Unbiased Transformation of Censored Survival Data

Eunju Lee : Master degree candidate, Department of Statistics, Pusan National University (dmswn1578@pusan.ac.kr)
Choongrak Kim : Professor, Department of Statistics, Pusan National University

Introduction

- **Goal** : Find unbiased transformation of censored observations to uncensored observations
- **Methods** : 3 methods are proposed
  (1) Method 1 : Buckley-James type transformation
  (2) Method 2 : Koul-Susarla-Van Ryzin type transformation
  (3) Method 3 : Mean imputation
- **Comparisons**
  (1) MISE : Mean Integrated Squared Error of distribution function
  (2) MSE : Mean Squared Error of median survival time
  (3) *c*-index : Harrell's concordance index

Basic Notations and Definition

- **Basic Functions**
  - $T_1, \cdots, T_n$: *iid* true survival times with cdf $F(\cdot)$ and pdf $f(\cdot)$
  - $C_1, \cdots, C_n$: *iid* censored times with cdf $G(\cdot)$ and pdf $g(\cdot)$
  - We observe ordered data $(Y_i, \delta_i), i = 1, \cdots, n$,
    where $\delta_i = I(T_i \le C_i)$ is censoring indicator and (T_i, C_i) are independent.

$$P(Y_i, \delta_i) = \begin{cases} P(Y_i = T_i, C_i > Y_i) = f_\theta(Y_i)(1 - G_\gamma(Y_i)) & \text{if } \delta_i = 1 \\ P(Y_i = C_i, T_i > Y_i) = g_\gamma(Y_i)(1 - F_\theta(Y_i)) & \text{if } \delta_i = 0 \end{cases}$$

- **Kernel Estimator**

$$\hat{f}(t) = \frac{1}{h}\sum_{i=1}^{n} s_i K\left(\frac{t - y_i}{h}\right), \quad \hat{F}(t) = \frac{1}{h}\sum_{i=1}^{n} s_i W\left(\frac{t - y_i}{h}\right)$$

where K : kerenel function, W : cumulative kernel function
s_i : jump size at y_i in Kaplan-Meier estimator
h : smoothing parameter, bandwidth

- **Kaplan-Meier Estimator**

$$\hat{S}(t) = \prod_{i:y_i \le t}\left(1 - \frac{1}{n - i + 1}\right)^{\delta_i}$$

- ***c*-index**

$$c = \frac{\sum_{i=1}\sum_{j \ne i} I(y_i > y_j) I(\hat{y}_i > \hat{y}_j)\delta_j}{\sum_{i=1}\sum_{j \ne i} I(y_i > y_j)\delta_j}$$

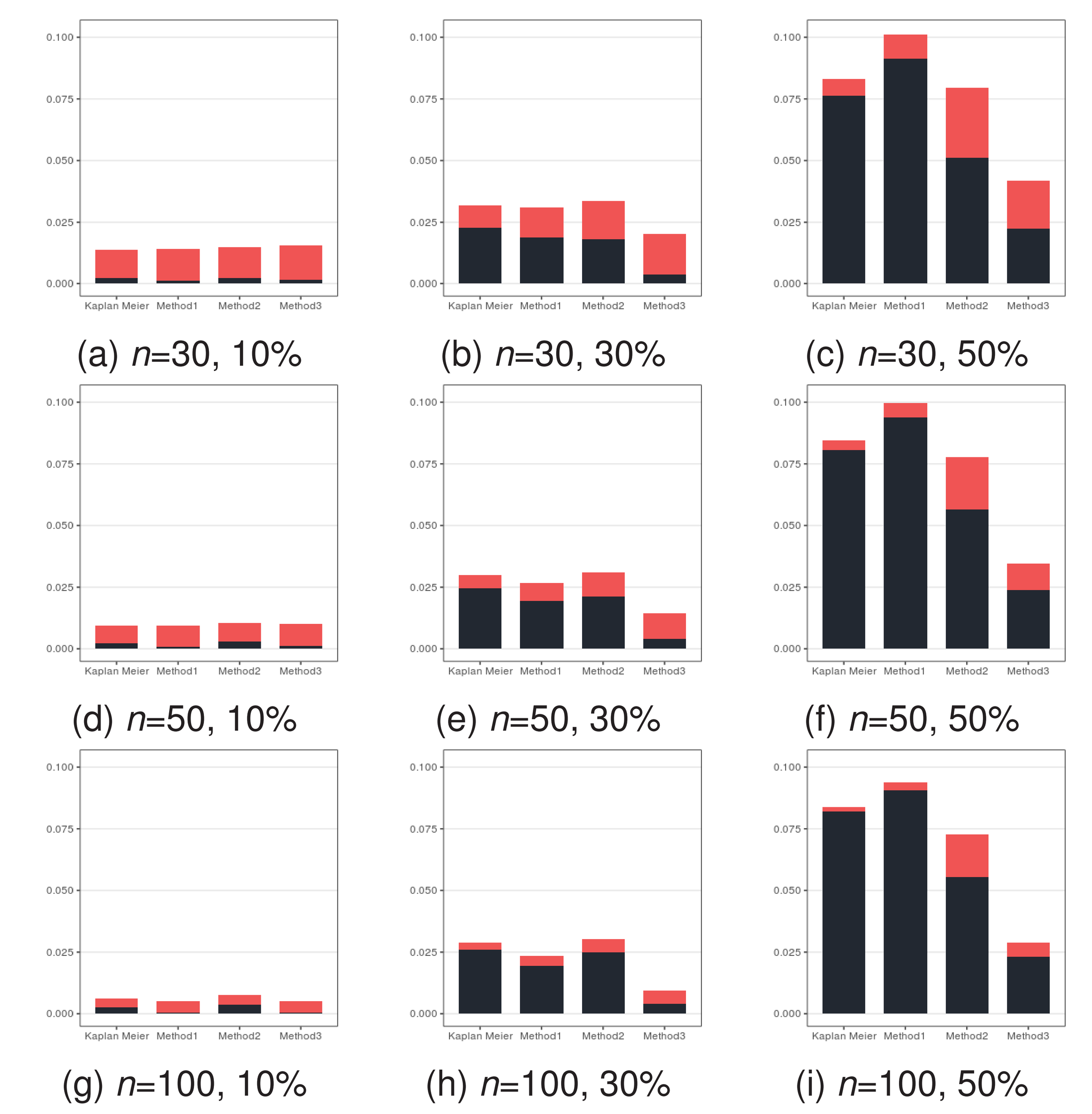
The Proposed Method

The proposed method
Method 1. $\hat{y}_i^{(1)} = \delta_i y_i + (1 - \delta_i)\frac{\hat{f}(y_i)}{\hat{g}(y_i)(1 - \hat{F}(y_i))} y_i$
Method 2. $\hat{y}_i^{(2)} = \delta_i y_i + (1 - \delta_i)\sum_{k:y_k > y_i} s_k y_k / \{1 - \hat{F}(y_i)\}$
Method 3. $\hat{y}_i^{(3)} = \delta_i y_i + (1 - \delta_i)\sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j$

- **Method 1 : Buckley-James type transformation**
  - Motivation : $E(Y^{(1)}) = E(T)$, $Y^{(1)} = \delta Y + (1 - \delta)\frac{f(Y)}{g(Y)(1 - F(Y))} Y$
  - Estimate $\hat{f}(y)$, $\hat{F}(y)$, and $\hat{g}(y)$ by kernel estimator using Epanechnikov kernel function.
  - Calculate bandwidth h using likelihood cross validation.
- **Method 2 : Koul-Susarla-Van Ryzin type transformation**
  - Motivation : $E(Y^{(2)}) = E(T)$, $Y^{(2)} = \delta Y + (1 - \delta)\, E(T|T > Y)$
  - Estimate $\hat{E}(T|T > y)$ as $\int_y^\infty \frac{tdF(t)}{1 - F(y)}$.
  - Estimate $F(y)$ by kernel estimator like Method 1.
- **Method 3 : Mean imputation**
  - Motivation : $E(Y^{(3)}) = E(T)$, $Y^{(3)} = \delta Y + (1 - \delta)\, E(T|T > Y)$
  - Estimate $\hat{E}(T|T > y)$ as $\sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j$.
  - Assume that y_n is uncensored data.

Numerical Result

Comparison of Distribution Function



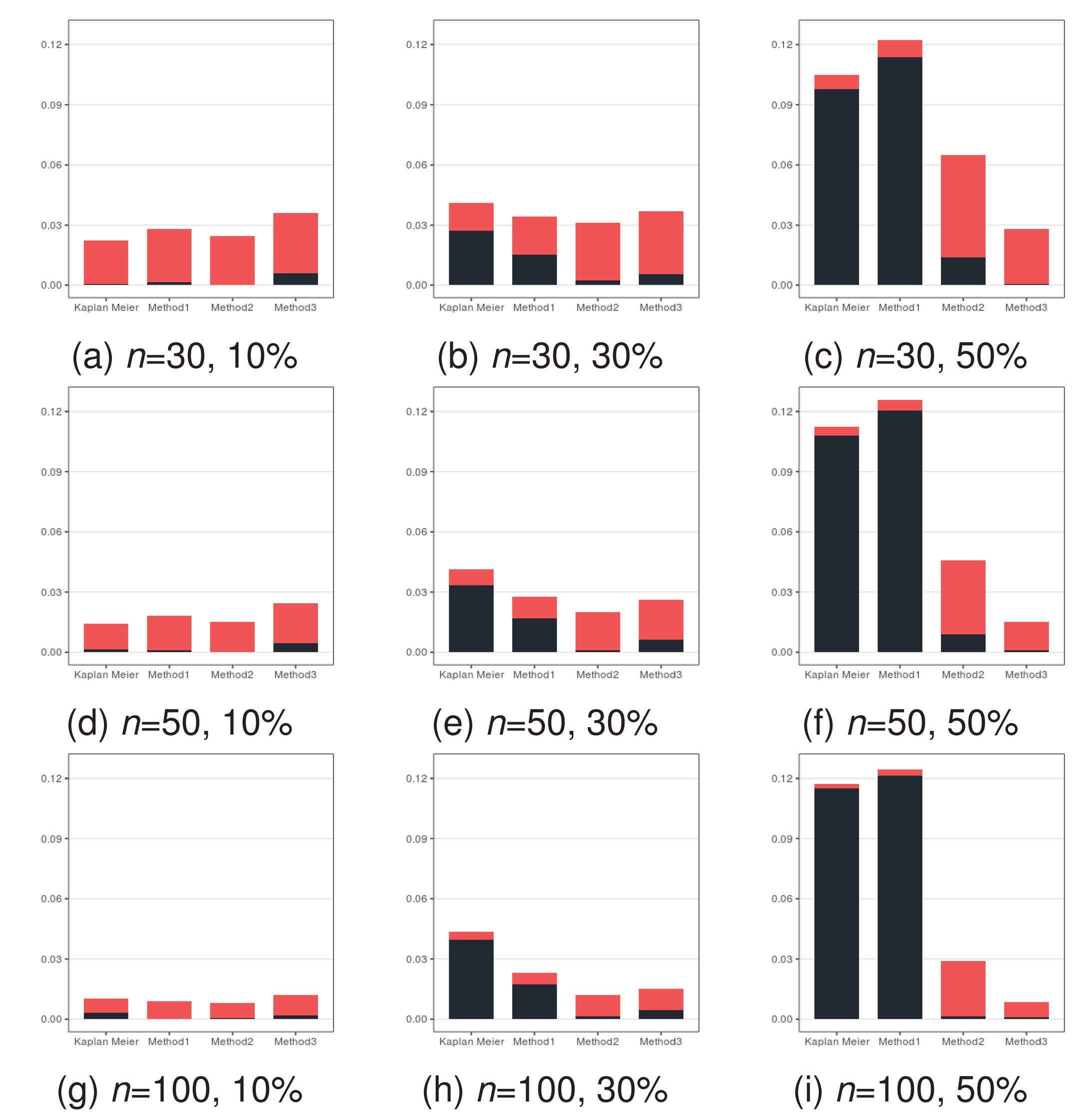
(a) n=30, 10% (b) n=30, 30% (c) n=30, 50%
(d) n=50, 10% (e) n=50, 30% (f) n=50, 50%
(g) n=100, 10% (h) n=100, 30% (i) n=100, 50%

Figure 1. The MISE (IV + ISB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100. The above and the below of each historgram denote IV and ISB, respectively.

Table 1. The MISE of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	MISE 10%	30%	50%
Kaplan-Meier	30	0.013	0.032	0.083
	50	0.009	0.029	0.085
	100	0.007	0.029	0.084
Method 1	30	0.014	0.031	0.102
	50	0.010	0.027	0.100
	100	0.006	0.023	0.094
Method 2	30	0.015	0.033	0.079
	50	0.011	0.031	0.077
	100	0.008	0.031	0.072
Method 3	30	0.015	0.021	0.042
	50	0.010	0.014	0.035
	100	0.006	0.009	0.029

Comparison of the Median Survival Time



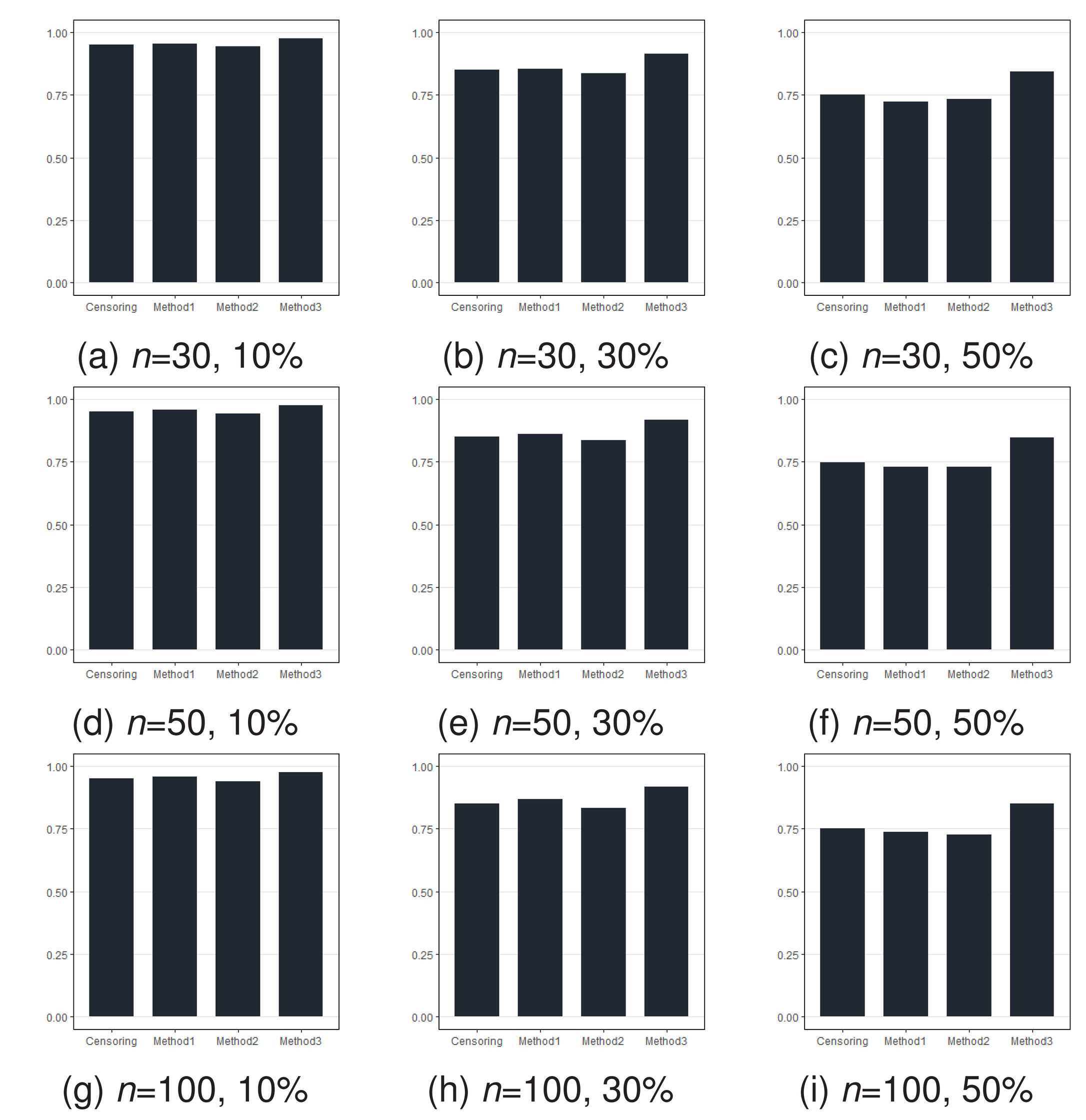
(a) n=30, 10% (b) n=30, 30% (c) n=30, 50%
(d) n=50, 10% (e) n=50, 30% (f) n=50, 50%
(g) n=100, 10% (h) n=100, 30% (i) n=100, 50%

Figure 2. The MSE (V + SB) of four estimators of the median survival time with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100. The above and the below of each historgram denote V and SB, respectively.

Table 2. The MSE (V + SB) of four estimators of the median survival time with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	MSE 10%	30%	50%
Kaplan-Meier	30	0.023	0.041	0.105
	50	0.014	0.041	0.122
	100	0.010	0.044	0.117
Method 1	30	0.029	0.034	0.122
	50	0.018	0.027	0.125
	100	0.009	0.024	0.124
Method 2	30	0.024	0.031	0.065
	50	0.015	0.020	0.046
	100	0.008	0.012	0.029
Method 3	30	0.036	0.037	0.028
	50	0.025	0.026	0.015
	100	0.012	0.015	0.009

Comparison of the Concordance



(a) n=30, 10% (b) n=30, 30% (c) n=30, 50%
(d) n=50, 10% (e) n=50, 30% (f) n=50, 50%
(g) n=100, 10% (h) n=100, 30% (i) n=100, 50%

Figure 3. The *c*-index of the censored data and the three data transformed by the proposed methods with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100.

Table 3. The *c*-index of the censored data and the three transformed data with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	*c*-index 10%	30%	50%
Censoring	30	0.951	0.851	0.750
	50	0.949	0.849	0.748
	100	0.950	0.850	0.749
Method 1	30	0.954	0.855	0.724
	50	0.955	0.860	0.729
	100	0.957	0.868	0.736
Method 2	30	0.943	0.836	0.732
	50	0.941	0.835	0.731
	100	0.938	0.831	0.728
Method 3	30	0.973	0.916	0.844
	50	0.974	0.918	0.846
	100	0.973	0.917	0.849

Concluding Remarks

- **Results** : The mean imputation showed the best results under various censoring percentages.
- **Future Research** : We might extend our results to the regression problem in censored survival data.