Available online at www.sciencedirect.com
SCIENCE DIRECT®

Statistical Methodology 2 (2005) 65–69

Statistical Methodology
Official Journal of the International Indian Statistical Association

www.elsevier.com/locate/stamet

Teachers' corner

Estimating the mean life time using right censored data

Somnath Datta[*]

Department of Statistics, University of Georgia, Athens, GA 30602, United States

Received 13 May 2004

Abstract

Estimation of the population mean based on right censored observations is considered. The naive sample mean will be an inconsistent and asymptotically biased estimator in this case. An estimate suggested in textbooks is to compute the area under a Kaplan–Meier curve. In this note, two more seemingly different approaches are introduced. Students' reaction to these approaches was very positive in an introductory survival analysis course the author recently taught.

Keywords: Right censored data; Mean; Imputation; Kaplan–Meier estimator; Reweighting

1. Area under the Kaplan–Meier curve

Estimation of a population mean by its sample counterpart can perhaps be regarded as the most basic statistical method students learn in their elementary statistics courses. However, the sample mean will no longer work as an estimate of the mean life time if the sample is right censored. The topic is generally discussed in a first course on survival analysis (say, at an advanced undergraduate/masters level). It is easy to argue that the sample mean of the censored sample estimates the mean of the minimum of the failure and censoring variables and is therefore biased. Since the students in such a course will be familiar with the Kaplan–Meier estimator, it can instead be properly estimated

[*] Tel.: +1 706 542 3324; fax: +1 706 542 3391.
E-mail address: datta@stat.uga.edu.

1572-3127/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2004.11.003

(see Section 4) by computing the area under the Kaplan–Meier curve $\widehat{\mu} = \int_0^\infty \widehat{S}(t)\,\mathrm{d}t$. This is the approach taken, e.g., in the book by Klein and Moeschberger [2]. In evaluating this integral it is necessary to make a tail correction to the Kaplan–Meier estimator if the largest observations $\tau_{\max}$ correspond to a censored event. One such correction, due to Efron [1] is to treat the largest observations $\tau_{\max}$ as observed failures so that the Kaplan–Meier estimator would drop to zero at $\tau_{\max}$ and beyond; in that case, $\widehat{\mu}$ can be computed by adding a finite number of rectangular areas. Consider, for example, a sample of right censored life times 1, 1, 1+, 2.5, 5+, 7+. The censored observation 7+ will be treated as a 7 (true failure) for computing the Kaplan–Meier estimator $\widehat{S}$ so that $\widehat{S}(7) = 0$ (Table 1). This leads to an estimated mean of

$$\widehat{\mu} = 1 \times (1 - 0) + \frac{2}{3} \times (2.5 - 1) + \frac{4}{9} \times (7 - 2.5) = 4.$$

The following two seemingly very different approaches would produce the same correct answer and provide additional insight into the problem of right censored data. Perhaps they can all be taught for a better understanding of the role of censoring and different ways to account for it.

2. Data imputation

In this approach, $\widehat{\mu}$ can be computed using the familiar sample mean formula provided the censored values are imputed. If a life time T is censored by C then the censored value is replaced by the conditional expectation of T given C and that $T > C$, which equals $\{-\int_C^\infty t\,\mathrm{d}S(t)\}/S(C)$ and can be estimated by

$$\frac{\sum\limits_{\tau_i > C} \tau_i \Delta\widehat{S}(\tau_i)}{\widehat{S}(C)}.$$

Once again, the largest observations will have to be treated as true failures for this calculation.

For our example data set, we set $7+ = 7$ and the only remaining censored observations 1+ and 5+ will be changed to

$$1+ \rightarrow \frac{2.5 \times \left(\frac{2}{3} - \frac{4}{9}\right) + 7 \times \left(\frac{4}{9} - 0\right)}{\frac{2}{3}} = \frac{33}{6},$$

$$5+ \rightarrow \frac{7 \times \left(\frac{4}{9} - 0\right)}{\frac{4}{9}} = 7.$$

Therefore the sample mean of the modified (imputed) sample is

$$\widehat{\mu} = \frac{1}{6}\left(1 + 1 + \mathbf{\frac{33}{6}} + 2.5 + \mathbf{7} + 7\right) = 4.$$

Table 1
Kaplan–Meier estimator of failure times

Time	1	2.5	7
Number of failures	2	1	1
Number at risk	6	3	1
$\widehat{S}$	$\frac{2}{3}$	$\frac{4}{9}$	0

Table 2
Kaplan–Meier estimator of censoring times

Time	1	5
Number censored	1	1
Number at risk	4	2
$\widehat{S}_c$	$\frac{3}{4}$	$\frac{3}{8}$

3. Reweighting

This approach is perhaps more complicated than the other two in this context. However it might provide a very different perspective into the problem of right censoring.

Let S_c be the survival function of the censoring variable C, i.e., $S_c(t) = P\{C > t\}$, $t \geq 0$. Consider the following identity (see, e.g., [4,5] for more general versions) for right censored life times $\widetilde{T} = T \wedge C$ and censoring status indicators $\delta = I(T \leq C)$:

$$E\left\{\frac{\delta\widetilde{T}}{S_c(\widetilde{T}-)}\right\} = E\left\{E\left(\frac{\delta T}{S_c(T-)}\middle| T\right)\right\} = E\left\{\frac{T}{S_c(T-)}E\ (I(C \geq T)|\ T)\right\}$$
$$= E\{T\},$$

where the first equality follows since $\widetilde{T} = T$ on the set $\delta \neq 0$, and the last equality follows from $E(I(C \geq T)|T) = S_c(T-)$, if the censoring time C is independent of the life time T.

The above identity can be taken as the theoretical basis for estimating the mean lifetime $\mu = E\{T\}$ by the (weighted) sample average $\widehat{\mu} = (\sum_{i=1}^{n} w_i \widetilde{T}_i)/n$, where the weights are $w_i = 0$, for a censored life time and $\{\widehat{S}_c(\widetilde{T}_i-)\}^{-1}$, for an observed life time. Here, $\widehat{S}_c$ will be calculated using the Kaplan–Meier product limit formula, with the status indicators $1 - \delta$. However, the following modification is necessary for the "risk set" calculation, as observed in [6]. In the case where some censored life times are tied with observed life times, the observed failures will be excluded from the risk set at that time.

This is illustrated for the example data by first computing the Kaplan–Meier for censoring $\widehat{S}_c$ (Table 2). As before, the last observation 7+ will be treated as a 7. Note that the observed life times 1 and 1 have been excluded from the "at risk" set at time 1. This leads to

$$\widehat{\mu} = \frac{1}{6}\left(\frac{1}{\widehat{S}_c(1-)} \times 1 + \frac{1}{\widehat{S}_c(1-)} \times 1 + \mathbf{0 \times 1}\right.$$
$$\left. + \frac{1}{\widehat{S}_c(2.5-)} \times 2.5 + \mathbf{0 \times 5} + 7 \times \frac{1}{\widehat{S}_c(7-)}\right)$$

$$= \frac{1}{6}\left(\frac{1}{1} \times 1 + \frac{1}{1} \times 1 + \frac{1}{\frac{3}{4}} \times 2.5 + 7 \times \frac{1}{\frac{3}{8}}\right) = 4,$$

as before.

4. Proofs

Of course, these three methods are equivalent generally, and depending on the level of the course, short proofs can be provided. These proofs are based on a self-consistency property of the Kaplan–Meier estimator first observed by Efron [1] and further exploited by Satten and Datta [6]. Since all three methods would treat the largest observations as true failures, we may assume that to be the case for the equivalence proofs presented here.

4.1. Equivalence of Methods 1 and 2

Let $\tau_1 < \cdots < \tau_m$ be the distinct event (failure or censoring) times. Also let f_i and g_i denote the number of failures and censoring events respectively at time τ_i. Using Method 1,

$$\widehat{\mu} = \sum_{i=1}^{m} \Delta\tau_i \widehat{S}(\tau_{i-1}),$$

where $\tau_0 = 0$, and $\Delta\tau_i = \tau_i - \tau_{i-1}$. Using the self-consistency equation of the Kaplan–Meier estimator (see e.g. [3], page 54, equation (5), with t set to τ_{i-1})

$$\widehat{S}(\tau_{i-1}) = n^{-1}\left\{Y(\tau_i) + \sum_{j<i} \frac{g_j \widehat{S}(\tau_{i-1})}{\widehat{S}(\tau_j)}\right\},$$

where $Y(\tau_i) = \sum_{j \ge i}\{f_j + g_j\}$ is the size of the "at risk" set at time τ_i, we get

$$\begin{aligned}
\widehat{\mu} &= n^{-1}\sum_{i=1}^{m} \Delta\tau_i \sum_{j\ge i}\{f_j + g_j\} + n^{-1}\sum_{i=1}^{m}\Delta\tau_i \sum_{j<i}\frac{g_j\widehat{S}(\tau_{i-1})}{\widehat{S}(\tau_j)},\\
&= n^{-1}\sum_{j=1}^{m} f_j\tau_j + n^{-1}\sum_{j=1}^{m} g_j\left\{\frac{\tau_j\widehat{S}(\tau_j) + \sum_{i>j}\Delta\tau_i\widehat{S}(\tau_{i-1})}{\widehat{S}(\tau_j)}\right\}\\
&= n^{-1}\sum_{j=1}^{m} f_j\tau_j + n^{-1}\sum_{j=1}^{m} g_j\tau_j^*
\end{aligned}$$

where $\tau_j^* = \sum_{i>j}\tau_i\Delta\widehat{S}(\tau_i)/\widehat{S}(\tau_j)$ is the imputed value for an observation censored at τ_j. Clearly the right hand side of the above equation equals the definition of $\widehat{\mu}$ obtained by Method 2.

4.2. Equivalence of Methods 2 and 3

This immediately follows from Satten and Datta [6, Section 3] who show that the jump size (mass) of $\widehat{S}$ at τ_j equals $n^{-1} f_j \widehat{S}_c(\tau_j-)$.

5. Discussion

In all three methods, we are forced to treat the largest observations as failures. Clearly, such a step will not be necessary if, for a sample, the largest observations are indeed true failures; for such a sample, these methods would lead to the same answer as well. When the largest observations correspond to at least one censored observation, estimating the tail of the survival distribution becomes an issue since the Kaplan–Meier does not drop to zero. The same issue is inherited by the problem of estimation of the mean. One can take one of the following two views. One may declare that it is impossible to estimate the tail of the survival curve for such samples, in which case, it is not possible to estimate the mean using such samples as well. Or one makes a somewhat arbitrary choice, of distributing the remaining mass of the Kaplan–Meier beyond the largest observed failures in order to produce a legitimate survival curve and to calculate a mean from it. The choice we have used to demonstrate the equivalence of the three approaches is to place all the remaining mass at the largest (censored) time in the study, as in [1].

Acknowledgments

This research was supported in part by a grant from the National Security Agency. I thank an anonymous reviewer for a useful suggestion.

References

[1] B. Efron, The two sample problem with censored data, in: Proc. Fifth Berkeley Symp. Math. Statist. Probab., vol. 4, Prentice-Hall, New York, 1967, pp. 831–853.

[2] J.P. Klein, M.L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, second ed., Springer Verlag, New York, 2003.

[3] R.G. Miller Jr., Survival Analysis, John Wiley & Sons, New York, 1981.

[4] J.M. Robins, A. Rotnitzky, Recovery of information and adjustment for dependent censoring using surrogate markers, in: N. Jewell, K. Dietz, V. Farewell (Eds.), AIDS Epidemiology—Methodological Issues, Birkhauser, Boston, 1992, pp. 297–331.

[5] G.A. Satten, S. Datta, J.M. Robins, An estimator for the survival function when data are subject to dependent censoring, Statistics and Probability Letters 54 (2001) 397–403.

[6] G.A. Satten, S. Datta, The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average, American Statistician 55 (2001) 207–210.