STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 387—395

ESTIMATION OF LIFE-TABLES UNDER RIGHT-CENSORING

Agnieszka Rossa[1]

ABSTRACT

The paper deals with a class of non-parametric estimators of conditional probabilities of failure prior *x+y* given survival to *x* under the random and observable right-censorship model. The proposed estimators are based on a specific sequential sampling scheme. Application of the estimators in life-table analysis is presented.

Key words: Life-table analysis; non-parametric estimation; right-censored data; sequential sampling.

1. Introduction

Let T be a non-negative random variable representing a duration time between two well-defined events, i.e. an initial and a final event (the second event is usually called a failure). Let T be defined on a probability space (Ω, $\mathcal{A}$, **P**). The survival function of T is then defined as

$$\overline{F}(x) = \mathbf{P}(T > x), \quad x \in \mathbf{R},$$

with $\overline{F}(x) = 1$ for $x \le 0$.

Consider a probability that the failure occurs in a time interval $(x,x+y]$ given $T>x$. Using the actuarial notation, such a probability is usually denoted by ${}_y q_x$. Thus, we have

$$ {}_y q_x = \mathbf{P}(T \le x + y \mid T > x). \tag{1}$$

It can be also expressed in terms of $\overline{F}$ as

[1] For correspondence: Institute of Statistics and Demography, University of Łódź, ul. Rewolucji 1905, 41,90-214 Łódź, Poland; e-mail: agrossa@uni.lodz.pl.

$$
{}_y q_x = \frac{\overline{F}(x) - \overline{F}(x+y)}{\overline{F}(x)} = 1 - \frac{\overline{F}(x+y)}{\overline{F}(x)}. \tag{2}
$$

Denoting

$$
{}_y p_x = \frac{\overline{F}(x+y)}{\overline{F}(x)} \tag{3}
$$

we receive from (2) the obvious formula ${}_y q_x = 1 - {}_y p_x$.

Thus, ${}_y q_x$ is a probability of failure prior $x+y$ given survival to x, and ${}_y p_x$ is a probability of surviving beyond $x+y$ given survival to x (for $y \geq 0$).

In many applications, especially in demographic and actuarial studies, a table of numerical values of ${}_y q_x$ and ${}_y p_x$ for a certain selected values of x and y are considered, most commonly the integers. Typically, a complete table shows values of ${}_y q_x$ and ${}_y p_x$ for all integer values of $x = 0,1,\ldots,\omega - 1$, and for y=1, where ω stands for a maximal possible lifetime. In such cases probabilities ${}_1 q_x$ and ${}_1 p_x$, denoted hereafter by q_x, p_x, are organized in a table of numbers (see Table 1).

Table 1. Tabular survival model

x	q_x	p_x
0	q_0	p_0
1	q_1	p_1
$\vdots$	$\vdots$	$\vdots$
$\omega-1$	$q_{\omega-1}$	$p_{\omega-1}$

From probabilities ${}_y q_x$ and ${}_y p_x$ various life-table characteristics are derived, such as numbers of lives and failures in a cohort group, expected lifetimes, times of exposure etc.

2. Two censorship models

In many real-life situations duration times $T_1, T_2, \ldots, T_n$ for individuals cannot be fully observed. There can be one or more random events (other than the failure), occurring of which terminates the observation of some individuals. In such cases it is said that the observations are right-censored. For instance, very often it is impossible to keep a study open until all the individuals in the sample have experienced the failure. Instead, a period of observation (follow-up) is chosen, and individuals are observed until the end of that period. If an *i*-th individual fails by the time of the analysis, then it yields a true duration time T_i, otherwise a censoring time Z_i, say, is observed, such that $Z_i < T_i$. If individuals enter the study at random times then the Z_i's are random variables, independent of the T_i's.

Under *the random censorship model* [see Efron (1967)] it is usually assumed, that the T_i's are iid, and independent of the Z_i's. It is also assumed that T_i is observed whenever $T_i \le Z_i$, and Z_i is observed whenever $Z_i < T_i$. In other words, a pair (X_i, δ_i) is observed, where

$$X_i = \min(T_i, Z_i), \tag{4}$$

$$\delta_i = \mathbf{1}(T_i \le Z_i), \tag{5}$$

and $\mathbf{1}(\cdot)$ denotes an indicator function, indicating whether X_i is right-censored ($\delta_i = 0$) or not ($\delta_i = 1$).

However, if censoring mechanism is only due to the termination of the observation period, then each Z_i represents the time elapsed from entering an *i*-th subject into the study to the end of the observation period. In such cases the Z_i 's are fully observed and one observes (X_i, Z_i) for *i*=1,2,…. We will refer to this special type of the model (4)-(5) as *the random and observable censorship model*.

Throughout the rest of the paper we will assume that the duration times T_i, as well as the censoring times Z_i are mutually independent, continuous random variables in the probability space (Ω, $\mathcal{A}$, **P**) and that the T_i's have a common cumulative distribution function (cdf) *F*, whereas the Z_i's have a common cdf *G*. These assumptions imply that variables X_i defined in (4) have a common cdf expressed as

$$H = 1 - \overline{F} \cdot \overline{G},$$

where $\overline{F} = 1 - F$, $\overline{G} = 1 - G$.

3. Estimation of conditional probabilities under the random censorship model

Let us consider a random censored sample of a fixed size n

$$(X_1, \delta_1), (X_2, \delta_2), \ldots, (X_n, \delta_n). \tag{6}$$

Hereafter, observations $(X_i, \delta_i = 1)$ will be called failures, whereas $(X_i, \delta_i = 0)$ will be called losses. Let

$$0 = x_0 < x_1 < \ldots < x_J < x_{J+1} = y_0 \tag{7}$$

be a partition of the range $[0, y_0]$ of observation into J+1 intervals $I_j = (x_j, x_{j+1}]$ for $j = 0,1,\ldots,J$. We will consider the following statistics, defined for each $j = 0,1,\ldots,J$.

$$D_j = \sum_{i=1}^{n} \mathbf{1}(X_i \in I_j, \delta_i = 1), \tag{8}$$

$$L_j = \sum_{i=1}^{n} \mathbf{1}(X_i \in I_j, \delta_i = 0), \tag{9}$$

$$M_j = \sum_{i=1}^{n} \mathbf{1}(X_i > x_j), \tag{10}$$

where D_j and L_j represent numbers of failures and losses in an interval $(x_j, x_{j+1}]$, respectively, whereas M_j represents number of items at risk at the beginning of the interval (i.e. numbers of individuals surviving beyond x_j). Note that $M_j = 0$ implies $D_j = 0$ and $L_j = 0$.

Table 2. Life-Table Statistics Under Random Censorship Model

Intervals $I_j = (x_j, x_{j+1}]$	No. of survivors M_j	No. of failures D_j	No. of losses L_j
$(0, x_1]$	M_0	D_0	L_0
$(x_1, x_2]$	M_1	D_1	L_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(x_J, x_{J+1}]$	M_J	D_J	L_J

The well-known estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$, where $y_j = x_{j+1} - x_j$, are usually defined by means of the statistics (8)-(10). For instance, the Standard Life-Table Estimators, encountered in the literature [e.g. Berkson and Gage (1950), Gehan (1965), Breslow and Crowley (1974), Daya (2005)], are defined as

$$ {}_{y_j}\hat{q}_{x_j} = \frac{D_j}{M_j - \frac{1}{2}L_j} \quad \text{and} \quad {}_{y_j}\hat{p}_{x_j} = 1 - {}_{y_j}\hat{q}_{x_j} \tag{11} $$

Both ${}_{y_j}\hat{q}_{x_j}$ and ${}_{y_j}\hat{p}_{x_j}$ are not defined if $M_j = 0$. In such cases it is usually assumed that ${}_{y_j}\hat{q}_{x_j} = 1$, ${}_{y_j}\hat{p}_{x_j} = 0$.

It is also well-known that the Standard Life-Table Estimators ${}_{y_j}\hat{q}_{x_j}$ are negatively biased, what indicates that ${}_{y_j}\hat{p}_{x_j}$ are biased positively. In general, both estimators are asymptotically biased and not consistent [see Breslow and Crowley (1974)].

Taking into account all these disadvantages mentioned above we will propose a new class of estimators of probabilities (2) and (3), which are unbiased and consistent. The proposed estimators are derived under the model of random and observable censorship and are based on a special type of sequential sampling.

4. Estimation of conditional probabilities under random and observable censoring

Consider the sequence of observations $(X_1, Z_1), (X_2, Z_2), \ldots$. Let *k* be a fixed integer ($k \geq 2$) and y_0 be a fixed real value such that $0 < y_0 < \sup\{y : H(y) < 1\}$, where $H(\cdot)$ denotes a common cdf of the X_i's.

Suppose that individuals arrive at random into the study and the observation period terminates if for *k* individuals one observes $X_{i_j} > y_0$, $j = 1,2,\ldots,k.$ Such a sequential sampling leads to a sample

$$ (X_1, Z_1), (X_2, Z_2), \ldots, (X_{N_k}, Z_{N_k}) \tag{12} $$

where the sample size N_k is a random variable distributed according to the negative binomial distribution with parameters *k* and $p = 1 - H(y_0)$.

Let us consider the partition (7) of the time interval $[0, y_0]$, and define the following statistics

$$R_{k,j} = \sum_{i=1}^{N_k} \mathbf{1}(X_i > x_{j-1}, Z_i > x_j), \quad j = 1,2,\ldots,J,J+1, \tag{13}$$

$$M_{k,j} = \sum_{i=1}^{N_k} \mathbf{1}(X_i > x_j), \quad j = 0,1,\ldots,J,J+1. \tag{14}$$

In the special case there is $M_{k,0} = N_k$ and $M_{k.J+1} = k$.

Definition.

Under the model of random and observable censorship the life-table estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$ ($j = 0,1,\ldots,J$) take the form

$$ {}_{y_j}\widetilde{q}_{k,x_j} = \frac{R_{k,j+1} - M_{k,j+1}}{R_{k,j+1} - 1} \tag{15}$$

and

$$ {}_{y_j}\widetilde{p}_{k,x_j} = 1 - {}_{y_j}\widetilde{q}_{k,x_j} \,. \tag{16}$$

where $y_j = x_{j+1} - x_j$.

Proposition.

The estimators (15) and (16) are unbiased and consistent estimators of the conditional probabilities ${}_{y_j}q_{x_j}$ and ${}_{y_j}p_{x_j}$. Their variances satisfy the following equivalence

$$\mathbf{V}\left({}_{y_j}\widetilde{q}_{k,x_j} \right) = \mathbf{V}\left({}_{y_j}\widetilde{p}_{k,x_j} \right) = {}_{y_j}q_{x_j}\,\mathbf{E}\left(\frac{1}{R_{k,j+1}} \right) - {}_{y_j}q^2_{x_j}\,\mathbf{E}\left(\frac{1}{R_{k,j+1} + 1`} \right) \tag{17}$$

where the expectations $\mathbf{E}\left(\frac{1}{R_{k,j+1}} \right)$ and $\mathbf{E}\left(\frac{1}{R_{k,j+1} + 1} \right)$ can be expressed as

$$\mathbf{E}\left(\frac{1}{R_{k,j+1}} \right) = \left(\frac{p_j}{q_j} \right)^k \int_0^{q_j} u^{k-1} (1-u)^{-k}\, du \,,$$

$$\mathbf{E}\left(\frac{1}{R_{k,j+1} + 1} \right) = \left(\frac{p_j}{q_j} \right)^k \frac{1}{q_j} \int_0^{q_j} u^{k} (1-u)^{-k}\, du \,,$$

and

$$ {}_{y_j}p_{x_j} = \frac{\overline{F}(x_{j+1})}{\overline{F}(x_j)}, \qquad {}_{y_j}q_{x_j} = 1 - {}_{y_j}p_{x_j}, $$

$$ p_j = \frac{\overline{H}(y_0)}{\overline{F}(x_j)\overline{G}(x_{j+1})}, \quad q_j = 1 - p_j. $$

The proof of the proposition was given by Rossa (2005), pp. 80-82.

It follows from (17) that variances of ${}_{y_j}\widetilde{q}_{k,x_j}$ and ${}_{y_j}\widetilde{p}_{k,x_j}$ can be estimated by means of the following expression

$$ \hat{\mathbf{V}}\left({}_{y_j}\widetilde{q}_{x_j} \right) = \hat{\mathbf{V}}\left({}_{y_j}\widetilde{p}_{x_j} \right) = {}_{y_j}\widetilde{q}_{x_j} \frac{1}{R_{k,j+1}} - {}_{y_j}\widetilde{q}^{\,2}_{x_j} \frac{1}{R_{k,j+1}+1}. \tag{18} $$

5. A numerical example

To illustrate in detail the sampling scheme and the estimators proposed, a hypothetical study will be considered. Assume that the subject of observation is the time T_i elapsed from the issuance of a life-insurance policy up to the death of an *i*-th insured person (in years), and let Z_i denote the time elapsed from the issuance of his/her policy up to the termination of the follow-up study.

Due to right-censoring variables $T_1, T_2, \ldots$ are possibly unobserved. However, variables $X_i = \min(T_i, Z_i)$, as well as censoring variables Z_i for $i = 1,2,\ldots$ are fully observed.

We will assume that persons arrive at random into the study and the follow-up period terminates when for *k*=15 insured persons we observe $X_i > y_0$. Suppose here that $y_0 = 5$ (years).

Consider a partition of the range $[0, y_0]$ into subintervals $I_j = (x, x+1]$, where $x = 0,1,\ldots, y_0 - 1$. Table 3 (columns 2 and 3) contains exemplary values of statistics $R_{k,x+1}, M_{k,x+1}, x = 0,1,\ldots,4$ which can be observed under such a sampling scheme. The next three columns present estimates of the conditional probabilities (2) and (3) derived from the formulae (15) and (16) for $y_j = 1$, as well as estimates of their variances $\mathbf{V}\left(\widetilde{q}_{k,x}\right)$ obtained from the formula (18).

Table 3. Estimates $\widetilde{q}_{15,x}$, $\widetilde{p}_{15,x}$ of probabilities q_x, p_x and $\mathbf{V}\left(\widetilde{q}_{15,x}\right)$

x	$R_{15,x+1}$	$M_{15,x+1}$	$\tilde{q}_{15,x}$	$\tilde{p}_{15,x}$	$\hat{\mathbf{V}}(\tilde{q}_{15,x})$
0	63	60	0,0484	0,9516	0,000805
1	48	45	0,0638	0,9362	0,001412
2	37	34	0,0833	0,9167	0,002434
3	28	24	0,1481	0,8519	0,006046
4	19	15	0,2222	0,7778	0,014163

6. Discussion

In the paper two classes of non-parametric estimators of conditional probabilities ${}_y q_x = \mathbf{P}(T \le x + y \mid T > x)$ and ${}_y p_x = \mathbf{P}(T > x + y \mid T > x)$ are proposed, derived under the so-called random and observable censorship model. The proposed estimators are unbiased and consistent, as opposed to the well-known Standard Life-Table Estimators.

Both classes of estimators are based on a special sequential sampling scheme. In this scheme an integer $k \ge 2$ and a positive value y_0 such that $y_0 < \sup\{y : H(y) < 1\}$ have to be fixed in advance, where H is a cdf of $X = \min(T, Z)$. Note that usually it is not difficult to choose a proper value of y_0, even if the function H is unknown. It is sufficient to know the maximal possible values t and z, say, of the duration and censoring times, respectively. Then, for any $y_0 \in (0, \min(t, z))$ and the condition $y_0 < \sup\{y : H(y) < 1\}$ is satisfied.

REFERENCES

BERKSON, J. & GAGE, R. 1950. Calculation of Survival Rates for Cancer. *Proc. of the Staff Meetings of the Mayo Clinic*, 25, pp. 270–286.

BRESLOW, N. E. & CROWLEY, J. J. 1974. Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. *Annals of Statistics,* 2, pp. 437–454.

DAYA, S. 2005. Life Table (Survival) Analysis to Generate Pregnancy Rates in Assisted Reproduction: Are We Overestimation Our Success Rates? *Human Reproduction*, 20, 1135–1143.

EFRON, B. 1967. The Two-Sample Problem with Censored Data. *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4, pp. 831–853, University of California Press, Berkeley.

GEHAN, E. A. 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Single-Censored Samples. *Biometrika*, 52, pp. 203–234.

ROSSA, A. 2005. Estimation of Survival Distributions under Right-Censoring and Applications, University of Łódź (in Polish).