A Study on Imputing Censored Observations for Exponential Distribution Based on Random Censoring

Kuo-Ching Chiou
Department of Finance
Chaoyang University of Technology
Taichung County, Taiwan, R. O. C.
E-mail: kcchiou@mail.cyut.edu.tw

Abstract

Censoring models are frequently employed in reliability analysis to reduce experimental time. There are three censoring model: type-I, type-II and random censoring. In this study, we focus on the right-random censoring model. In the previous literature, an imputation of the censored observation is considered as the censoring time (Miller (1981), Lawless (1982), Lee (1992) and among others). Clearly, the censored observation is imputed by the censoring time to underestimate the original failure time. In this paper, we consider the failure time to follow an exponential distribution, and alternatively we attempt to propose three measures to impute the censored observations. By Monte Carlo simulation, the goodness of fit test is employed to compare the four methods of imputing censored observations. It is found that the method of imputing censored data by censoring time little outperforms the other three imputing methods.

Keywords: Right-random censoring, failure time, censoring time, imputation value, exponential distribution, goodness of fit test

1. Introduction

Three censoring cases, type-I, type-II and random censoring, are frequently employed in reliability analysis to save experimental time. Both type-I and type-II censoring cases are commonly used in engineering applications and the random censoring case is often implemented in medical studies involving animals or clinical trials [5].

In this study, the right-random censoring setting is considered. The right-random censoring process is one in which an individual is assumed to have a failure time T and censoring time C, where T and C are independent continuous random variables. Assume that n individuals are considered and the *i*th individual has a failure time T_i and a censoring time C_i, for i = 1, 2, ..., n. Allow $Y_1, Y_2, ..., Y_n$ to be the data from a right-random censoring setting. Researchers considered $Y_i = \min(T_i , C_i)$ for i = 1, 2, ..., n. Data in such a setting can be conveniently represented as (Miller [5], Lawless [2], Lee [3, 4]). Random variables (Y_i , δ_i), i= 1,2…n, $\delta_i = 1$ (uncensored) if $T_i \le C_i$,$\delta_i = 0$ (censored) if $T_i > C_i$ for i = 1, 2, ..., n. Therefore, $Y_i = \delta_i T_i + (1 - \delta_i) C_i$ indicate whether the failure time T_i is censored or not. If Y_i is a censored observation, denoted by T_i^+, Researches considered the censoring time C_i to be an imputation of the censored observation T_i^+ ((Miller[5], Lawless[2], Lee[3, 4]). Clearly, the T_i^+ imputed by the censoring time C_i would likely underestimate the original failure time T_i. Buckley [1] proposed the pseudo random variables, in which $Y_i = T_i \cdot\delta_i + E(T_i \mid T_i > C_i) \cdot(1 - \delta_i)$ where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n. Tong and Chiou [6] proposed two imputations methods of T_i^+ to estimate the quantiles for an exponential distribution. By simulation, their results confirm the estimates quantiles under the two imputing methods are superior to that under the censoring time imputation for mediate and high quantiles.

The exponential distribution is widely used to model lifetimes in both the theoretical study of reliability and practical reliability engineering. In this study, we consider that the failure time T follows an exponential distribution distribution. Furthermore, compare the right-random censored data $(y_1, y_2, \ldots, y_n)$, which distributes to original data (the exponential distribution) by goodness of fit test.

In this paper, we compare four methods to impute T_i^+, given the relevant parameters: sample size n, censoring rate p (p = r / n, r is the number of the uncensored data), scalar parameter θ and number of replications N. By Monte Carlo simulation, we employ the goodness of fit test (i.e. chi-square test) [3], to assess which method collect the right-random censored data to be approximated to the original data (the exponential distribution).

The organization of this paper is as follows: In section 2, the conditional expectations of an empirical distribution and an exponential distribution are derived, respectively. Section 3 introduces the four imputing methods for censored observations. We construct four steps for Monte Carlo simulation study and provide the best imputing method for censored observations in section 4. Section 5 presents conclusions.

2. Deriving the conditional expectations

2.1 Deriving the approximation of conditional expectation for the empirical distribution

Given T > k, the conditional cumulative distribution function (c.d.f.) of a continuous random variable (r.v.) T is

$$\Pr\{T \le t \mid T > k\} = \frac{\Pr\{k < T \le t\}}{\Pr\{T > k\}} = \frac{\int_k^t f(x)dx}{1 - F(k)}$$

, $t > k$. (1)

Given T > k the conditional probability density function (p.d.f.) of a continuous random variable (r.v.) T can be obtained by differentiating (1) with respect to t as follows:

$$f(t \mid T > k) = \frac{f(t)}{1 - F(k)} \quad , t > k. \tag{2}$$

The conditional expected value of a continuous r.v. T given T > k is

$$E(T \mid T > k) = \int_k^\infty t \cdot f(t \mid T > k)dt = \frac{\int_k^\infty t \cdot f(t)dt}{1 - F(k)}. \tag{3}$$

For most reliability distributions, no simple closed form generally exists as Eq. (3). However, Eq. (3) can be approximated by a nonparametric empirical distribution as follows [3].

Let $t_{1:n}$, $t_{2:n}$, ... , $t_{n:n}$ be the ordered observations of $T_{1:n}$, $T_{2:n}$, ... , $T_{n:n}$, respectively, then

$$\int_{t_{i:n}^+}^\infty f(x)dx \approx \frac{n-i}{n}, \tag{4}$$

$$\int_{t_{i:n}^+}^\infty x \cdot f(x)dx \approx \sum_{j=i+1}^{n} \frac{t_{j:n}}{n}. \tag{5}$$

Combining Eq.(4) and Eq.(5), we obtain

$$E(T \mid T > t_{i:n}) \approx \sum_{j=i+1}^{n} \frac{t_{j:n}}{n-i}. \tag{6}$$

2.2 Deriving the conditional expectation for an exponential distribution

If a random variable T follows an exponential distribution with mean θ, then the p.d.f. of T is

$$f(t) = \frac{1}{\theta} e^{-t/\theta}, \; t > 0, \tag{7}$$

where the scale parameter θ is positive.

The conditional p.d.f. and the expected value of T, given T > k, are

$$f(t \mid T > k) = \frac{1}{\theta} \cdot e^{-(t-k)/\theta} , \; t > k, \tag{8}$$

$$E(T \mid T > k) = k + \theta. \tag{9}$$

The estimator of θ is as follows:

$$\hat{\theta} = \overline{T} \tag{10}$$

where $\hat{\theta}$ a maximum likelihood estimator of θ and $\overline{T}$ is the sample mean.

3. Imputation of censored observations based on right-random censoring

In a right-random censoring, researchers [2, 3, 4, 5] contend that for a situation in which observation y_i is a censored datum, then the censoring time c_i is an imputation of the censored observation T_i^+. To conquer underestimation, Tong and Chiou applied the pseudo random variables [1], which are utilized to construct the two imputations of each of censored observations. The two imputations proposed herein can be obtained by substituting the values of $E(T_i \mid T_i > C_i)$ calculated by the following two methods [6]:

(1) Non-parametric method: By Eq.(6), $E(T_i \mid T_i > C_i) \approx \sum_{\substack{j=1 \\ j \ne i}}^{n} \{T_j \mid T_j \ge C_i$ and T_j is an uncensored datum$\} / n_{ui}$, where n_{ui} denotes the number of $\{T_j \mid T_j \ge C_i$ and T_j represents an uncensored datum, j = 1, ... , n, $j \ne i\}$.

(2) Parametric method: If the continuous r.v. T follows an exponential distribution, then the conditional expected value $E(T_i \mid T_i > C_i)$ is equal to $C_i + \theta$, where the estimator θ is $\hat{\theta}$ is formulated as [2]:

$$\hat{\theta} = \sum_{i=1}^{n} Y_i \Big/ r, \tag{11}$$

where r is the uncensored data and $Y_i = \min(T_i, C_i)$, i = 1, 2, …, n.

In this paper, we propose another imputation measure that the censored observation T_i^+ is imputed by Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n $\}$.

In right-random censoring, the experimental data, $y_1, y_2, \ldots, y_n$, are collected from an exponential distribution, the imputing methods are denoted by symbols as follow. (i) $Y_i = \min(T_i , C_i)$ for i = 1, 2, ..., n. The method is denoted by "M1" that the censoring time C_i is an imputation of the censored observation T_i^+. (ii) $Y_i = T_i \cdot \delta_i + E(T_i \mid T_i > C_i) \cdot (1 - \delta_i)$ where $\delta_i = 1$ if $T_i \le C_i$ and $\delta_i = 0$ if $T_i > C_i$ for i = 1, 2, ..., n. By non-parametric method, the two methods of imputations ($E(T_i \mid T_i > C_i) \approx \sum_{\substack{j=1 \\ j \ne i}}^{n} \{T_j \mid T_j \ge C_i$ and T_j is an uncensored datum$\} / n_{ui}$, and Median$\{T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …,

n }) are denoted by "M2" and "M4", respectively. By parametric method, the method is denoted by "M3" that the estimate ($C_i + \hat{\theta}$) (by Eq. (11)), $E(T_i \mid T_i > C_i)$, is a imputation of the censored observation T_i^+.

In right-random censoring, the experimental data, $y_1, y_2, \ldots, y_n$ (containing uncensored and censored data) are drawn from an exponential distribution for four imputing methods, respectively. And then we use the goodness of fit test to assess which data distribute to be closed to original data (the exponential distribution).

4. Simulation study and results

4.1 Simulation

In right-random censoring, given the failure time T following an exponential distribution, a Monte Carlo simulation was conducted to compare the performances of the four imputing methods ("M1", "M2", "M3" and "M4"). The results that deciding the optimum imputing measure is independent of the scale parameter θ are obtained. Therefore, the relevant parameters are given as follows: sample size n = 30, 50, 100, censoring rate p = 0.1 (0.1) 0.5, and θ = 1.

The combinations of (n, p and θ), N = 1000 replications are generated by using IMSL STAT/LIBRARY (C Functions for Statistical Analysis). The simulation procedure is described as:

Step 1: Generate the data failure time T_i which follows an exponential distribution with mean θ =1, for i = 1, 2, ..., n.

Step 2: Determine the censoring time C for given censoring rate p.

In this paper, we choose two random variables, the failure time T and the censoring time C, to follow an exponential distribution, because the exponential distribution is applied broadly for reliability analysis. Let the failure time T follow an exponential distribution with mean θ (given), and the censoring time C follow an exponential distribution with mean θ_c (unknown). Therefore, θ_c can be obtained by the following equation.

$$p = P_r\{T > C\} = \int_0^\infty e^{-z/\theta} \frac{1}{\theta_c} e^{-z/\theta_c} dz \qquad (12)$$

By the Eq. (12), $\theta_c = \theta(1-p)/p$, θ = 1. So, θ_c is (1-p)/p. Next, we generate the censoring time C from an exponential distribution with mean (1-p)/p, for i = 1, 2, ..., n.

Step 3: Accumulate the data Y_i for i = 1, 2, ..., n in a right-random censoring.

(i) $Y_i = \min(T_i, C_i)$ for i = 1, 2, ..., n.

(ii) $Y_i = T_i \cdot \delta_i + E(T_i \mid T_i > C_i) \cdot (1 - \delta_i)$ for i = 1, 2, ..., n. If $\delta_i = 0$ then $Y_i = E(T_i \mid T_i > C_i)$. The value of $E(T_i \mid T_i > C_i)$ is then replaced by method (1) (non-parametric method), method (2) (parametric method) in Section 3.1.

(iii) $Y_i = T_i \cdot \delta_i$ + Median{$T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n }$\cdot(1 - \delta_i)$ for i = 1, 2, ..., n.

In this Step, the maximum datum of (Y_1, Y_2, ...,Y_n) must be constrained to be an uncensored datum. Otherwise, Eq. (6) can not be used and the data are discarded.

Step 4: The experimental data, $y_1, y_2, \ldots, y_n$, are obtained by four imputing methods ("M1", "M2", "M3" and "M4"), respectively. By goodness of fit test, the null hypothesis is H_0: the experimental data, $y_1, y_2, \ldots, y_n$ follow an exponential distribution with mean θ =1. By 1000 replications, we can obtain the number that the test conclusion does not reject H_0 for given the significant level $\alpha = 0.05$. If we fail to reject H_0, we suggest the best of imputing method that the collected number is the largest.

4.2. Results

As mentioned earlier in Section 3, the two imputations ($E(T_i \mid T_i > C_i)$ and Median{$T_j \mid T_j > C_i$ and T_j is an uncensored datum, j = 1, 2, …, n }) to impute censored observation are obviously larger than the imputation censoring time C_i.

In Table 1, the results indicate the best imputing measure as follows:

(1) For censoring rate $p \leq 0.3$, the imputing method would be "M3".

(2) For censoring rate $p \geq 0.4$, employing imputing "M2" is preferred.

(3) In n = 100 and p = 0.5, the collected number is too smaller that the conclusions fail to reject H_0, denoted as "*", because the censoring rate p is too large to distribute, $y_1, y_2, \ldots, y_n$, more different from the original data (the exponential distribution).

5. Conclusions

It is apparent that the censored observation is underestimated by censoring time. Tong and Chiou [6] proposed that the censored observation T_i^+ was imputed by $E(T_i \mid T_i > C_i)$. Given by the imputation $E(T_i \mid T_i > C_i)$, the estimates of moderate and high quantiles are superior to the imputation censoring time C_i. In this study, we intend whether the three imputing methods ("M2", "M3" and "M4") are superior to imputing method "M1" or not for distributing of data ($y_1, y_2, \ldots, y_n$,). By goodness of fit test, the results indicate the three imputing

methods ("M2", "M3" and "M4") could outperform the imputing method "M1" (in Table 1). As shown in Table 1, the imputing method "M3" ($p \leq 0.3$) and "M2" ($p \geq 0.4$) are preferred to implement under the different setting, respectively.

Table1. Collected numbers of not reject H_0 (N=1000)

n	p	M1	M2	M3	M4	best method
30	0.1	934	936	946	914	M3
	0.2	896	880	917	781	M3
	0.3	777	777	840	618	M3
	0.4	555	616	687	426	M3
	0.5	308	435	330	232	M2
50	0.1	947	939	951	875	M3
	0.2	875	882	912	728	M3
	0.3	731	735	768	480	M3
	0.4	453	523	487	265	M2
	0.5	129	267	94	87	M2
100	0.1	937	928	937	826	M3
	0.2	838	827	856	517	M3
	0.3	589	568	620	214	M3
	0.4	216	273	147	38	M2
	0.5	12	36	0	3	*

"": do not propose the best imputing method*

For engineers, they get the experimental data in right-random censoring. By goodness of fit test or hazard plot [3], if the experimental data follow an exponential distribution and then, our results suggest some suitable imputing measures to obtain the reliable data (see Table 1) to analyze parametric estimating, quantile estimating, and so on.

References

[1] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, 66, pp. 429-436, 1979.

[2] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley, New York, 1982.

[3] E. T. Lee, *Statistical Methods for Survival Data Analysis (2_{nd} ed)*, John Wiley and Sons, New York, 1992.

[4] J. B. Lee and E. Max, *Statistical Analysis of Reliability and Life-testing Models (2_{nd} ed)*, Marcel Dekker, New York , 1991.

[5] R. G. Miller, *Survival Analysis*, John Wiley and Sons, New York , 1981.

[6] L. I. Tong and K. C. Chiou, "Estimating the censored observations under random censoring model for the Exponential distribution," *Journal of the Chinese Institute of Industrial Engineers*, 16, pp. 85-92, 1999.