Interdiscip Sci Comput Life Sci (2017) 9:153–172
DOI 10.1007/s12539-015-0132-9

ORIGINAL RESEARCH ARTICLE

Evaluating Left-Censored Data Through Substitution, Parametric, Semi-parametric, and Nonparametric Methods: A Simulation Study

Mustafa Agah Tekindal[1] · Beyza Doğanay Erdoğan[2] · Yasemin Yavuz[2]

Received: 24 June 2015 / Revised: 12 October 2015 / Accepted: 11 November 2015 / Published online: 21 November 2015

Abstract In this study, an attempt was made to determine the degrees of bias in particular sampling sizes and methods. The aim of the study was to determine deviations from the median, the mean, and the standard deviation (SD) in different sample sizes and at different censoring rates for log-normal, exponential, and Weibull distributions in the case of full and censored data sampling. Thus, the concept of "censoring" and censoring types was handled in the first place. Then substitution, parametric (MLE), nonparametric (KM), and semi-parametric (ROS) methods were introduced for the evaluation of left-censored observations. Within the scope of the present study, the data were produced uncensored based on the different parameters of each distribution. Then the datasets were left-censored at the ratios of 5, 25, 45, and 65 %. The censored data were estimated through substitution (LOD and LOD/$\sqrt{2}$), parametric (MLE), semi-parametric (ROS), and nonparametric (KM) methods. In addition, evaluation was made by increasing the sample size from 20 to 300 by tens. Performance comparison was made between the uncensored dataset and the censored dataset on the basis of deviations from the median, the mean, and the SD. The results of simulation studies show that LOD/$\sqrt{2}$ and ROS methods give better results than other methods in deviation from the mean in different sample sizes and at different censoring rates, while ROS gives better results than other methods in deviation from the median in almost all sample sizes and at almost all censoring rates.

Keywords Left-censor · Kaplan–Meier estimator · Measurement uncertainty · Regression on order statistics · Limit of detection · Limit of quantification

✉ Mustafa Agah Tekindal
matekindal@gmail.com

Beyza Doğanay Erdoğan
beyzadoganay@gmail.com

Yasemin Yavuz
genc@medicine.ankara.edu.tr

[1] Department of Biostatistics, School of Medicine, Izmir University, Izmir, Turkey

[2] Department of Biostatistics, School of Medicine, Ankara University, Ankara, Turkey

1 Introduction

Study designs containing censored measurements are used in most of research carried out in applied disciplines. It may not always be possible to observe the decay time of all constituents of a system or a process while making inferences about its reliability. For example; in a life test conducted to obtain information about the lifetime of an expensive electronic part, it may be not be favorable to observe the decay of all parts because it may increase cost and duration of testing, or data related to patients treated in a clinic may not be observed perfectly. In such cases, post-experiment or observation censored data are obtained. Censored data are confronted in medicine, biology, food, engineering, quality control, and many other fields.

Limit of detection (LOD) is simply defined by Strobel and Heineman [1] as the smallest analyte concentration of a variable that can be defined as statistically different. Two sets of measurement are needed to determine the limit. Besides a regular set, several measurements (10–20 are recommended) need to be carried out on the draft having the same conditions. As an equation, LOD can be defined as in the Eq. 1 provided by Miler and Miller [2]:

$$y - y_B = ks_B \quad (1)$$

y is the response produced by the analyte existing in the limit of detection concentration; y_B is "draft response"; and s_B is the standard deviation (SD) of the draft response (sample). k is the correction value to ensure the acceptable levels of wrong positive or wrong negative situations in the process of determining the limit of detection. k must never be less than +1 as any actual value, and k +3 value must be used as recommended by the International Union of Pure and Applied Chemistry (IUPAC).

Since the limit of detection is definitely a positive value, the definition of limit of detection given in the Eq. 2 is obtained:

$$\mathrm{TL} \equiv (\mathrm{LOD}) \equiv \frac{ks_{\mathrm{draft}}}{|b|} \tag{2}$$

In this equation, $|b|$ is the size of slope, and s_{draft} is the sample SD of the draft match responses.

Since the limit of detection defined in the Eq. 2 is a sample test statistic, the Eq. 3 is defined for the population.

$$\mathrm{TL}(\Lambda) \equiv \mathrm{trueLOD} \equiv \frac{k\sigma_{\mathrm{draft}}}{|\beta|} \tag{3}$$

Here, $|\beta|$ is the size of the actual slope of the underlying linear measurement model (analytical size sensitivity) and is the population SD of the draft response. The actual limit of detection defined in the Eq. 3 is faultless and is always bigger than zero. However, the sample limits of detection defined in the Eq. 2 are the functions of two random variables: the calculation of the analytical sensitivity of the measuring tool and uncertainty measured as draft. Just like any quantity calculated from one or more measured results, any measured result is a random variable and a sample obtained from certain distributions of possible results. As a result, the limits of detection obtained from the sample are random variables and are characterized by probability density functions that determine their own statistical properties. As defined in the Eq. 2 in particular, the probability density function of the test statistic of a limit of detection determines its estimation value (e.g., the mean of the population), its accuracy (e.g., the SD of the population or its "significant figures" in general terms), and its confidence limits (e.g., 95 %).

The LOD value is determined as the value slightly exceeding the SD value +3 around 0 [3].

The LOQ value is determined as the SD value +10 around 0. Censoring limit can be taken as LOD or LOQ depending on the researcher's or the research's purpose. However, the values between LOD and LOQ are neither absolutely reliable nor absolutely unreliable. However, those values which are above LOQ value are completely reliable [3].

The statistical methods used for the left-censored observations of chemical contaminants are divided into four general categories: substitution methods [4, 5], semi-parametric methods [regression on order statistics (ROS)] [6–8], parametric methods [maximum-likelihood estimation (MLE) methods] [9], and nonparametric methods [Kaplan–Meier estimation (KM)] [10, 11]. Hewett and Ganser conducted a study comparing different approaches to processing censored data [12].

Table 1 summarizes the studies evaluating the performance of different approaches to processing left-censored data statistically, as focused on in Hewett and Ganser [12]. These studies vary in terms of the statistical methods evaluated, sample sizes, the number of repetitions, evaluation statistics (mean, the 95th percentile, median, other descriptive statistics), deviation, which refers to the difference between the calculation and the unknown actual value, and mean squared error, which gives information about the accuracy of a specific quantity to calculate the relevant parameter. The results obtained from the evaluation studies are quite different, while ROS and/or MLE were preferred in some studies, KM was recommended in others (see Table 1). Hewett and Ganser compared substitution, ROS, MLE, and two nonparametric methods by producing data with log-normal distribution or mixes of data with log-normal distribution at random in simulation datasets involving simple and more complex censoring mechanisms. Hewett and Ganser concluded that there is no method which is inarguably superior to other methods in all scenarios. In their study, they defined scenario as the combination of concentration values and different number of LOD values in datasets. Hewett and Ganser found that nonparametric methods are not reliable in the scenarios where log-normal distribution mixes are produced. In addition, in their study, complex data sets having only three LOD values were produced. If the number of LOD values increases, the performance of KM method may improve.

In their evaluation on the statistical methods used for processing undetected values, Antweiler and Taylor [13] did not use left-censored data produced in computer. Instead, they employed two different analytical techniques having different LOD values and thus censoring at different levels, thereby making use of environmental samples on which metal analysis was made. While the dataset with the highest LOD value was deemed censored, the other dataset involved actual uncensored values. They evaluated five different methods: substitution, MLE, ROS, nonparametric methods, and methods produced by use of tools (the signal-to-noise ratios which are below 3:1, which is the generally accepted value). This method is referred to as "lab" in Table 1. The result of that study indicates that sample type and the number of samples have very little effect on the quality of results obtained through different methods, but the degree of censoring has a big influence on it. In general, when the percentage of censoring was less than 70 %, deviation value was found to be low. In these cases, the best general technique was seen to be the KM method. ROS

Table 1 The summary of the evaluation studies on statistically handling left-censored data

Authors	Methods	Sample size	The number of repetitions	Distribution produced	Censoring (%)	Evaluation parameter	Evaluation statistic	Results
Gilliom and Helsel [22]	LOD/2, LOD, MLE	10, 20, 50	500	Log-normal, the mixture of two log-normals	0; 20, 40, 60, 80	MSE	Mean, median, interquartile range	MLE was preferred
Helsel and Cohn [23]	LOD/2, LOD, ROS, MLE	25	500	Log-normal, the mixture of two log-normals	20, 40, 80	MSE	Mean, median, interquartile range	When the distribution used is separated from log-normal distribution, ROS is superior
Kroll and Stedinger [25]	ROS, MLE	10, 25, 50		Log-normal, the mixture of two log-normals, gamma, delta	0; 10, 20, 40, 60, 80	MLE		It is difficult to interpret the results. When censoring was <60 %, the methods yielded similar results. However, MLE was superior when censoring was done at a higher rate
Schmoyer et al. [10]	MLE, KM	10	500	Log-normal, Log-normal, gamma	0; 25, 50, 75		Mean	In general, KM performed better than MLE
She [11]	LOD/2, ROS, MLE, KM	21	1000		At random between 10 and 80	Deviation, MSE		KM performs better than MLE, ROS, and LOD/2
Shumway [24]	ROS, MLE	20,50		Log-normal	50, 80		Mean, variance	No method is consistently better. The choice depends on censoring percentage and deviation from log-normal distribution
Hewett and Ganser [12]	LOD/2, ROS, MLE, KM		100	Log-normal	At random between 10 and 80	Deviation, MSE	Mean, 95 % confidence levels	Only MLE and ROS were successful in all scenarios. KM did not perform well
Antweiler and Taylor [13]	0 (Zero), LOD/2, LOD, MLE, ROS, KM, Lab	34–841	44 (a total of 5000 samples)	No distribution was produced. Actual samples were analyzed through two different techniques	At random between 14 and 95	Deviation	Mean, median, 25th percentile, 75th percentile, SD, interquartile range	KM was preferred when censoring percentage was lower than 70 %. However, ROS, LOD/2, and random value are appropriate alternatives. MLE performed badly. No method yielded a good result when censoring rate was higher than 70 %

and putting LOD/2 and a random value between 0 and LOD, which are two substitution techniques, were determined to be adequate alternatives. The lab method performed worse than the above-mentioned methods. When the censoring rate was over 70 %, none of the methods was able to make a correct and precise estimation of the actual value.

On the other hand, the study of Hewett and Ganser involved two important limitations. They mostly used a range of normal distribution and log-normal distribution regardless of the above-mentioned other distributions.

In addition, a study was carried out on bone leaf concentrations exposed to random left censoring. They were measured in vivo via X-ray fluorescence. In the analysis, the inverse-variance weighting of the measurements was used for estimating the mean and the standard error of the dataset. The methods were applied to the six datasets obtained from epidemiologic studies, and it was concluded that the proposed statistical analysis methods could allow making valid inferences about bone lead concentrations [14]. A comparative simulation study was carried out in order to evaluate perfect estimations in different censoring cases involving different sample sizes, and it was concluded that the appropriateness of different methods varies depending on different conditions [15].

1.1 Substitution Methods

The method of substituting other values for values that cannot be detected in studies including left-censored data is widely used. Some suggestions have been made for substituting LOD/2, LOD/$\sqrt{2}$, 0 (Zero), and LOD values for the samples that cannot be detected depending on the rate of value that cannot be detected. Similar principles are recommended for nonquantitative values. The substitution methods are deemed quite biased. In data, such biasness manifests itself as actual diversity function in the dimension of sample and in the percentage of censored observation [16]. Another disadvantage of the substitution methods is that the complete distribution of the detected samples is not taken into consideration. In other words, whether the detectable data rate of a sample is 1 or 90 % does not affect the way the undetected data are processed. These two different sample types may have different distributions in practice. The most critical situations for the substitution methods are those situations in which there is more than one LOD value (Helsel 2005). This is because the substituted values depend on the conditions that determine the limit of detection such as laboratory prevision or sample matrix interferences. These factors may not always have something to do with the actual value (Helsel 2005).

Despite their disadvantages, the substitution methods are still widely used because they are mostly easy to apply; they are understood by many people; and they allow making estimations (high average, low diversity) in upper limit exposure evaluation calculations.

1.1.1 Parametric Methods

It is possible to use models appropriate to certain non-negative probability distributions such as exponential, log-normal, normal, and Weibull. The maximum-likelihood estimation (MLE) is used as a parametric method.

The maximum-likelihood estimation (MLE) consists of the estimations that make the probability function maximum after the parametric distribution that fits the data best is defined. This method is based on the studies of Fisher [17].

In the MLE method, three types of information are used [8]:

1. Values above the LOD;
2. The rate of the data below the LOD;
3. Assumption about the way positive values are distributed.

It is assumed that the data below and above the LOD have a specific statistical distribution. It is estimated that parameters related to the selected distribution accommodate themselves to the values observed above the limit of detection perfectly and to the values below the limit of detection adequately. The estimated parameters are the parameters that make the probability function maximum.

Thus, in left-censored data;

$$L = \prod p(x_i)^{\delta_i} \times F(x_i)^{1-\delta_i} \tag{4}$$

MLE is expressed as indicated in the Eq. 4 refers to the censoring situation.

$$\delta_i = \begin{cases} 0 & \text{Censored observation} \\ 1 & \text{Uncensored observation} \end{cases} \tag{5}$$

The maximum-likelihood method will be used most for parameter estimation.

Let us assume that ti is the observation time of the *i*th data in a study where the number of observations is *n*. If all the observations are censored at ti, the probability function of the *i*th data can be expressed as follows:

$$L_i = f(t_i) = h(t_i)S(t_i) \tag{6}$$

If the data are censored, the probability function of the *i*th data will be equal to the observation function as follows:

$$L_i = S(t_i) \tag{7}$$

The probability function of the distribution where censored and uncensored data are together is as follows:

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-_i\delta} = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} [S(t_i)] \quad (8)$$

Here, as showed in the Eq. 5, censored observation is treated as =0, while uncensored observation is treated as =1. If the natural logarithm of the Eq. 8 is taken, the logarithmic probability function will be as follows:

$$\log L = \ell = \sum_{i=1}^{n} \{\delta_i \log h(t_i) + \log S(t_i)\}$$
$$= \sum_{i=1}^{n} \{\delta_i \log h(t_i) - H(t_i)\} \quad (9)$$

The maximum-likelihood estimator can be calculated under various distribution (normal, log-normal, exponential, Weibull) assumptions. The fact that this method involves assumptions for left-censored data brings about certain limitations.

1.1.2 Nonparametric Methods

The reason why nonparametric methods are called nonparametric is that they do not involve the calculation of "parameters" like the mean or the SD of a specific distribution. Instead, related data size (sequence) is used. Standard nonparametric technique used for censored data is the Kaplan–Meier (KM) method [18]. In this method, empirical cumulative distribution function is taken as basis. It was initially developed to calculate the average of the (right) censored survival data (e.g., it can be used for measuring the rate of patients who survive for a specific period after treatment in medical research). A series of descending horizontal steps are followed while the survival function is being calculated via the Kaplan–Meier method. When an adequately large sample is taken from it, the actual survival function of the relevant population is approached. It is assumed that the survival function value remains stable in different examinations on different samples taken one after another.

The KM method was proposed to be used for analyzing left-censored concentration data in environmental research [10, 11]. The advantage of this approach is that when there are undetectable values, mean can be calculated based on median and other distributions without having to rely upon distribution assumptions [19]. With the KM method, the weight of censored data is distributed to different observation data below censored data (LOD, LOQ, and Zero). Therefore, if there is a single LOD value, there is no need to apply the KM method. This is because it would be equal to substituting zero or the biggest observed data below the LOD for censored data.

Since it is not parametric, the KM method is not sensitive to outliers that are frequently encountered in environmental data [13]. This method is available in a lot of statistical packages, but since it was initially intended for right-censored data, concentration data have to be reversed prior to analysis (i.e., left-censored dataset has to be converted into right-censored dataset).

The Kaplan–Meier (K–M) method has always been taken as a standard method for the estimation of the summary statistics of censored survival data. Though this method has mainly been used for right-censored data, it can also be applied to left-censored data. Several authors taking the basic algorithm as basis have revealed various algorithms of the Kaplan–Meier method. The results obtained from all algorithms are almost the same. Summary statistics for left-censored data can be calculated through the Kaplan–Meier method in two different ways:

- Converting censored data from left to right [8].
- Directly using left-censored data [14].

To Helsel, left-censored data have to be reversed through a conversion process prior to the use of the KM method, and then survival probabilities have to be calculated in a similar way to right-censored data [8]. Finally, results have to be reconverted while making estimations about left-censored data (mean, median, and other percentages). Here, the value found to be highest (left-censored value) is going to be at the bottom (Eq. 10). S is the probability of survival, and such probability is the product of the increasing probabilities that indicate the probability of survival until the next lowest limit of detection (considering the number of data in such limit of detection or below it).

For each subject, $i = 1, \ldots, n$ (considering both censored and observed values), right-censored data can be produced by reversing left-censored data. After all of the left-censored values are arranged in a descending order, all values of the left-censored data are subtracted from a value that is bigger than the biggest value of the dataset, thereby turning into right-censored data. They are assigned in this way. Then all of the right-censored data will automatically be assigned in an ascending order. After that, all the observations (both the observed ones and the censored ones) are arranged from the lowest to the highest. At this point, the right-censored value will be as follows:

$$\text{Reversed } j = M - x_i \quad (10)$$

Here, M refers to a constant that is bigger than the observed biggest value of the dataset, while x_i refers to the observed data.

The number of observations under risk shows both detected and censored data at each observed value or below such value.

$$b_j = n - r_j + 1 \quad (11)$$

Here, n refers to the total number of observations concerning both observed and censored data, while r_j indicates only the sequence of the observed data.

When the increasing probabilities are calculated,

$$p_j = \frac{b_j - d_j}{b_j} \tag{12}$$

Here, d_j shows the number of observations at the jth value. These dependent values are above 1.

The KM estimator is the product of the probabilities increasing from $j = 1$ to k.

k refers to sequencing from high values to low values for detected observations.

$$\hat{S}(x) = \prod_{j=1}^{k} p_j \tag{13}$$

The mean survival time is as follows:

$$\hat{\mu}(x_j) = \hat{S}(x_0)x_1 + \hat{S}(x_1)(x_2 - x_1) + \cdots + \hat{S}(x_{k-1}) \times (x_k - x_{k-1}) \tag{14}$$

It is generally considered as follows:

$$\hat{S}(x_0) = 1 \quad \text{and} \quad \hat{S}(x_n) = 0 \tag{15}$$

Position estimations regarding reversed data (mean, median, and other percentages) have to be converted into the original scale again through subtraction from the constant M. Therefore, the mean survival time for the original data will be as follows:

$$\hat{\mu}(x_i) = M - \hat{\mu}(x_j) \tag{16}$$

The median is the value corresponding to the 50th percentile in the K–M survival curve reversed after the reconversion procedure.

$$S(x_{50}) = 0.5 \tag{17}$$

As to the other method;

The algorithm related to this process was basically developed by Popovic et al. [14] in order to estimate the survival function based on Kaplan and Meier [18], Hosmer et al. [20], and Ware and Demets [21].

For each subject, $i = 1, \ldots, n$, sequencing is done in an ascending order (concerning both censored and observed data). For each, a censoring level like $\delta_i = 1$ is assigned if the subject is observed data. If the subject is not censored, a censoring level like $\delta_i = 0$ is assigned. For that reason, when there is dependence, censored inputs have to come before observed events.

In such a case, subjects like $\delta_i = 1$ have to be chosen because there are only detected events, and all censored observations are ignored. For each input, calculation in Eq. 18 is made.

$$p_i = \frac{r_i - \delta_i}{r_i} \tag{18}$$

In the survival analysis, the survival function $S(x)$ yields the ratio of the subjects that are expected to survive for at least x units is the estimation of $S(x)$. Calculation for starts with the input at the top.

$$\hat{S}(x) = \prod_{i=n}^{1} p_i \tag{19}$$

The mean survival time (x) is calculated as follows:

$$\mu(x^*) = \int_{b_1}^{b_2} S(u)du \tag{20}$$

x^* shows that the mean of the x variable is a function of the selected range like. The parameter refers to the lower limit selected for measuring set.

The mean estimator for the range given above is as follows:

$$\hat{\mu}(x^*) = \hat{\mu}(b_2) - \sum_{i=n}^{1} \hat{S}(x)(x_i - x_{i-1}) \tag{21}$$

Here, $x_0 = b_1$.

Since the survival function for left-censored data is equal to a concentration unit higher than the highest measured event value, $\hat{\mu}(b_2)$ is equal to the highest concentration in the dataset. Thus, the probability of detecting that all concentrations are bigger than the highest value of the dataset is 1. As x gets closer to, probability decreases (there are precisions in every event measured).

The median survival time is where the probability of survival is 0.5:

$$S(x_{50}) = 0.5 \tag{22}$$

1.1.3 Semi-parametric Methods

In the regression on order statistics (ROS) method (also called log probit regression), data are categorized; it is assumed that there is a linear relationship between the logarithm of probability of observation and inverse cumulative normal distribution belonging to the graph position of observations; and such relationship is mostly determined based on Blom's formula [8]. This is a linear equality solved for each undetectable observation. The ROS method has been suggested for the analysis of left-censored data [6]. For the ROS approach, certain variations have been proposed for separations from log-normal distribution [8, 12]. However, it has been reported that ROS or the methods used instead of it should not be applied to complex datasets (i.e., datasets with more than one LOD value) [12].

The probability of exceedance is used for each limit of detection for drawing positions to be calculated in both censored and uncensored data.

$$E_j = E_{j+1} + \frac{A_j}{A_j + B_j}(1 - E_{j+1}) \tag{23}$$

Here, E_j =the probability of exceeding the j limit of detection. A_j =the total number of uncensored observations in the range of $[j, j+1)$. B_j =the total number of the censored and the uncensored observations that are lower than or equal to the j limit of detection.

For the highest limit of detection, $E_{j+1} = 0$ and $A_j + B_j = n$.

The number of those which are undetected below the j limit of detection is defined as:

$$C_j = B_j - B_{j-1} - A_{j-1} \tag{24}$$

A Weibull-type drawing position (p) for a specific uncensored observation can be calculated by taking into consideration the probability of exceeding the limit of censoring below the observation (E_j), the probability of exceeding the limit of censoring above the observation (E_{j+1}), and the order of observation among all values within j and $j+1$ limit of detection. In general, the Weibull-type drawing positions for uncensored observation are as follows:

$$p(i) = (1 - E_j) = \frac{E_j - E_{j+1}}{A_j + 1} r_i \tag{25}$$

Here, =the ith observation rank among the observations in the range of $(j, j+1]$.

Likewise, the Weibull-type drawing positions for censored observations are as indicated in Eq. 26 below:

$$p(i) = \frac{(1 - E_j)}{C_j + 1} r_i \tag{26}$$

Here, r_i =the total number of censored data in the range of $(j, j+1]$.

At this point, the linear regression of uncensored observations occurs against the normal distributions of uncensored drawing positions. The normal distributions of drawing positions are known as the order statistics of the ROS method. For that reason, the regression equality for the estimation of unobserved data is as indicated in Eq. 27.

$$\text{Estimated log value} = \beta + \alpha \tag{27}$$

Censored concentrations are modeled by use of the parameters of linear regression and the normal distributions of censored data. These modeled censored observations are used together with the uncensored observations in order to model the distribution of the population. The specified values are not treated as values to be available in the case of lack of individual censoring.

Then the observed uncensored data are combined with the modeled censored values in order to estimate the summary statistics of the whole population. This method prevents biasedness resulting from conversion by combining the two value types.

2 Materials and Methods

Datasets were derived from different distributions (log-normal, exponential, and Weibull). Such derived datasets were arranged in different sample sizes (20–300) and at different censoring rates (5–65 %). Within distributions, the locations of the censored observations were changed to be from left, and their performance was evaluated on the basis of deviations from the mean, the median, and the SD. Firstly, uncensored datasets were produced from 3 different distributions, namely log-normal, exponential, and Weibull. Then these datasets were left-censored at the ratios of 5, 25, 45, and 65 %. Substitution method (limit of detection and limit of detection [LOD and LOD/$\sqrt{2}$]), parametric method (MLE), semi-parametric method (ROS), and non-parametric method (KM) were applied to the left-censored datasets in order to estimate the censored data.

To evaluate the performance of left-censored datasets, datasets were derived from log-normal and Weibull distributions in such a way that the coefficient of variance would be 0.5 and 2. Since no different coefficient of variance can be produced for exponential distribution because of its probability distribution function, data were produced with reference to the coefficient of variance of 1. The datasets were produced as follows: when the coefficient of variance was taken as 0.5, the mean would be 1 and SD would be 0.473 for log-normal distribution, while shape parameter would be 2.1 and scale parameter would be 1 for Weibull distribution; and when the coefficient of variance was taken as 2, the mean would be 1 and SD would be 1.27 for log-normal distribution, while shape parameter would be 0.542 and scale parameter would be 1 for Weibull distribution. On the other hand, a dataset whose parameter was 0.05 was obtained for exponential distribution.

The data were firstly produced uncensored based on the above-mentioned parameters of each distribution specified. Then the data were left-censored at the ratios of 5, 25, 45, and 65 %. After censoring, the censored data were estimated through the application of substitution (LOD and LOD/$\sqrt{2}$), parametric (MLE), semi-parametric (ROS), and nonparametric (KM) methods via NADA (Nondetects And Data Analysis: Statistics for Censored Environmental Data), which was written by Lopaka Lee on R (version i386 3.0.2) and whose most recent version was updated on the 2nd of July 2014. The number of repetitions was adopted as 10,000 for the simulation made. In addition, evaluations were made by increasing sample sizes from 20 to 300 by tens.

Evaluations were made in different sample sizes and distributions and at different censoring rates in order to compare the performance of the above-mentioned methods

used in the analysis of left-censored data with their performance in the case of uncensored datasets and with one another. Firstly, uncensored datasets were obtained in sample sizes varying from 20 to 300 and in different distributions. Then these datasets were arranged in an ascending order and censored at the above-mentioned censoring rates. The censored parts were estimated through substitution, parametric, semi-parametric, and nonparametric methods. With such estimated values, deviations from the mean, the median, and the SD relative to uncensored datasets were found, and performance comparison was made.

Study designs involving censored measurement are used in most of research conducted in applied sciences. Censored measurements are divided into two: right-censored and left-censored. Estimation values may be used instead of censored observations through different methods in studies in which left-censored data are used. The use of parametric or nonparametric tests is prescribed in the analysis of data of this sort based on deviations from the mean, the median, and the SD relative to uncensored data. The analysis methods to be used vary depending on the results obtained.

The results of simulation from Tables 2, 3, 4, 5, and 6 provide important information.

For log-normal distribution, Gillom and Hensel [22] took the number of repetitions as 500, sample sizes as 10, 20, and 50, and censoring rates as 20, 40, 60, and 80 %, compared substitution methods and MLE in terms of deviations from the mean and the median, and determined that MLE displayed the best performance. Helsel and Cohn [23] kept the sample size fixed at 25 when they took the number of repetitions as 500. They took censoring rates as 20, 40, 60, and 80 % to compare substitution methods, ROS, and MLE in terms of deviations from the mean and the median and found out that the best performance was achieved by ROS. Shumway et al. [24] took the number of repetitions as 20 and 50 and censoring rates as 50 and 80 % to compare ROS and MLE in terms of deviations from the mean and the median and concluded that neither of these two methods consistently performed better than the other. Hewett and Ganser took the number of repetitions as 100 and censoring rates as 10, 20, 40, 60, and 80 % to compare substitution methods, KM, ROS, and MLE in terms of deviations from the mean and the median and found out that only MLE and ROS performed well in all scenarios. KM did not perform well. The simulation studies conducted proved the number of repetitions inadequate. As the number of repetitions was taken as 10,000 in the present study, the inferences provided below can be considered more reliable [12].

Though LOD substitution, KM, MLE, and ROS perform similarly in deviations from the mean and the SD at the censoring rates of 5, 25, 45, and 65 % for the dataset derived from the log-normal distribution with a mean of 1 and a SD of 0.473, the lowest deviation is achieved by ROS among all methods. It is again ROS method which performs best in deviation from the median. In general, LOD and LOD/$\sqrt{2}$ substitution methods display, similarly to each other, the highest deviations from the mean, the median, and the SD when sample sizes are 20, 30, 70, and 110. As sample size increases in these methods, a systematical decrease occurs in deviations from the mean, the median, and the SD. That is, these methods perform worse when the sample size goes below 110 in comparison with when the sample size is over 110. KM, on the other hand, displays the highest deviations from the mean, the median, and the SD when sample sizes are 20, 30, and 70. Higher deviations are observed in the case of KM when sample size is below 70, though it does not follow a systematical pattern. Regardless of sample size, MLE displays deviations in different sample sizes. However, deviations from the SD increase more in comparison with deviations from the mean and the median in MLE as sample size increases. Among the methods, it is MLE which displays the highest deviation. Deviations in this method substantially increase especially when the censoring rate is raised.

ROS performs best in deviations from the mean, the median, and the SD at the censoring rates of 5, 25, and 45 % for the dataset derived from the log-normal distribution with a mean of 1 and a SD of 1.2. Though LOD/$\sqrt{2}$ substitution and ROS perform similarly in deviations from the mean and the SD at the censoring rate of 65 %, the lowest deviation is achieved by ROS. LOD/$\sqrt{2}$ performs best in deviations from the median. LOD, LOD/$\sqrt{2}$, KM, MLE, and ROS similarly display the highest deviations from the mean, the median, and the SD when the sample size is usually below 50. However, in the case of censored data, deviations from the SD increase as samples size increases. As sample size increases, deviations from the mean and the SD gradually decrease in ROS and LOD/$\sqrt{2}$. By the nature of log-normal distribution, deviations substantially increase as censoring rate rises in all methods. However, deviations in MLE are even worse than the censored situation. That clearly indicates the effect of censoring rate on the methods.

Chowdhury and Gulshan [15] took the number of repetitions as 1000, sample sizes as 25, 35, 80, and 200, and censoring rates as 5, 25, 35, and 50 % to compare KM, ROS, and MLE in terms of deviations from the mean and the median and found out that it was KM which performed best at the censoring rate of 5 %, while it was ROS which performed best at other censoring rates. Considering that the coefficient of variance of the exponential distribution is always 1, scenario diversity is seen to be limited in that study. Since the number of repetitions is bigger in the present study, the results presented below can be considered more reliable.

Table 2 The performance evaluation of the methods at different censoring rates for the dataset derived from the log-normal distribution with a mean of 1 and a SD of 0.473

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.098935	−0.08135	1.134796	−0.013247	0	0.017939	0.007528	0	−0.010196
	80	−0.103571	−0.082835	1.336209	−0.011167	0	0.014651	0.007784	0	−0.010293
	140	−0.104301	−0.081616	1.385103	−0.010924	0	0.014249	0.007761	0	−0.01019
	200	−0.104588	−0.081506	1.40843	−0.010777	0	0.014041	0.007776	0	−0.010183
	260	−0.104782	−0.081855	1.423282	−0.010695	0	0.013918	0.007804	0	−0.010217
25	20	−0.485527	−0.440333	1.114148	−0.124041	0	0.132796	0.027755	0	−0.020341
	80	−0.499634	−0.442853	1.323699	−0.115275	0	0.116975	0.031228	0	−0.023993
	140	−0.501533	−0.44217	1.375245	−0.113741	0	0.114262	0.031944	0	−0.02462
	200	−0.502487	−0.441877	1.400094	−0.113299	0	0.113386	0.032004	0	−0.024695
	260	−0.502981	−0.441716	1.415912	−0.113075	0	0.112923	0.03219	0	−0.024845
45	20	−0.932847	−0.884422	1.072775	−0.343745	NA	0.304285	0.006472	NA	0.031414
	80	−0.960053	−0.889537	1.299018	−0.323923	NA	0.270681	0.016787	NA	0.018288
	140	−0.963787	−0.889402	1.35586	−0.321368	NA	0.265328	0.018086	NA	0.016572
	200	−0.965506	−0.889091	1.383704	−0.320245	NA	0.263392	0.018527	NA	0.015967
	260	−0.966574	−0.888001	1.401437	−0.319515	NA	0.262187	0.019102	NA	0.015356
65	20	−1.53524	−1.50808	0.997159	−0.765526	−0.645978	0.553111	−0.121193	0.345302	0.179346
	80	−1.588481	−1.513531	1.251891	−0.721567	−0.57025	0.490199	−0.094579	0.394357	0.146433
	140	−1.596497	−1.511626	1.318586	−0.714287	−0.558045	0.479209	−0.090121	0.402222	0.140888
	200	−1.599584	−1.510378	1.352064	−0.712119	−0.553523	0.475663	−0.089018	0.4051	0.139304
	260	−1.601682	−1.512855	1.373428	−0.711095	−0.551479	0.473605	−0.088161	0.406879	0.138218
Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.013252	0.081349	0.015934	−0.004104	0.035914	−0.086284	0.006601	−0.000001	−0.007636
	80	−0.011166	0.020279	0.014155	−0.003726	0.02786	−0.08565	0.000785	0.000002	−0.00022
	140	−0.010926	0.011367	0.013965	−0.003609	0.026408	−0.083032	−0.000235	0.000003	0.000978
	200	−0.010774	0.007996	0.013842	−0.003659	0.026734	−0.083134	−0.000656	−0.000002	0.001472
	260	−0.010698	0.006146	0.013765	−0.003631	0.02639	−0.08276	−0.000896	0.000003	0.00174
25	20	−0.124041	0.081349	0.120974	0.044619	0.224257	−0.454266	0.016457	−0.000001	−0.003143
	80	−0.115276	0.020279	0.114037	0.045811	0.221643	−0.45991	−0.002663	0.000002	0.012755
	140	−0.113744	0.011367	0.112581	0.046091	0.220985	−0.458078	−0.005979	0.000003	0.015235
	200	−0.113303	0.007996	0.11221	0.045981	0.221612	−0.458741	−0.007744	−0.000002	0.016628
	260	−0.113073	0.006146	0.112018	0.046151	0.221495	−0.458422	−0.008517	0.000003	0.017222

Table 2 continued

Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
45	20	−0.343745	0.081349	0.278508	0.212009	0.596304	−0.971064	0.040365	−0.000019	0.009522
	80	−0.323926	0.020279	0.264239	0.216325	0.603394	−0.980224	0.002756	0.000002	0.032235
	140	−0.321366	0.011367	0.261639	0.217046	0.604071	−0.97878	−0.004252	0.000003	0.036073
	200	−0.320244	0.007996	0.260809	0.217052	0.60527	−0.979801	−0.008278	−0.000002	0.038584
	260	−0.319515	0.006146	0.260199	0.217403	0.60554	−0.979539	−0.010278	0.000003	0.03974
65	20	−0.765526	NA	0.505108	0.594379	1.248922	−1.801748	0.109421	0.700094	0.039506
	80	−0.721568	NA	0.478057	0.604178	1.270881	−1.796725	0.042344	0.604998	0.066202
	140	−0.714287	NA	0.472231	0.605403	1.273632	−1.794328	0.027078	0.582524	0.072215
	200	−0.712116	NA	0.470774	0.605658	1.27559	−1.794617	0.018574	0.570616	0.076143
	260	−0.711095	NA	0.469841	0.606372	1.276599	−1.793793	0.01401	0.564063	0.07809

Table 3 The performance evaluation of the methods at different censoring rates for the dataset derived from the log-normal distribution with a mean of 1 and a SD of 1.27

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.302303	−0.224188	−0.133778	−0.010117	NA	0.007932	−0.00264	NA	0.002177
	80	−0.307074	−0.229555	−0.18186	−0.00698	NA	0.004329	−0.001406	NA	0.000899
	140	−0.307705	−0.225793	−0.196902	−0.006496	NA	0.00379	−0.001233	NA	0.000744
	200	−0.308996	−0.228045	−0.203275	−0.006394	NA	0.003605	−0.001198	NA	0.000698
	260	−0.308354	−0.225718	−0.207599	−0.006304	NA	0.003487	−0.001181	NA	0.000675
25	20	−1.779013	−1.430918	−0.76205	−0.16533	NA	0.117145	−0.064475	NA	0.048939
	80	−1.807517	−1.374842	−1.053099	−0.138766	NA	0.078748	−0.050305	NA	0.030165
	140	−1.811362	−1.362056	−1.144024	−0.135367	NA	0.072218	−0.048708	NA	0.027402
	200	−1.819493	−1.366479	−1.182184	−0.133652	NA	0.068962	−0.047668	NA	0.025901
	260	−1.814999	−1.362934	−1.208792	−0.133398	NA	0.067635	−0.047678	NA	0.025444
45	20	−3.969636	−3.279142	−1.580329	−0.659584	NA	0.419394	−0.308083	NA	0.214195
	80	−4.049018	−3.13675	−2.247885	−0.573521	NA	0.293896	−0.257026	NA	0.142138
	140	−4.060276	−3.101329	−2.457704	−0.562036	NA	0.270964	−0.250669	NA	0.130092
	200	−4.080815	−3.119045	−2.544632	−0.557198	NA	0.260307	−0.247637	NA	0.124423

Table 3 continued

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
	260	−4.068573	−3.100217	−2.608243	−0.556032	NA	0.255142	−0.246949	NA	0.121875
65	20	−7.731627	−6.621525	−2.696355	−2.109678	−2.244347	1.151999	−1.132023	−0.740222	0.687841
	80	−7.950754	−6.311192	−4.056648	−1.818809	−1.834623	0.810789	−0.944464	−0.489474	0.464559
	140	−7.982451	−6.240108	−4.489387	−1.780431	−1.778341	0.747613	−0.920484	−0.455354	0.425793
	200	−8.028282	−6.255427	−4.666225	−1.773072	−1.768233	0.72175	−0.9158	−0.449339	0.410293
	260	−8.00204	−6.220965	−4.802475	−1.761933	−1.751323	0.704428	−0.908262	−0.43799	0.399831
Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.010132	0.224191	−0.004666	−0.531132	0.138732	−6.606075	0.004673	0	−0.003089
	80	−0.006981	0.05513	0.000815	−0.371258	0.083128	−3.896922	0.001092	−0.000004	−0.000517
	140	−0.006498	0.03046	0.001708	−0.34982	0.080111	−3.321002	0.000468	−0.000001	−0.000167
	200	−0.006397	0.021811	0.002127	−0.343356	0.072788	−3.134803	0.000292	0.000001	−0.000088
	260	−0.006303	0.016724	0.002339	−0.343646	0.075613	−2.942453	0.00005	0	0.000038
25	20	−0.165318	0.224191	0.03743	−2.338463	0.622502	−27.750213	0.026922	0	−0.010144
	80	−0.138762	0.05513	0.056589	−2.010601	0.56814	−19.553648	0.009879	−0.000004	−0.002392
	140	−0.135363	0.03046	0.059099	−1.971988	0.564721	−18.449121	0.006417	−0.000001	−0.00107
	200	−0.133656	0.021811	0.059645	−1.964681	0.559067	−18.110927	0.005404	0.000001	−0.000874
	260	−0.133397	0.016724	0.0604	−1.95583	0.561175	−17.746557	0.003628	0	−0.000076
45	20	−0.659578	0.224191	0.226566	−6.463617	1.410139	−184.51537	0.091216	−0.000031	−0.020911
	80	−0.573527	0.05513	0.240376	−5.360753	1.352442	−93.543732	0.05603	−0.000004	−0.0125
	140	−0.562046	0.03046	0.239267	−5.240963	1.347365	−86.608419	0.047071	−0.000001	−0.009992
	200	−0.557196	0.021811	0.23779	−5.222007	1.343956	−84.512265	0.043162	0.000001	−0.009098
	260	−0.556031	0.016724	0.237656	−5.171865	1.345075	−82.505976	0.039548	0.000001	−0.007706
65	20	−2.109682	NA	0.733347	−27.54229	2.322146	−50501.3036	0.263554	1.406784	−0.01443
	80	−1.818807	NA	0.69427	−17.54358	2.245965	−1681.89798	0.227914	1.314176	−0.033124
	140	−1.780431	NA	0.678473	−16.66263	2.236152	−1283.59225	0.2109	1.285251	−0.031734
	200	−1.773077	NA	0.672608	−16.46282	2.234475	−1179.73344	0.199771	1.266922	−0.030071
	260	−1.761932	NA	0.666237	−16.15530	2.233894	−1106.96751	0.191801	1.25442	−0.027754

Table 4 The performance evaluation of the methods at different censoring rates for the dataset derived from the exponential distribution with a parameter of 0.05

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.998352	−1.00301	0.043701	−0.053052	NA	0.057599	−0.022763	NA	0.025725
	80	−1.020253	−1.016865	0.009829	−0.032628	NA	0.03246	−0.013767	NA	0.013962
	140	−1.022982	−1.028444	0.006329	−0.029443	NA	0.029006	−0.012408	NA	0.012355
	200	−1.02268	−1.024458	0.005309	−0.02843	NA	0.027885	−0.0119	NA	0.011928
	260	−1.023804	−1.029171	0.002922	−0.027974	NA	0.027229	−0.011738	NA	0.011613
25	20	−5.575697	−5.70895	0.263477	−0.925564	NA	0.865794	−0.416569	NA	0.423171
	80	−5.715854	−5.731593	0.060031	−0.797255	NA	0.690288	−0.353602	NA	0.327498
	140	−5.73553	−5.780476	0.040526	−0.77816	NA	0.666766	−0.344463	NA	0.314972
	200	−5.735264	−5.749035	0.034326	−0.76906	NA	0.656332	−0.34005	NA	0.309393
	260	−5.740907	−5.765419	0.018864	−0.765834	NA	0.65088	−0.338226	NA	0.306342
45	20	−11.524003	−11.941641	0.622778	−3.381402	NA	2.708496	−1.614569	NA	1.461305
	80	−11.864063	−11.973364	0.145164	−3.058136	NA	2.272488	−1.435223	NA	1.193226
	140	−11.9128	−11.990148	0.101155	−3.012682	NA	2.216214	−1.411063	NA	1.159092
	200	−11.909835	−11.948871	0.080753	−3.004528	NA	2.199991	−1.408525	NA	1.150911
	260	−11.929395	−11.954898	0.050559	−2.987051	NA	2.180224	−1.398027	NA	1.138152
65	20	−20.017915	−20.88745	1.308584	−8.96504	−8.581236	5.957656	−4.588187	−1.84766	3.583259
	80	−20.792103	−21.011542	0.32782	−8.217476	−7.46621	5.072491	−4.131006	−1.179377	2.98613
	140	−20.8906	−21.023657	0.22724	−8.14311	−7.358549	4.973079	−4.088763	−1.120984	2.921678
	200	−20.886928	−20.962705	0.171633	−8.091182	−7.271187	4.919466	−4.05706	−1.064686	2.886484
	260	−20.931533	−21.008739	0.113096	−8.080142	−7.26533	4.894459	−4.051936	−1.068056	2.870762
Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.053056	1.002884	0.031269	−7.955578	2.939071	−86.215774	−0.015698	−0.000046	0.020456
	80	−0.032632	0.248131	0.025866	−7.24146	2.891078	−44.81852	−0.033768	0.000008	0.034221
	140	−0.029478	0.143183	0.025245	−7.179737	2.883419	−43.083068	−0.036728	−0.000008	0.036383
	200	−0.02838	0.100769	0.025258	−7.161678	2.898996	−42.514492	−0.038146	0.000004	0.0375
	260	−0.027957	0.07623	0.025206	−7.173423	2.882565	−42.357063	−0.038939	0.000009	0.037926
25	20	−0.925604	1.002884	0.704347	−17.36106	4.888587	−809.567064	−0.248446	−0.000046	0.291375
	80	−0.797318	0.248131	0.649804	−14.37351	4.91368	−121.818497	−0.372644	0.000008	0.350096
	140	−0.778205	0.143183	0.643672	−14.14393	4.917255	−112.833177	−0.398835	−0.000008	0.364058
	200	−0.769099	0.100769	0.640206	−14.06372	4.934818	−109.983513	−0.412358	0.000004	0.372036
	260	−0.765791	0.07623	0.638454	−14.05852	4.922623	−108.998585	−0.416104	0.000009	0.372888

Table 4 continued

Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
45	20	−3.381428	1.002884	2.345343	−39.40301	8.090107	−109.287971	−0.584427	−0.00072	0.720601
	80	−3.058174	0.248131	2.180914	−25.69240	8.174528	−449.428225	−0.884212	0.000008	0.817284
	140	−3.012677	0.143183	2.163962	−24.93775	8.187855	−379.734965	−0.964027	−0.000008	0.85424
	200	−3.004474	0.100769	2.163518	−24.66175	8.205456	−359.770747	−1.000552	0.000004	0.872193
	260	−2.987049	0.07623	2.152102	−24.61907	8.195826	−352.59439	−1.018455	0.000009	0.878839
65	20	−8.965031	NA	5.267869	−254.1371	11.745466	−1466.92785	−0.849747	3.90379	1.402755
	80	−8.217529	NA	4.896041	−53.41336	11.744771	−5309.47252	−1.439605	2.961311	1.499885
	140	−8.143099	NA	4.872404	−50.03284	11.744794	−3207.17320	−1.638433	2.648665	1.575192
	200	−8.091247	NA	4.849131	−48.87845	11.751815	−2771.99149	−1.721383	2.528627	1.604948
	260	−8.080173	NA	4.840215	−48.57636	11.739687	−2613.83101	−1.771919	2.439685	1.622227

Table 5 The performance evaluation of the methods at different censoring rates for the dataset derived from the Weibull distribution with a shape parameter of 1 and a scale parameter of 2.1

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.035347	−0.028243	0.019371	−0.005212	NA	0.008248	−0.000571	NA	0.001244
	80	−0.037331	−0.028971	0.021095	−0.004358	NA	0.006941	−0.0005	NA	0.001027
	140	−0.037564	−0.028923	0.021222	−0.004148	NA	0.006612	−0.000414	NA	0.000872
	200	−0.0377	−0.029041	0.021339	−0.004089	NA	0.006524	−0.00041	NA	0.000856
	260	−0.037788	−0.029202	0.021449	−0.00407	NA	0.006503	−0.000423	NA	0.000876
25	20	−0.166532	−0.149156	0.073743	−0.051135	NA	0.062563	−0.00799	NA	0.015225
	80	−0.172198	−0.150517	0.075281	−0.048187	NA	0.057918	−0.007052	NA	0.013337
	140	−0.17287	−0.151133	0.075072	−0.047649	NA	0.057054	−0.006813	NA	0.012907
	200	−0.173284	−0.150908	0.075249	−0.047517	NA	0.05685	−0.006811	NA	0.012857
	260	−0.173432	−0.150935	0.075329	−0.047458	NA	0.056766	−0.00681	NA	0.012848
45	20	−0.303773	−0.28537	0.11718	−0.135339	NA	0.134975	−0.02808	NA	0.045639
	80	−0.312956	−0.288039	0.116518	−0.129673	NA	0.126367	−0.025383	NA	0.04125
	140	−0.314323	−0.28939	0.115968	−0.128569	NA	0.124759	−0.024809	NA	0.040315
	200	−0.314959	−0.289103	0.116018	−0.128211	NA	0.124298	−0.02466	NA	0.040087

Table 5 continued

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
	260	−0.315229	−0.288862	0.116132	−0.128177	NA	0.124176	−0.02469	NA	0.040062
65	20	−0.467694	−0.4553	0.16087	−0.275231	−0.213874	0.223207	−0.07421	0.095385	0.100318
	80	−0.482914	−0.459961	0.155522	−0.264309	−0.191	0.208654	−0.067796	0.111323	0.091815
	140	−0.485727	−0.463037	0.154811	−0.262774	−0.188642	0.206474	−0.066982	0.112578	0.090404
	200	−0.486725	−0.463093	0.154553	−0.262125	−0.186833	0.205706	−0.066619	0.113937	0.089961
	260	−0.48701	−0.463199	0.154429	−0.261733	−0.185957	0.205308	−0.066398	0.114566	0.089738

Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.005213	0.028241	0.007653	−0.033882	0.084732	−0.193897	−0.002287	−0.000001	0.004252
	80	−0.004358	0.007247	0.006796	−0.036528	0.088083	−0.210027	−0.004644	0	0.007477
	140	−0.004148	0.004021	0.006529	−0.036498	0.087154	−0.210593	−0.004982	0	0.007898
	200	−0.004089	0.002896	0.006466	−0.036529	0.087777	−0.211224	−0.005161	0	0.008133
	260	−0.00407	0.002199	0.006459	−0.0369	0.088169	−0.212889	−0.005287	0	0.00831
25	20	−0.051135	0.028241	0.059236	−0.034616	0.150515	−0.383527	−0.019386	−0.000001	0.0296
	80	−0.048186	0.007247	0.057106	−0.036013	0.157398	−0.394817	−0.028062	0	0.037382
	140	−0.047649	0.004021	0.056591	−0.035745	0.156863	−0.394299	−0.029333	0	0.038373
	200	−0.047517	0.002896	0.056527	−0.03579	0.15765	−0.394893	−0.030124	0	0.039104
	260	−0.047459	0.002199	0.056517	−0.036126	0.158246	−0.396724	−0.03054	0	0.039498
45	20	−0.135337	0.028241	0.128259	0.010857	0.276741	−0.648061	−0.037721	−0.000017	0.055798
	80	−0.129673	0.007247	0.124721	0.012223	0.289147	−0.641956	−0.054195	0	0.066827
	140	−0.128568	0.004021	0.12382	0.01272	0.289251	−0.639155	−0.057066	0	0.068553
	200	−0.128215	0.002896	0.123641	0.012769	0.290336	−0.639275	−0.058692	0	0.069722
	260	−0.128176	0.002199	0.123671	0.012487	0.291186	−0.641207	−0.059752	0	0.070513
65	20	−0.275228	NA	0.211936	0.129485	0.474087	−1.109028	−0.051101	0.130936	0.086508
	80	−0.264304	NA	0.205866	0.134444	0.492088	−1.032447	−0.080438	0.091871	0.100491
	140	−0.262775	NA	0.204882	0.135095	0.492999	−1.022574	−0.087311	0.081305	0.103759
	200	−0.26212	NA	0.204592	0.135273	0.494365	−1.020711	−0.090514	0.077176	0.105479
	260	−0.261738	NA	0.204452	0.135165	0.495394	−1.021674	−0.09249	0.074426	0.10659

Table 6 The performance evaluation of the methods at different censoring rates for the dataset derived from the Weibull distribution with a shape parameter of 0.542 and a scale parameter of 1

Censoring rate (%)	*N*	Censored			LOD			LOD/$\sqrt{2}$		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.090918	−0.073116	−0.049698	−0.000686	NA	0.000489	−0.000386	NA	0.000276
	80	−0.091452	−0.072709	−0.060402	−0.000237	NA	0.000135	−0.000126	NA	0.000075
	140	−0.091306	−0.072404	−0.062212	−0.000189	NA	0.000103	−0.0001	NA	0.000057
	200	−0.091513	−0.072198	−0.063574	−0.000172	NA	0.000092	−0.000093	NA	0.000051
	260	−0.091692	−0.072458	−0.064568	−0.000167	NA	0.000086	−0.000094	NA	0.000047
25	20	−0.561442	−0.484124	−0.267522	−0.027491	NA	0.018474	−0.01577	NA	0.010849
	80	−0.567778	−0.463447	−0.333198	−0.019103	NA	0.010502	−0.010823	NA	0.006021
	140	−0.567267	−0.462105	−0.344563	−0.018205	NA	0.00961	−0.010297	NA	0.005494
	200	−0.568729	−0.460818	−0.352964	−0.017777	NA	0.009189	−0.01005	NA	0.005248
	260	−0.569942	−0.460141	−0.359076	−0.01754	NA	0.008936	−0.009907	NA	0.005099
45	20	−1.298927	−1.164128	−0.522209	−0.164509	NA	0.10141	−0.096965	NA	0.062438
	80	−1.322347	−1.114491	−0.675629	−0.131533	NA	0.067433	−0.076691	NA	0.040542
	140	−1.322106	−1.103212	−0.702969	−0.126897	NA	0.062637	−0.073768	NA	0.037487
	200	−1.32611	−1.100742	−0.722759	−0.125531	NA	0.060703	−0.072964	NA	0.036297
	260	−1.329319	−1.100822	−0.737098	−0.124645	NA	0.059461	−0.072434	NA	0.03553
65	20	−2.592098	−2.41746	−0.817492	−0.668635	−0.781815	0.359681	−0.408316	−0.381325	0.235721
	80	−2.668289	−2.318934	−1.136129	−0.55456	−0.629648	0.251685	−0.334403	−0.290945	0.160925
	140	−2.671247	−2.299469	−1.195046	−0.540342	−0.612255	0.23652	−0.325232	−0.281327	0.150709
	200	−2.680955	−2.293046	−1.237279	−0.535005	−0.603869	0.229489	−0.321748	−0.275784	0.146002
	260	−2.688658	−2.295865	−1.267367	−0.53209	−0.600212	0.225281	−0.319875	−0.273729	0.143213
Censoring rate (%)	*N*	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
5	20	−0.000689	0.073116	−0.003595	−308.413721	0.20804	−1.16E+09	−0.000085	−0.000001	0.000084
	80	−0.000238	0.017417	−0.000961	−7.059293	0.185261	−7.50E+02	−0.000222	0	0.00013
	140	−0.000187	0.009712	−0.00053	−6.139736	0.180529	−2.76E+02	−0.000241	0	0.000135
	200	−0.00017	0.006799	−0.000356	−5.888855	0.180772	−2.24E+02	−0.000248	0	0.000135
	260	−0.000163	0.005308	−0.000262	−5.894265	0.180604	−2.10E+02	−0.000254	0	0.000135
25	20	−0.027491	0.073116	−0.007518	−2.68E+04	3.15E−01	−3.25E+13	−0.004121	−0.000001	0.003714
	80	−0.019104	0.017417	0.003566	−3.30E+01	2.90E−01	−9.35E+04	−0.006391	0	0.003745
	140	−0.018203	0.009712	0.005607	−2.46E+01	2.85E−01	−8.00E+03	−0.006895	0	0.003797
	200	−0.017781	0.006799	0.006358	−2.27E+01	2.86E−01	−5.06E+03	−0.007133	0	0.003811
	260	−0.017534	0.005308	0.00674	−2.24E+01	2.86E−01	−4.04E+03	−0.007272	0	0.003808

Table 6 continued

Censoring rate (%)	N	K–M			MLE			ROS		
		Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD	Deviations from the mean	Deviations from the median	Deviations from the SD
45	20	−0.164507	0.073116	0.038188	−2.73E+07	4.55E−01	−5.52E+20	−0.011746	−0.000052	0.01328
	80	−0.131537	0.017417	0.050626	−3.58E+02	4.22E−01	−3.01E+08	−0.020226	0	0.012596
	140	−0.126894	0.009712	0.052944	−1.75E+02	4.16E−01	−2.87E+06	−0.022418	0	0.012788
	200	−0.125528	0.006799	0.053847	−1.47E+02	4.16E−01	−1.38E+06	−0.023638	0	0.012953
	260	−0.124646	0.005308	0.054144	−1.39E+02	4.15E−01	−6.17E+05	−0.024342	0	0.013002
65	20	−0.668639	NA	0.223375	−5.33E+12	5.57E−01	−1.06E+33	−0.012858	0.22708	0.036737
	80	−0.554562	NA	0.215355	−6.49E+04	5.08E−01	−1.37E+15	−0.029402	0.17829	0.028065
	140	−0.540337	NA	0.215568	−8.01E+03	5.00E−01	−5.45E+11	−0.03723	0.161754	0.028801
	200	−0.535006	NA	0.214659	−5.23E+03	4.99E−01	−2.64E+11	−0.040646	0.156686	0.028932
	260	−0.532088	NA	0.213773	−4.14E+03	4.98E−01	−2.72E+10	−0.043135	0.15202	0.029179

Though LOD substitution, KM, LOD/$\sqrt{2}$ substitution, and ROS perform similarly in deviations from the mean at the censoring rates of 5, 25, and 45 % for the dataset derived from the exponential distribution with a parameter of 0.05, it is LOD/$\sqrt{2}$ which displays the lowest deviation. It is ROS which performs best in deviations from the median. ROS performs best in deviations from the mean at the censoring rate of 65 %, while it is LOD/$\sqrt{2}$ substitution which performs best in deviations from the median. LOD and LOD/$\sqrt{2}$ similarly display the highest deviations from the mean, the median, and the SD when sample sizes are 20, 30, 70, and 110. When sample size is above 70, a systematical decrease occurs in deviations from the mean, the median, and the SD in these methods. It is KM which displays the biggest deviations from the mean, the median, and the SD when sample size is 20, 30, and 70. When sample size is 20, 70, and 110 in MLE, deviations from the mean and the median increase. However, deviations from the SD increase more, though not systematically, as sample size increases. Again though not systematically, deviations are higher in KM when sample size is below 70. When sample size is above 230 in ROS, deviations from the mean, the median and the SD are much bigger in comparison with uncensored data. A systematical increase occurs in deviations from the SD in particular when sample size is over 200.

For Weibull distribution, Krol and Stedinger [25] took the number of repetitions as 500, kept sample size fixed at 25, and took censoring rates as 10, 20, 40, and 60 % to compare ROS and MLE in terms of deviations from the mean and the median and found out that it was MLE which performed best. Schmoyer et al. took the number of repetitions as 1000 and censoring rates as 20, 40, 60, and 80 % to compare substitution, KM, ROS, and MLE in terms of deviations from the mean and the median and concluded that KM performed best [10]. Chowdhury and Gulshan [15] took the number of repetitions as 1000, sample sizes as 25, 35, 80, and 200, and censoring rates as 5, 25, 35, and 50 % to compare KM, ROS, and MLE in terms of deviations from the mean and the median and found out that ROS performed best. The comments made below can be considered more reliable in terms of the number of repetitions and sample size for cases involving different coefficients of variance for Weibull distribution.

Though LOD substitution, KM, LOD/$\sqrt{2}$ substitution, and ROS perform similarly at the censoring rates of 5, 25, 45, and 65 % in terms of deviations from the mean and the SD for the dataset derived from Weibull distribution with a shape parameter of 0.542 and a scale parameter of 1, it is LOD/$\sqrt{2}$ which displays the lowest deviation. ROS performs best in deviations from the median. LOD, LOD/$\sqrt{2}$, KM, and MLE similarly display the highest deviations

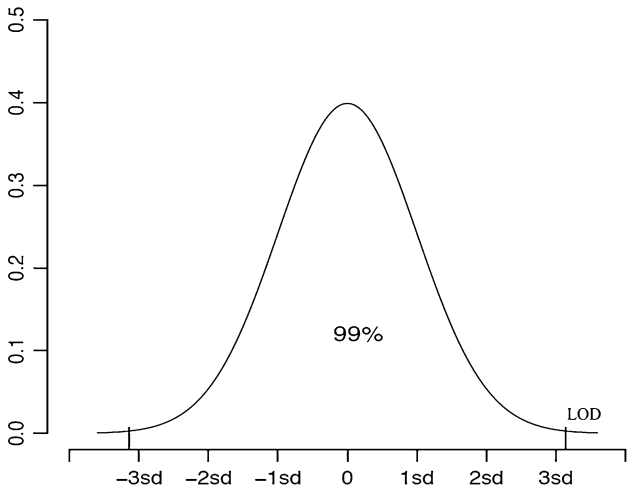

Fig. 1 Limit of detection

from the mean, the median, and the SD when sample size is below 50. In all these four methods, there is a systematical decrease in deviations from the mean, the median, and the SD as sample size increases. This being the case, it can be said that a worse performance is displayed when sample size is below 50 in comparison with when it is above 50. Among all methods, MLE displays the highest deviations. At the censoring rates of 5, 25, 45, and 65 %, MLE displays higher deviations from the mean and the SD relative to censored observation. ROS performs much better in deviations from the mean, the median, and the SD when censoring rate is high relative to lower censoring rates in comparison with other methods.

Censored data display the highest deviations from the mean, the median, and the SD in different sample sizes (30, 70, 110) without any increase or decrease on the basis of sample size at the censoring rates of 5, 25, and 65 % for the dataset derived from Weibull distribution with a shape parameter of 1 and a scale parameter of 2.1. LOD, LOD/$\sqrt{2}$, and KM similarly displays the highest deviations from the mean, the median, and the SD when sample size is below 70. It can be said that as sample size increases, a systematical decrease occurs in deviations from the mean, the median, and the SD in these three methods. In MLE and ROS, data display the highest deviations from the mean, the median, and the SD in different sample sizes (30, 160, 190, 270) without any increase or decrease on the basis of sample size. Though LOD substitution, KM, LOD/$\sqrt{2}$ substitution, and ROS perform similarly in deviations from the mean and the SD, it is LOD/$\sqrt{2}$ which displays the lowest deviation. It is ROS which performs best in deviations from the median. Though MLE, LOD/$\sqrt{2}$ substitution, and ROS perform similarly in deviations from the mean and the SD at the censoring rate of 45 %, it is MLE which displays the lowest deviation. ROS performs best in deviations from the median.

3 Conclusions

The present study mainly aimed to determine deviations from the mean, the median, and the SD in different sample sizes and at different censoring rates for log-normal, exponential, and Weibull distributions in complete and censored data samples. Thus, attention was focused on the concept of "censoring" and censoring types in the first place. Then substitution, parametric (MLE), nonparametric (KM), and semi-parametric (ROS) methods suggested for the evaluation of left-censored observations were introduced. In the present study, the maximum-likelihood method was used for making estimations in parametric survival models. It may be guiding for researchers when left-censored data structure is confronted in the field of health that distribution be determined firstly, and then diagrams provided in Figs. 1, 2, and 3 be used.

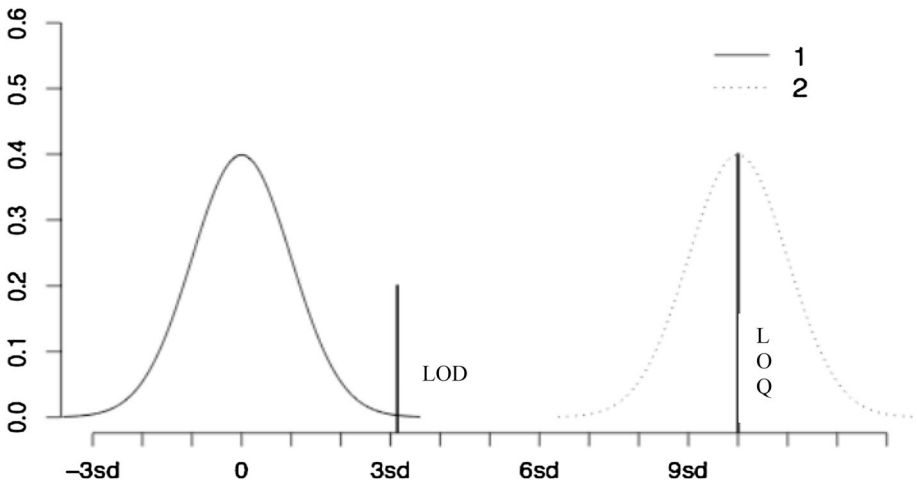

Fig. 2 Limit of quantification (LOQ)

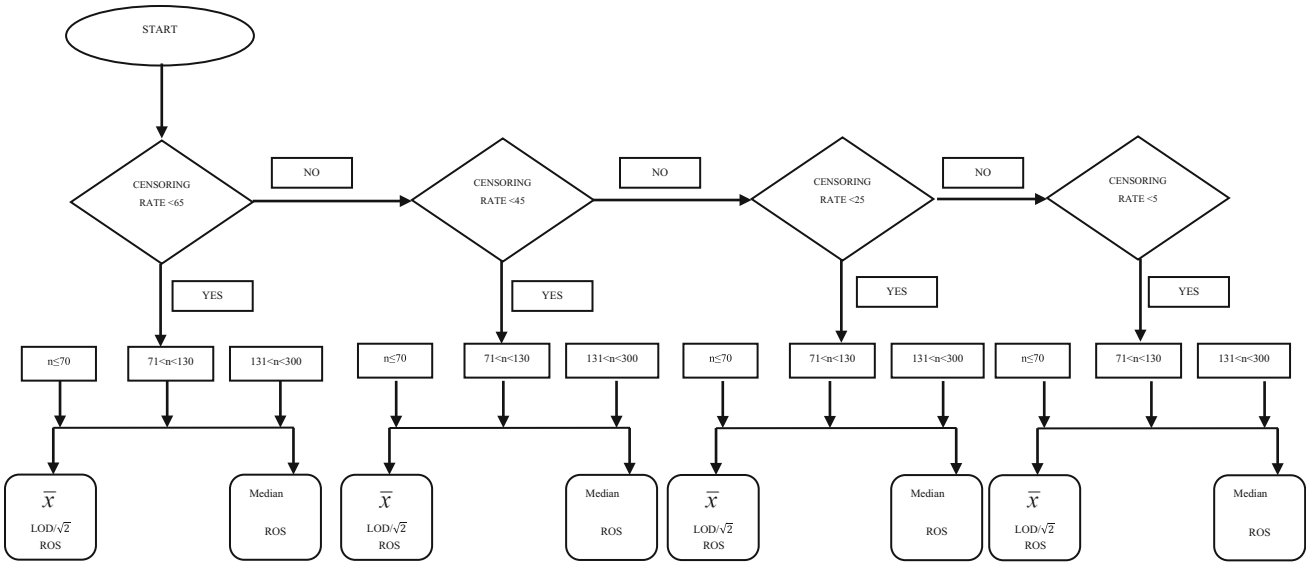

Fig. 3 Suggestion diagram belonging to the estimation methods for the summary statistics belonging to the lognormal distribution

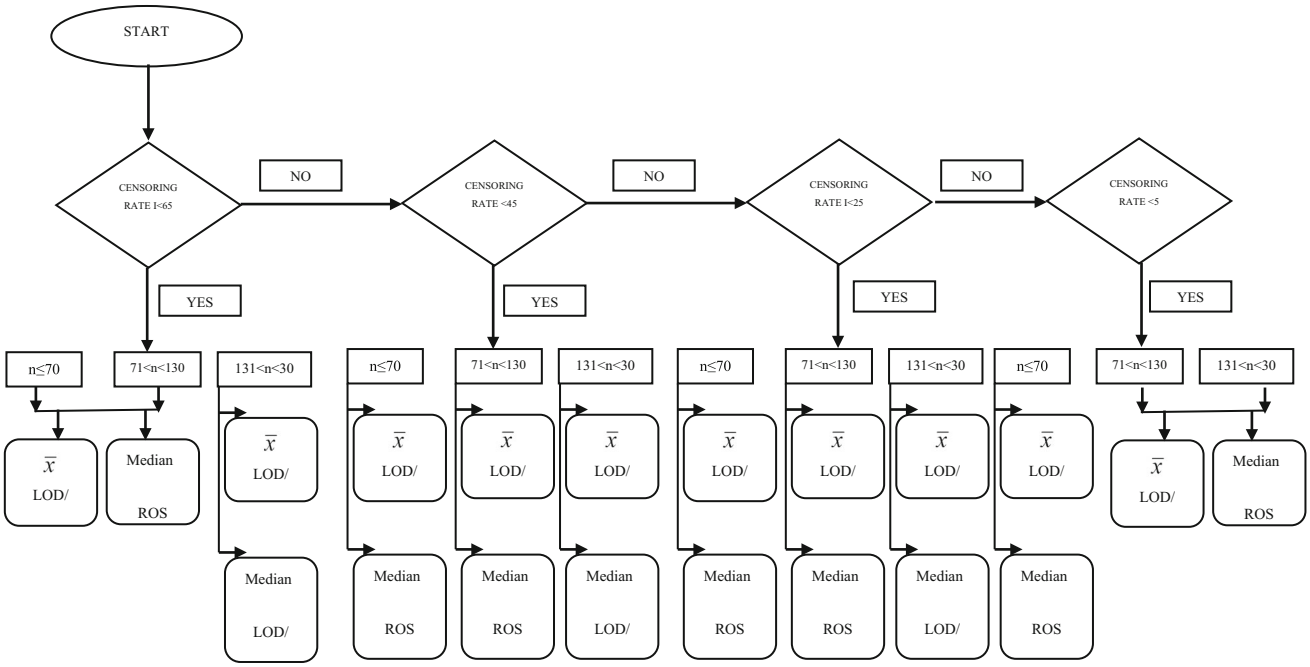

Fig. 4 Suggestion diagram belonging to the estimation methods for the summary statistics belonging to the exponential distribution

Within the scope of the present study, the data were produced uncensored based on different parameters of each distribution. Then the data sets were left-censored at the ratios of 5, 25, 45, and 65 %. The censored data were estimated through substitution (LOD and LOD/$\sqrt{2}$), parametric (MLE), semi-parametric (ROS), and non-parametric (KM) methods. In addition, evaluation was made by increasing the sample size from 20 to 300 by tens. The methods used in the analysis of the left-censored data were evaluated in different sample sizes and at different censoring rates in order to determine their performance in comparison with the uncensored dataset and one another. Performance comparison was made between the uncensored dataset and the censored dataset on the basis of deviations from the median, the mean, and the SD.

At the end of the simulation studies conducted, the best methods for three distributions at different censoring rates and in different sample sizes are indicated in diagrams between Figs. 3, 4, and 5.

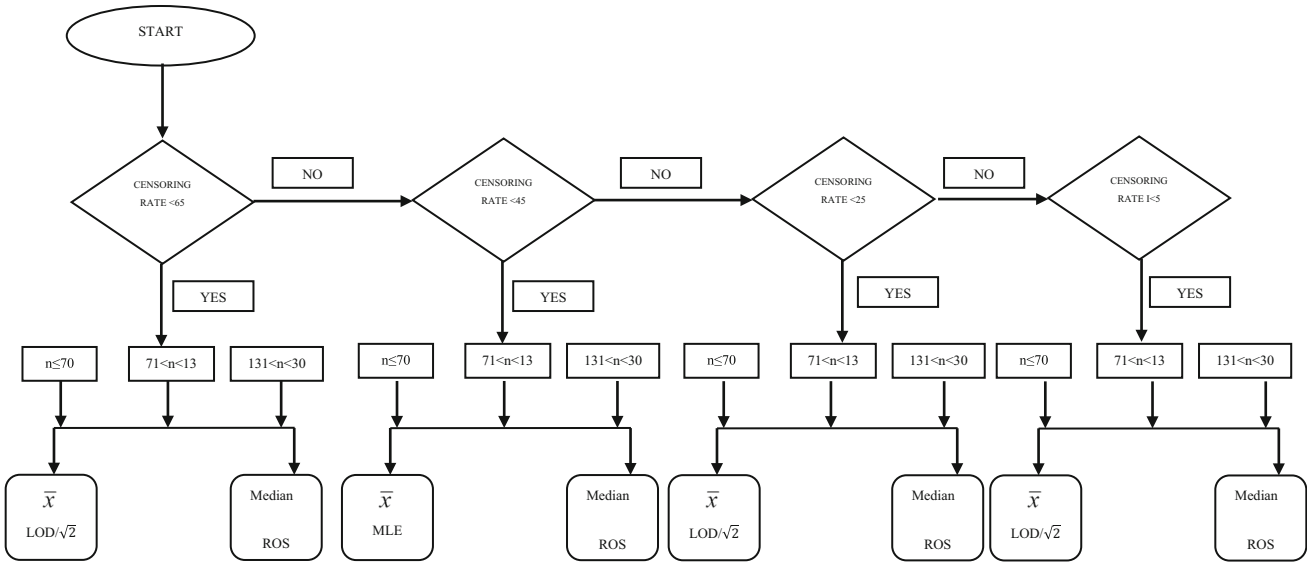

Fig. 5 Suggestion diagram belonging to the estimation methods for the summary statistics belonging to the Weibull distribution

To conclude, designs associated with the methods explained in this thesis study provide several advantages for use in applied fields. These advantages are as follows;

- Substitution methods generally perform well.
- Summary statistics indicate that when parametric test assumptions are not fulfilled, ROS method, which is a nonparametric method, may be used because it displays the best performance in deviations from the median in all distributions and at all censoring rates. This being the case, censored observations may be estimated through the ROS method when a nonparametric test is to be used.
- Substitution methods may be used in analyses when the dataset contains any missing data.
- As to the shortcomings of these designs, even if analyses can be made in case of incomplete observations, the obtained results may involve certain deviations. In such cases, it may be problematic to interpret and understand the results clearly and easily. Since it is not possible to conduct simulation studies about all probable scenarios, those scenarios which are close to one another are generally preferred instead of the others. All methods display substantial deviations when the censoring rate is over 65 %. That implies that no dataset should be used when the censoring rate is over 65 %.

References

1. Strobel HA, Heineman WR (1989) Chemical instrumentation. Wiley, New York
2. Miller JC, Miller JN (1993) Statistics for analytical chemistry. Ellis Horwood, New York
3. Huston C, Juarez-Colunga E (2009) Guidelines for computing summary statistics for data-sets containing non-detects. Bulkley Valley Research Center with Assistance from the B.C. Ministry of Environment
4. Hornung R, Reed L (1990) estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg 5:46–51
5. Glass D, Gray C (2001) Estimating mean exposures from censored data: exposure to benzene in the Australian Petroleum Industry. Ann Occup Hyg 45:275–282
6. Hawkins N, Norwood S, Rock J (1991) A strategy for occupational exposure assessment. Fairview. American Industrial Hygiene, Fairfax
7. Mulhausen J, Damiano J (1998) A strategy for assessing and managing occupational exposures. American Industrial Hygiene Association, Fairfax
8. Lee L, Helsel DR (2005) Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. Comput Geosci 31:1241–1248
9. Finkelstein M, Verma D (2001) Exposure estimation in the presence of nondetectable values: another look. Ind Hyg Assoc J 62:195–198
10. Schmoyer R, Beaucamp J, Brandt C (1996) Difficulties with the log-normal model in mean estimation and testing. Environ Ecol Stat 3:81–97
11. She N (1997) Analyzing censored water quality data using a non-parametric approach. J Am Water Resour Assoc 33:615–624
12. Hewett P, Ganser G (2007) A comparison of several methods for analyzing censored data. Ann Occup Hyg 51:611–632
13. Antweiler R, Taylor H (2008) evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. Environ Sci Technol 42:3732–3738
14. Popovic M, Nie H, Chettle D, Mcneill F (2007) random left censoring: a second look at bonelead concentration measurements. Phys Med Biol 52:5369–5378
15. Chowdhury F, Gulshan J (2012) Comparison of estimation methods for left censored data. In: International conference on statistical data mining for bioinformatics health agriculture and environment, 132–141
16. El-Shaarawi A, Esterby S (1992) Replacement of censored observations by a constant: an evaluation. Water Resour Res 26:835–844

17. Fisher R (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc London Ser A 222:309–368
18. Kaplan E, Meier P (1958) Non parametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
19. Tressou J, Leblanc J, Feinberg M, Bertail P (2004) Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to ochratoxin A. Regul Toxicol Pharmacol 40:252–263
20. Hosmer D, Lemeshow S, May S (2008) Applied survival analysis: regression modeling of time to event data 618. Wiley, New York
21. Ware J, Demets D (1976) Reanalysis of some baboon descent data. Biometrics 32:459–464
22. Gilliom R, Helsel D (1986) Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques. Water Resour Res 22:135–146
23. Helsel DR, Cohn TA (1988) Estimation of descriptive statistics for multiply censored water quality data. Water Resour Res 24(12):1997–2004
24. Shumway R, Azari R, Kayhanian M (2002) Statistical approaches to estimating mean water quality concentrations with detection limits. Environ Sci Technol 36:3345–3353
25. Kroll C, Stedinger J (1996) Estimation of moments and quantiles using censored data. Water Resour Res 32:1005–1012