Available online at www.dergipark.gov.tr/beuscitech

Journal of Science and Technology

E-ISSN 2146-7706

Nonparametric estimation of a renewal function in the case of censored sample

Çiğdem Cengiz[a,*]

[a] Bitlis Eren University, , Faculty of Science, Department of Statistics, TR-13000, Bitlis Turkey

ARTICLE INFO

Article history:
Received 20 April 2019
Received in revised form 09 December 2019
Accepted 09 December 2019

Keywords:
Renewal function,
random right-censored,
Nonparametric,
Kaplan-Meier

ABSTRACT

A renewal process is a counting process which counts the number of renewals that occurs as a function of time, wherein the durations between successive renewals are random variables independent of one another, with identical F distributions. The mean value function data is frequently needed in applications of renewal processes. For the renewal function, open expressions depending on distribution function F can be calculated from each other. However, even though the distribution function F is known, the renewal function cannot be obtained analytically except for a few distributions. In this study, in the case that F is totally unknown, life table management and Kaplan-Meier estimator were used depending on random right-censored sampling for the estimation of F value. Then, for the estimation of the renewal function value in the random right-censored data, nonparametric estimators were proposed and the problem of how to calculate these estimators were discussed.

1. Introduction

The renewal theory has numerous applications, in a broad range of applied probability problems and is used for the modelling of subsequent repairs of a broken machine or relocations in reliability problems. The renewal theory is also proven to be a powerful tool in the planning problems of the number of workers. The renewal process wherein is used for modelling the sequence of the departure of a job given by assignment.

Let be $\{X_n, n=1, 2, ...\}$ a non-negative, independent set of random variables having the same F distribution function, and let $F(0)= P(X_n=0)<1$, that is X_n be not equal to zero with a possibility.

Let X_n: $(n-1)$ be the time interval passing until renewals n after the renewal.

$S_o=0$, $S_n= X_1+....+ X_n$, $n\geq1$, where S_n is the random variable and n is the time interval until the renewal is made.

Let $N(t)=sup\{n:S_n\leq t\}$, for each $t\geq0$. $N(t)$ is the number of renewals up to time t, that is, the number of those between the time interval $[0, t]$. The renewal random variable of $N(t)$ and the stochastic process of $\{N(t), t\geq0\}$, which are defined so, are also called a "renewal process" [1-6].

$\{N(t), t\geq0\}$ being a renewal process, the M mean value function given by $M(t)=E(N(t)), t\geq0$, is called a renewal function [7, 8]. Where, $M(t)$ is the number of renewals made in time interval [0, t].

$$I_k = \begin{cases} 1, & S_k \leq t \\ 0, & S_k > t \end{cases} \text{ and } N(t) = \sum_{k=1}^{\infty} I_k \text{ so when}$$

$$\mathrm{E(N(t))}=\left(\sum_{k=1}^{\infty} I_k\right) = \sum_{k=1}^{\infty} E(I_k) = \sum_{k=1}^{\infty} P(S_k \leq t) = \sum_{k=1}^{\infty} F^{k^*}(t) \quad (1)$$

Then

$$M(t) = \sum_{k=1}^{\infty} F^{k^*}(t), \qquad t\geq 0 \quad (2)$$

By using equality (2), it is easily obtained that M is a right-continuous and non-decreasing function. However, when $\lim_{\infty} M(t) = \infty$, where renewal function M has all the features of a distribution function except not converging to 1 for $t \to \infty$.

Theorem 1 Let be a finite non-arithmetic distribution function with F and a second momentum μ_2. So [9],

* *Corresponding author.*
E-mail address: cigdemcengiz44@gmail.com

$$\lim_{t\to\infty}(M(t)+\frac{t}{\mu})=\frac{\mu_2}{2\mu^2} \quad (3)$$

Theorem 2. "The Blackwell Theorem" is very useful for calculating *M(t)* for the big values of the asymptotic expression *t* given for *M(t)* and also for the estimation of *M(t)* when F is known. Let μ and σ^2 be $c^2=\frac{\sigma^2}{\mu^2}$ to show the expected values and variance of *F* distribution, where *c*, is the variation coefficient of *F*.

In this case, $M(t)\approx\frac{t}{\mu}+\frac{1}{2}(c^2-1)$ can be derived from the asymptotic expression (3) for the sufficiently large *t*. When *F* is a mean exponential distribution function μ, that is, *c=1*, the above asymptotic expansion is equal to *M(t)* itself for every $t\geq 0$. If c^2 is not too large or too small, this asymptotic expansion is true for some values of t in practice. Quantitative studies show that the expansion $\frac{t}{\mu}+\frac{1}{2}(c^2-1)$ can be used for *M(t)* in practice for $t\geq t_0$, where [9]

$$t_o=\begin{cases}\frac{3}{2}c^2\mu, & c^2>1\\ \mu, & 0.2<c^2\leq 1\\ \frac{\mu}{2c}, & 0<c^2\leq 0.2\end{cases}$$

When $c^2\to 0$, the expansion deteriorates. The relative error in the approach is typically below 5% for $t\geq t_0$ and mostly below 2%.

2. Right random censoring

Censoring is to ignore the data that cannot be known exactly and observed for any reasons due to a number of limitations, such as time and cost [10]. For example, the death of an individual from a different cause such as traffic accident within the period of his/her treatment.

Different kinds of censoring are given in the literature [11]. Some of them are as follows:

- Right Censoring
- Type I Censoring (censoring of time)
- Random Censoring (random censoring of time)
- Type II Censoring (partial censoring)
- Left Censoring
- Dual Censoring
- Range Censoring

We are only concerned with random right censoring of the above kinds of censorings in this study. This censoring status is discussed in detail in the following section.

2.1. Life table methods and product-limit (Kaplan-Meier) estimate

Let's suppose that *F* is unknown for a renewal process whose distribution function of time ranges passed between renewals is *F*. Let's consider a sample of n units that was randomly right censored from the *F* distribution, let the observations be classified so as to know from what ranges the components are deteriorated or censored and let their lives and censoring ranges be not known. As the time axis being an upper limit on $a_o=0$, $a_k=t$, $a_{k+1}=\infty$ and *t* observation , be divided *k+1* pieces of $I_j=[a_{j-1},a_j)$ $j=1,2,.....k+1$ ranges. Therefore, the observations consist of the numbers of lives and censoring ranges for each *k+1* range. I_{k+1}, the last range, will be able to be approached as the range of life intervals only, all the components not deteriorated up to *t* time, have to deteriorate in a time in I_{k+1}.

n_j is the number of components (working and uncensored) under risk at time a_{j-1}

d_j is the number of deteriorated ones at I_j .

w_j is the number of the ones censored.

Since the number of components living at the beginning of I_j is n_j ;

$n_1=n$ and $n_j=n_{j-1}-d_{j-1}-w_{j-1}$, $j=1,2,.....k+1$

Let denote a random variable with an *F* distribution function.

$\bar{F}(a_j)=P(Y>a_j)$, $j=1,2,\dots\dots k+1$

$p_j=P(Y>a_j|Y>a_{j-1})$ and q=1-p_j, j=1, 2, ..., k+1

Since;

$\bar{F}(a_o)=1$

$\bar{F}(a_1)=P(Y>a_1)$

$\bar{F}(a_2)=P(Y>a_2)=P(Y>a_2|Y>a_1)\,P(Y>a_1)$

.

.

.

$\bar{F}(a_k)=P(Y>a_k)=P(Y>a_k|Y>a_{k-1})P(Y>a_{k-1}|Y>a_{k-2})\dots P(Y>a_1)$

Then

$$\bar{F}(a_j)=p_1p_2\dots.,k+1 \qquad j=1,2,\dots,k+1 \quad (4)$$

Now, let's consider the estimation problem of $\bar{F}(a_j)$. If the data are not censored, there will be no difficulties while performing this. So, an open estimation is the best probability estimator $\frac{n_{j+1}}{n}$ for $\bar{F}(a_j)$ $\frac{n_{j+1}}{n}$, where a_j is the ratio of working observations. If the ranges comprise back-off that is censorings, it will not be this way, because n_{j+1} is not necessarily the number of working components in time a_j. Since it is likely that some of the censored components will still work in a_j, $\frac{n_{j+1}}{n}$ will tend to estimate in most cases from below. This problem can be solved with the life table method given below.

If there is no censoring in range I_j, $\hat{q}_j=\frac{d_j}{n_j}$ is a significant estimation of $\hat{q}_j$, because when component $\hat{q}_j$ is known to be working at the beginning of I_j, this is the conditional possibility of deterioration of that component in I_j. However, if $w_j>0$, $\frac{d_j}{n_j}$

can be expected to estimate q_j from below, because some of the censored components in I_j can be deteriorated before the end of I_j. Additionally, those previously censored but deteriorated in I_j are not already within n_j. Therefore, it is desirable to make an arrangement for censored observations. The most commonly used method is the estimation made by

$$\hat{q}_j=\frac{d_j}{n_j-\frac{w_j}{2}} \quad (5)$$

which calls the number of q_j as the standard life table estimation.

The expression 5, suggests that $n_j>0$. When $n_j=0$, it is defined by $\hat{q}_j=1$ for compatibility reasons. $n_j' = n_j - \frac{w_j}{2}$ can be considered as an efficient number of the components under risk for I_j range; with this expression, it is , to some extent, accepted that a censored component work under risk for half of the range. This arrangement is arbitrary, but mostly reasonable. In some cases, other estimators of q_j are preferred. For example, if all the censorings in I_j take place on the right at the end of I_j, the estimation of $\hat{q}_j=\frac{d_j}{n_j}$ will be suitable, while all the censorings do not take place at the beginning of I_j, $\hat{q}_j=\frac{d_j}{n_j-w_j}$ will be suitable. After estimations of $\hat{q}_j$ and $\hat{p}_j = 1 - \hat{q}_j$ are calculated, with the help of $\bar{F}(a_j)$, equation 4

$$\bar{F}(a_j) = \hat{p}_1 \ldots\ldots \hat{p}_j, \qquad J = 1,2,\ldots.k+1 \quad (6)$$

can be estimated.

The life table itself is a table showing the data, $\hat{q}_j$ and $\hat{p}_j$ estimations. This table includes columns giving the values of n_j d_j, w_j, $\hat{q}_j$ and $\hat{p}_j$ and $\hat{p}_j$ for each. n'_j and $\hat{p}_j$ are sometimes placed in the columns giving the estimations of other characteristics of the distribution. For all particular cases where $w_j=0$, $\hat{p}_j$ is reduced to aforementioned $\frac{n_{j+1}}{n}$ estimation for an uncensored case. Suppose that the deterioration intervals in the sample of n units and the censoring intervals of the observations whose deterioration times are not observed are known, then the estimator of $\bar{F}(t)$ is;

$$\hat{\bar{F}}(t) = \frac{Equal\ to\ or\ greater\ number\ of\ observations\ on\ lifetime\ t}{n} \quad (7)$$

In the case that the sample is censored, since the number of the observations whose lives are equal to or longer than t cannot be known exactly, the estimator will have to be re-arranged from equation 7. With the arrangement given below, the estimation of $\bar{F}(t)$ called product-limit or Kaplan-Meier can be reached [13].

Let's suppose that k pieces of n components deteriorate at different times of $t_1<t_2<\ldots\ldots<t_k$, and that at t_j $(j=1,2,\ldots., k)$, more than one deterioration are allowed, and let d_j be the number of deteriorations at t_j. In addition to the deterioration times in $t_1, t_2, \ldots\ldots, t_k$, there are also censoring times for the components whose lives are not observable. Then, the Kaplan-Meier (product-limit) estimation of $F(t)$ is denoted with [11, 12];

$$\hat{F}_{KM}(t) = 1 - \prod_{j:t_j<t} \frac{n_j-d_j}{n_j} \quad (8)$$

Example 1 the deterioration times (in months) for a unit of a device are given Table 1 below. Here, the censoring for the deterioration times are claimed to be random and independent and the deterioration times are given in Table 1 below.

Table 1. The deterioration times (in months) for a unit of a device are

Uncensored Observations								
36.3	41.7	43.9	49.9	50.1	50.8	51.9	58.9	60.7
52.1	52.3	52.3	52.4	52.6	52.7	53.1	59.0	
53.6	53.6	53.9	53.9	54.1	54.6	54.8	59.1	
54.8	55.1	55.4	55.9	56.0	56.1	56.5	59.6	
56.9	57.1	57.1	57.3	57.7	57.8	58.1	60.4	

Censored Observations
26.8 29.6 33.4 35.0 40.0 41.9 42.5

Table 2. Kaplan-Meier estimators of the device at the given t points

T	$\hat{F}_{KM}(t)$	T	$\hat{F}_{KM}(t)$
36.3	0.0227	55.1	0.5597
41.7	0.0460	55.4	0.5842
43.9	0.0705	55.9	0.6086
49.9	0.0949	56.0	0.6331
50.1	0.1194	56.1	0.6575
50.8	0.1438	56.5	0.6820
51.9	0.1683	56.9	0.7065
52.1	0.1928	57.1	0.7554
52.3	0.2417	57.3	0.7798
52.4	0.2662	57.7	0.8043
52.6	0.2906	57.8	0.8288
52.7	0.3151	58.1	0.8532
53.1	0.3395	58.9	0.8777
53.6	0.3885	59.0	0.9022
53.9	0.4374	59.1	0.9266
54.1	0.4618	59.6	0.9511
54.6	0.4863	60.4	0.9755
54.8	0.5352	60.7	1.0000

3. Material and Method

In applications regarding renewal processes, generally the data of renewal function, which is the mean value function of this process, is needed. While the distribution function is known as modal in practice, some of its parameters are not known or is totally unknown. In such cases, depending on the sample of n units taken from distribution, the values of need to be

estimated, in this case, a non-parametric estimation of is made [14].

Let $X_1,\ldots, X_n$ be a sample of n units randomly censored from F distribution Depending on this sample, let's consider the Kaplan-Meier estimator of;

$$\hat{F}_{KM}(t) = 1 - \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}$$

In the convolution set equation 2, expression of $M(t)$,

$$\widehat{M}_{KM}(t) = \sum_{k=1}^{\infty} \hat{F}_{KM}^{k^*}(t) \quad (9)$$

defined by taking the estimator $\hat{F}_{KM}(t)$ for $F(t)$, can be considered as a non-parametric estimator of $M(t)$, where $\hat{F}_{KM}^{k^*}(t)$ is the n times Stieltjes convolution of $\hat{F}_{KM}(t)$ by itself. The value of estimator $\widehat{M}_{KM}(t)$ for each constant value of the sample $(X_1,\ldots, X_n)$ cannot be easily obtained from equation (9). For each constant value of the sample $(X_1,\ldots, X_n)$, $\widehat{M}_{KM}(t)$ from renewal equation $M(t) = F(t) + \int_0^t F(t-x)dM(x), \quad t \geq 0$ can be denoted as;

$$\widehat{M}_{KM}(t) = \hat{F}_{KM}(t) + \int_0^t \widehat{M}_{KM}(t-x)\ d\widehat{M}_{KM}(x) \quad (10)$$

Thus, $M_{KM}(t)$ is obtained by solving this equation. Schneider, Lin and O'Cinneide proposed a numerical solution of an integral equation like equation 10 [6]. Now, let's briefly examine this method. By dividing the $(0,t)$ range, X_i, $I=1, 2,\ldots, n$ values are rescaled multiplying by an appropriate h integer and rounded down to the closest number. And as a result, they are converged to the function F_{KM} with an arithmetic function for the calculation of the renewal function. Similar to RS-method approach, by the replacement of the integer in equation 10 with a finite sum, $\widehat{M}_{KM}(t)$ can be calculated subsequently. $t=k$ and F_a as the abovementioned arithmetic distribution function, $\widehat{M}_{KM}$ is calculated subsequently from the sum of

$$M_{KM}^a(i) = F_a(i) + \sum_{j=1}^{i} M_{KM}^a(i)\big(F_a(j) - F_a(j-1)\big), i = 1,2, \ldots k \quad (11)$$

When $i=1,2,\ldots, k$ $M_{KM}(i/h)=M_{KM}^a(i)$.

4. Conclusions

In this study, some concepts forming basis for renewal processes are introduced, and the renewal theory is discussed, then the non-parametrical Kaplan-Meier estimation of F is explained in the case of random right-censored sampling, which is a type of censoring commonly-used in censored renewal processes.

Acknowledgements

I would like to thank Halil Aydoğdu for their support.

References

[1] Ross, M.S., 1983. *Stochastic Processes*. John Wiley&Sons, 510, New York.

[2] Aydoğdu, H. 1997. Yenileme Süreçlerinde Tahmin. Doktora Tezi. Ankara Üniversitesi , 153 s, Ankara.

[3] Barlow, E. R. and Proschan, F. 1965. Mathematical Theory of Reliability. John Willey& Sons, 290. New York.

[4] Binshan, L. 1988. Estimation of the Renewal Function. The İnterdepartmental Program in Business Administration. 84. Lousiana.

[5] Frees, E. W. 1986b. Nonparemetrik Renawal Function Estimation. Ann. Statist., Vol. 14 (4), 1366-1378.

[6] Frees, E. W. 1986a. Warranty and Renewal Function Estimation. Nov. Res. Logist. Q.,33,361;372

[7] Karlin, S., and Taylor H. M., 1975. A First Course in Stochastic Processes, Second edition. Acedemic Press, 557, New York, 1975.

[8] Parzen, E. 1962. Stochastic Processes. Holden-Day, San Fransisco.

[9] Tijms, 1994. *Stochastic Models: An Algorithmic Approach*, John Wiley&Sons, New York, 1994.

[10] Tamam D., 2008. *Tam ve Sansürlü Örneklem Durumlarında Weibull Dağılımı için Bazı İstatistiki Sonuç Çıkarımları,* Yüksek Lisans Tezi. Ankara Üniversitesi, Ankara.

[11] Lawless, J. F., 2003. *Statistical Models and Methods for Lifetime Data*, John Wiley&Sons, 630, Canada.

[12] Schneider H., Lin B. S., and O'Cinneide C., 1990. Comparison of Nonparametrik Estimators for the Renewal Function. *Journal Roy. Statist. Soc. Ser. C.,* 39, 55-61.

[13] Kaplan, E. L. and Meier, P. Nonparametric Estimation From Incomplete Observations. Journal of the American Statistical Association, Vol. 53 (282), 457-481.

[14] Cengiz Ç., Yapıcı İ., Cengiz M.S. 2018. Fourier Analysis in Rail Systems. International Conference on Multidisciplinary, Science, Engineering and Technology, Oct 25-27, 2018, Dubai.