ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 164, 2003

*Agnieszka Rossa**

PARTIALLY PARAMETRIC ESTIMATION OF SURVIVAL FUNCTION IN THE RIGHT–CENSORED DATA

Abstract. In many medical, biological or economic follow-up studies the subject of observation is survival, failure or duration time, that is the length of time elapsed from a specific starting point to an event of interest. In engineering applications it may be the time to failure of piece of equipment, in medical trials – time to occurrence of a particular disease or time to death of a patient due to some specific disease, in economic studies – time of being unemployed and so on.

In the analysis of survival-type variables one is often faced with right-censored observations. Sometimes it is impossible to measure the true failure time of an individual due to previous occurrence of some other event called competing event, which result in interruption of observation before the event occurs. It may be withdrawal of the subject from the study or failure from some causes other the one of interest or simply limitation on the length of study. If we are only interested in failure time, then the competing events can be regarded as right-censoring the event of interest. It means that for each individual we observe either the time to failure or the time to censoring and for censored individuals we know only that the time to failure is greater then the censoring time.

In reliability studies censoring is often planned in order to obtain information sooner than it is otherwise possible. Instead of testing m units until they fail, the Type I censoring design is employed in which more then m units are tested but observation is terminated earlier at the end of some specified period x^*. Those units, which failed before this time yield complete observations and the rest of them is right-censored. Despite such incompleteness of the data it is often desired to estimate survival function that is the probability $P(X > x)$ that the true failure time X in the population of individuals exceeds x.

The paper deals with a problem of estimating survival function in the right-censored data. Some improvements of the well-known Kaplan-Meier estimator are discussed and their properties are studied.

Key words: censored data, survival function.

* Ph. Doctor, Institute of Econometrics and Statistics, University of Łódź.

1. MODEL

Nearly all the statistical methods for censored survival data are based on the assumption that censoring mechanism is not related to mechanism causing failures. A common example of this type of censoring occurs in a clinical trials where surviving is measured from entry into the study and one observes true survival times of those patients who fail by the time of analysis and censored times for those who do not. Thus, the usual model for censored survival analysis assumes independent random censoring, which can be expressed in the following form

$$T = \min(X, Z), \quad \delta = \begin{cases} 1, & \text{if } X \leqslant Z \\ 0, & \text{if } X > Z, \end{cases}$$

where Z and X are independent, non-negative random variables, X represents the true failure time with strictly increasing cumulative distribution function F_X and Z represents the censoring time with a cumulative distribution function F_Z.

In the random censoring model the total number of items n under study is known in advance and for i-th item one observes only the minimum of the failure time and the censoring time.

The problem of estimating the survival function S in the presence of right censoring has been extensively studied. One of the most popular non-parametric estimators is a distribution-free product limit estimator suggested by K a p l a n and M e i e r (1958). The general idea of construction of any product limit estimator of survival function is based on a partition of time axis. Let

$$0 = x_0 < x_1 < \cdots < x_k < \cdots \tag{1}$$

be some distinct values such that the intervals $[x_{k-1}, x_k)$ constitute a partition of $[0, \infty)$. Let us consider the survival probability $S(x_k) = P(X > x_k)$. It is obvious that

$$P(X > x_k) = P(X > x_k | X > x_{k-1}) \cdot P(X > x_{k-1}).$$

Let us denote by $p(x_k)$ the conditional probability

$$p(x_k) = P(X > x_k | X > x_{k-1}) \tag{2}$$

Thus the survival probability $S(x) \equiv P(X > x)$ may be expressed as a product of conditional probabilities $p(x_k)$

$$S(x_k) = \prod_{j=1}^{k} p(x_j), \quad \text{for} \quad k = 1, 2, \ldots \tag{3}$$

with an initial assumption $S(x_0) = 1$. Any estimator $\hat{S}(x_k)$ of $S(x_k)$ can be constructed in a similar way, in terms of partition (1) from the product (3) of estimators of conditional probabilities (2).

2. PRODUCT LIMIT ESTIMATOR

Let us assume further the random censoring model. Let $x_1, x_2, \ldots, x_q$, $q \leqslant n$, denote the ordered sequence of distinct failure times, observed in the censored sample such that $0 = x_0 < x_1 < \ldots < x_q < \infty$ constitute a random partition of the half-line $[0, \infty)$ (it is assumed that no failure occurs in the time zero). Denote by n_k the number of individuals still alive and under observation just after x_k and by d_k the number of failures occurred at x_k. Let l_k be a number of individuals in a subinterval (x_{k-1}, x_k), $k = 1, 2, \ldots, q$, censored or failed, respectively. Here the usual convention is adopted that failures occurred in a time x are treated as if they appeared slightly before x, and censored observations occurred in a time x are treated as if they appeared just after x. It is worth also noting, that in the case of no ties $d_k = 1$ otherwise $d_k \geqslant 1$, and in the case of no censoring $l_k = 0$ otherwise $l_k \geqslant 0$ for each subinterval (x_{k-1}, x_k), where $k = 1, 2, \ldots, q$. As estimators of the conditional probabilities $p(x_k)$ Kaplan and Meier proposed

$$\hat{p}(x_k) = \frac{n_{n-1} - l_k - d_k}{n_{k-1} - l_k}, \quad k = 1, 2, \ldots, q \tag{4}$$

where $n_0 = n$. On the basis of (3) and (4) we obtain

$$S_{KM}(x) = \prod_{x_k \leqslant x} \frac{n_{n-1} - l_k - d_k}{n_{k-1} - l_k}, \quad k = 1, 2, \ldots, q,$$

with $S_{KM}(x_0) = 1$. The Kaplan-Meier estimator S_{KM} is a step function with jumps at those observations for which $\delta = 1$. If no censoring occurs it reduces to the step function with jumps of height $1/n$ at each x_k which is the usual empirical distribution function. The estimator has been shown by Kaplan and Meier to maximise the likelihood func-

tion of the observations in the class of all possible distributions. The Kaplan-Meier estimator is originally undefined for x beyond the largest observation, when this observation is censored. Efron (1967) proposed the convention of defining the estimator to be always zero for all $x > x_n$, where x_n denotes the last observation in the sample. Gill (1980) considered another modification, defining the estimator to be equal to $S_{KM}(x_n)$ for all $x > x_n$.

Klein *et al.* (1990) described a method for improving the Kaplan-Meier estimator by treating the uncensored observations non-parametrically and using a parametric model only for the censored observations. Their estimator is constructed by analogy to the complete data problem where an estimator of $S(x)$ is computed as the proportion of observations which exceed x. Let us define

$$H(x) \equiv P(T > x) \quad \text{and} \quad G_0(x) \equiv P(Z \leqslant x,\ \delta = 0).$$

Klein *et al.* considered the following representation of the survival probability $S(x)$

$$S(x) = H(x) + \int_0^x \frac{S(x)}{S(z)}\, dG_0(z). \tag{5}$$

Let $\hat{H}(x)$ and $\hat{G}_0(x)$ be estimators of probabilities $H(x)$ and $G_0(x)$, defined as

$$\hat{H}(x) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > x) \quad \text{and} \quad \hat{G}_0(x) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leqslant x,\ \delta = 0),$$

where I denotes the characteristic function. Thus an estimator of (5) can be expressed in the following form

$$\hat{S}(x) = \hat{H}(x) + \int_0^x \frac{S(x)}{S(z)}\, d\hat{G}_0(z). \tag{6}$$

Unfortunately, probabilities $S(x)$, $S(z)$ given on right hand side of (6) are not known. Klein *et al.* proposed to assume a reasonable family of distributions and estimate these unknown probabilities from this parametric model. In their study the Weibull family of distributions was taken under consideration. The Weibull survival function is of the form

$$S_W(x) = \exp(-\beta x^{\gamma}), \quad \beta,\ \gamma > 0 \tag{7}$$

Thus replacing in (6) unknown probabilities $S(x)$, $S(z)$ by parametric estimators of $S_W(x)$, $S_W(z)$ Klein *et al.* obtained a partially parametric estimator of survival probability (5)

$$S_{KL}(x) = \hat{H}(x) + \int_0^x \exp\left(\hat{\beta}(z^{\hat{\gamma}} - x^{\hat{\gamma}})\right) d\hat{G}_0(z), \tag{8}$$

where $\hat{\beta}$, $\hat{\gamma}$ are ML – estimators of β, γ.

Rossa (2002) proposed an estimator of survival probability, constructed by analogy to (3), as the product of estimators of conditional probabilities $p(x_k)$. However each $p(x_k)$ was estimated in a way suggested by Klein *et al.* The modification proposed by Rossa lead to the following partially parametric estimator (*pKM* estimator) of survival probability $S(x_j)$, $j = 1, 2, \ldots, q$.

$$S_{pKM}(x_j) = \prod_{k=1}^{j} \left(\frac{\hat{H}(x_k)}{\hat{H}(x_{k-1})} + \frac{1}{\hat{H}(x_{k-1})} \int_{x_{k-1}}^{x_k} \exp\left(\hat{\beta}(z^{\hat{\gamma}} - x^{\hat{\gamma}})\right) d\hat{G}_0(z) \right) \tag{9}$$

with an initial assumption $S_{pKM}(x_0) = 1$.

3. SIMULATION STUDY

Both the Kaplan-Meier estimator and its partially parametric versions (8) and (9) are very difficult for the theoretical analysis, because their distributions depend in a very complicated way on the survival distribution F_X and the censoring distribution F_Z. Thus, to assess the accuracy of the estimators a number of simulations was conducted.

It is well known that the mean squared error *MSE* or mean absolute deviation *MAD* are generally accepted measures of accuracy of an estimator when the estimator is biased.

Let $MSE_{KM}(p, F_X, F_Z)$ and $MAD_{KM}(p, F_X, F_Z)$ denote a mean squared error and mean absolute deviation, respectively, of the Kaplan-Meier estimator evaluated for the fixed survival probability $p \in (0,1)$ at the point $x = F_X^{-1}(p)$. Similarly, let $MSE_{KL}(p, F_X, F_Z)$ and $MAD_{KL}(p, F_X, F_Z)$ be a mean squared error and a mean absolute deviation of the Klein *et al.'s* estimator. By the analogy let $MSE_{pKM}(p, F_X, F_Z)$ and $MAD_{pKM}(p, F_X, F_Z)$ be a mean squared error and a mean absolute deviation of the *pKM* estimator. In the simulations the *BIAS* of the three estimators and the following ratios were studied

$$R_{MSE}^{KL} = \frac{MSE_{KL}(p, F_X, F_Z)}{MSE_{KM}(p, F_X, F_Z)}, \quad R_{MAD}^{KL} = \frac{MAD_{KL}(p, F_X, F_Z)}{MAD_{KM}(p, F_X, F_Z)}$$

and

$$R_{MSE}^{pKM} = \frac{MSE_{pKM}(p, F_X, F_Z)}{MSE_{KM}(p, F_X, F_Z)}, \quad R_{MAD}^{pKM} = \frac{MAD_{pKM}(p, F_X, F_Z)}{MAD_{KM}(p, F_X, F_Z)}$$

for some survival distributions F_X and censoring distributions F_Z. The survival distributions were simulated from the following distributions:

– The Weibull distribution $Wei(\beta, \gamma)$ with survival function $S_W(x)$ defined in (7); a special case of this family is the exponential distribution $Exp(\beta)$, with $\gamma = 1$.

– The log-logistic distribution $LogL(\beta, \gamma)$ with survival function equal to $\frac{1}{1+\beta x^{\gamma}}$

– The log-normal distribution $LogN(\mu, \sigma)$.

– The gamma distribution $Gam(\alpha, \beta)$ with density function equal to $\beta^{\alpha} x^{\alpha-1} \exp(-\beta x)/\Gamma(\alpha)$.

– The Gompertz distribution $Gomp(\beta, \gamma)$ with survival function equal to $\exp\left[\beta \frac{1-\exp(\gamma x)}{\gamma}\right]$.

– The Pareto distribution $Par(\beta, \gamma)$ with survival function $\left(\frac{\beta}{\beta+x}\right)^{\gamma}$.

These families of distributions are typical distributions usually used in the survival analysis.

Censored data were simulated by using an exponential distribution $Exp(\beta)$, with various values of the mean time to censoring β, yielding an assumed censoring fraction.

Each separate simulation comprised $N = 10\,000$ samples of size 10, 20, 30 and 50 consisting of n independent pairs (X_i, Z_i) such that the independent random variables X_i and Z_i were distributed according to the assumed survival and censoring distributions, respectively. For each sample, the sequence (T_i, δ_i), $i = 1, 2, \ldots, n$ was determined and the estimators S_{KM}, S_{KL}, S_{pKM} and their characteristics at points $x_j = F_X^{-1}(p_j)$ for equally spaced probabilities p_j, $j = 1, 2, \ldots, 19$ were calculated.

The representative small sample results obtained for various types of survival distributions are plotted on Figures 1, 2 and 3.

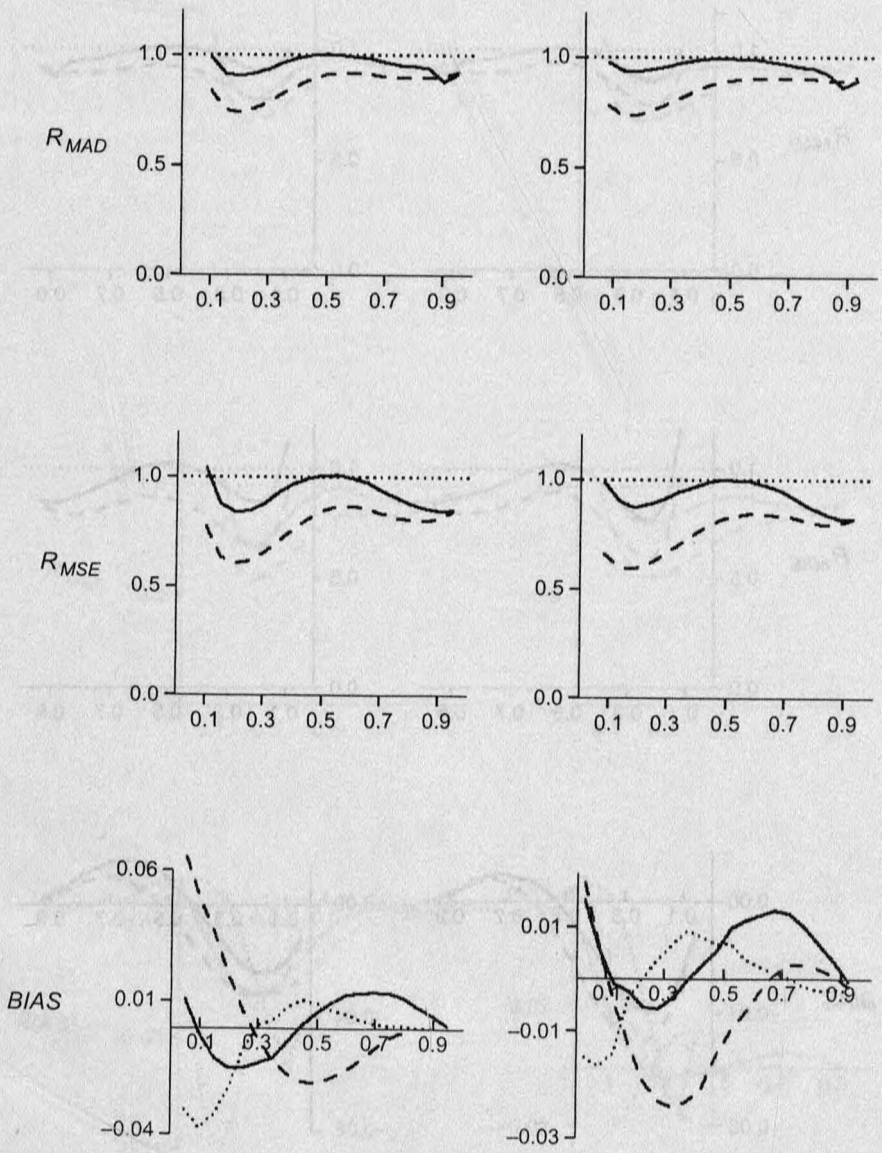

Fig. 1. Ratios of *MAD* and *MSE* for S_{pKM} (solid line) and S_{KL} (dashed line), *BIAS* for S_{pKM} (solid line), S_{KL} (dashed line) and S_{KM} (dotted line), $N = 10000$ repetitions, sample size $n = 10$

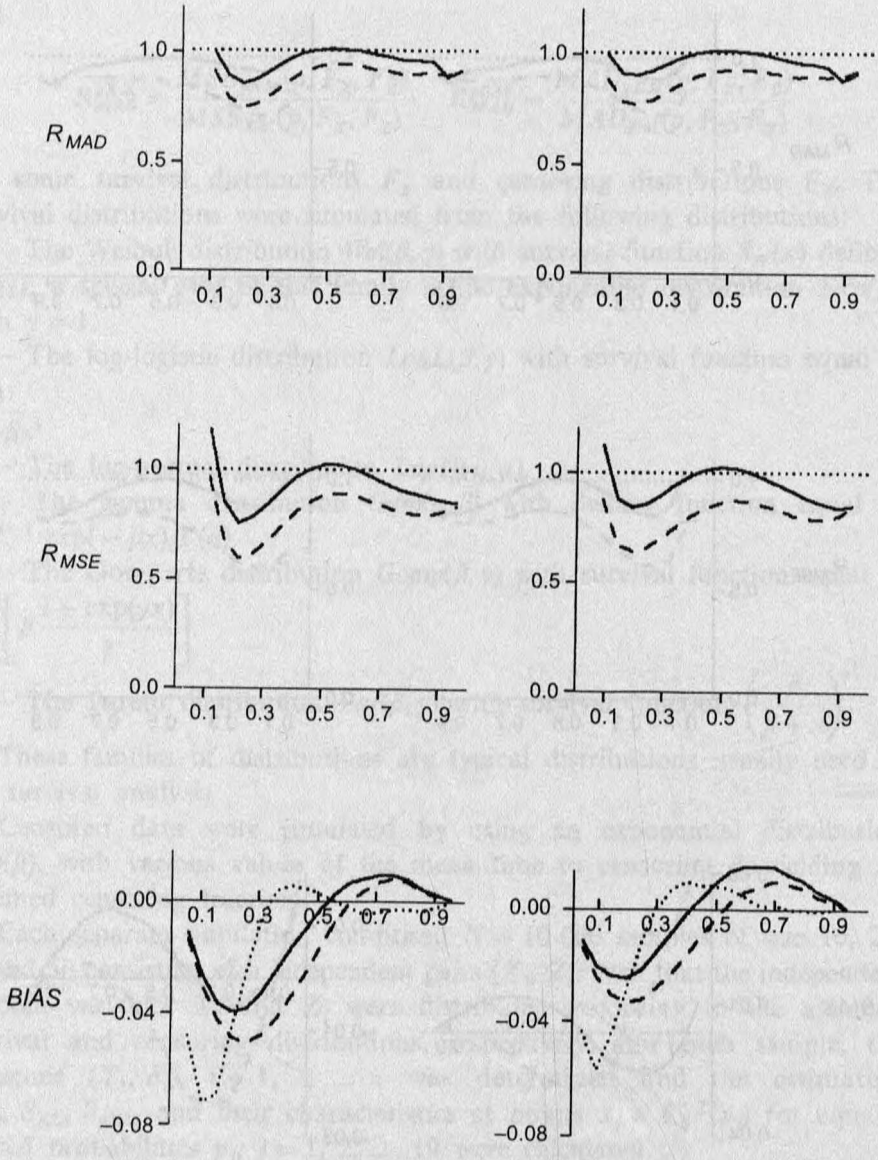

Fig. 2. Ratios of *MAD* and *MSE* for S_{pKM} (solid line) and S_{KL} (dashed line), *BIAS* for S_{pKM} (solid line), S_{KL} (dashed line) and S_{KM} (dotted line), $N = 10000$ repetitions, sample size $n = 10$

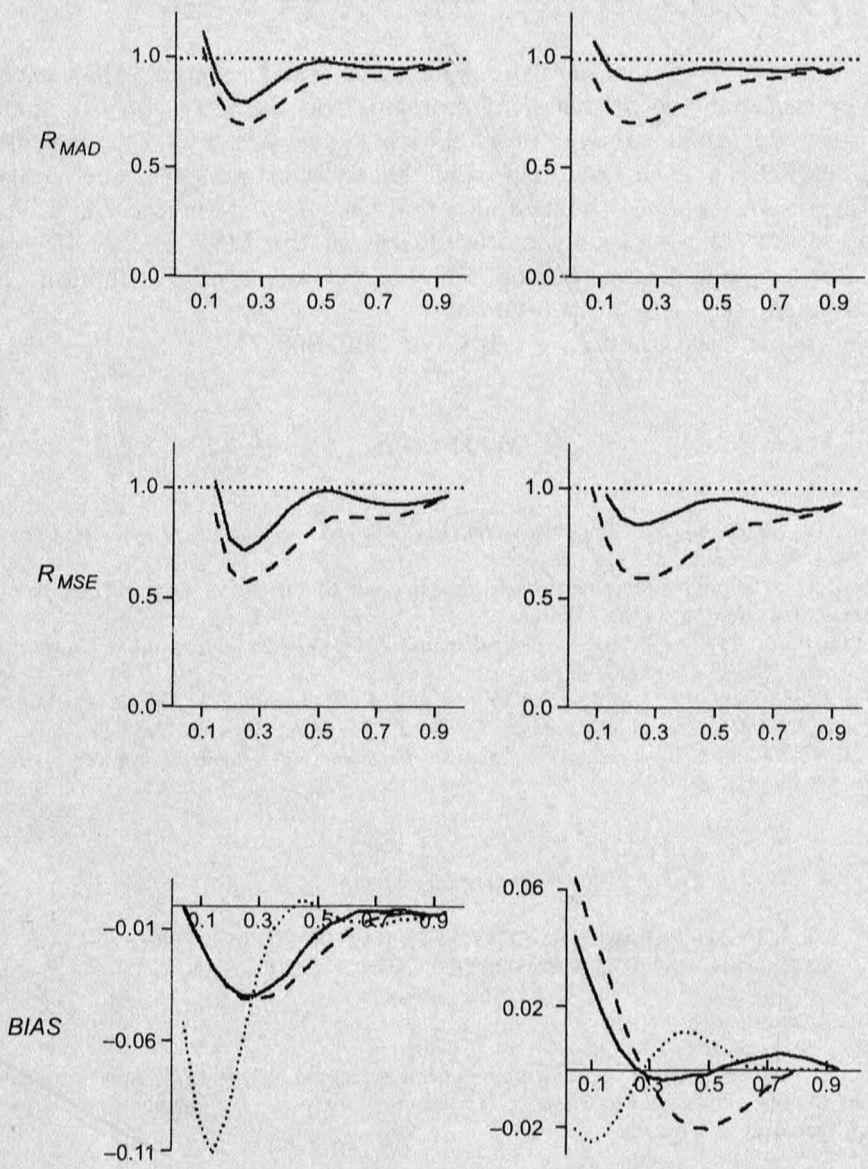

Fig. 3. Ratios of *MAD* and *MSE* for S_{pKM} (solid line) and S_{KL} (dashed line), *BIAS* for S_{pKM} (solid line), S_{KL} (dashed line) and S_{KM} (dotted line), $N = 10\,000$ repetitions, sample size $n = 10$

4. RESULTS AND CONCLUSIONS

It appears from the simulations that the *pKM* estimator has usually smaller bias than the original *KM* estimator and Klein *et al.'s* estimator, especially for small survival probabilities $1-p \leqslant 0.2$. For $1-p > 0.2$ the *BIAS* of Klein *et al.'* estimator and the *pKM* estimator is not regular.

The results obtained indicate also that the *pKM* estimator and Klein *et. al.'* estimator are usually more efficient in the *MSE* and *MAD* sense than the original *KM* estimator. For some families of distributions this improvement seems to be substantial.

The paper was granted by KBN No 5H02B00921.

REFERENCES

Efron, B. (1988), *Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve*. J. Amer. Stat. Assoc., **82**, 414–422.

Gill, R. D. (1980), *Censoring and Stochastic Integrals*, Mathematical Centre Tract No 124, Amsterdam: Mathematisch Centrum.

Kaplan E. L., Meier P. (1958), *Nonparametric Estimation From Incomplete Observations*, J. Amer. Statist. Assoc. **53**, 457–481.

Klein J. P., Lee S.-C., Moeschberger (1990), *A Partially Parametric Estimator of Survival in the Presence of Randomly Censored Data*, Biometrics **46**, 795–811.

Rossa A. (2002), *On the Estimation of Survival Function Under Random Censorship*, Comm. in Statistics **31**.

Agnieszka Rossa

CZĘŚCIOWO PARAMETRYCZNY ESTYMATOR FUNKCJI PRZEŻYCIA DLA DANYCH PRAWOSTRONNIE CENZUROWANYCH

(Streszczenie)

W pracy omówione są dwa estymatory funkcji przeżycia, będące modyfikacją estymatora Kaplana-Meiera. Podstawowe własności statystyczne estymatorów zostały porównane za pomocą metod symulacyjnych.