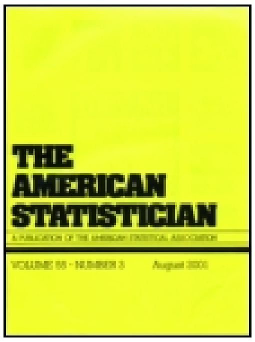

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

The Kaplan–Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average

Glen A Satten & Somnath Datta

To cite this article: Glen A Satten & Somnath Datta (2001) The Kaplan–Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average, The American Statistician, 55:3, 207-210, DOI: 10.1198/000313001317098185

To link to this article: https://doi.org/10.1198/000313001317098185

Published online: 01 Jan 2012.

The Kaplan–Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average

Glen A. SATTEN and Somnath DATTA

The Kaplan–Meier (product-limit) estimator of the survival function of randomly censored time-to-event data is a central quantity in survival analysis. It is usually introduced as a nonparametric maximum likelihood estimator, or else as the output of an imputation scheme for censored observations such as redistribute-to-the-right or self-consistency. Following recent work by Robins and Rotnitzky, we show that the Kaplan–Meier estimator can also be represented as a weighted average of identically distributed terms, where the weights are related to the survival function of censoring times. We give two demonstrations of this representation; the first assumes a Kaplan–Meier form for the censoring time survival function, the second estimates the survival functions of failure and censoring times simultaneously and can be developed without prior introduction to the Kaplan–Meier estimator.

KEY WORDS: Horvitz–Thompson estimator; Product-limit estimator; Survival analysis.

1. INTRODUCTION

The Kaplan–Meier (product-limit) estimator for the survival function of randomly censored time-to-event data (Kaplan and Meier 1958) is often introduced as the maximizer of a nonparametric maximum likelihood (Kalbeisch and Prentice 1978). Because data are subject to censoring, estimating the survival function can be thought of as a missing data problem. There are two general approaches to missing data problems: imputation and weighting. Alternate presentations of the Kaplan–Meier estimator, including the redistribution-to-the-right algorithm of Efron (1967), the self-consistency property (Efron 1967), or the EM algorithm approach (Turnbull 1976) are all examples of the imputation approach. In a series of papers, Robins and coworkers have shown that the weighting approach to missing data problems has a number of advantages over the imputation approach (Robins and Rotnitzky 1992; Robins 1993; and Robins and Finkelstein 2000 relate directly to survival analysis). An outcome of their approach applied to survival analysis is an inverse-probability-of-censoring representation of the Kaplan–Meier estimator. We give two simple demonstrations of this representation. The first, found in Section 3, is more straightforward but uses as weights the Kaplan–Meier estimator for censoring times, and hence does not stand alone. For this reason, we give a second approach in Section 4 that simultaneously estimates the cumulative distribution functions of survival and censoring times using coupled inverse-probability-weighted sums. The weighted average form given in this article is convenient for asymptotic theory and it leads to an interesting variance decomposition for the Kaplan–Meier estimator {not shown here; see Satten, Datta, and Robins (in press) or Robins and Finkelstein (2000) for examples of this type of result}.

2. NOTATION AND PRELIMINARY RESULTS

For $i = 1, \ldots, N$ let T_i^* be the random variable denoting the (possibly unobserved) failure time and C_i be the random variable denoting the (possibly unobserved) censoring time for the ith person. We adopt the usual convention that realizations of random variables are denoted by lower-case letters. Let $T_i = \min(T_i^*, C_i)$ and let $\Delta_i = I[T_i^* \leq C_i]$. The observed data consist of iid replicates of (T_i, Δ_i). We assume "random censoring;" that is, that T_i^* and C_i are independent. The goal is to estimate the survival function $S(t) = \Pr[T_i^* > t]$ or, equivalently, the cumulative distribution function $F(t) = 1 - S(t)$.

Let the ordered failure or censoring times be $\tau_j, j = 1, \ldots, J$ and let n_j be the number of persons who fail at time τ_j and m_j be the number of persons censored at time τ_j. We assume that no person can have a failure time equal to their censoring time (i.e., such persons are taken to be uncensored with failure time τ_j). Then, the risk set (number of persons at risk for failure at time t) can be written as

$$Y(t) = \sum_{j=1}^{J} (n_j + m_j) I\left[\tau_j \geq t\right]. \tag{1}$$

The Kaplan–Meier estimator $\widehat{S}_{\mathrm{km}}(t)$ of $S(t)$ is

$$\widehat{S}_{\mathrm{km}}(t) = \prod_{\{j \mid \tau_j \leq t\}} \left(1 - \frac{n_j}{Y(\tau_j)}\right). \tag{2}$$

We can also estimate the survival function for censoring times, $K(t) = \Pr[C_i > t]$ using the Kaplan–Meier approach but considering failure events as "censored" observations and censored observations as "failures." The Kaplan–Meier estimator of $K(t)$ is thus

$$\widehat{K}(t) = \prod_{\{j \mid \tau_j \leq t\}} \left(1 - \frac{m_j}{Y(\tau_j)}\right). \tag{3}$$

Glen A. Satten is Mathematical Statistician, Division of HIV/AIDS Prevention, Surveillance, and Epidemiology, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30333 (E-mail: GSatten@cdc.gov). Somnath Datta is Professor, Department of Statistics, University of Georgia, Athens, GA 30602.

If there were no censoring, we could estimate $F(t)$ by the empirical cumulative distribution function

$$F^*(t) = \frac{1}{N}\sum_{i=1}^{N} I[t_i^* \le t] \ , \tag{4}$$

which, considered as a random variable for each t, is an average of iid terms. The inverse-probability-of-censoring estimator analogous to $F^*(t)$ is also an average of iid terms $I[t_i^* \le t]$, each multiplied by $\delta_i = I[t_i^* \le c_i]$ and weighted inversely by the probability that the failure time is observed; that is, by $\Pr[C_i \ge t_i^*] \equiv K(t_i^*-)$. Of course we do not know $K(t)$ so we must use an estimate; we use the Kaplan–Meier estimator of $K(t)$ given in (3). Because this estimator was first proposed by Robins and Rotnitzky (1992) we denote the resulting estimator by $\widehat{F}_{\rm rr}(t)$; it is given by

$$\widehat{F}_{\rm rr}(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{I[t_i \le t]\delta_i}{\widehat{K}(t_i-)}. \tag{5}$$

Note that we have used $I[t_i \le t]$ rather than $I[t_i^* \le t]$ in (4) to emphasize that $\widehat{F}_{\rm rr}(t)$ can be calculated using the observed data; this replacement is possible as $I[t_i \le t]\delta_i = I[t_i^* \le t]\delta_i$.

3. EQUIVALENCE OF $\widehat{\mathbf{F}}_{\mathbf{rr}}(\mathbf{t})$ AND $\widehat{\mathbf{F}}_{\mathbf{km}}(\mathbf{t})$

Note that both $\widehat{F}_{\rm rr}(t)$ and $\widehat{F}_{\rm km}(t) := 1 - \widehat{S}_{\rm km}(t)$ are right-continuous step functions with possible jumps at times τ_j. Thus, $\widehat{F}_{\rm rr}$ and $\widehat{F}_{\rm km}$ are the same if the magnitudes of the jumps in the two functions are equal. The jump in $\widehat{F}_{\rm km}$ at time τ_j is given by

$$\widehat{S}_{\rm km}(\tau_j-) - \widehat{S}_{\rm km}(\tau_j) = \widehat{S}_{\rm km}(\tau_j-)\frac{n_j}{Y(\tau_j)}, \tag{6}$$

while the jump in $\widehat{F}_{\rm rr}(\tau_j)$ is given by

$$\widehat{F}_{\rm rr}(\tau_j) - \widehat{F}_{\rm rr}(\tau_j-) = \frac{1}{N}\frac{n_j}{\widehat{K}(\tau_j-)}$$

The jumps are equal provided

$$\frac{1}{N}\frac{1}{\widehat{K}(\tau_j-)} = \frac{\widehat{S}_{\rm km}(\tau_j-)}{Y(\tau_j)}$$

or

$$\widehat{S}_{\rm km}(\tau_j-)\widehat{K}(\tau_j-) = \frac{1}{N}Y(\tau_j). \tag{7}$$

As long as there is no time τ_j for which $n_j m_j > 0$ (i.e., no ties between deaths and censored values), then

$$\begin{aligned}
&\widehat{S}_{\rm km}(\tau_j-)\widehat{K}(\tau_j-) \\
&\quad = \prod_{j'<j}\left(1 - \frac{n_{j'}}{Y(\tau_{j'})}\right)\prod_{j'<j}\left(1 - \frac{m_{j'}}{Y(\tau_{j'})}\right) \\
&\quad = \prod_{j'<j}\left(1 - \frac{n_{j'} + m_{j'}}{Y(\tau_{j'})}\right);
\end{aligned}$$

but

$$\begin{aligned}
&\prod_{j'<j}\left(1 - \frac{n_{j'} + m_{j'}}{Y(\tau_{j'})}\right) \\
&= \left(1 - \frac{n_1 + m_1}{n_1 + m_1 + \cdots + n_J + m_J}\right) \\
&\quad \times \left(1 - \frac{n_2 + m_2}{n_2 + m_2 + \cdots + n_J + m_J}\right) \\
&\quad \cdots \left(1 - \frac{n_{j-1} + m_{j-1}}{n_{j-1} + m_{j-1} + \cdots + n_J + m_J}\right) \\
&= \left(\frac{n_2 + m_2 + \cdots + n_J + m_J}{n_1 + m_1 + \cdots + n_J + m_J}\right) \\
&\quad \times \left(\frac{n_3 + m_3 + \cdots + n_J + m_J}{n_2 + m_2 + \cdots + n_J + m_J}\right) \\
&\quad \cdots \left(\frac{n_j + m_j + \cdots + n_J + m_J}{n_{j-1} + m_{j-1} + \cdots + n_J + m_J}\right) \\
&= \left(\frac{n_j + m_j + \cdots + n_J + m_J}{n_1 + m_1 \cdots + n_J + m_J}\right) = \frac{Y(\tau_j)}{N}
\end{aligned}$$

since $n_1 + m_1 \cdots + n_J + m_J = N$, so that equation (7) holds.

For the case where $n_j m_j > 0$ for some j, the argument above breaks down because

$$\left(1 - \frac{n_j}{Y(\tau_j)}\right)\left(1 - \frac{m_j}{Y(\tau_j)}\right) \ne \left(1 - \frac{n_j + m_j}{Y(\tau_j)}\right).$$

We can ask, what function $K'(t)$ of the form $K'(t) = \prod_{\{j|\tau_j \le t\}}(1 - d_j)$ would make $\widehat{F}_{\rm rr}(t)$ equal to $\widehat{F}_{\rm km}(t)$ even in the presence of ties. The appropriate choice of d_j solves

$$\left(1 - \frac{n_j}{Y(\tau_j)}\right)(1 - d_j) = \left(1 - \frac{n_j + m_j}{Y(\tau_j)}\right)$$

for each j, from which we obtain $d_j = m_j/\{Y(\tau_j) - n_j\}$ and hence

$$K'(t) = \prod_{\{j|\tau_j \le t\}}\left(1 - \frac{m_j}{Y(\tau_j) - n_j}\right).$$

Note that K' is the Kaplan–Meier estimator of censoring times we would obtain if we broke the ties between failures and censored observations by assuming that the failures had occurred just before the censored observations. This coincides with the usual convention when calculating the Kaplan–Meier estimator of failure times with data where there are ties between failure and censoring times (Kaplan and Meier 1958, p. 461).

4. COUPLED ESTIMATION OF THE DISTRIBUTION OF FAILURE AND CENSORING TIMES

The results in Section 3 are somewhat unsatisfactory in that the definition of $\widehat{F}_{\rm rr}(t)$ uses a Kaplan–Meier estimator (for the censoring times, $\widehat{K}$). Hence, these results would be unsuitable for an a priori development of the Kaplan–Meier estimator. In this section, we introduce a "new" inverse-probability-of-censoring weighted estimator of $F(t)$ that makes no reference to the Kaplan–Meier estimator of the censoring times. We then show that this "new" estimator is identical to the Kaplan–Meier estimator. Our approach is to simultaneously estimate $F(t) = \Pr[T_i^* \le t]$ and $G(t) = 1 - K(t) = \Pr[C_i \le t]$ using coupled inverse-probability-of-censoring weighted estimators. Let $\widehat{F}(t)$ and $\widehat{G}(t)$ be given by

$$\widehat{F}(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{I[t_i \le t]\delta_i}{1-\widehat{G}(t_i-)},$$

and

$$\widehat{G}(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{I[t_i \le t]\overline{\delta}_i}{1-\widehat{F}(t_i)},$$

where $\overline{\delta}_i = I[c_i < t_i^*]$. Then $\widehat{F}(t)$ is a step function with jumps at times τ_j for which $n_j > 0$ and $\widehat{G}(t)$ is a step function with jumps at times τ_j for which $m_j > 0$. The asymmetry in definitions of $\widehat{F}(t)$ and $\widehat{G}(t)$ reflects the choice that when failure and censoring times are tied, the censored observations are considered to have been lost to follow-up after the failures had occurred. Denoting the jumps in $\widehat{F}(\tau_j)$ and $\widehat{G}(\tau_j)$ by f_j and g_j we have

$$f_j = \frac{1}{N}\frac{n_j}{\left(1-\sum\limits_{j'<j} g_{j'}\right)}, \tag{8}$$

and

$$g_j = \frac{1}{N}\frac{m_j}{\left(1-\sum\limits_{j'\le j} f_{j'}\right)}, \tag{9}$$

where the sum $\sum_{j<1} g_j$ is to be interpreted as 0. Note that these equations are easily uncoupled to yield

$$f_j = \frac{n_j}{N - \sum\limits_{j'<j} \frac{m_{j'}}{\left(1-\sum\limits_{j''\le j'} f_{j''}\right)}}, \tag{10}$$

and

$$g_j = \frac{m_j}{N - \sum\limits_{j'\le j} \frac{n_{j'}}{\left(1-\sum\limits_{j''<j'} g_{j''}\right)}}. \tag{11}$$

Equations (10)–(11) for the f_j and g_j are triangular; that is, the right side of the equation (10) expresses f_j in terms of $f_{j'}, j' < j$. Hence, the f_j and hence $\widehat{F}(t)$ can be calculated recursively using (10). Similarly, $\widehat{G}(t)$ can be calculated using (11), if desired.

Although it is not immediately obvious, the fact is that $\widehat{F}(t) = \widehat{F}_{\mathrm{km}}(t)$. To see this recall that the masses in the Kaplan–Meier estimator are the maximizers of the likelihood

$$L = \prod_{j=1}^{J} f_j^{n_j}\left(\sum_{j'>j} f_{j'}\right)^{m_j} \tag{12}$$

subject to $\sum_{j=1}^{J+1} f_j = 1$. Following Turnbull (1976), note that $\{f_j, 1 \le j \le J+1\}$ solves this maximization problem if

$$D_j \doteq \frac{\partial \ln L}{\partial f_j} - \sum_{j=1}^{J+1} f_j \frac{\partial \ln L}{\partial f_j} = 0,$$

and $\sum_{j=1}^{J+1} f_j = 1$. Some algebra shows that the condition $D_j = 0$ can be rewritten as

$$\frac{n_j}{f_j} + \sum_{j'<j} \frac{m_{j'}}{\left(\sum\limits_{j''>j'} f_{j''}\right)} - N = 0\,; \tag{13}$$

solving (13) for f_j yields Equation (10), establishing the equivalence of $\widehat{F}(t)$ and $\widehat{F}_{\mathrm{km}}(t)$.

5. DISCUSSION

The Kaplan–Meier estimator is a fundamental tool in survival analysis. It is usually introduced as a nonparametric maximum likelihood estimator. The likelihood-based approach is useful, leading to useful generalizations when data are subject to interval censoring (Turnbull 1976), truncation (Woodroofe 1985; Wang, Jewell, and Tsai 1986) or both (Frydman 1994). We have shown that the Kaplan–Meier estimator can also be expressed as an inverse-probability-of-censoring weighted estimator.

The weighted average form given in this article with the true K is an average of iid terms under the random censoring model. Even under the model when censoring times are regarded fixed (Meier 1975), it is an average of independent (but not necessarily identically distributed) terms and is therefore subject to appropriate laws of large numbers and central limit theorems. Thus, the inverse-probability-of-censoring weighted estimator is also convenient for asymptotic theory.

Since the inverse-probability-of-censoring approach in survival analysis was introduced by Robins and Rotnitzky (1992) it has also led to useful generalizations, primarily to more general censoring models where the censoring hazard may depend on an observable covariate history {see, e.g., Robins and Finkelstein (2000); Satten and Datta (in press); and Satten, Datta, and Robins (in press)}. We have given two demonstrations of the equivalence of the inverse-probability-of censoring weighted sum and product-limit representations of the Kaplan–Meier estimator. The first, given in Section 3, is designed to achieve the result quickly, but requires the availability of the Kaplan–Meier estimator of censoring times. The second, given in Section 4, is less direct, but constructs the weighted estimator without making any reference to the Kaplan–Meier estimator.

[Received February 2000. Revised December 2000.]

REFERENCES

Efron, B. (1967), "The Two Sample Problem With Censored Data," in *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability* (vol. IV), Berkeley, CA: University of California Press, pp. 831–853.

Frydman, H. (1994), "A Note on Nonparametric Estimation of the Distribution Function From Interval-Censored and Truncated Observations," *Journal of the Royal Statistical Society*, Series B, 56, 71–74.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Kalbeisch, J., and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Meier, P. (1975), "Estimation of a Distribution Function From Incomplete Observations," in *Perspectives in Probability and Statistics*, ed. J. Gani, Sheffeld, England: Applied Probability Trust. *Also* Miller, R. G., Jr. (1981), *Survival Analysis*, New York: Wiley.

Robins J. M. (1993), "Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers," in *Proceedings of the American Statistical Association—Biopharmaceutical Section*, Alexandria, VA: American Statistical Association, pp. 24–33.

Robins, J., and Finkelstein, D. (2000), "Correcting for Non-compliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests," *Biometrics*, 56, 779–788.

Robins J. M., and Rotnitzky A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology–Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhauser, pp. 297–331.

Satten, G. A., and Datta, S. (in press), "Marginal Estimation for Multistage Models: Waiting Time Distributions and Competing Risk Analyses," *Statistics in Medicine*.

Satten G. A., Datta S., and Robins J. M. (in press), "Estimating the Marginal Survival Function in the Presence of Time Dependent Covariates," *Statistics and Probability Letters*.

Turnbull (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society*, Series B, 38, 290–295.

Wang, M.-C., Jewell, N., and Tsai, W.-Y. (1986), "Asymptotic Properties of the Product Limit Estimator Under Random Truncation," *The Annals of Statistics*, 124, 1597–1605.

Woodroofe, M. (1985), "Estimating a Distribution Function With Truncated Data," *The Annals of Statistics*, 13, 163–177; Correction, 15, 883 (1987).