ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Agnieszka Rossa**

UNBIASED ESTIMATION OF SURVIVAL PROBABILITIES FOR CENSORED DATA WITH KNOWN CENSORING TIMES

Abstract

A class of unbiased estimators of survival probability $P(T_i > t)$ under random and independent censorship model is considered, where the potential survival times T_i are possibly unobserved, but the censoring times Z_i and $\min(T_i, Z_i)$ are observed and the sample size is random.

Key words: survival analysis, censored data, non-parametric estimation, Reduced-Sample Estimator.

I. INTRODUCTION

Suppose that we observe survival times of some individuals, i.e. times measured from entry into the study to an event of interest, usually called a failure. Very often, for ethical as well as practical reasons the observation terminates after a predetermined follow-up period. This causes that not all the failures can be observed. Individuals who failed during this period provide the true survival times, other individuals who did not fail yield the so-called censored times.

In many studies individuals arrive at random during the given period of observation, thus the length of time elapsed from their entry into the study to the end of the follow-up period is a random variable. Throughout the paper this variable will be called a monitoring or censoring variable. It is worth noting that such a variable is fully observed and independent of the survival time.

* Ph.D., Chair of Statistical Methods, University of Łódź.

The type of censorship, which assumes random, observable censoring times, independent of the survival times, appears often in applications (see Cox, Oakes, 1984, p. 5). Such a model of censoring was firstly considered by Kaplan and Meier (1958), who proposed an estimator of the survival probability called the Reduced-Sample Estimator (*RSE*). They investigated also a more general model, when the censoring times are not observed and proposed a product-limit estimator. This estimator is well-known as the Kaplan-Meier Estimator. Its properties have been widely studied in the literature (e.g. Breslow, Crowley, 1974; Peterson, 1977; Chen et al., 1982; Efron, 1988, Lumley, Heagerty, 2000). For instance, it was shown that it is negatively biased.

Gajek and Gather (1991) considered estimation of survival distribution F_θ being an element of some scale family of distributions. They derived the lower bound of the minimax value of the weighted mean squared error for estimating $\theta^s (\theta \in (0, \infty), s \neq 0)$ and showed that under Type I censoring this bound is independent of the sample size and equals to 1. They showed that Type II censoring does not lead to such anomalies. Mizera (1996) as well as Gajek and Mizera-Florczak (1998) considered also sequential estimation of θ^s, where θ was a parameter indexing a general family of distributions $\{P_\theta, \theta \in \Theta\}$. They derived lower bounds for the minimax value of a modified risk and applied this result to construct a minimax sequential estimator of a failure rate in an exponential model under Type II censoring. A conclusion which can be drawn from these papers is that under Type I censoring there does not exist an unbiased estimator of a survival probability. However Type II censoring or a sequential sampling scheme can be used instead.

In this paper we will consider some sequential estimators of survival probability assuming Type II censoring, in which sample size is a random variable distributed according to a negative binomial distribution. These estimators will be called Sequential-Type Reduces Sample Estimators.

The paper is organized as follows. Section II introduces basic notation. In section III the Reduced-Sample Estimator proposed by Kaplan and Meier is considered. In section IV a model of Type II censoring is considered and class of Sequential-Type Reduced Sample Estimators is proposed. Section V contains final remarks.

II. NOTATION

Let $T_1, T_2, ..., T_N$ denote iid survival times with a common continuous cumulative distribution function $F(t)$, $F(0) = 0$ and a survival function $\overline{F}(t) = 1 - F(t)$. Similarly, let $Z_1, Z_2, ..., Z_N$ denote iid censoring times with a common continuous cumulative distribution function $G(t)$, $G(0) = 0$. We will assume that the sample size N is random, the sequence $\{T_i\}$, is independent of the sequence $\{Z_i\}$, $i = 1, 2, ..., N$ and the variables $T_1, T_2, ..., T_N$ and $Z_1, Z_2, ..., Z_N$ are defined on the same probability space $(\Omega, \mathfrak{F}, P)$.

Although some of the survival times T_i may be unobserved, we will assume, that all the $\min(T_i, Z_i)$ for $i = 1, 2, ..., N$ are observed. We will also assume that all the censoring times Z_i for $i = 1, 2, ..., N$ are observed. In other words, we assume that the following random sample is observed

$$(X_1, Z_1), (X_2, Z_2), ..., (X_N, Z_N) \tag{1}$$

where

$$X_i = \min(T_i, Z_i), \quad i = 1, 2, ..., N. \tag{2}$$

Let us assume the following notation

$$\overline{F}(t) \equiv P(T_1 > t), \quad F(t) = 1 - \overline{F}(t), \tag{3}$$

$$\overline{G}(t) \equiv P(Z_1 > t), \quad G(t) = 1 - \overline{G}(t), \tag{4}$$

$$\overline{H}(t) \equiv P(X_1 > t) = P(T_1 > t, Z_1 > t) = \overline{F}(t)\overline{G}(t), \quad H(t) = 1 - \overline{H}(t). \tag{5}$$

Our goal is the pointwise estimation of a life-time distribution of T_i, represented by a cumulative distribution function F or a survival function $\overline{F}$, on the basis of the sequence (1).

If all the T_i's were observed with probability one and the sample size were fixed (i.e. $N = n = \text{const}$) then an estimator of $\overline{F}$ known as the empirical distribution function (*EDF*) could be applied

$$EDF(t) = \frac{1}{n} \sum_{i=1}^{n} 1(T_i > t),$$

where $1(\cdot)$ denotes the indicator variable.

It is well-known that $EDF(t)$ is an unbiased estimator of $\overline{F}(t)$ and its variance equals to $F(t)\overline{F}(t)/n$. Unfortunately, in the case of right-censoring it could not be evaluated, because not all the T_i's can be observed.

III. THE REDUCED-SAMPLE ESTIMATOR

The Reduced-Sample Estimator of the survival probability $\overline{F}(t)$ proposed by Kaplan and Meier (1958) was derived from the sample (1) for a fixed sample size.

Let us consider a random sample (1), where $N = n$ with probability one. Let $Z_{n:n} = \max\{Z_1, Z_2, ..., Z_n\}$. The Reduced-Sample Estimator $RSE(t)$ (Kaplan, Meier, 1958) is constructed as the ratio

$$RSE(t) = \frac{\sum_{i=1}^{n} 1(X_i > t)}{\sum_{j=1}^{n} 1(Z_j > t)}, \quad \text{for} \quad t \in [0, Z_{n:n}) \tag{6}$$

Let us notice that the estimator (6) is defined on the random time interval $[0, Z_{n:n})$ and it takes values from the interval [0,1]. It follows also that (6) is a truncated estimator, for it is not defined when $Z_{n:n} \leqslant t$. A disadvantage of using RSE as a function is that it is not monotonic and, as a consequence, it is not a distribution function. Let $N_1(t)$ and $N_2(t)$ be the following sums

$$N_1(t) = \sum_{j=1}^{n} 1(Z_j > t), \quad N_2(t) = \sum_{i=1}^{n} 1(X_i > t), \quad t \in R.$$

We will find the expected value and the variance of $RSE(t)$, i.e. $\mathrm{E}(RSE(t))$ and $\mathrm{Var}(RSE(t))$ with respect to the probability measure P. It is obvious that for $t = 0$ there is $RSE(0) = \overline{F}(0) = 1$ with probability one. Thus,

$$\mathrm{E}(RSE(0)) = \overline{F}(0) \quad \text{and} \quad \mathrm{Var}(RSE(0)) = 0.$$

Throughout the rest of this section we will assume that $t \in (0, Z_{n:n})$. It is easy to see, that the expected value of $RSE(t)$ equals to the conditional expectation of the ratio $N_2(t)/N_1(t)$ given $Z_{n:n} > t$. The variables $N_2(t)$, $N_1(t) - N_2(t)$ and $n - N_1(t)$ have the multinomial distribution with parameters n and $\mathbf{p}^T = [\overline{H}(t), \overline{G}(t) - \overline{H}(t), G(t)]$. Thus, the expected value of $RSE(t)$ equals

$$\mathrm{E}(RSE(t)) = \mathrm{E}\left(\frac{N_2(t)}{N_1(t)} | Z_{n:n} > t\right) =$$

$$= \frac{1}{\mathrm{P}(Z_{n:n} > t)} \sum_{n_1=1}^{n} \sum_{n_2=0}^{n_1} \frac{n_2}{n_1} \frac{n!}{(n_1 - n_2)!(n - n_1)! n_2!} \overline{H}^{n_2}(t)(\overline{G}(t) - \overline{H}(t))^{n_1 - n_2} G^{n - n_1}(t) \tag{7}$$

In order to transform the double sum on the right-hand side of (7) we will need the following equality

$$\sum_{n_2=0}^{n_1} \frac{n_2}{n_1}\binom{n_1}{n_2}\overline{F}^{n_2}(t)F^{n_1-n_2}(t) = \overline{F}(t),$$

which is valid for $\overline{F}(t) \in [0, 1)$ and any integer $n_1 \geqslant 1$ and flows from properties of any binomial distribution. We will also use the equality of the form

$$\overline{F}(t) = \frac{\overline{H}(t)}{\overline{G}(t)}, \quad \text{if} \quad \overline{G}(t) \neq 0,$$

what flows directly from (5). The probability $P(Z_{n:n} > t)$ in (7) can be expressed as follows

$$P(Z_{n:n} > t) = 1 - G^n(t).$$

From (7) and from the last three equalities the expected value of $RSE(t)$ takes the form

$$E(RSE(t)) = \frac{1}{P(Z_{n:n} > t)} \sum_{n_1=1}^{n} \binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) \sum_{n_2=0}^{n_1} \frac{n_2}{n_1}\binom{n_1}{n_2}\overline{F}^{n_2}(t)F^{n_1-n_2}(t) =$$

$$= \frac{\overline{F}(t)}{1-G^n(t)} \sum_{n_1=1}^{n} \binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) = \frac{\overline{F}(t)}{1-G^n(t)}(1 - G^n(t)) = \overline{F}(t).$$

The variance of $RSE(t)$ can be derived in a similar way

$$Var(RSE(t)) = E\left(\left(\frac{N_2(t)}{N_1(t)} - \overline{F}(t)\right)^2 | Z_{n:n} > t\right) =$$

$$= \frac{1}{1-G^n(t)} \sum_{n_1=1}^{n} \binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) \sum_{n_2=0}^{n} \left(\frac{n_2}{n_1} - \overline{F}(t)\right)^2 \binom{n_1}{n_2}\overline{F}^{n_2}(t)F^{n_1-n_2}(t).$$

From properties of any binomial distribution the following equality holds

$$\sum_{n_2=0}^{n_1} \left(\frac{n_2 - n_1\overline{F}(t)}{n_1}\right)^2 \binom{n_1}{n_2}\overline{F}^{n_2}(t)F^{n_1-n_2}(t) = \frac{\overline{F}(t)F(t)}{n_1},$$

where $n_1 \geqslant 1$. Hence,

$$\mathrm{Var}(RSE(t)) = \frac{\overline{F}(t)F(t)}{1-G^n(t)} \sum_{n_1=1}^{n} \frac{1}{n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t). \tag{8}$$

We will transform the sum on the right-hand side of (8), i.e.

$$\sum_{n_1=1}^{n} \frac{1}{n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) \tag{9}$$

using the following equality

$$\frac{1}{n_1} = \frac{1}{n} + \frac{n-n_1}{n(n-1)} + \frac{(n-n_1)(n-n_1-1)}{n(n-1)(n-2)} + \ldots + \frac{(n-n_1)(n-n_1-1)\cdot\ldots\cdot 1}{n(n-1)(n-2)\cdot\ldots\cdot n_1}.$$

Thus, the sum (9) can be expressed as follows

$$\begin{aligned}
\sum_{n_1=1}^{n} \frac{1}{n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) &= \sum_{n_1=1}^{n} \frac{1}{n}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \sum_{n_1=1}^{n-1} \frac{n-n_1}{n(n-1)}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \\
&+ \sum_{n_1=1}^{n-2} \frac{(n-n_1)(n-n_1-1)}{n(n-1)(n-2)}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \ldots + \\
&+ \sum_{n_1=1}^{1} \frac{(n-n_1)(n-n_1-1)\cdot\ldots\cdot 1}{n(n-1)(n-2)\cdot\ldots\cdot n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) = \\
&= \frac{1}{n}\sum_{n_1=1}^{n}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \frac{1}{n-1}\sum_{n_1=1}^{n-1}\binom{n-1}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \\
&+ \frac{1}{n-2}\sum_{n_1=1}^{n-2}\binom{n-2}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) + \ldots + \overline{G}(t)G^{n-1}(t) = \\
&= \frac{1}{n}(1-G^n(t)) + \frac{1}{n-1}(G(t)-G^n(t)) + \frac{1}{n-2}(G^2(t)-G^n(t)) + \\
&+ \ldots + G^{n-1}(t) - G^n(t) = \sum_{s=0}^{n-1}\frac{G^s(t)-G^n(t)}{n-s},
\end{aligned}$$

and we have

$$\sum_{n_1=1}^{n} \frac{1}{n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) = \sum_{s=0}^{n-1}\frac{G^s(t)-G^n(t)}{n-s}$$

or equivalently

$$\sum_{n_1=1}^{n} \frac{1}{n_1}\binom{n}{n_1}\overline{G}^{n_1}(t)G^{n-n_1}(t) = \sum_{s=1}^{n} \frac{G^{n-s}(t) - G^n(t)}{s}. \tag{10}$$

Finally, the variance of $RSE(t)$ takes the form

$$\mathrm{Var}(RSE(t)) = \frac{\overline{F}(t)F(t)}{1-G^n(t)} \sum_{s=1}^{n} \frac{G^{n-s}(t) - G^n(t)}{s}, \tag{11}$$

and satisfies the inequality

$$\mathrm{Var}(RSE(t)) = \frac{\overline{F}(t)F(t)}{n} + \frac{\overline{F}(t)F(t)}{1-G^n(t)} \sum_{s=1}^{n-1} \frac{G^{n-s}(t) - G^n(t)}{s} \geqslant \frac{\overline{F}(t)F(t)}{n}.$$

It follows that $RSE(t)$ is less effective than the empirical distribution function $EDF(t)$ usually used in the uncensored case.

IV. SEQUENTIAL-TYPE REDUCED-SAMPLE ESTIMATORS

In this section we will consider a class of the Reduced-Sample Estimators of $\overline{F}(t)$ defined for $t \in [0, t_0]$ under Type II censoring. For this reason the special type of experiments will be defined.

Let us assume that individuals enter the follow-up study at random time points. For an i-th individual we observe a pair of random variables (X_i, Z_i), where $X_i = \min(T_i, Z_i)$ and Z_i represents a random, observable censoring time, independent of T_i, where T_i represents an unobserved potential survival time.

Assume that the observation of individuals terminates when for k items ($k \geqslant 1$ is an integer chosen in advance) we obtain $Z_{i_j} > t_0$, $j = 1, 2, ..., k$, where $t_0 > 0$ is a fixed time point, such that $\overline{G}(t_0) > 0$. In the experiment the number N_k of individuals is a random variable distributed according to the negative binomial distribution with parameters k and $p = \overline{G}(t_0)$. Thus, the probability distribution function of N_k takes the form

$$\mathrm{P}(N_k = n) = \binom{n-1}{k-1}\overline{G}^k(t_0)G^{n-k}(t_0), \quad n = k, k+1, k+2, ... \tag{12}$$

It is worth noting, that it is usually possible to fix such a t_0 for which $\overline{G}(t_0) > 0$, even if the distribution function G is unknown.

Let us consider a class of estimators of the form

$$RSE_k(t) = \frac{\sum_{i=1}^{N_k} 1(X_i > t)}{\sum_{j=1}^{N_k} 1(Z_j > t)}, \quad \text{for} \quad t \in [0, t_0], \ k \in \mathbf{N}. \tag{13}$$

We will study the expected value and the variance of the estimators (13). Let $N_{1,k}(t)$, $N_{2,k}(t)$ denote the following sums defined for $t \in R$

$$N_{1,k}(t) = \sum_{j=1}^{N_k} 1(Z_j > t), \quad N_{2,k}(t) = \sum_{i=1}^{N_k} 1(X_i > t).$$

If $t = 0$ then $RSE_k(0) = \overline{F}(0) = 1$ with the probability 1 and

$$\mathrm{E}(RSE_k(0)) = \overline{F}(0) \quad \text{and} \quad \mathrm{Var}(RSE_k(0)) = 0.$$

Further we will assume that $t \in (0, t_0]$.

Given $N_k = n$ and for $t \in (0, t_0)$ the variable $N_{1,k}(t) - k$ has the binomial distribution with parameters $n - k$ and $p = (\overline{G}(t) - \overline{G}(t_0))/G(t_0)$. For $t = t_0$ the variable $N_{1,k}(t_0)$ takes the value k with probability one. Thus,

$$\mathrm{P}(N_{1,k}(t) = n_1 | N_k = n) = \begin{cases} \binom{n-k}{n_1-k} p^{n_1-k}(1-p)^{n-n_1}, & t \in (0, t_0), \ n_1 = k, k+1, \ldots, n \\ \begin{cases} 1, & n_1 = k, \\ 0, & n_1 \neq k, \end{cases} & t = t_0 \end{cases} \tag{14}$$

Given $N_{1,k}(t) = n_1$, $N_k = n$ and for $t \in (0, t_0]$ the variable $N_{2,k}(t)$ has the binomial distribution with parameters n_1 and $\overline{F}(t) = \overline{H}(t)/\overline{G}(t)$ (notice that for $t \in (0, t_0]$ there is $\overline{G}(t) \neq 0$). Thus,

$$\mathrm{P}(N_{2,k}(t) = n_2 | N_{1,k}(t) = n_1, \ N_k = n) = \binom{n_1}{n_2} \overline{F}^{n_2}(t) F^{n_1-n_2}(t), \tag{15}$$

where $n_2 = 0, 1, 2, \ldots, n_1$.

The joint distribution of $(N_k, N_{1,k}(t), N_{2,k}(t))$ can be obtained from (12), (14) and (15). The expectation and the variance of (13) will be derived from the joint distribution of $(N_k, N_{1,k}(t), N_{2,k}(t))$.

We have

$$\mathrm{E}(RSE_k(t)) = \sum_{n=k}^{\infty} \sum_{n_1=k}^{n} \sum_{n_2=0}^{n_1} \frac{n_2}{n_1} \mathrm{P}(N_k = n, N_{1,k}(t) = n_1, N_{2,k}(t) = n_2) =$$

$$= \sum_{n=k}^{\infty} \mathrm{P}(N_k = n) \sum_{n_1=k}^{n} \mathrm{P}(N_{1,k}(t) = n_1 | N_k = n) \sum_{n_2=0}^{n_1} \mathrm{P}(N_{2,k}(t) =$$

$$= n_2 | N_{1,k}(t) = n_1,\ N_k = n)$$

and it follows that

$$\mathrm{E}(RSE_k(t)) = \sum_{n=k}^{\infty} \mathrm{P}(N_k = n) \sum_{n_1=k}^{n} \frac{1}{n_1} \mathrm{P}(N_{1,k}(t) = n_1 | N_k = n) \cdot n_1 \cdot \overline{F}(t) = \overline{F}(t).$$

Thus the estimators defined in (13) are unbiased estimators of the survival probability $\overline{F}(t)$ for $t \in [0, t_0]$. The variance of $RSE_k(t)$ is equal to

$$\mathrm{Var}(RSE_k(t)) =$$

$$= \sum_{n=k}^{\infty} \sum_{n_1=k}^{n} \sum_{n_2=0}^{n_1} \left(\frac{n_2}{n_1} - \overline{F}(t) \right)^2 \mathrm{P}(N_k = n, N_{1,k}(t) = n_1, N_{2,k}(t) = n_2) =$$

$$= \sum_{n=k}^{\infty} \sum_{n_1=k}^{n} \mathrm{P}(N_k = n, N_{1,k}(t) = n_1) \times$$

$$\times \sum_{n_2=0}^{n_1} \frac{(n_2 - n_1 \overline{F}(t))^2}{n_1^2} \mathrm{P}(N_{2,k}(t) = n_2 | N_{1,k}(t) = n_1,\ N_k = n) =$$

$$= \overline{F}(t) F(t) \sum_{n=k}^{\infty} \sum_{n_1=k}^{n} \frac{1}{n_1} \mathrm{P}(N_k = n, N_{1,k}(t) = n_1).$$

For $t = t_0$ the double sum in the last expression reduces to $1/k$ and the variance is equal to

$$\mathrm{Var}(RSE_k(t_0)) = \frac{\overline{F}(t_0) F(t_0)}{k}.$$

For $t \in (0, t_0)$ this double sum can be expressed in the form

$$\sum_{n=k}^{\infty}\sum_{n_1=k}^{n}\frac{1}{n_1}\mathrm{P}(N_k=n, N_{1,k}(t)=n_1)=$$

$$=\sum_{n=k}^{\infty}\binom{n}{k}\overline{G}^k(t_0)G^{n-k}(t_0)\sum_{n_1=k}^{n}\frac{1}{n_1}\binom{n-k}{n_1-k}\times$$

$$\times\left(\frac{\overline{G}(t)-\overline{G}(t_0)}{G(t_0)}\right)^{n_1-k}\left(\frac{G(t)}{G(t_0)}\right)^{n-n_1}=$$

$$=\sum_{n_1=k}^{\infty}\frac{1}{n_1}\binom{n_1-1}{k-1}\left(\frac{\overline{G}(t_0)}{\overline{G}(t)}\right)^k\times$$

$$\times\left(1-\frac{\overline{G}(t_0)}{\overline{G}(t)}\right)^{n_1-k}\sum_{n=n_1}^{\infty}\binom{n-1}{n_1-1}\overline{G}^{n_1}(t)G^{n-n_1}(t)=$$

$$=\sum_{n_1=k}^{\infty}\frac{1}{n_1}\binom{n_1-1}{k-1}\left(\frac{\overline{G}(t_0)}{\overline{G}(t)}\right)^k\left(1-\frac{\overline{G}(t_0)}{\overline{G}(t)}\right)^{n_1-k}.$$

Let $p(t)$ denote the ratio $\overline{G}(t_0)/\overline{G}(t)$. Then we have

$$\sum_{n=k}^{\infty}\sum_{n_1=k}^{n}\frac{1}{n_1}\mathrm{P}(N_k=n, N_{1,k}(t)=n_1)=\sum_{n_1=k}^{\infty}\frac{1}{n_1}\binom{n_1-1}{k-1}p^k(t)(1-p(t))^{n_1-k}. \quad (16)$$

The expression on the right-hand side of (16) can be treated as the expectation of $1/N^*(t)$, where $N^*(t)$ is a random variable with the negative binomial distribution with parameters k and $p(t)$.

When $k=1$ and $t\in(0, t_0)$ this expression reduces to the form

$$\sum_{n_1=1}^{\infty}\frac{1}{n_1}\binom{n_1-1}{0}p(t)(1-p(t))^{n_1-1}=\frac{p(t)}{1-p(t)}\sum_{n_1=1}^{\infty}\frac{(1-p(t))^{n_1}}{n_1}. \quad (17)$$

In the general case, for $k\geqslant 2$ and $t\in(0, t_0)$ there is

$$\sum_{n_1=k}^{\infty}\frac{1}{n_1}\binom{n_1-1}{k-1}p^k(t)\overline{p}^{n_1-k}(t)=$$

$$=\sum_{s=1}^{k-1}(-1)^{s-1}p^s(t)\frac{(s-1)!(k-s-1)!}{(k-1)!}+$$

$$+(-1)^{k-1}\frac{p^k(t)}{\overline{p}^k(t)}\left(\sum_{n_1=1}^{\infty}\frac{\overline{p}^{n_1}(t)}{n_1}-\sum_{n_1=1}^{k-1}\frac{\overline{p}^{n_1}(t)}{n_1}\right) \quad (18)$$

where $\overline{p}(t)=1-p(t)$.

On the right-hand side of (17) and (18) there is a series $\sum_{n_1=1}^{\infty} \frac{\overline{p}^{n_1}(t)}{n_1}$. To find this sum it is worth noting that for any $x \in (0, 1)$ there is

$$\sum_{n=1}^{\infty} \frac{x^n}{n} = \sum_{n=1}^{\infty} \int_0^x v^{n-1} dv.$$

For $x \in (0, 1)$ the series $\sum_{n-1}^{\infty} v^{n-1}$ is uniformly convergent on $[0, x]$. Thus,

$$\sum_{n=1}^{\infty} \frac{x^n}{n} = \sum_{n=1}^{\infty} \int_0^x v^{n-1} dv = \int_0^x \sum_{n=1}^{\infty} v^{n-1} dv = \ln \frac{1}{1-x}. \tag{19}$$

Finally, from (16), (17) and (19) the variance of $RSE_1(t)$ for $t \in (0, t_0)$ is equal to

$$\mathrm{Var}(RSE_1(t)) = \overline{F}(t)F(t) \frac{p(t)}{\overline{p}(t)} \ln \frac{1}{p(t)}, \tag{20}$$

and in the general case, for $k \geqslant 2$, from (16), (18) and (19) , for $t \in (0, t_0)$ we have

$$\mathrm{Var}(RSE_k(t)) = \overline{F}(t)F(t)A_k(t), \tag{21}$$

where

$$A_k = \sum_{s=1}^{k-1} (-1)^{s-1} p^s(t) \frac{(s-1)!(k-s-1)!}{(k-1)!} + (-1)^{k-1} \frac{p^k(t)}{\overline{p}^k(t)} \left(\ln \frac{1}{p(t)} - \sum_{n_1=1}^{k-1} \frac{\overline{p}^{n_1}(t)}{n_1} \right), \tag{22}$$

and $p(t) = \overline{G}(t_0)/\overline{G}(t)$, $\overline{p}(t) = 1 - p(t)$.

Finally, the variance of (13) takes the form

$$\mathrm{Var}(RSE_k(t)) = \begin{cases} 0, & \text{for } t = 0, \\ \begin{cases} \overline{F}(t)F(t) \dfrac{p(t)}{\overline{p}(t)} \ln \dfrac{1}{p(t)}, & k = 1, \\ \overline{F}(t)F(t)A_k(t), & k \geqslant 2, \end{cases} & \text{for } t \in (0, t_0), \\ \overline{F}(t)F(t)/k, & \text{for } t = t_0, \end{cases}$$

where $A_k(t)$ is defined in (22).

V. DISCUSSION

In the paper we have considered the non-parametric, unbiased and pointwise estimation of a survival probability under fully observed, random and independent censoring. Some detailed results concerning the variances of the proposed estimators have been presented.

It is worth noting that in many applications it is more satisfying to regard the censoring variables as fixed numbers rather than as random variables. In such studies censoring variables can be treated in terms of values actually observed rather than in terms of their unknown distribution. The results obtained in this paper may be also extended to such a type of censoring.

REFERENCES

Breslow N., Crowley J. (1974), A large sample study of the life table and product limit estimates under random censorship, *Annals of Statistics*, **2**, 437–453,

Chen Y.Y., Hollander M., Langberg N.A. (1982), Small-sample results for the Kaplan-Meier Estimator, *Journal Am. Stat. Assoc.*, **77**, 141–144.

Cox D.R., Oakes D. (1984), *Analysis of Survival Data*, Chapman & Hall, London.

Efron B. (1988), Logistic regression, Survival analysis, and the Kaplan-Meier curve, *J. Amer. Statist. Assoc.*, **82**, 414–422.

Gajek L., Gather U. (1991), Estimating a scale parameter under censorship, *Statistics*, **22**, 529–549.

Gajek L., Mizera-Florczak B. (1998), Information inequalities for the minimax risk of sequential estimators (with applications), *Applicationes Mathematicae*, **25**, 85–100.

Kaplan E.L., Meier P. (1958), Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457–481.

Lumley T., Heagerty P. (2000), Graphical exploratory analysis of survival data, *J. Graph. Statist.*, **9**, 738–749.

Mizera B. (1996), Lower bounds on the minimax risk of sequential estimators, *Statistics*, **28**, 123–129.

Peterson A.V. (1977), Expressing the Kaplan-Meier Estimator as a function of empirical subsurvival functions, *J. Am. Stat. Assoc.*, **72**, 854–858.

Agnieszka Rossa

NIEOBCIĄŻONA ESTYMACJA PRAWDOPODOBIEŃSTW PRZEŻYCIA W MODELU Z OBSERWOWALNYMI CZASAMI CENZUROWANIA

Streszczenie

W pracy zaproponowana została klasa nieobciążonych estymatorów prawdopodobieństwa przeżycia $P(T_i > t)$ w modelu losowego, niezależnego i obserwowalnego cenzurowania, w którym potencjalne czasy życia T_i są nieobserwowalnymi zmiennymi losowymi, natomiast zmienne cenzurujące Z_i oraz zmienne $\min(T_i, Z_i)$ są obserwowalne.