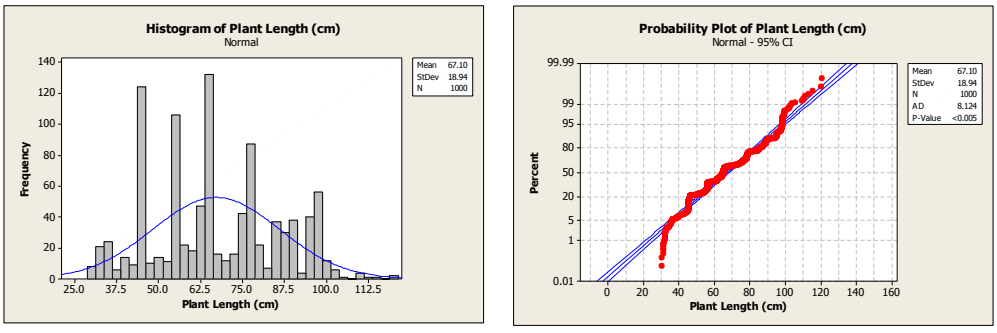

Figure 2. Histogram and normal probability plot of the lengths of peas plants

At first we check the normality assumption for the data. Figure 2 presents a histogram and normal probability plot of the length of peas plants and apparently this data does not look normal. Now we employ the rescaled moments test here. The sample skewness and kurtosis for this data are 0.155719 and 2.19055 respectively that yield the value of the RM 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. We tried to fit this data with some traditional growth distributions such as exponential, lognormal, gamma, Weibull etc but the details results are not presented for brevity. But none of these distribution adequately fits the data.

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 as shown in the following box plot.

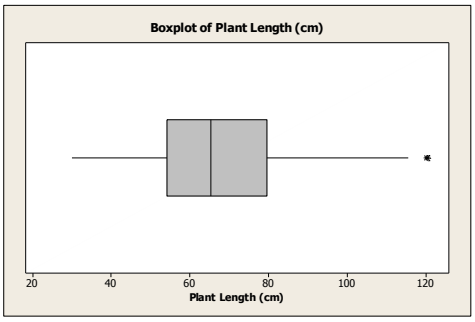

Figure 3. Box plot of the lengths of peas plants

To understand the effect of outliers we remove observations 39 and 74 and repeat all steps that we did before. Figure 4 presents the histogram of the length of peas plants without outliers. The plot looks very similar to Figure 2 and their normal probability plot shows that there is not much improvement in the results when outliers are omitted. For the full data set the value of the RM statistic was 22.24 and now after the omission of outliers it becomes 23.71. This means that the omission of outliers did not improve the normality pattern of the data.

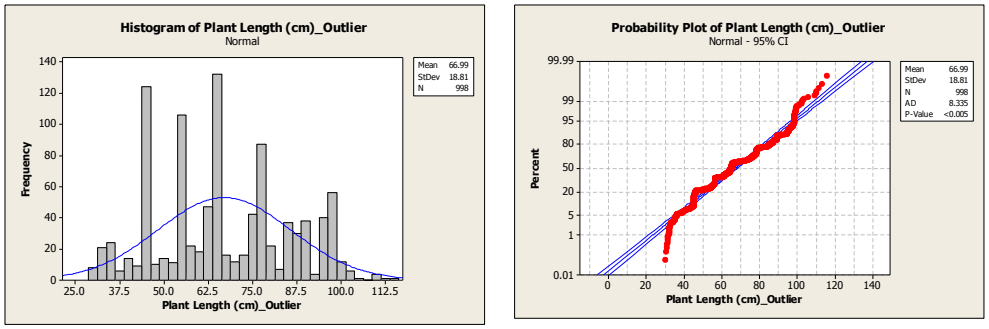

Figure 4. Histogram and normal probability plot of the lengths of peas plants without outliers

Next we move to checking another important assumption, that is the assumption of the randomness. Here we employ a very popular randomness test, called run test, as described by Hogg *et al*. (2015). 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 $r \le c_1$ or $r \ge c_2$. 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

$$\mu = \frac{2n_1\, n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma^2 = \frac{2n_1 n_2 \left(2n_1 n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2 \left(n_1 + n_2 - 1\right)} \tag{7}$$

The test statistic is

$$z = \frac{r - \mu}{\sigma} \tag{8}$$

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. When we employ the run test as mentioned in (7) and (8) we 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.

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.

. Table 1. Summary statistics for lengths of peas plants using different estimation methods

Estimation Method	Mean	Standard deviation	95% Confidence/ Bootstrap interval	95% Confidence/ Bootstrap 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 bootstrap	61.069	16.960	(60.915, 61.224)	0.309

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/bootstrap interval of the mean and also the length of the confidence/bootstrap interval. We compare five 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 method and bootstrap on bias corrected 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.

Later we employ the bias correction techniques as they were described in (1) – (6). Using (3) 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 16.96 cm which is about 1.94 cm less than the original standard deviation, which is significantly different from the original value. It is worth mentioning that we use bootstrap technique to estimate the standard deviation since a plausible estimator of the true standard deviation is not known. The above results make much sense. Since the data was collected by 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. It is intersting to note that the bootstrap mean is exacly equal to the bias

corrected mean. But the most intersting feature of this method is the confidence interval for the mean. Here the 95% bootstrap interval is (60.915, 61.224) with the length 0.309 cm. This length is much smaller than other four intervals and clearly indicates that how precisely bootstrap estimates the mean length of peas plants.

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 highly significant for a sample of size 1000 and it clearly reemphasizes our concern that when the data are length biased no matter how shophsticated method we use, unless we correct the bias we will not get the correct results. But we get a very intersting result when we look at the 95% confidence/bootstrap lengths for the means. This length is only 0.039 cm for the bootstrap method whereas they are more than 2 cm for all other methods, even for the robust and bias corrected classical methods. Although these results may look surprising, it makes much practical sense.

We must not forget the fact the first four methods use normal assumption while constructing the confidence interval. But there is clear evidence that this particular data do not follow a normal distribution. Bootstrap confidence interval is computed based on only the empirical 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.

4. Conclusions

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.

References

Barnett, V. (2004). Environmental statistics: Theory and methods. New York: Wiley.

Barnett, V. & Lewis, T.B. (1994). Outliers in statistical data, 2nd ed. New York: Wiley.

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

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

Hampel, F.R., Ronchetti, E.M., Rousseeuw P.J. & Stahel, W. (1986). Robust statistics: The approach based on influence function, New York: Wiley.

Hogg, R.V., Tanis, E.A. & Zimmerman, D.L. (2015). Probability and statistical inference, 9th 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.

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