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

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 \( \{X_n, n=1, 2, \ldots\} \) a non-negative, independent set of random variables having the same \( F \) distribution function, and let \( F(0) = P(X_0 = 0) < 1 \), that is \( X_0 \) be not equal to zero with a possibility.

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

\( S_0 = 0, S_n = X_1 + \ldots + X_n, n \geq 1 \), 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 \( t \geq 0 \). \( N(t) \) 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 \( \{N(t), t \geq 0\} \), which are defined so, are also called a “renewal process” [1-6].

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

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

\[
E(N(t)) = \sum_{k=1}^{\infty} E(I_k) = \sum_{k=1}^{\infty} P(S_k \leq t) = \sum_{k=1}^{\infty} F^*(t)
\]

(1)

Then

\[
M(t) = \sum_{k=1}^{\infty} F^*(t), \quad t \geq 0
\]

(2)

By using equality (2), it is easily obtained that \( M \) is a right-continuous and non-decreasing function. However, when \( \lim_{t \to \infty} M(t) = \infty \), 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 \( \mu_2 \). So [9],
\[
\lim_{t \to \infty} (M(t) + \frac{1}{\mu}) = \frac{\mu_1}{2\mu^2}
\] (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 \( \mu \) and \( \sigma^2 \) 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, 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 \( t_0 \) is the number of components (working and uncensored) under risk at time \( a_j \).

\( n_j \) is the number of components living at the beginning of \( I_j \); \( n_2 = n \) and \( n_j = n_{j-1} - d_j \), \( j = 1, 2, ..., k + 1 \). Let denote a random variable with an \( F \) distribution function. \( \tilde{F}(a_j) = P(Y > a_j) \) for \( j = 1, 2, ..., k + 1 \) and \( q = 1 - p_j, j = 1, 2, ..., k + 1 \). Since;

- \( \tilde{F}(a_1) = 1 \)
- \( \tilde{F}(a_2) = P(Y > a_1) \)
- \( \tilde{F}(a_3) = P(Y > a_2) = P(Y > a_2 | Y > a_1)P(Y > a_1) \)
- \( \tilde{F}(a_4) = P(Y > a_3) = P(Y > a_3 | Y > a_2)P(Y > a_2) \)
- \( \tilde{F}(a_k) = P(Y > a_k) = P(Y > a_k | Y > a_{k-1})P(Y > a_{k-1})P(Y > a_{k-2}) \)

Then

\[
\tilde{F}(a_j) = p_1p_2 \ldots p_{j+1} \quad j = 1, 2, ..., k + 1
\] (4)

Now, let’s consider the estimation problem of \( \tilde{F}(a_j) \). 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 \( \tilde{F}(a_j) \), where \( a_j \) is the ratio of working observations. If the ranges comprise back-off that is censurings, 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.

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. 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 where 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 \( a_2 = 0, a_0 = t \), \( a_{k+1} = \infty \) and \( t \) observation, be divided \( k+1 \) pieces of \( I_j = [a_j, a_{j+1}], j = 1, 2, ..., k+1 \) 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 \( t \) have to deteriorate in a time in \( I_{k+1} \).

\( n_i \) is the number of components (working and uncensored) under risk at time \( a_j \).

\( d_i \) 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 \( n_j = n_{j-1} - d_j \), \( w_j, j = 1, 2, ..., k + 1 \).

Let denote a random variable with an \( F \) distribution function.

\[
\tilde{F}(a_j) = P(Y > a_j), \quad j = 1, 2, ..., k + 1
\]

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

Since;

\[
\tilde{F}(a_j) = 1
\]

\[
\tilde{F}(a_j) = P(Y > a_1)
\]

\[
\tilde{F}(a_2) = P(Y > a_2) = P(Y > a_2 | Y > a_1)P(Y > a_1)
\]

\[
\tilde{F}(a_3) = P(Y > a_3) = P(Y > a_3 | Y > a_2)P(Y > a_2)P(Y > a_1)
\]

\[
\tilde{F}(a_k) = P(Y > a_k) = P(Y > a_k | Y > a_{k-1})P(Y > a_{k-1})P(Y > a_{k-2}) \ldots P(Y > a_1)
\]

Then

\[
\tilde{F}(a_j) = p_1p_2 \ldots p_{j+1} \quad j = 1, 2, ..., k + 1
\] (4)

Now, let’s consider the estimation problem of \( \tilde{F}(a_j) \). 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 \( \tilde{F}(a_j) \), where \( a_j \) is the ratio of working observations. If the ranges comprise back-off that is censurings, 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 \( I_j \), \( \tilde{q}_j = \frac{d_j}{n_j} \) is a significant estimation of \( \tilde{q}_j \), because when component \( \tilde{q}_j \) 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 \( I_t \) can be deteriorated before the end of \( I_t \). Additionally, those previously censored but deteriorated in \( I_t \) are not already within \( n_i \). Therefore, it is desirable to make an arrangement for censored observations. The most commonly used method is the estimation made by

\[
\hat{q}_j = \frac{d_j}{n_j - w_j}
\]

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

The expression 5, suggests that \( n_i \geq 0 \). When \( n_i = 0 \), 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 \( I_t \) 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 \( I_t \) take place on the right at the end of \( I_t \) 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 \( I_t \), \( \hat{q}_j = \frac{d_j}{n_j - w_j} \) will be suitable. After estimations of \( \hat{q}_j \) and \( \hat{p}_j = 1 - \hat{q}_j \) are calculated, with the help of \( F(a_j) \), equation 4

\[
F(a_j) = \hat{p}_1 \cdots \hat{p}_j, \quad j = 1, 2, \ldots, k + 1
\]

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 \( n_i \), \( d_j \), \( w_j \), \( \hat{q}_j \) and \( \hat{p}_j \) for each \( n_i \) 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 \( \frac{n_i}{n} \) 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 \( F(t) \) is:

\[
\hat{F}(t) = \text{Equal to or greater number of observations on lifetime} \frac{t}{\text{n}}
\]

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 \( F(t) \) called product-limit or Kaplan-Meier can be reached [13].

Let’s suppose that \( k \) pieces of \( n \) components deteriorate at different times of \( t_1 < t_2 < \ldots < t_k \) and at that \( j \) \( j = 1, 2, \ldots, k \), 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, \( t_k \), 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];

\[
F_{KM}(t) = 1 - \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}
\]

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

<table>
<thead>
<tr>
<th>Uncensored Observations</th>
<th>Censored Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.3</td>
<td>41.7</td>
</tr>
<tr>
<td>52.1</td>
<td>52.3</td>
</tr>
<tr>
<td>53.6</td>
<td>53.6</td>
</tr>
<tr>
<td>54.8</td>
<td>55.1</td>
</tr>
<tr>
<td>56.9</td>
<td>57.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Censored Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.8</td>
</tr>
</tbody>
</table>

### Table 2. Kaplan-Meier estimators of the device at the given \( t \) points

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

### 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 \( n \) need to be
estimated in this case, a non-parametric estimation of \( M \) 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:

\[
P_{KM}(t) = 1 - \prod_{j, t_j \leq t} \frac{n_j - d_j}{n_j}
\]

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

\[
\hat{M}_{KM}(t) = \sum_{k=1}^{\infty} P_{KM}^k(t)
\]

defined by taking the estimator \( \hat{P}_{KM}(t) \) for \( F(t) \), can be considered as a non-parametric estimator of \( M(t) \), where \( P_{KM}^k(t) \) is the \( n \) times Stieljes convolution of \( P_{KM}(t) \) by itself. The value of estimator \( \hat{M}_{KM}(t) \) 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 \( (X_1, \ldots, X_n) \), \( \hat{M}_{KM}(t) \) from renewal equation \( M(t) = F(t) + \int_0^t F(t-x) dM(x), \ t \geq 0 \) can be denoted as:

\[
\hat{M}_{KM}(t) = \hat{P}_{KM}(t) + \int_0^t \hat{M}_{KM}(t-x) d\hat{M}_{KM}(x)
\]  

(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 integer and rounded down to the closest number. And as a result, they are converted 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, \( \hat{M}_{KM}(t) \) can be calculated subsequently. \( t=k \) and \( F_0 \) as the abovementioned arithmetic distribution function, \( \hat{M}_{KM} \) is calculated subsequently from the sum of

\[
M_{KM}^k (i) = F_0 (i) + \sum_{j=1}^{\infty} M_{KM}^j (i) (F_0 (j) - F_0 (j-1)), i=1,2, \ldots k
\]

(11)

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

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


