Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2009 *Paper* 66

Mean Survival Time from Right Censored Data

Ming Zhong* Kenneth R. Hess†

*Department of Statistics; Texas A & M University, zhongming@gmail.com

†Department of Biostatistics; The University of Texas M. D. Anderson Cancer Center, khess@mdanderson.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/cobra/art66

Copyright ©2009 by the authors.

Mean Survival Time from Right Censored Data

Ming Zhong and Kenneth R. Hess

Abstract

A nonparametric estimate of the mean survival time can be obtained as the area under the Kaplan-Meier estimate of the survival curve. A common modification is to change the largest observation to a death time if it is censored. We conducted a simulation study to assess the behavior of this estimator of the mean survival time in the presence of right censoring.

We simulated data from seven distributions: exponential, normal, uniform, log-normal, gamma, log-logistic, and Weibull. This allowed us to compare the results of the estimates to the known true values and to quantify the bias and the variance. Our simulations cover proportions of random censoring from 0% to 90%.

The bias of the modified Kaplan-Meier mean estimator increases with the proportion of censoring. The rate of increase varied substantially from distribution to distribution. Distributions with long right tails (log-logistic, log normal, exponential) increased the quickest (i.e., at lower censoring proportions). The other distributions are relatively unbiased until around 60% censoring. The Normal distribution remains unbiased up to 90% censoring.

Thus, the behavior of the modified Kaplan-Meier mean estimator depends heavily on the nature of the distribution being estimated. Since we rarely have knowledge of the underlying true distribution, care must be taken when estimating the mean from censored data. With modest censoring, estimates are relatively unbiased, but as censoring increases so does the bias. With 30% or more censoring the bias may be too high. This is in contrast to the Kaplan-Meier estimator of the median which is relatively unbiased.

Introduction

In the presence of right censoring, the usual estimate of the mean survival time is not appropriate [1]. The censoring leads to an underestimate of the true mean which worsens as the censoring increases. Alternatively, the mean survival time can be defined as the area under the survival curve, $S(t)$ [2, 3]. In the absence of censoring, this is equivalent to the usual estimate of the mean.

A nonparametric estimate of the mean survival time can be obtained by substituting the Kaplan-Meier estimator for the unknown survival function.

$$\hat{\mu} = \int \hat{S}(t)dt$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator [2]. When the largest observed time is censored, the Kaplan-Meier estimator is undefined beyond the largest observed time. Thus, this estimator is only appropriate when the largest observed time is a death time [2, 3].

One approach to overcome this limitation is to change the largest observation to a death time if it is censored [4]. This modification is used by several statistical programs in computing the mean survival time [5]. While the estimator has been shown to be consistent and asymptotically normal [6, 7], the behavior of the estimator has not been studied. We conducted a simulation study to assess the behavior.

Methods

We assume that we have n independent, identically distributed lifetimes (that is, non-negative random variables), T_i, with continuous distribution function F, and n independent, identically distributed censoring times, C_i, with continuous distribution function G. We also assume that C_i and T_i are independent for i =1, 2,…, n. The actual observations consist of (x_i, d_i), where $x_i = \min(T_i, C_i)$ and $d_i = \mathrm{I}[T_i \leq C_i]$ is an indicator of the censoring status of x_i. This set of assumptions is often referred to as the random censorship model.

For our comparisons, we simulated data from known distributions. This allowed us to compare the results of the estimates to the known true values and to quantify the bias and the variance. We chose seven distributions for simulation: exponential, normal, uniform, lognormal, gamma, log-logistic, and Weibull (Figure 1). The density functions of the seven distributions are listed in Table 1 as well as their true means and medians.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

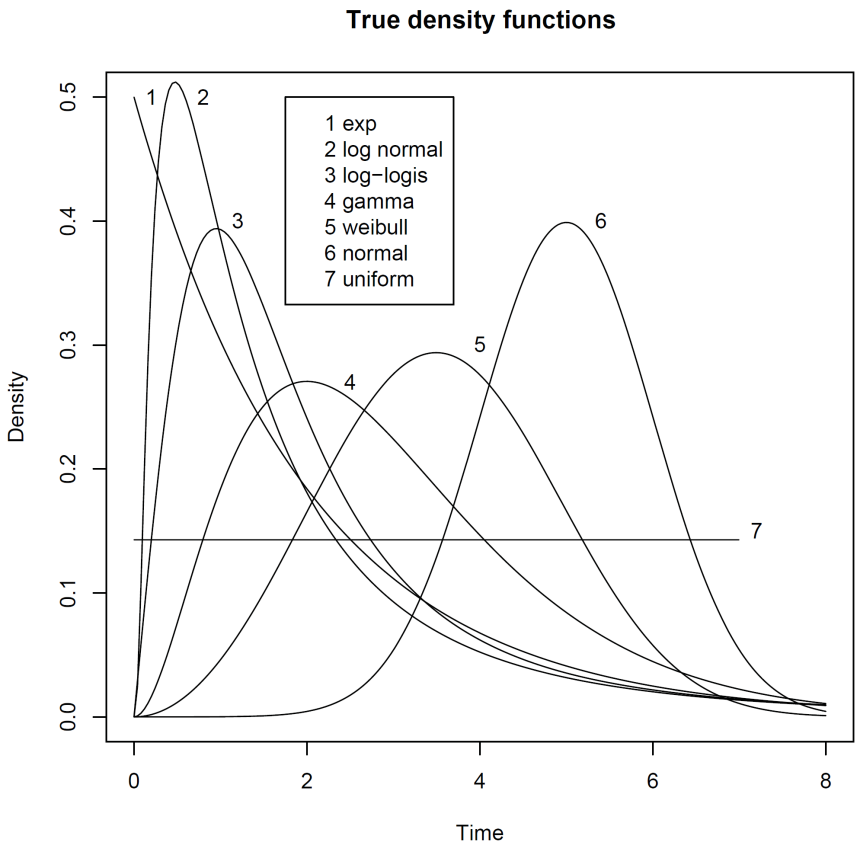

Figure 1: The seven distributions used in the simulation study

Distribution	Density	Mean	Median
Exponential	$\frac{1}{2}\exp(-\frac{x}{2})$	2	1.39
Normal	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-5)^2}{2})$	5	5
Uniform	$\frac{1}{7}I[0 \leq x \leq 7]$	3.5	3.5
Log normal	$\frac{1}{\sqrt{2\pi}}\frac{\exp(-(\log x-0.25)^2/2)}{x}$	2.12	1.28
Gamma	$\frac{1}{\Gamma(3)}x^2\exp(-x)$	3	2.67
Log-logistic	$\frac{2x/e}{(1+x^2/e)^2}$	2.59	1.65
Weibull	$\frac{3}{64}x^2\exp(-\frac{x^3}{64})$	3.57	3.54

Table 1: The underlying true distributions with their means and medians

To implement random censorship, we independently generated uniform censoring times on the interval [0, U], where U was selected to achieve a given proportion of censoring and solved analytically. Our simulations cover proportions of random censoring from 0% to 90%.

http://biostats.bepress.com/cobra/art66

Results

Figure 2 shows how the bias of the Kaplan-Meier mean estimator increases with the proportion of censoring. The rate of increase varied substantially from distribution to distribution. Distributions with long right tails (log-logistic, log normal, exponential) increased the quickest (i.e., at lower censoring proportions). The other distributions are relatively unbiased until around 60% censoring. The Normal distribution remains unbiased up to 90% censoring.

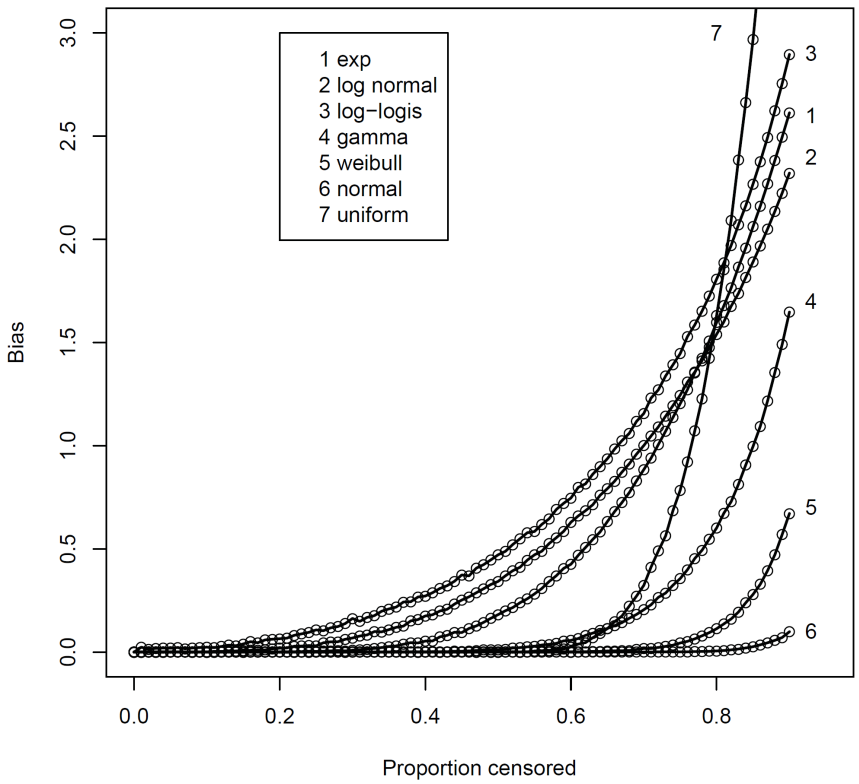

Figure 2: Bias vs. proportion censored. Lowess smooths superimposed.

Figure 3 shows the variance of the modified Kaplan-Meier mean estimator vs. the proportion of censoring. The variances are much smaller than the biases in general and do not increase with the proportion censoring.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

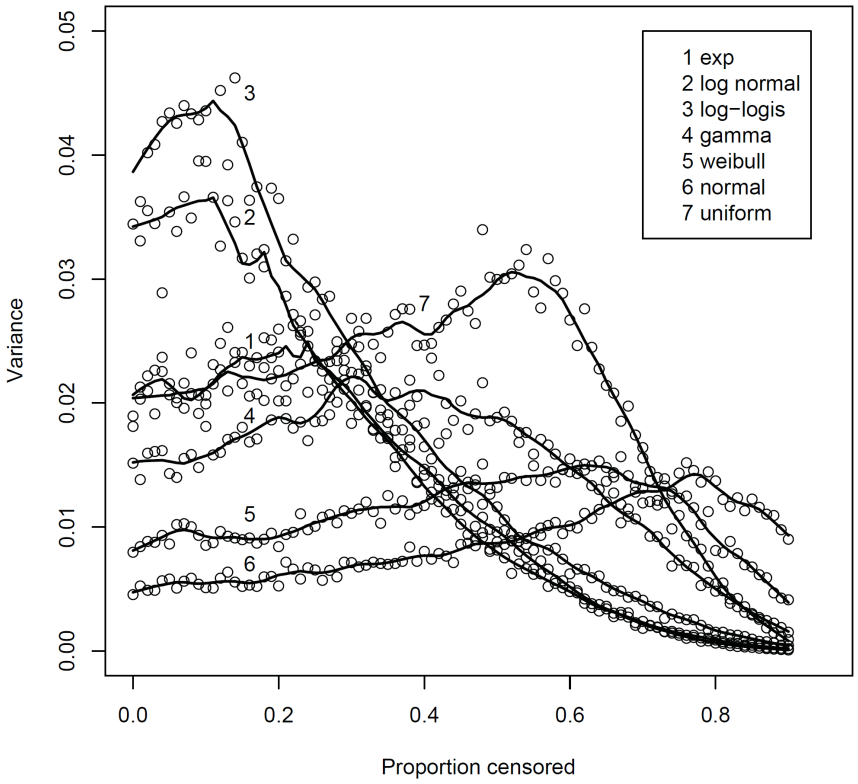

Figure 3: Variance vs. proportion censored. Lowess smooths superimposed.

We also studied the bias and variance of the Kaplan-Meier estimator of the median (time at which the Kaplan-Meier estimator crosses 50% survival). Because this estimator is undefined when the Kaplan-Meier estimator does not fall below 50%, we were able to study its behavior from 0% to 60% censoring (for the survival time distributions we studied). As expected, the estimator remains relatively unbiased (mean bias = 0.0001). The variance increases slowly with increasing censoring.

Discussion

The behavior of the modified Kaplan-Meier mean estimator (i.e., area under the Kaplan-Meier curve with the last observation changed to uncensored if originally uncensored) depends heavily on the nature of the distribution being estimated. Since we rarely have knowledge of the underlying true distribution, care must be taken when estimating the mean from censored data. With modest censoring, estimates are relatively unbiased, but as censoring increases so does the bias. With 30% or more censoring the bias may be too high. This is in contrast to the Kaplan-Meier estimator of the median which is relatively unbiased. Given that most survival-time distributions are skewed with longer right tails, it would seem prudent to report an unbiased estimate of the median rather than a biased estimate of the mean.

COBRA
Collection of Biostatistics
Research Archive

References

1. Datta S. Estimating the mean life time using right censored data. *Statistical Methodology*. 2005; **2**:65-69.

2. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; **53**:457-481.

3. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag. 1997; pgs 109-114. Section 4.5 Point and interval estimates of the mean and median survival time.

4. Efron B. The two sample problem with censored data. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. New York: Prentice Hall. 1967; **4**:831-853.

5. Barker C. The mean, median, and confidence intervals of the Kaplan-Meier Survival Estimate – Computations and Applications. *The American Statistician*. 2009; **63**:78-80.

6. Gill R. Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*. 1983; **11**:49-58.

7. Susaral V, Van Ryzin J. Large sample theory for an estimator of the mean survival time for censored samples. *The Annals of Statistics*. 1980; **8**:1002-1016.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press