International Journal of Mathematical Archive-4(6), 2013, 215-221
Available online through www.ijma.info ISSN 2229 – 5046

A COMPARATIVE STUDY OF THE METHODS FOR ESTIMATION OF DISTRIBUTIONAL PARAMETERS FOR LEFT CENSORED DATA WITH SINGLE NON-UNIFORM DETECTION LIMITS

Brij Kumar*, M Pandey and D. Datta

Health Physics Division, Bhabha Atomic Research Center, Mumbai-400085, India

(Received on: 17-04-12; Revised & Accepted on: 17-05-13)

ABSTRACT

A difficult step in dietary exposure assessment, which is a very important part of radiological/chemical risk analysis, is the handling of concentration data that has been reported below the detection limit (DL). These data are known as censored values or non-detects and therefore the resulting distribution of concentration values is left-censored. Handling left-censored data represents a challenge for statistical analysis of chemical/radiological data. Non detects have been so far treated with widely used substitution methods recommended by international organizations. Based on simulation a comparative study has been carried out to assess the performance of different statistical methods to handle non-detects, i.e. parametric Maximum likelihood (ML) methods, and the log-probit regression method. Monte Carlo simulations were used to evaluate statistical methods for estimating mean and standard deviation of left-censored concentration data with non-uniform detection limits. Sample size and the percentage of censoring were allowed to vary randomly to generate a variety of left-censored data sets. The log probit regression was the method that yielded high correlation coefficient (r^2= 0.92) between mean calculated using log probit method and that of the mean calculated using uncensored samples, similar were the results for the standard deviation.

Key Words: *Censored data, ROS, UMLE, RMLE.*

INTRODUCTION

Non-detects are concentration values somewhere between zero and the laboratory's detection/reporting limits. Measurements are considered too imprecise to report, so the value is commonly reported as being less than an analytical threshold and complicate the familiar computations of descriptive statistics. The worst practice when dealing with nondetects is to exclude or delete them. This produces a strong upward bias in all subsequent measures of location, such as means and medians. Whenever available, nondetects should be replaced by laboratory estimated values, which could fall below the laboratory's reporting limit for a given constituent but above the method detection limit. However, if best estimates for nondetects are not available, statistical methods for left-censored data sets must be applied. The U.S. EPA [7, 8] suggests different approaches for replacing nondetects on the basis of their percentage in the data set (e.g., < 15%, 15–50%, 51–80%, > 80%). The two extremes of this range (< 15% and > 80%) are the easiest to manage. In the situation where the percentage of nondetects is < 15%, the U.S. EPA recommends the replacement of nondetects with one-half of the detection limit (DL). When the percentage of NDS is > 80%, it is likely that no statistical method will provide a reliable measure. U.S. EPA [2] recommends the Cohen's maximum likelihood estimation method for addressing nondetects in left-censored data sets. Cohen's method is relatively unbiased and has low mean square errors [4], but it requires that the data be normally distributed with uniform detection limits. Unfortunately, environmental data sets rarely meet these conditions. They are often log-normally distributed and have nonuniform detection limits that result from sample dilution or aggregation of data collected at different times from the same site. Thus, the primary goal of this research is to test several methods that might be used to replace censored data for left-censored, log-normal data sets containing 10 to 80% nondetects with nonuniform detection limits.

MATERIALS AND METHODS

A few statistical methods are evaluated for data sets that contain between 10% and 80% non-detected/left censored values: Unbiased Cohen's Maximum Likelihood Estimation (UCMLE) method, Unbiased Restricted Maximum Likelihood Estimation (URMLE) method and the regression on order statistics (ROS) or Log Probit Regression (LPR) method. The data sets used to evaluate these methods were generated from an arbitrary parent log-normal distribution ***LN(5,1)***, where 5 is mean and 1 is standard deviation of log-transformed values. Samples of size, ***n*** (between 20 and 100) were randomly drawn from the parent distribution and ranked. The percentage of data to be designated as censored (between 10 and 80%) was randomly selected. The detection limits (***DL***) corresponding to each data set were

Corresponding author: *Brij Kumar**
Health Physics Division, Bhabha Atomic Research Center, Mumbai-400085, India

allowed to be non-uniform and were determined randomly. Each data set generated in this way was evaluated by the methods described in the next sections and then a new data set was generated with new random sample sizes, percent censoring, and non uniform detection limit. A total of 10,000 data sets were generated in this way.

Cohen's Maximum Likelihood Estimators

Consider an ordered data set $y_1 \le y_2 \le y_3 \ldots \le y_n$, where the first k observations out of n are censored. Assume that the variable y can be described adequately by a lognormal distribution. Let $x_i = \ln(y_i)$ for $i = k+1, k+2 \ldots, n$ and x_L be the natural logarithm (ln) of the detection limit DL. The likelihood function L for the data is given by:

$$L(\mu_x, \sigma_x) = \frac{n!}{(n-k)!k!}\left(\Phi\left(\frac{x_L - \mu_x}{\sigma_x}\right)\right)^k \frac{1}{\sqrt{\left(2\pi\sigma_x^2\right)^n}} \exp\left(\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{-2\sigma_x^2}\right) \quad (1)$$

with Φ the cumulative distribution function of a standard normal variate, μ_x the mean and σ_x the standard deviation of the log transformed data. The maximum likelihood estimates, $\hat{\mu}_x$ and $\hat{\sigma}_x$ of μ_x and σ_x can be found by calculating the values of μ_x and σ_x that maximize the function L. By taking natural logarithm of (1) and setting the partial derivatives with respect to μ_x and σ_x to zero, maximum likelihood estimators $\hat{\mu}_x$ and $\hat{\sigma}_x$ can be calculated. The maximum likelihood method has been studied extensively. The most useful results are those of Cohen [1,2], who gives the following maximum likelihood estimators in terms of a tabulated function of two arguments:

$$\hat{\mu}_{MLE} = \bar{x}_o - (\bar{x}_o - x_L) * \lambda(\gamma, h) \quad (2)$$

$$\hat{\sigma}_{MLE} = s_o^2 + (\bar{x}_o - x_L)^2 * \lambda(\gamma, h) \quad (3)$$

where, $\bar{x}_o$ and s_o^2 are mean and variance calculated from the observed data i.e. the values which are above the limit of detection (x_L), $\gamma = s_o^2/(\bar{x}_o - x_L)^2$ and $h = k/n$ is the fraction of samples that has been truncated (i.e., the number of non quantifiable observations divided by the total sample size) . In his original development, Cohen [2] provided tables of the function $\lambda(\gamma, h)$ however, these are difficult to use for computer purposes. Therefore, we have used the following power series expansion of this function developed by Hass and Scheff [5], that fits the tabulated values of the function to within a 6% relative error:

$$\begin{aligned} \ln\lambda(\gamma,h) \;=\; & 0.182344 - 0.3756/(\gamma+1) + 0.10017*\gamma + 0.78079*\beta \\ & - 0.00581*\gamma^2 - 0.06642*\beta^2 - 0.0234*\gamma*\beta + 0.000174*\gamma^3 \\ & + 0.001663*\gamma^2*\beta - 0.00086*\gamma*\beta^2 - 0.00653*\beta^3 \end{aligned} \quad (4)$$

where $\beta = \ln[h/(1-h)]$

Bias-Corrected Maximum Likelihood Estimators:

The estimates $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}$ given by equations (2) and (3) are biased. Saw [3] computed the first-order bias correction terms to the maximum likelihood estimates of the mean and standard deviation of a type II censored normal sample. A type II censored normal sample is one in which a constant proportion of observations are censored, rather than all observations below a fixed value. Schneider [4] reduced these results to the simple computational formulas

$$B(\hat{\mu}) = -\exp\{2.692 - 5.439*(n-k)/(n+1)\} \quad (5)$$

$$B(\hat{\sigma}) = -\{0.312 + 0.859*(n-k)/(n+1)\}^{-2} \quad (6)$$

In practice, the bias corrections given by equations (5) and (6) are also used for Type I censored data. The bias-corrected MLEs or unbiased MLEs, denoted by UMLE, are given as follows.

$$\hat{\mu}_{UMLE} = \hat{\mu}_{MLE} - \hat{\sigma}_{MLE} * B(\hat{\mu})/(n+1) \quad (7)$$

$$\hat{\sigma}_{UMLE} = \hat{\sigma}_{MLE} - \hat{\sigma}_{MLE} * B(\hat{\sigma})/(n+1) \quad (8)$$

Note that, if no censoring exists, then the ordinary values of sample mean and variance are used. Also, if less than two uncensored data points exist, the value of S_o^2 is not computable, hence, neither are the maximum likelihood estimators by this method. It should be noted that the tables in Cohen [2] only provide values of λ up to a value of $\gamma = 1$. Schneider [4] extended these tables to values of $\gamma = 1.48$, and Schmee *et. al.* (10) extended these tables to $\gamma = 10$. The adequacy of the numerical approximation beyond this bound cannot be ascertained. It should also be observed that the maximum likelihood method can be obtained in other ways as well.

Bias-Corrected Restricted Maximum Likelihood Estimators:

Although less frequently considered for calculating summary statistics from censored environmental data, the one-step restricted maximum likelihood estimators developed by Persson and Rootzen [9], are somewhat simpler to compute. The method provides the following explicit solution to maximize the likelihood function Eq. (1) for the mean and standard deviation by imposing an assumption that the number of observations below the censoring limit follows a binomial distribution. Estimators of the mean and standard deviation for censored normally distributed data x_i are given by:

$$\hat{\mu}_{RML} = x_L - \Phi^{-1}(h) * \hat{\sigma}_{RML} \tag{9}$$

$$\hat{\sigma}_{RML} = 0.5\left(\frac{a\,\Phi^{-1}(h)}{(n-k)} + \sqrt{\left(\frac{a\,\Phi^{-1}(h)}{(n-k)} \right)^2 + 4\frac{b}{(n-k)}} \right) \tag{10}$$

where $\Phi^{-1}(h)$ is the inverse cumulative normal distribution function evaluated at *h*, the proportion of censored data. The parameters *a* and *b* are calculated as

$$a = \sum_{i=k+1}^{n} (x_i - x_L) \quad \text{and} \quad b = \sum_{i=k+1}^{n} (x_i - x_L)^2 \tag{11}$$

Since the estimators are not asymptotically unbiased at low degrees of censoring, the following bias-corrected estimators of the mean and standard deviation were suggested by Haas and Scheff [5],

$$\hat{\mu}_{URML} = \frac{a'}{(n-k)} - \xi\,\hat{\sigma}_{RML} \tag{12}$$

$$\hat{\sigma}_{URML} = \sqrt{\frac{b'}{(n-k)} - \left(\frac{a'}{(n-k)} \right)^2 - \left(\xi\,\Phi^{-1}(h) - \xi^2 \right) \hat{\sigma}^2_{RML}} \tag{13}$$

where $a' = \sum_{i=k+1}^{n} x_i$, $b' = \sum_{i=k+1}^{n} x_i^2$ and the correction term is given by

$$\xi = \frac{n}{(n-k)\sqrt{2\pi}} \exp\left(-0.5\left(\Phi^{-1}(h) \right)^2 \right) \tag{14}$$

Log Probit Regression (LPR) or Regression on Order Statistics (ROS):

A number of potentially robust methods are available using the normal scores for the order statistics. We take the one here recommended by Gilliom and Helsel [6]. A logarithmic transformation has yielded ***k*** observations x_i, ***i = 1, 2, ….k*** each below a common transformed detection limit x_L and ***(n-k)*** observations x_i, ***i = k+1, k+2, ….(n-k)*** that are observed and greater than x_L. The observations having common mean μ_x and variance σ_x^2 will satisfy the equation

$$x_i = \mu_x + \sigma_x \Phi^{-1}(P_i) \tag{15}$$

where $P_i = Prob\{Y_i \le y_i\}$ and $\Phi^{-1}(.)$ denotes the inverse of the cumulative normal distribution function. This suggests that the intercept and slope from a regression on the normal scores would yield the mean and variance of the transformed observations. The regression is performed on the inverse transformed adjusted order statistics. It should be

noted here that if the procedure is truly robust to departures from normality, the original untransformed observations y_i and detection limits ***DL*** could be used. Our simulations, presented later in the later section, suggest that regressing on the order statistics is robust for log-normal populations. The commonly accepted procedure is to replace the probabilities by the adjusted ranks, so that the regression equation becomes

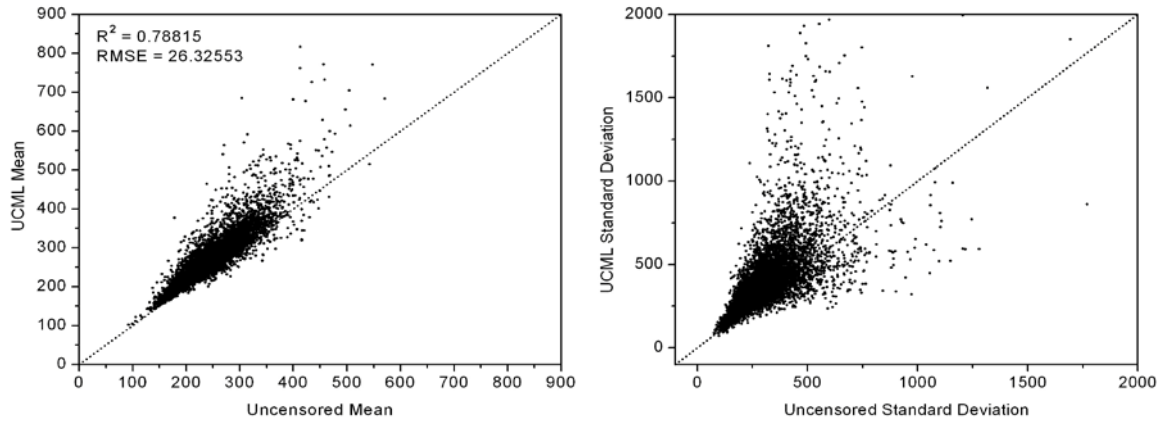
$$x_i = \mu_x + \sigma_x \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right) + \varepsilon_i \qquad (16)$$

where, $i = k+1, k+2, \ldots, (n-k)$, with the estimators for μ_x and σ_x estimated by least squares. This implies that the residual errors ε_i are assumed to have equal variance and are uncorrelated. Helsel and Gilliom [6] recommend using ordinary least squares as an easier computational alternative. Their procedure uses the predicted values from the regression model (16) for $i = 1, 2, \ldots, k$ for the censored values. Then, transform back to the original observations y_i and compute the usual mean and variance from the resulting sample consisting of predicted values for the k censored/BDL observations from the model (16) and the $(n-k)$ observed/detected values from the original sample. Note that the procedure produces estimators for the censored values based on extrapolation from a normal model for the transformed values and then back-transforming to the original raw observations. We have applied the bootstrap for estimating the variance of the parameter estimators produced by the ROS/LPR procedure since no theoretical arguments have been given in the literature for preferring another method.

RESULTS AND DISCUSSIONS

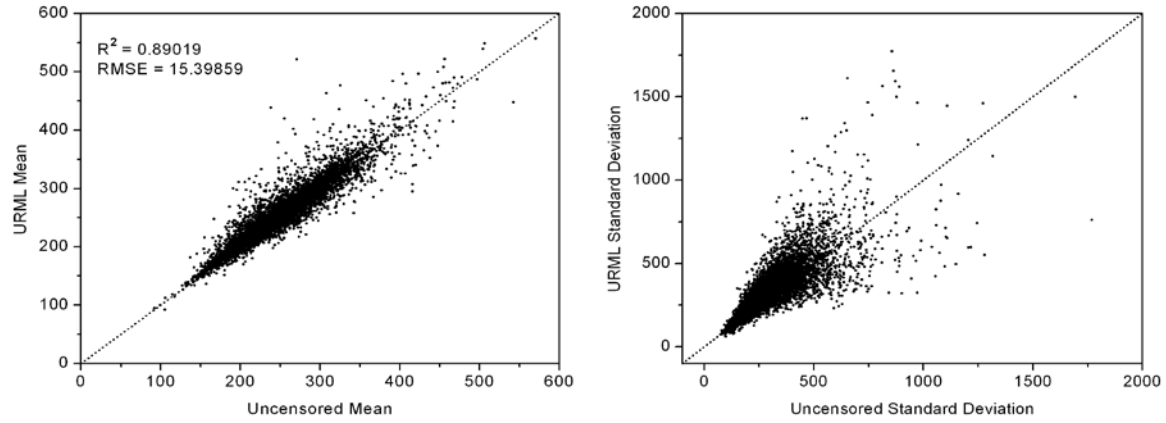
The statistical methods described herein were evaluated by comparing means for the left-censored data sets to both the mean for the uncensored data sets and to the true mean of the data. Correlation coefficients (R^2) between the means of left-censored versus uncensored data were calculated along with the root mean square error. Each of the above mentioned statistical methods for left-censored data sets were applied to the 10,000 data sets randomly generated from the arbitrary parent log-normal distribution LN (5, 1). The means calculated by the above methods for left censored sample were compared with the means calculated for the original non-censored sample to assess the effectiveness of the statistical methods in dealing with left-censored data (Figs. 1 to 3).

The UCML method on average tends to slightly over-estimate the uncensored mean (Fig. 1A). The correlation coefficient between the means of UCML means and uncensored means is found to be $R^2 = 0.78815$, which shows a poor correlation as well as root mean square error is also large (RMSE = 26.32553). The correlation between standard deviations from UCLM method and that of uncensored standard deviation is also very large as evident from (Fig 1B). URML method shows a little improvement over the UCML method (Fig 2A). The correlation coefficient between URML mean and uncensored mean ($R^2 = 0.89019$) is more, indicating a stronger correlation between the two, also the root mean square is less (RMSE = 15.39859). The correlation between standard deviations of URML method and that of uncensored samples is also improved (Fig 2B). The LPR/ROS method however supersedes over the other two methods as can be seen from (Fig 3A), because this method uses a regression to derive the relationship between the detected data and the underlying distribution, the data imputed for the non detects also fits the underlying distribution. Consequently, both the mean and standard deviation of a LPR/ROS data set is generally similar to that of the corresponding uncensored data set. The correlation coefficient between the LPR/ROS means and uncensored sample means is very high ($R^2 = 0.92228$), also the root mean square error is very less (RMSE = 12.66111). The correlation between the standard deviations of LPR/ROS method and uncensored sample standard deviation is very strong ($R^2 = 0.99238$), however on average the LPR/ROS method slightly underestimates the standard deviation as evident from (Fig 3B). LPR/ROS method yields statistical summaries for left censored data that are nearly the same as if all the data had been detected. However, it should be noted that data sets evaluated in this simulation suit the LPR/ROS method because the data sets are drawn from a parent log-normal distribution. The LPR/ROS method might not be quite as effective for data sets that are not strictly log-normal or are drawn from more than one parent distribution. We have also examined the relationship between means from these methods and other study variables, i.e. % censoring and sample size (Figs. 4 and 5). Despite the occasional outliers, most means calculated by these methods are relatively insensitive to % censoring, with most means ranging between 150 and 350. As can be seen from Fig 4 and Fig 5, LPR/ROS method is the one that behaves very similar to that of the uncensored samples both for sample size and % censoring.



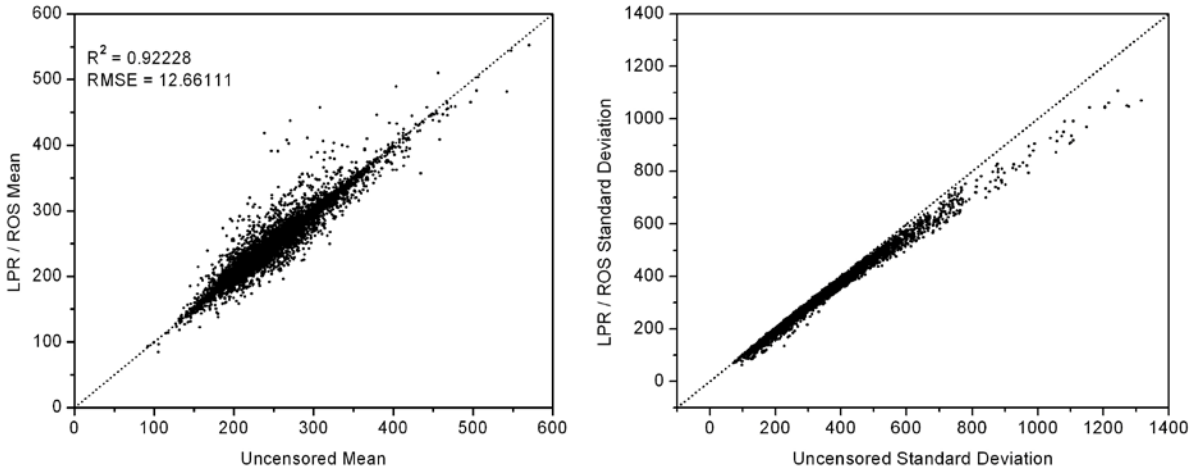
(A) (B)

Figure 1: (A) Unbiased Cohen's Maximum Likelihood (UCML) mean vs noncensored mean. (B) Unbiased Cohen's Maximumu Likelihood (UCML) standard deviation vs noncensored standard deviation. Correlation coefficients (R^2) shows the relationship between noncensored and UCML means and the dotted line indicates the theoretical 1:1 fit



(A) (B)

Figure 2: (A) Unbiased Restricted Maximum Likelihood (URML) mean vs noncensored mean. (B) Unbiased Restricted Maximumu Likelihood (URML) standard deviation vs noncensored standard deviation. Correlation coefficients (R^2) shows the relationship between noncensored and URML means and the dotted line indicates the theoretical 1:1 fit



(A) (B)

Figure 3: (A) Log Probit or Regression on Oredr Statistics (LPR/ROS) mean vs noncensored mean. (B) Log Probit or Regression on Oredr Statistics (LPR/ROS) standard deviation vs noncensored standard deviation. Correlatio n coefficients (R^2) shows the relationship between noncensored and LPR/ROS means and the dotted line indicates the theoretical 1:1 fit

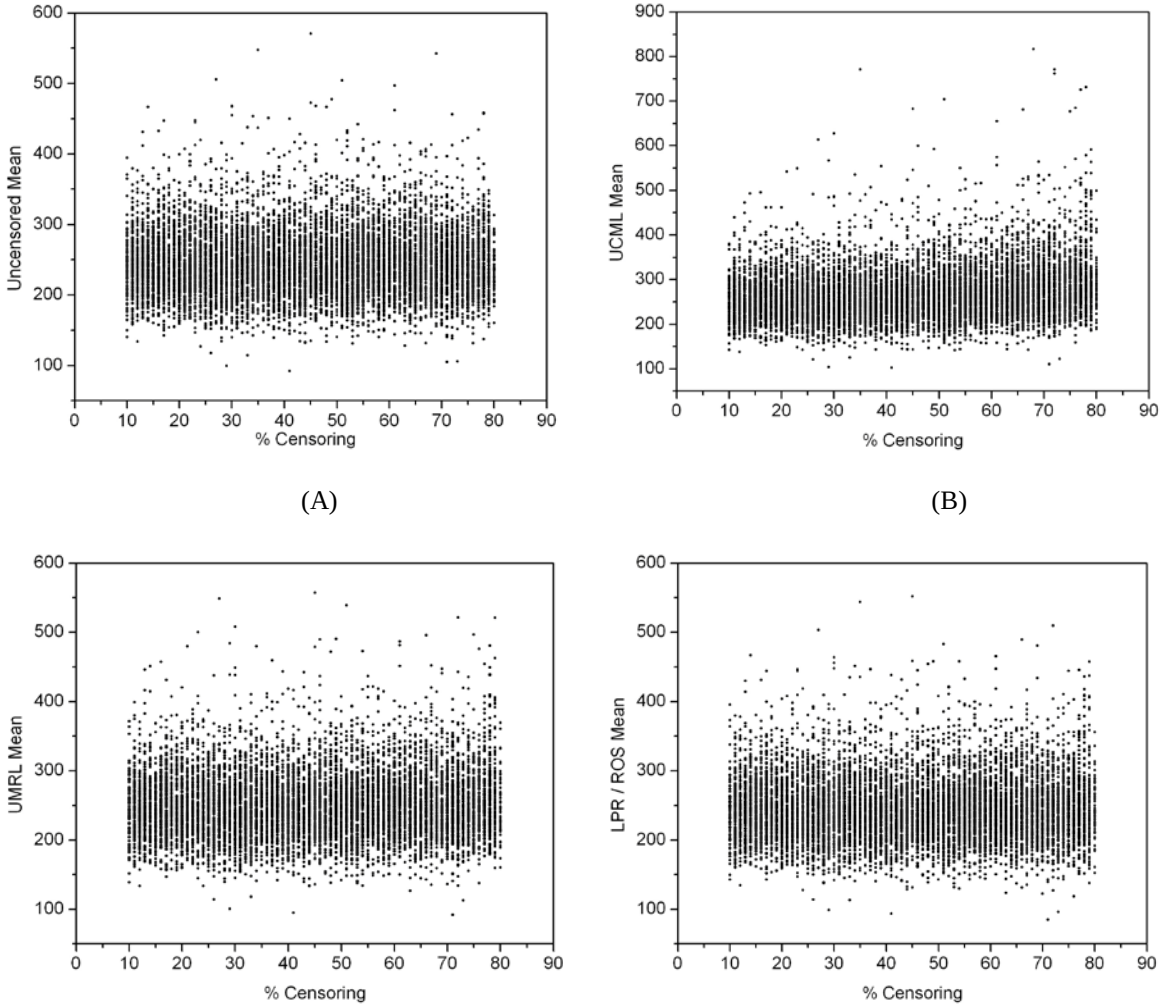
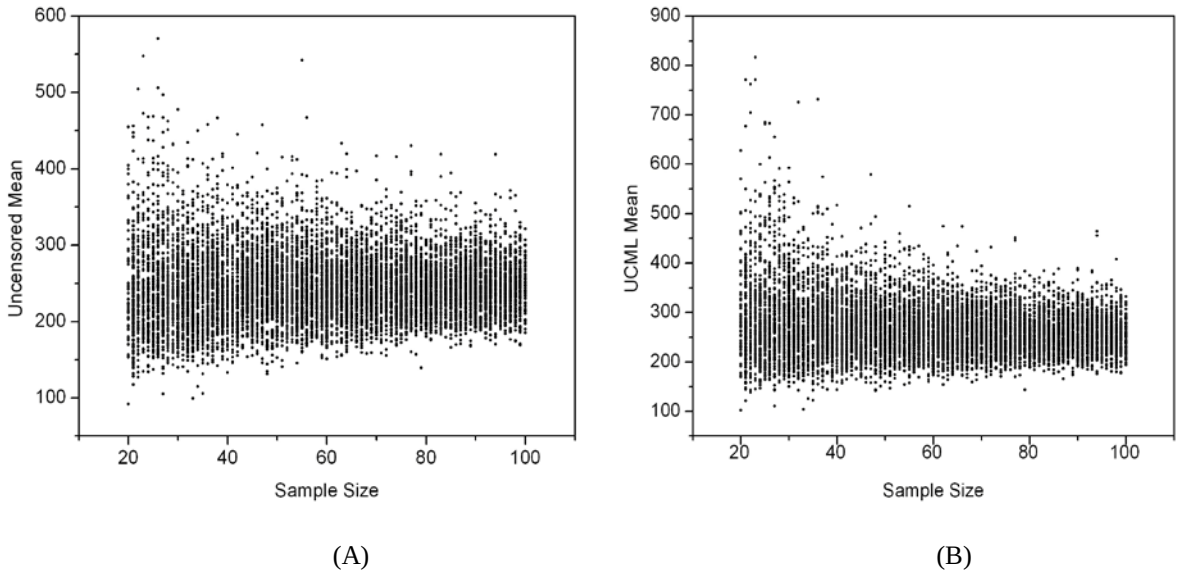
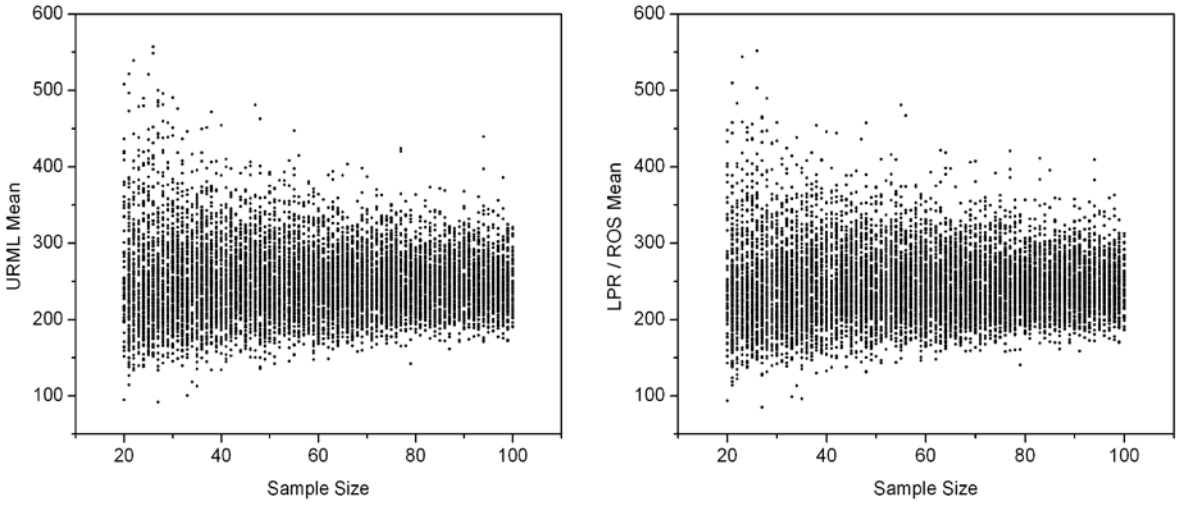


Figure 4: (A) % Censoring vs uncensored mean. (B) % Censoring vs UCML mean (C) % Censoring vs URML mean (D) % Censoring vs LPR/ROS mean





(C) (D)

Figure 5: (A) Sample size vs uncensored mean. (B) Sample size vs UCML mean (C) Sample size vs URML mean (D) Sample size vs LPR/ROS mean

CONCLUSION

Log probit method or regression on order statistics method (LPR/ROS) performs better for estimation of summary statistics even in the cases of high censoring (upto 80%), than the other two methods i.e. unbiased Cohen's maximum likelihood method (UCLM) and unbiased restricted maximum likelihood method (URML), for the data that comes from a parent Log normal distribution. For the data that comes from other than log normal distribution or from mixed distributions, other alternatives can be explored for estimation of the descriptive statistics.

REFERENCES

1. Cohen, A.C. *Simplified estimators for the normal distribution when samples are singly censored or truncated.* Technometrics, 1:217-237, 1959.
2. Cohen, A.C. *Tables for maximum likelihood estimates: singly truncated and singly censored samples.* Technometrics, 3:535-541, 1961.
3. Saw J. G., *Estimation of the Normal Population Parameters Given a Type I Censored Sample*, Vol. 48, No. 3/4 (Dec., 1961), pp. 367-377
4. Schneider, H. *Truncated and Censored Samples from Normal Populations*; Marcel Dekker: New York, 1986.
5. Haas C. N. and Scheff P. A. *Estimation of averages in truncated samples*. Environmental Science and Technology, 24:912-919, 1990.
6. Gilliom, J.R. and Helsel, D.R. *Estimation of Distributional Parameter for censored Trace Level Water Quality Data 1. Estimation Techniques*, Water Resources Research22 (2), 135–146, 1986.
7. U.S. Environmental Protection Agency. 2000. *Guidance for data quality assessment: Practical methods for data analysis.* EPA/600/R-96/084. Office of Environmental Information, Washington, DC.
8. U.S. Environmental Protection Agency. 2002. *Calculating upper confidence limits for exposure point concentrations at hazardous waste sites.* OSWER 9285.6-10. Office of Solid Waste and Emergency Response, Washington, DC
9. Persson T and Rootzen H. *Simple and highly efficient estimators for a type I censored normal sample.* Biometrika, 64:123-128, 1977.

Source of support: Nil, Conflict of interest: None Declared

