Journal of Modern Applied Statistical Methods

Volume 14 | Issue 1 Article 17

5-1-2015

A Comparison of Semi-Parametric and Nonparametric Methods for Estimating Mean Time to Event for Randomly Left Censored Data

Farzana Chowdhury
Department of Business Administration, Northern University Bangladesh, fchowdhury@isrt.ac.bd

Jahida Gulshan
Institute of Statistical Research and Training, University of Dhaka, gulshan@isrt.ac.bd

Syed Shahadat Hossain
Institute of Statistical Research and Training, University of Dhaka, shahadat@isrt.ac.bd

Recommended Citation
Chowdhury, Farzana; Gulshan, Jahida; and Hossain, Syed Shahadat (2015) "A Comparison of Semi-Parametric and Nonparametric Methods for Estimating Mean Time to Event for Randomly Left Censored Data," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 1 , Article 17.
DOI: 10.22237/jmasm/1430453760

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 196-207.

ISSN 1538 − 9472

A Comparison of Semi-Parametric and Nonparametric Methods for Estimating Mean Time to Event for Randomly Left Censored Data

Farzana Chowdhury
Northern University Bangladesh
Dhaka, Bangladesh

Jahida Gulshan
University of Dhaka
Dhaka, Bangladesh

Syed Shahadat Hossain
University of Dhaka
Dhaka, Bangladesh

The aim of this study was to make a comparison among existing estimation methods (Kaplan-Meier, Nelson-Aalen and Regression on Ordered Statistics (ROS)) for randomly left censored time to event data under selected distributions and for different level of censoring and sample sizes in order to determine the strength of these methods based on simulated data. Comparisons among the methods are made on the basis of unbiasedness and Monte Carlo Standard Error of the summary statistics (mean time to event) obtained by those methods under different conditions.

Keywords: Time to event data, Left censoring, detection limit, bias, Monte Carlo Standard Error

Introduction

Time to event data arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, demography, actuarial science and many other scientific areas in which time to the occurrence of some event is of interest for some population of individuals. The most typical characteristic of time to event data is incompleteness where it arises either by censoring or by truncation. Censoring, a very common feature of time to event data broadly indicates the situation that some events are known to have occurred only within certain intervals but the exact time of occurrence is unknown (Klein & Moeschberger, 2003). Among different censoring situations, left censoring provides information indicating only that the event of interest has occurred prior

Farzana Chowdhury is a Masters graduate in Applied Statistics and currently working as a lecturer in Department of Business Administration. Email her at: fchowdhury@isrt.ac.bd. Jahida Gulshan is an Associate Professor. Email her at: gulshan@isrt.ac.bd. Dr. Hossein is a Professor. Email him at: shahadat@isrt.ac.bd .

to entry into the study (Klein & Moeschberger, 2003). In other words, left censored data are commonly encountered as values below a detection limit and hence are often termed as non-detects. A detection limit is a threshold below which measured values are not considered significantly different from a blank value, at a specified level of probability (Helsel, 2005a).

Although the analysis of left-censored data has important applications in various fields of study, very few studies focused on left censoring. Owen and DeRouen (1980) used Monte Carlo simulation techniques for estimating the average exposure of industrial workers to an air contaminant. Another study on water-quality data containing multiple detection limits used a robust approach to estimate the summary statistics and model the distributions of trace-level environmental data (Lee & Helsel, 2005). Popovic, Nie, Chettle, and McNeill (2007) used inverse variance weighting (IVW) of measurements to estimate the mean and standard error of the randomly left censored data on bone lead concentrations in order to provide valid inference about bone lead concentrations. A comparison based simulation study was done by Annan, Liu, and Zhang (2009) to compare a non-parametric, a semi parametric and a parametric approach to obtain estimates of summary statistics in different censoring situations and varying sample sizes

The Kaplan-Meier (Kaplan & Meier, 1958), Nelson-Aalen (Nelson, 1972 and Aalen, 1978), Maximum Likelihood (Cohen, 1959) and the Regression on Order Statistics (ROS) (Helsel & Cohn, 1988) are the methods available in literature for computing summary statistics on data with non-detects. The objective of this study is to compare three nonparametric and one semi-parametric estimation methods for finding summary statistics.

In this study, two different algorithms of Kaplan-Meier (1958) methods, one (denoted as KM-I in the rest of this paper) proposed by Helsel (2005a) and the other one (KM-II) by Popovic et al. (2007), was compared with another non parametric method based on modified Nelson Aalen method proposed by Popovic et.al (2007) and a semi parametric method based on Regression on Order Statistics (denoted as ROS) suggested by Helsel and Cohn (1988). A Monte Carlo simulation study was conducted to determine the efficiency of these methods for analyzing left-censored data under different distributions in terms of Bias and Monte Carlo Standard Error of the mean time to event in which the methods were employed for different sample sizes and different censoring levels.

Non-parametric Estimation of Mean

Let $S(x)$ be the survivorship function that gives the proportion of subjects expected to live at least x units of time. The survival probability is a product of incremental probabilities indicating the probabilities of surviving to the next lowest detection limit, given the number of observations at and below that detection limit. The mean of survival time x is calculated by

$$\mu\left(x^*\right) = \int_{b_1}^{b_2} S\left(u\right) du \tag{1}$$

where $\mu(x^*)$ signifies that the mean of variable x is a function of the chosen interval $x_i : \{b_1 \le x_i \le b_2\}$. Parameter b_1 is the chosen lower boundary for the set of measurements.

Kaplan-Meier (KM) method

The Kaplan-Meier (KM) method proposed by Kaplan and Meier (1958) is a nonparametric method frequently considered as a standard method for estimating summary statistics of censored time to event data. The method has primarily been used for right-censored data. However, for calculation of summary statistics of left-censored data, the basic algorithm of Kaplan Meier method (used for right-censored data) has been modified. The modifications suggested are:

i. to transform left censored data to right censored one (Helsel, 2005b)
ii. to directly use left censored data with modified formulae (Popovic et al. 2007).

Formulation of KM method 1 According to the transformation method suggested by Helsel (2005b), the following steps are carried out to obtain the KM estimator of the survival probability:

i. All left-censored values are first arranged in descending order and subtracted from an arbitrarily chosen value larger than maximum value of the data set. Consequently, the left-censored data will automatically be transformed into right-censored data arranged in ascending order. All observations are then ranked from lowest to highest. For each subject $i = 1, \ldots, n$ (considering both censored and observed values), the transformed value will be

$$\omega = A_i - x_i \tag{2}$$

where A_i is an arbitrary constant, greater than the maximum observed value of the data set and x_i is the i^{th} observed value.

ii. The number of both detected and censored data that are at and below each observed value (observations at risk) are then computed as

$$b_j = n - r_j + 1 \tag{3}$$

where n is the total number of observations regarding both observed and censored and r_j is the rank of observed values only.

iii. If d_j denotes the number of observations at the j^{th} value (for tied values it is greater than 1), the incremental probabilities are given by

$$\frac{b_j - d_j}{b_j}, \; j = 1, \ldots, k \;, \tag{4}$$

and the product of the k incremental probabilities, going from high to low values for the k detected observations will give the KM estimator

$$\hat{S}(x) = \prod_{j=1}^{k} p_j \tag{5}$$

iv. The mean survival time is then estimated as

$$\begin{aligned}\hat{\mu}(x_j) = \hat{S}(x_0)x_1 + \hat{S}(x_1)(x_2 - x_1) \\ + \ldots + \hat{S}(x_{k-1})(x_k - x_{k-1}).\end{aligned} \tag{6}$$

Generally we consider $\hat{S}(x_0)=1$ and $\hat{S}(x_n)=0$.

v. The estimated mean survival time for original data will thus be

$$\hat{\mu}(x_j)=A_j-\hat{\mu}(x_j) \tag{7}$$

Formulation of KM method 2 The algorithm of this process was developed by Popovic et al. (2007) for estimating the survival function based, primarily, on the work of Kaplan and Meier (1958), Hosmer and Lemeshow (1999) and Ware and Demets (1976). According to this method, the following steps are to be carried out for obtaining this estimator:

i. For each subject $i = 1, \ldots, n, x_i$ is ordered in ascending order regarding both censored and observed data, and a censoring level δ_i is assigned such that $\delta_i = 1$, if the subject is observed and $\delta_i = 0$ if it is censored. Therefore, in case of a tie, censored entries should precede the observed events. Only the observed values along with their rank order r_i and censoring level δ_i from previous step will be considered. Thus the subjects with $\delta_i = 1$ are selected. For each entry, the incremental probabilities are calculated as

$$p_i=\frac{r_i-\delta_i}{r_i} \tag{8}$$

ii. Conventionally, $\hat{S}(x)$ is computed starting with the highest ranked entry X_n which is given as

$$\hat{S}(x)=\prod_{i=n}^{1}p_i \tag{9}$$

and the estimator of the mean for the given range $\{ x_i : \{b_1 \le x_i \le b_2\}$is given by

$$\hat{\mu}(x^*)=\hat{\mu}(b_2)-\sum_{i=n}^{1}\hat{S}(x)(x_i-x_{i-1}),\text{ where } x_0=b_1 \tag{10}$$

Since the survivorship function for left censored data equals unity for observations greater than the maximum observed event, $\hat{\mu}\left(b_2\right)$ is equal to the maximum observation in the set. As a result, the probability of having detected all observations greater than the maximum value of the data set is one. The probability decreases as x becomes progressively closer to b_1, with discontinuities at each measured event.

Nelson-Aalen method According to Popovic et al. (2007), computation method of Nelson-Aalen estimator (Nelson, 1972 and Aalen, 1978) for left-censored data set is similar to the KM method that uses left censored data directly. The basic difference between these two methods lies in the process of computing the survival probability, which instead of equation (7), is computed as

$$p_i = \frac{\delta_i}{r_i} \tag{11}$$

Semi-parametric Method (Regression on Order Statistics (ROS))

The algorithm of Regression on Order Statistics (ROS) method, developed by Helsel and Cohn (1988) can be summarized into following steps:

i. Let E_j be the probability of exceeding the j^{th} detection limit, by A_j the total number of uncensored observations in the range $[j, j + 1)$ and by B_j the total number of observations, censored and uncensored, less than or equal to the j^{th} detection limit. Note that for highest detection limit, $E_{j+1} = 0$ and $A_j + B_j = n$. The exceedance probability E_j for each detection limit can be utilized for the computation of plotting positions for both censored and uncensored data using the relation

$$E_j = E_{j+1} + \frac{A_j}{A_j + B_j}\left(1 - E_{j+1}\right) \tag{12}$$

and the number of non-detects below the j^{th} detection limit is defined as

$$C_j = B_j - B_{j-1} - A_{j-1} \tag{13}$$

ii. A Weibull-type plotting position p can be calculated for a given uncensored observation by

$$p(i) = \left(1 - E_j\right) + \frac{\left(E_j - E_{j+1}\right)}{A_j + 1} r_i \tag{14}$$

where, E_j is the exceedance probability of the censoring limit below the observation, E_{j+1} is the exceedance probability of the censoring limit above the observation and r_i is the rank of the observation falling within the j^{th} and $(j + 1)^{th}$ detection limit.

iii. The Weibull-type plotting positions for censored observations are generally given by

$$p(i) = \frac{\left(1 - E_j\right)}{C_j + 1} r_i \tag{15}$$

iv. The normal quantiles of the plotting positions are known as the order statistics of the ROS method. A linear regression of the uncensored observations against the normal quantiles of the uncensored plotting positions is formed and the regression equation for predicting the unobserved data can be obtained as

$$\text{Predicted log-value} = \beta + \alpha \times \text{ normal scores of the plotting positions} \tag{16}$$

v. The censored concentrations are modeled using the parameters of the linear regression and normal quantiles of the censored data. These modeled censored observations are used along with the uncensored observations, to model the distribution of the sample population. Individually, they are not considered the values that would have existed in the absence of censoring. The observed uncensored values are then combined with modeled censored values to corporately estimate summary statistics of the entire population. By combining both types of values this method avoids transformation bias.

Methodology

Simulation study

In this study, randomly left censored time to event data was simulated from exponential, Weibull and lognormal distribution where 1000 simulations were conducted for different combinations of sample sizes and censoring levels. The levels of censoring were considered to be 15%, 25% and 50% and the sizes of samples considered are small (25), moderately large (80) and large (200).

Results and Findings

A comparison of the methods by this simulation is made on the basis of the performances of the four methods, KM-I, KM-II, N-A and ROS in terms of absolute bias and MCSE of the estimates. Note that the performances of the four methods according to the two criteria have a nested factorial structure of its own, the factors that are taken under consideration of the simulation are:

1. Three different populations, namely exponential ($\lambda = 0.5$), Weibull ($\lambda = 1$, $k = 2$) and lognormal distribution ($\mu = 0.19$ and $\sigma = 1$)
2. Three different sample sizes 25, 80 and 200,
3. Three different levels (15%, 25% and 50%) of censored observations, and
4. Any possible interaction between the above factors.

The major findings of the simulation studies are summarized in Table 1. From these findings, it can be observed that when the populations mean is estimated using a sample drawn from an exponential (0.5) distribution, the ROS method performs the best in terms of absolute bias for all sample sizes and censoring levels considered in the study. For sample size 80, with 15%, 25% and 50% censored observations, the ROS method produced an absolute bias of 0.017, 0.037 and 0.112 respectively, which are lowest among the four methods, whereas the corresponding highest (among the four methods) absolute biases, 0.028, 0.083 and 0.412 respectively are observed for the KM-I method. Similar observations can be made for sample sizes 25 and 200 from exponential population, where ROS method produced the least absolute bias for estimate of mean for each of the censoring levels 15%, 25% and 50% and KM-I method produced the corresponding highest absolute bias.

Table 1. Comparison of Bias and Monte Carlo Standard Error (MCSE) of mean time to event for KM-I, KM-II, N-A and ROS methods under three different distributions (exponential with $\lambda = 0.5$, Weibull with $\lambda = 1$, $k = 2$ and lognormal with $\mu = 0.19$ and $\sigma = 1$)

Distribution	Sample size	Cens. level	Absolute Bias				MCSE			
			KM-I	KM-II	N-A	ROS	KM-I	KM-II	N-A	ROS
Exponential	25	0.15	0.047	0.163	0.206	0.025	0.397	0.422	0.414	0.401
		0.25	0.112	0.232	0.282	0.042	0.390	0.415	0.409	0.401
		0.50	0.486	0.194	0.301	0.108	0.367	0.401	0.387	0.418
	80	0.15	0.028	0.174	0.188	0.017	0.216	0.229	0.228	0.217
		0.25	0.083	0.234	0.252	0.037	0.211	0.225	0.225	0.217
		0.50	0.412	0.157	0.215	0.112	0.190	0.208	0.203	0.221
	200	0.15	0.025	0.169	0.175	0.017	0.141	0.148	0.148	0.142
		0.25	0.077	0.233	0.242	0.038	0.138	0.146	0.146	0.142
		0.50	0.395	0.127	0.164	0.120	0.124	0.134	0.131	0.146
Weibull	25	0.15	0.046	0.155	0.198	0.025	0.388	0.408	0.401	0.392
		0.25	0.104	0.239	0.290	0.035	0.395	0.416	0.411	0.409
		0.50	0.476	0.209	0.313	0.094	0.377	0.414	0.339	0.440
	80	0.15	0.033	0.157	0.172	0.023	0.219	0.228	0.227	0.221
		0.25	0.092	0.230	0.249	0.046	0.221	0.236	0.255	0.227
		0.50	0.416	0.155	0.215	0.119	0.192	0.210	0.207	0.227
	200	0.15	0.027	0.168	0.173	0.019	0.137	0.145	0.144	0.138
		0.25	0.079	0.231	0.240	0.041	0.133	0.140	0.140	0.137
		0.50	0.392	0.133	0.169	0.118	0.122	0.134	0.131	0.143
Lognormal	25	0.15	0.029	0.147	0.183	0.001	0.427	0.418	0.411	0.428
		0.25	0.070	0.218	0.260	0.004	0.425	0.402	0.396	0.426
		0.50	0.302	0.273	0.353	0.020	0.422	0.371	0.359	0.427
	80	0.15	0.032	0.133	0.145	0.009	0.247	0.246	0.245	0.247
		0.25	0.065	0.200	0.216	0.008	0.245	0.236	0.235	0.245
		0.50	0.265	0.228	0.271	0.001	0.237	0.208	0.204	0.242
	200	0.15	0.022	0.136	0.141	0.002	0.155	0.151	0.151	0.155
		0.25	0.055	0.203	0.211	0.001	0.154	0.147	0.146	0.154
		0.50	0.248	0.213	0.239	0.003	0.148	0.135	0.133	0.151

In case of Weibull (1, 2) population, the absolute bias produced by the ROS method is, again, the least among those of the four methods for each of the sample sizes and each of the censoring levels considered in the simulation. In comparison between methods, we can observe that for sample size 25 with 25% censored observations, absolute bias for the KM-I, KM-II, N-A and ROS methods are 0.104, 0.239, 0.289 and 0.035 respectively. For sample size 80, the computed absolute bias for the ROS method for 15%, 25% and 50% censored observations are 0.023, 0.046 and 0.119 respectively.

Considering the lognormal (0.19, 1) population, the absolute bias produced by the ROS method is still the least among those of the four methods for each of the sample sizes and each of the all censoring levels considered in the simulation. In comparison between methods, we observe for sample size 80 with 25% censored observations, absolute bias for the KM-I, KM-II, N-A and ROS methods are 0.065, 0.200, 0.216 and 0.008 respectively. For sample size 25, the computed absolute bias for the ROS method for 15%, 25% and 50% censored observations are 0.001, 0.004 and 0.020 respectively.

For all the methods and for all the sample sizes from lognormal (0.19, 1) population, the simulation results conform to the almost obvious affirmation that the absolute bias decreases as the censoring levels increases. When the samples are drawn from an exponential (0.5) or Weibull (1, 2) population, the same observation, that is, the absolute bias decreases as the censoring level increases, can be made for the KM-I and ROS methods and for all the sample sizes. The KM-II and N-A methods in cases of both exponential (0.5) or Weibull (1, 2) population, however, surprisingly showed inconsistency where the absolute bias decreases for 50% censoring levels.

The effect of increasing sample size on the absolute bias of the estimate of mean for the three methods other than the ROS method seems to be apparent for all the parent populations. For example, with exponential (0.5) population, the ROS method produces an absolute bias of 0.025, 0.017 and 0.017 for the sample sizes 25, 80 and 200 respectively at a censoring level of 15%. This eventually is indicating evidence of ROS method being insensitive to the increase of sample size from 80 to 200. The method has also been observed to be robust to the change of sample sizes with 25% and 50% of censoring levels and with Weibull (1, 2) and lognormal (0.19, 1) populations.

Although, the four methods differ substantially in terms of the bias of the estimated mean, it is noticeable that for lognormal (0.19, 1) population, the Monte Carlo Standard Error (MCSE) of the estimated mean is almost the same for the methods for same sample size and level of censoring. However, for exponential (0.5) and Weibull (1, 2) populations, slight differences in MCSEs is observed, and these differences reveal that the KM-I and ROS methods have a marginal advantage over the KM-II and N-A method. For example, for Weibull (1, 2) population, the MCSE for the four methods, KM-I, KM-II, N-A and ROS, for sample size 80 with 15% censoring level are 0.054, 0.057, 0.057 and 0.054 respectively. The difference of MCSE for different methods is seemingly higher for smaller sample sizes and higher level of censoring. The generally anticipated feature that the MCSE would be smaller for larger sample has been observed throughout.

Conclusion

The discussion in the earlier section can be summarized to reach to the following conclusions:

1. The ROS method produces the least absolute bias among those of the four methods for all sample sizes, all level of censoring for exponential (0.5), Weibull (1, 2) and lognormal (0.19, 1) populations.
2. The ROS method is more robust to the level of censoring. For increasing level of censoring, absolute bias of the estimate of mean increase for all sample sizes and all methods except for the ROS method.
3. For larger sample sizes, the MCSE of the estimate of mean of ROS method is the least among those of the four methods, although the differences of MSE are trivially small.
4. The ROS method is more robust to the change of sample size. For increasing sample size, absolute bias of both the estimates of mean increase for all levels of censoring and all methods except for the ROS method.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics, 6*(4), 701-726. doi:10.1214/aos/1176344247

Annan, S. Y., Liu, P. & Zhang, Y. (2009). *Comparison of the Kaplan-Meier, Maximum Likelihood, and ROS Estimators for left-censored data using simulation studies*. http://homepage.divms.uiowa.edu/~kcowles/s166_2009/Annan.pdf.

Cohen, A. C. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, *1*(3), 217-237. doi:10.1080/00401706.1959.10489859

Helsel, D. R. & Cohn, T. A. (1988). Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, *24*(12), 1997-2004. doi:10.1029/WR024i012p01997

Helsel, D. R. (2005a). More than obvious: Better methods for interpreting nondetect data. *Environmental science & technology*, *39*(20), 419A-423A. doi:10.1021/es053368a

Helsel, D. R. (2005b). *Nondetects and data analysis: Statistics for censored environmental data*. Hoboken, NJ: Wiley-Interscience.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied survival analysis: Regression modelling of time to event data*. New York, NY: Wiley.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457-481. doi:10.1080/01621459.1958.10501452

Klein, J. P. & Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. New York: Springer-Verlag.

Lee, L. & Helsel, D. (2005). Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. *Computers & Geosciences*, *31*(10), 1241-1248. doi:10.1016/j.cageo.2005.03.012

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945-966. doi:10.1080/00401706.1972.10488991

Owen, W. J. & DeRouen, T. A. (1980). Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants. *Biometrics*, *36*(4), 707-719. doi:10.2307/2556125

Popovic, M., Nie, H., Chettle, D. R. & McNeill, F. E. (2007). Random left censoring: A second look at bone lead concentration measurements. *Physics in Medicine and Biology*, *52*(17), 5369-5378. doi:10.1088/0031-9155/52/17/018

Ware, J. H. & Demets, D. L. (1976). Reanalysis of some baboon descent data. *Biometrics,* 32(2), 459-463. doi:10.2307/2529516