A bias-corrected Nelson-Aalen estimator

R Jiang
Professor, Changsha University of Science and Technology, Changsha, China
E-mail: jiang@csust.edu.cn

Abstract. The starting point to model an incomplete dataset of a non-negative random variable is usually to estimate its reliability function using a non-parametric reliability estimator. Two such estimators are Kaplan-Meier estimator and Nelson-Aalen estimator. Their performances have been compared and some drawbacks (e.g., biasness) have been identified in the literature. However, improvements on them are scarce. This paper aims to fill this gap by proposing a simple and almost unbiased estimator. This estimator is a weighted average of the Nelson-Aalen reliability estimates at two successive time points. An application of the proposed estimator in model selection is discussed and a real-world dataset is analyzed to illustrate the appropriateness and usefulness of the proposed estimator.

1. Introduction.
Reliability estimation problem is a basic problem of reliability analysis of non-repairable items. It deals with fitting a given dataset to an appropriate life distribution model. Typical approaches for this estimation problem can be roughly divided into two categories: parametric methods and non-parametric methods. The parametric methods need to assume a distribution model and the resulting model depends on this assumption. A potential disadvantage of the parametric methods is possible model misspecification [1]. The non-parametric methods can avoid the misspecification problem and hence are usually used as the starting point of reliability data analysis.

Data for reliability analysis are often incomplete and typically contain right censored observations. Two classical non-parametric estimators for such data are Kaplan-Meier estimator (KME, [2]) and Nelson-Aalen estimator (NAE, [3-4]). In spite of their wide applications, they have some drawbacks. For example, the estimators are biased and only depend on the number of items at risk at each failure time so that the information of censored times is not utilized [5-9].

Several attempts have been made to address these issues in the following two directions: (a) performance comparison between the KME and the NAE, and (b) development of improved estimators.

Bohoris compares the KME with the NAE and shows that the reliability obtained from the NAE is always larger than the one from the KME [5-6]. However, relative comparisons cannot answer the question: which is the most accurate or most appropriate reliability estimator? Some authors use Monte Carlo simulations to quantitatively evaluate the performances of the KME and the NAE in terms of mean square error and conclude that the NAE is more accurate than the KME [7, 10].

Several improved estimators have been proposed in the literature and they can be roughly divided into three categories: (a) simple smoothing methods, (b) competing risk approaches and (c) total-time-on-test (TTT) based approaches.
A simple smoothing is obtained by connecting the midpoints of the steps in the KME [11]. This is also called as the linear interpolation, which has been applied to both the KME and the NAE and is used to estimate survival fractions [6-7, 12].

The competing-risk-based approaches need to explicitly model the distribution of censoring time [1, 10-11, 13]. Though they can reduce the bias of the KME, a common drawback of these approaches is that they are not simple. This hinders their practical applications.

The TTT-based approaches relate the cumulative failure number and the TTT to the failure rate function [9, 14], which can be used to estimate the cumulative hazard function (CHF) and reliability function. The TTT-based approaches can effectively utilize the information of censored times but some additional assumptions are needed to obtain the failure rate estimate.

The purpose of this paper is three-fold. Firstly, a new performance evaluation method for a non-parametric reliability estimator is proposed. The proposed evaluation method is based on the fact that for a complete dataset of sample size n the cumulative distribution function (CDF) at the ith failure follows the beta distribution with shape parameter i and n+1-i (i = 1, 2, ..., n). This allows a quantitative performance evaluation for a given estimator in case of complete data. Secondly, a bias-corrected NAE is proposed based on the performance evaluation results of several estimators. Finally, a model selection method is proposed. It is based on a match between the shapes of empirical and theoretical CHFs.

The paper is organized as follows. The classical non-parametric reliability estimators are presented and their performances are evaluated in Section 2. The performances of moving average estimators associated the KME and the NAE are evaluated and the bias-corrected NAE is proposed in Section 3. The model selection approach is presented and illustrated in Section 4. The paper is concluded in Section 5.

2. Performance evaluation method

2.1 Classical non-parametric reliability estimators

2.1.1 Data with right-censored observations

Consider an ordered dataset with both failed and right-censored observations, given by

\[ t_j(d_j) \leq t_{j+1}(d_{j+1}); 0 \leq j \leq n; t_0 = 0, t_{n+1} = \infty, d_0 = d_{n+1} = 0 \]  \hspace{1cm} (1)

Here, \( d_j = 1 \) if the observation at \( t_j \) is a failure event; otherwise, \( d_j = 0 \). This implies that each observation has itself value of \( j \). If a failure observation and a censored observation occur at the same time, the censored observation is always arranged after the failed observation.

The total number of failures is

\[ m = \sum_{j=1}^{n} d_j \leq n \]  \hspace{1cm} (2)

Clearly, for a complete dataset, \( m = n \). All the failure observations in equation (1) form a failure dataset, which is denoted as

\[ (t_i', t_i'; 1 \leq i \leq m) \]  \hspace{1cm} (3)

Let \( e_i \) denote the number of right-censored observations in interval \( [t_i', t_i') \) and \( n_i \) denote the number of items at risk (or the number of survival items) just prior to \( t_i' \). The \( j \)-value of \( t_i' \) is given by

\[ j = i + \sum_{k=1}^{i} e_k \]  \hspace{1cm} (4)
The value of $n_i$ is given by
\[ n_i = n - (i - 1 + \sum_{k=1}^{i} e_k) = n + 1 - j \tag{5} \]

\[ R(t_j) = \prod_{k=1}^{j} (1 - d_k / n_k) \tag{6} \]

It is a staircase function of $t$ with a jump at each failure time.

The NAE is given by
\[ H(t_j) = \sum_{k=1}^{j} d_k / n_k \times R(t_j) = \exp[-H(t_j)] = \prod_{k=1}^{j} e^{-d_k/n_k} \tag{7} \]

where $H(.)$ is the CHF and is also a staircase function of $t$. It is noted that above definitions extend the original definitions of the KME and the NAE to the censored observations. The main advantage of the extension is that it can be conveniently implemented in an Excel spreadsheet program.

Noting $e^{-x} = 1 - x + x^2 / 2 ... > 1 - x$ and letting $x = d_k / n_k$, it is easy to show from equations (6) and (7)
\[ R_{KME}(t_j) < R_{NAE}(t_j) \tag{8} \]

The life-table methods (LTM) can be viewed as a variant of the KME [15]. In the LTM the tied data are often merged as
\[ \{(t_{j-1}, t_j), d_j, 1 \leq j \leq m+1\}; \quad t_0 = 0, t_{m+1} = \infty; \quad d_j \geq 1 \tag{9} \]

The reliability at the right end of each interval is estimated as
\[ R(t_j) = \prod_{k=1}^{j} (1 - d_k / n_k^*) \sum_{k=1}^{j} e_k / 2 \] Clearly, the denominator of the KME is revised as the number of items at risk at the mid-point of the time interval under the assumption that the censored times are uniformly distributed over the interval. Relative to the KME, the increase in the denominator leads to an increase in the conditional reliability. As a result, the reliability estimate obtained from the LTM is always larger than the one from the KME, and this is desirable.

2.2 Performance evaluation of estimators in case of complete data
A good non-parametric reliability estimator can provide accurate reliability estimate not only for incomplete data but also for complete data, which can be viewed as a special case of incomplete data. It is much simpler to evaluate the performance of an estimator for complete data than for incomplete data because a complete dataset is characterized only by sample size $n$.

Denote a complete dataset as
\[ t_1 \leq t_2 \leq ... \leq t_i \leq ... \leq t_n \tag{11} \]

In this case, the CDF, $F(t_i) = 1 - R(t_i)$, is a random variable and follows the standard beta distribution with shape parameters $i$ and $n+1-i$ [16]. The mean and median estimates of $R(t_i)$ are given respectively by
\[ R_0(t_i) = 1 - i / (n + 1), \quad R_1(t_i) = 1 - B^{-1}(0.5; i, n + 1 - i) \]  
where \( B^{-1}(\cdot) \) is the inverse of the standard beta CDF. The mean and median estimates can be thought to be good estimates. The accuracy of an estimator, \( R(t_i) \), can be evaluated by examining the magnitudes of the following two curves:

\[ \delta_0(i) = R(t_i) - R_0(t_i), \quad \delta_1(i) = R(t_i) - R_1(t_i) \]

Table 1 shows three possible cases. The third case can be thought to be unbiased and hence is desired. The smaller \(|\delta_0(i)|\) and \(|\delta_1(i)|\) are, the more accurate the estimator is.

**Table 1.** Three possible cases for a non-parametric reliability estimator.

| Case | If | Then |
|------|----|------|
| 1    | \( \max(\delta_0, \delta_1) < 0 \) | Underestimate reliability |
| 2    | \( \min(\delta_0, \delta_1) > 0 \) | Overestimate reliability |
| 3    | \( \max(\delta_0, \delta_1) > 0 > \min(\delta_0, \delta_1) \) | Almost unbiased |

Figure 1 shows the plots of \( \delta_0(i) \) and \( \delta_1(i) \) associated with the KME and the NAE for \( n = 25 \).

As seen, the absolute value of bias generally increases as \( i \) increases. Therefore, the biasness can be measured by \( \delta_0(n) \) and \( \delta_1(n) \). Figure 2 shows the plots of \( \delta_0(n) \) and \( \delta_1(n) \) as functions of \( n \).

As can be seen from Figures 1 and 2, both the NAE and the KME underestimate the reliability; but the NAE is much more accurate than the KME. In addition, the estimators are closer to the median of the beta distribution and the bias generally decreases as \( n \) increases. Therefore, it is necessary to develop bias-corrected estimators.
3. Performances of moving average estimators

3.1 Moving average estimator of the KME
For the KME, the reliability estimate just prior to $t_j$ is $R(t_j-1)$ and the estimate just after $t_j$ is $R(t_j)$. Thus, a more reasonable estimate at $t_j$ is their average, i.e.,

$$R_{av}(t_j) = \frac{[R(t_{j-1}) + R(t_j)]}{2}$$  \hspace{1cm} (14)

Equation (14) is called as the averaged KME (AKME).

For the complete data given by equation (11), the KME is given by

$$R_i = \prod_{k=1}^{i-1} \left(1 - \frac{1}{n+1-j}\right) = \prod_{k=1}^{i-1} \frac{n-j}{n+1-j} = 1 - \frac{i}{n}.$$  \hspace{1cm} (15)

Using equation (15) into equation (14) yields

$$R_{av}(t_j) = 1 - \frac{i-0.5}{n} = R(t_j) + \frac{0.5}{n}.$$  \hspace{1cm} (16)

Figure 3 shows the plots of $\delta_i(i)$ and $\delta_i(i)$ for $n = 25$. Clearly, it overestimates [underestimates] the reliability for $i < (n+1)/2$ [i > (n+1)/2]. However, the AKME is much more accurate than the KME due to much smaller biasness.
3.2 Moving average estimators of the NAE
For the NAE, both \( R(t) \) and \( H(t) \) are staircase functions of \( t \). Therefore, two moving average estimators can be defined: (a) \( R(t) \)-based average, and (b) \( H(t) \)-based average. The former is defined by equation (14) and the latter is defined by

\[
R_{\omega}(t_j) = \exp\left[ -\frac{H(t_{j-1}) + H(t_j)}{2} \right] = \sqrt{R(t_{j-1})R(t_j)}
\]

Clearly, the \( H(t) \)-based moving average estimator is the geometric mean of \( R(t_{j-1}) \) and \( R(t_j) \), which is smaller than the \( R(t) \)-based average.

Figure 4 shows the plots of \( \delta_0(i) \) and \( \delta_1(i) \) for these two averaged NAEs (ANAE) associated with \( n = 25 \). As seen from the figure, the ANAEs overestimate the reliability; the \( H(t) \)-based ANAE (i.e., the dotted lines) is slightly better than the \( R(t) \)-based ANAE (i.e., the solid lines) in terms of biasness; and the plots of \( \delta_0(i) \) and \( \delta_1(i) \) of the \( R(t) \)-based ANAE are highly linear.

![Figure 4. Plots of \( \delta_0(i) \) and \( \delta_1(i) \) for the ANAE.](image)

3.3 Weighted average estimator
As can be seen from Figures 1 and 4, the NAE and the \( R(t) \)-based ANAE can be mutually compensatory. Thus, their arithmetic average is probably an unbiased estimator, which is given by

\[
R_{\omega}(t_j) = \frac{R(t_j) + R_{\omega}(t_j)}{2} = 0.25R(t_{j-1}) + 0.75R(t_j) = R(t_j) + \frac{\Delta R}{4}
\]

where \( \Delta R_j = R(t_{j-1}) - R(t_j) \). Equation (18) has the form of weighted average and hence is called as the weighted average estimator (WAE). Clearly, the WAE has an increment of \( \Delta R_j/4 \) relative to the NAE.

For \( n = 25 \), Figure 5 shows the plots of \( \delta_0(i) \) and \( \delta_1(i) \) for the WAE. From the figure, we have the following observations:

- The reliability estimate obtained from the WAE is generally in-between the mean and median estimates (i.e., the third case in Table 1) and hence is an almost unbiased estimator.
- The WAE is slightly closer to the median estimator.
- The largest absolute values of \( \delta_0(i) \) and \( \delta_1(i) \) usually appear at \( i = 1 \) and \( n \).

Table 2 shows the minimum, maximum and average of \([ \delta_0(i), \delta_1(i) ]\) for \( n = 5(5)50 \). These values are also displayed in Figure 6, which clearly illustrates the unbiasedness of the WAE.
4. A model selection method

4.1 WAE-based model selection

As mentioned earlier, one advantage of the non-parametric estimation methods is to avoid possible model misspecification. In this section, an approach used to select an appropriate distribution model is presented based on the WAE estimator. Specific details are outlined as follows.

For a given dataset, one first obtains the empirical reliability function using the WAE, and then calculates the empirical CHF using the following

\[ H_n(t_j) = -\ln[R_n(t_j)], \quad 1 \leq j \leq n. \]  

(19)
One or more optional models can be selected based on the shape of the plot of $H_w(t)$ vs. $t$ and referring to Table 3, which shows possible shapes of failure rate function (denoted as $r(t)$) and CHF of several distribution models.

Once the model is selected, the maximum likelihood method (MLM) can be used to estimate the distribution parameters. The fitted model is a smoothed estimator and can be used to predict the reliability at any time (including those censored times). The approach is illustrated as follows.

### Table 3. Shapes of $H(t)$ of typical distribution models.

| Model          | Shape of $r(t)$                                      | Shape of $H(t)$                                    |
|----------------|------------------------------------------------------|----------------------------------------------------|
| Weibull        | Decreasing & tending to zero, constant, increasing & tending to infinity | Concave, linear, convex                            |
| Gamma with scale parameter $\eta$ | Decreasing & tending to $1/\eta$, constant, increasing & tending to $1/\eta$ | Concave & asymptotically linear, convex & asymptotically linear, linear |
| Normal         | Increasing & asymptotically linear                   | Convex & asymptotically quadratic function         |
| Lognormal      | Unimodal & tending to zero                          | $S$-shaped                                         |
| Bathtub models | Bathtub-shaped                                       | Inverse $S$-shaped                                 |

#### 4.2 A real-world example

The data shown in the first three columns of Table 4 come from [5]. Reliability functions estimated from the KME, the NAE and the WAE are shown in the 4th-6th columns of the table.

Figure 7 displays the plot of $H_w(t)$ vs. $t$ associated with the WAE (i.e., those dots). As can be seen, the CHF looks like $S$-shaped. This implies that the Weibull and gamma distributions are not appropriate models for fitting the data and the potentially appropriate model is the lognormal distribution.

### Table 4. Estimated reliability functions.

| $j$ | $t_j$ | $d_j$ | KME  | NAE  | WAE  | Lognormal |
|-----|-------|-------|------|------|------|-----------|
| 1   | 69    | 1     | 0.9524 | 0.9535 | 0.9651 | 0.9906   |
| 2   | 176   | 1     | 0.9048 | 0.9070 | 0.9186 | 0.8853   |
| 3   | 196   | 0     | 0.9048 | 0.9070 | 0.9070 | 0.8577   |
| 4   | 208   | 1     | 0.8545 | 0.8580 | 0.8702 | 0.8406   |
| 5   | 215   | 1     | 0.8042 | 0.8090 | 0.8212 | 0.8306   |
| 6   | 233   | 1     | 0.7540 | 0.7600 | 0.7722 | 0.8045   |
| 7   | 289   | 1     | 0.7037 | 0.7109 | 0.7232 | 0.7237   |
| 8   | 300   | 1     | 0.6534 | 0.6619 | 0.6742 | 0.7082   |
| 9   | 384   | 1     | 0.6032 | 0.6129 | 0.6252 | 0.5969   |
| 10  | 390   | 1     | 0.5529 | 0.5639 | 0.5762 | 0.5896   |
| 11  | 393   | 0     | 0.5529 | 0.5639 | 0.5639 | 0.5859   |
| 12  | 401   | 1     | 0.4976 | 0.5103 | 0.5237 | 0.5763   |
| 13  | 452   | 1     | 0.4423 | 0.4566 | 0.4700 | 0.5182   |
| 14  | 567   | 1     | 0.3870 | 0.4029 | 0.4164 | 0.4081   |
| 15  | 617   | 0     | 0.3870 | 0.4029 | 0.4029 | 0.3685   |
| 16  | 718   | 0     | 0.3870 | 0.4029 | 0.4029 | 0.3009   |
| 17  | 782   | 1     | 0.3096 | 0.3299 | 0.3482 | 0.2655   |
| 18  | 783   | 1     | 0.2322 | 0.2569 | 0.2752 | 0.2650   |
| 19  | 806   | 1     | 0.1548 | 0.1841 | 0.2023 | 0.2535   |
| 20  | 1000  | 0     | 0.1548 | 0.1841 | 0.1841 | 0.1767   |
| 21  | 1022  | 0     | 0.1548 | 0.1841 | 0.1841 | 0.1699   |
To confirm the above observation, the data are fitted to the above-mentioned three models using the MLM. The estimated parameters and corresponding log-likelihood values are shown in Table 5. As seen, the lognormal distribution provides the best fitting to the data in terms of log-likelihood.

The reliability function estimated from the fitted lognormal distribution is shown in the last column of Table 4, and the plots of the reliability functions obtained from the three non-parametric methods and the fitted lognormal model are shown in Figure 8. As seen, the fitted lognormal distribution provides a smoothed reliability estimate at any time, including the range of \( t > t_{19} = 806 \).

To further confirm the appropriateness of the WAE. The TTT estimates at \( t = t_{19} \) obtained from the three non-parametric estimators are compared with the TTT value evaluated from the fitted lognormal distribution, which is viewed as the “true value”.

Noting that the KME and NAE are staircase functions of \( t \), their TTT estimate is given by

\[
\mu(t_i^*) = \frac{\sum_{j=1}^{i} R(t_{j-1}^*) (t_j^* - t_{j-1}^*)}{R(t_{j-1}^*)}
\]

where \( t_i^* \) is the time of the \( i \)th failure. Since \( R(t_{j-1}^*) > R(t_j^*) \), the estimate given by equation (20) is not necessarily an underestimate. On the other hand, since the WAE is not a staircase function, its TTT estimate is different from equation (20), and given by

\[
\mu(t_i) = \frac{\sum_{j=1}^{i} R(t_{j-1}) + R(t_j^*) (t_j^* - t_{j-1})}{2}
\]

The TTT associated with the lognormal distribution is given by

\[
\mu(t) = tR(t) + \exp(\mu + \frac{\sigma^2}{2}) \Phi(\frac{\ln(t) - \mu}{\sigma}; \sigma, l)。
\]

where \( F(\cdot; l, s) \) is the normal CDF with location parameter \( l \) and scale parameter \( s \). The results are shown in Table 6. The last row of the table shows the relative errors with the “true value”. As seen, the WAE provides the most accurate estimate for the TTT\((t_{19})\), whose relative error is smaller than 1%.

| Parameter | Weibull | Gamma | Lognormal |
|-----------|---------|-------|-----------|
| \( \beta \) or \( \mu \) | 1.5459  | 2.0853| 6.1509    |
| \( \eta \) or \( \sigma \) | 643.9   | 281.4 | 0.8157    |
| \( \ln(L) \) | -110.774| -110.493| -110.286 |

Figure 7. Plots of \( H_w(t) \) vs. \( t \) associated with the WAE.

Table 5. Maximum likelihood estimates of distribution parameters.
5. Conclusions

In this paper, we have developed a new method to quantitatively evaluate the performance of a non-parametric reliability estimator, and applied it to the KME, the NAE and their moving average estimators. The main findings or contributions have been

- Both the KME and the NAE are biased and the NAE has a smaller bias than the KME.
- The NAE and its moving average estimator are mutually compensatory. This finding leads to the proposal of the NAE-based weighted average estimator. The proposed WAE is almost unbiased and can be conveniently implemented in an Excel spreadsheet program.
- A CHF-shape-based model selection method has been proposed and illustrated.
- A bias-corrected KME has been proposed by the author [17]. A topic for future research is to develop new non-parametric estimators for grouped data.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (No. 71771029).

References

[1] Rossa A 2002 On the estimation of survival function under random censorship J. Commun. Stat.-Theory Methods 31 961-75
[2] Kaplan E and Meier P 1958 Nonparametric estimation from incomplete observations J. Am. Stat. Assoc. 53 457–81
[3] Nelson W 1969 Hazard plotting for incomplete failure data J. Qual. Technol. 1 27–52
[4] Aalen O 1978 Nonparametric inference for a family of counting processes J. The Annals of Statistics 6 701-26
[5] Bohoris G 1994 Comparison of the cumulative-hazard and Kaplan-Meier estimators of the survivor function J. IEEE Trans. Reliab. 43 230 –2
[6] Bohoris G 1994 Numerical differences in the survival probabilities obtained by the cumulative-hazard and Kaplan-Meier estimators of the reliability function J. Qual. Reliab. Eng. Int. 10 99-104
[7] Colosimo E, Ferreira F and Oliveira M 2002 Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators J. Stat. Comput. Simul. 72 299–308
[8] Walraven C and Hawken S 2016 Competing risk bias in Kaplan–Meier risk estimates can be corrected J. Clin. Epidemiol. 70 101-5
[9] Costella J A 2010 Simple alternative to Kaplan–Meier for survival curves (Peter MacCallum Cancer Centre Working Paper Sept/2010)
[10] Fleming T and Harrington D 1984 Nonparametric estimation of the survival distribution in censored data J. Communications in Statistics -Theory and Methods 13 2469–86
[11] Klein J, Lee S and Moeschberger M 1990 A partially parametric estimator of survival in the presence of randomly censored data J. Biometrics 46 795–811
[12] Lee E T 1992 Statistical Methods for Survival Data Analysis (New York: John Wiley)
[13] Stute W 1994 Improved estimation under random censorship Commun. J. Stat.-Theory Methods 23 2671–82
[14] Kulasekera K and White B 1996 Estimation of the survival function censored data: a method based on total time on test J. Communications in Statistics-Simulation 25 189–200
[15] Lawless J F 1982 Statistical Models and Methods for Lifetime Data (New York: Wiley)
[16] Jiang R 2015 Introduction to Quality and Reliability Engineering (London: Springer & Beijing: Science Press)
[17] Jiang R 2020 A bias-corrected Kaplan-Meier estimator. The 9th Asia-Pacific International Symposium on Advanced Reliability and Maintenance Modelling (Vancouver, Canada)