Intercepted Sampling and Residual Lifetimes

Maki Momma

— Contents —

Abstract

This article studies the problem of estimating a lifetime distribution based on data obtained by intercepted sampling. Of particular interest is the case where complete lifetime is not observable. Properties of nonparametric estimators based on observations of the residual lifetimes are studied, and a bias correction method using the empirical histogram is proposed.

1. INTRODUCTION

Survival analysis is used in areas such as biology, medicine, engineering, epidemiology and economics, to name just a few. When studying survival times or duration times, data are often taken from items that are "alive" or "in operation" at a particular time point. This sampling method, referred to as intercepted sampling, is in contrast to taking random samples over a period of time, and is employed for its convenience and cost effectiveness. The method is particularly useful in cases where a controlled study of the items is not feasible. An important feature of the method is that data obtained by intercepted sampling contain upward bias, since items with longer survival times have higher probability of being "in operation" at a particular time point, i.e. higher possibility of being a member of the intercepted population. It is therefore necessary to take proper care of this bias when analyzing data obtained by intercepted sampling.

There are many cases involving intercepted sampling, where, in addition to the problem of observation bias, data on survival or duration times are either unavailable or only partially available. When patients are screened for a certain disease, for example, it is only possible to observe the current condition of the patient, not the initiation time of the disease. As another example, consider the case where people on-line are sampled for the purpose of estimating the time spent on the internet. It is more than likely that at the time of sampling, each individual does not have an accurate record or memory of the length of time he/she has been on the internet. As such, the total time of their stay on-line cannot be observed. In such instances, it is still possible to observe the time spent on-line after the sampling time point.

The objective of this article is to discuss the estimation of lifetime distribution from data on the residual lifetimes (duration times) of the intercepted sample. The article is organized as follows. Section 2 introduces basic notations and results. Estimation is discussed in Section 3, and simulation studies are performed to assess the properties of nonparametric estimators in finite samples. Other problems pertaining to the intercepted sampling method are discussed briefly in Section 4. Section 5 concludes.

2. NOTATIONS AND PRELIMINARIES

It is assumed throughout the paper that data are obtained by intercepted sampling. The distribution of survival or duration time will be called the lifetime distribution. In particular, the lifetime distribution of the whole population will be called the population lifetime distribution. The survival time (current lifetime) of an item at the sampling time point is referred to as its age.

The notion of intercepted sampling was formally studied by Vardi (1988). He argued heuristically that when items are born randomly with individually and identically distributed lifetimes, the lifetime of a typical item in the intercepted population follows the length-biased distribution of the corresponding lifetime distribution of the whole population. Here, a length-biased density f^L associated with an arbitrary probability density f with domain $[0, \infty]$ takes the form $f^L(x)=\frac{xf(x)}{\int_0^\infty uf(u)du}$. Many authors have since argued heuristically that intercepted sampling results in length-biased lifetime distribution, and estimation problems based on observations from length-biased densities have been studied extensively, including cases where parts of the observations are censored. See Asgharsian, M'Lan, and Wolfson (2002) for a list of references. Some have claimed that "choosing a sampling time point randomly" justifies the use of length-biased densities. Random sampling of a time point, however, is not possible, since we clearly cannot go back in time. Moreover, if the sampling time point need to be "randomly selected" from within a long interval, intercepted sampling will lose its appeal as a convenient sampling scheme. It is therefore necessary to study the population existing at an arbitrary fixed time point.

Momma (1991) gave a rigorous proof that an item selected randomly from the intercepted population does indeed follow the length biased distribution, under the assumption that the population birth process in stationary Poisson and that the population at a given time point is finite. The proof is valid for an arbitrary fixed time point, provided the birth process of the whole population has started in the remote past. In addition, the distribution of age, along with the joint distribution of age and lifetime of an item randomly chosen from the intercepted population were derived.

Let X^* denote the age of an item in the intercepted population and Z^*, its lifetime. Further, let $f_Z(z)$ denote the lifetime density, and $F_Z(z)$ the corresponding lifetime distribution of the whole population (variables with * represent variables in the intercepted population, whereas variables without * represent variables in the whole population). The joint density of age X^* and lifetime Z^* of an item selected randomly from the intercepted population is given by

$$f_{X^*,Z^*}(x,z)=\frac{f_Z(z)}{\int_0^\infty uf_Z(u)du} \qquad 0\le x\le z, \tag{1}$$

provided the population at the sampling time point is finite, and the birth process of the whole population is stationary Poisson. The marginal densities of X^* and Z^* are given by

$$f_{Z^*}(z)=\frac{zf_Z(z)}{\int_0^\infty uf_Z(u)du} \tag{2}$$

and

$$f_{X^*}(x)=\frac{1-F_Z(x)}{\int_0^\infty uf_Z(u)du}, \tag{3}$$

while the conditional densities are given by

$$f_{X^*|Z^*}(x|z)=\frac{1}{z} \qquad 0\le x\le z \tag{4}$$

and

$$f_{Z^*|X^*}(z|x)=\frac{f_Z(z)}{1-F_Z(x)} \qquad 0\le x\le z, \tag{5}$$

respectively. The marginal density of age X^* given by (3) is known in renewal theory as the recurrence time density. The above relations serve as a basis when estimating the population lifetime distribution $F_Z(z)$ from data obtained by the intercepted sampling method. Note that the value of X^* does not enter into the joint distribution of X^* and Z^* in an explicit form, and that given the value z of Z^*, X^* is uniformly distributed.

3. ESTIMATION BASED ON RESIDUAL LIFETIMES

3.1 Distribution of Residual Lifetimes

When collecting data from the intercepted population, it is not always possible to observe items' ages and/or complete lifetimes. Consider, for example, the case where people in a shopping mall are asked how long they have been shopping. The answers will at best be approximate, not precise. As a result, accurate values of the shopping times (lifetimes) will not be observable.

A possible approach to estimate the length of times people spend shopping from this type of data, is to recognize that data collected by surveying on-site contain observation errors, and proceed with the errors-in-variables method. Instead of observing the value z of the lifetime Z^* of an item in the intercepted population, the value of $y=z+\varepsilon$ is recorded, where ε is the unobservable error component. Assuming that ε is independent and identically distributed with mean 0, the density of Z^* is obtained by integrating out the ε,

$$f_{Z^*}(z)=\int f_{Z^*}(z|\varepsilon)f_\varepsilon(\varepsilon)d\varepsilon. \tag{6}$$

An alternative approach is to estimate the lifetime distribution based on observable quantities. In the shopping mall example, it is relatively easy to observe the times spent shopping since the sampling time point, simply by asking the participants in the study to report the time they exit the mall. These observations could, in turn, be used to estimate the lifetime distribution of the whole population. This is the approach taken in the article.

Let Y^* denote the remaining time (residual lifetime) of an item after interception. In order to make use of the values of Y^* to estimate the population lifetime distribution, it is necessary to derive the relation between the distribution $F_{Y^*}(y)$ of Y^* and the population lifetime distribution $F_Z(z)$. But, by symmetry, $F_{Y^*}(y)$ takes exactly the same form as the density of age (current lifetime) of an item at interception. To see this, note that the distribution of $Y^*=Z^*-X^*$ is obtained by first conditioning on X^* and using the relation between X^* and Z^* given in (5). After some manipulation, it is seen that

$$f_{Y^*}(y|x)=\frac{f_Z(x+y)}{1-F_Z(x)}. \tag{7}$$

Inserting (7) and (3) into the relation

$$f_{Y^*}(y)=\int f_{Y^*}(y|x)f_{X^*}(x)dx, \tag{8}$$

it is seen that the density of the residual lifetime of an item in the intercepted population $f_{Y^*}(y)$ is given by

$$f_{Y^*}(y)=\frac{1-F_Z(y)}{\int_0^\infty uf_Z(u)du} \qquad y\geq 0. \tag{9}$$

A notable feature of the residual lifetime distribution given in (9) is that the density is monotone decreasing, regardless of the form of the population lifetime distribution. Moments of this distribution in relation to the population lifetime distribution $F_Z(z)$ is seen to be

$E\left(Y^{*k}\right)=\dfrac{\mu_{k+1}}{(k+1)\mu_1}$, where μ_k denotes the kth moment of the lifetime distribution of the whole population. Since $E\left(Z^{*k}\right)=\dfrac{\mu_{k+1}}{\mu_1}$ for the length-biased lifetime density, $E\left(Y^{*k}\right)=\dfrac{1}{k+1}E\left(Z^{*k}\right)$. In particular, $E\left(Y^{*}\right)=\dfrac{1}{2}E\left(Z^{*}\right)$. It can be seen that the variance of the length-biased lifetime Z^* exceeds the residual lifetime Y^* if and only if $8\mu_1\mu_3>9\mu_2^2$.

Some examples of the density of Y^* corresponding to the distributional forms of Z are given below.

Example 1. Weibul distribution

When the population lifetime follows the Weibul distribution, so that

$$f_Z(z)=1-e^{-\lambda z^{\alpha}},$$

the corresponding residual lifetime density becomes

$$f_{Y^*}(y)=\frac{e^{-\lambda y^{\alpha}}}{\lambda^{-\frac{1}{\alpha}}\Gamma\left(\frac{1}{\alpha}+1\right)}.$$

Note that when α=1, the distribution reduces to the exponential distribution, in which case the density of Y^* is identical to that of Z, due to the memoryless property of the exponential distribution.

Example 2. Gamma distribution

When the lifetime follows a Gamma distribution with density

$$f_Z(z)=\frac{\lambda^r z^{r-1}e^{-\lambda z}}{(r-1)!},$$

the corresponding density of Y^* is seen to be

$$f_{Y^*}(y)=\frac{\lambda}{r}e^{-\lambda y}\sum_{k=1}^{r-1}\frac{(\lambda y)^{r-k-1}}{(r-k-1)!}.$$

With proper interpretation, one can also think of discrete examples, such as number of visits to doctors.

Example 3. Poisson distribution

When Z follows the Poisson distribution,

$$f_Z(z)=\frac{\lambda^z e^{-\lambda}}{z!},$$

the corresponding distribution of Y^* is seen to be

$$f_{Y^*}(y)=e^{-\lambda}\sum_{k=y+1}^{\infty}\frac{\lambda^{k-1}}{k!}$$

Example 4. Negative Binomial distribution

For the negative binomial distribution,

$$f_Z(z)=\frac{\Gamma(\alpha+z)}{\Gamma(\alpha)\Gamma(z+1)}\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(\frac{\lambda}{\alpha+\lambda}\right)^{z},$$

the corresponding distribution of Y^* is seen to be

$$f_{Y^*}(y)=\frac{1}{\lambda}\sum_{k=y+1}^{\infty}\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(\frac{\lambda}{\alpha+\lambda}\right)^{k}.$$

3.2 Estimation

Momma (1996) studied the problem of estimating a continuous lifetime distribution based on observations of items' ages at interception. Since the distributional form of the residual lifetime is identical to that of age, the arguments hold valid for this case also. The difference is that age is observed immediately upon interception, so that estimates of lifetime distribution are obtained without delay, whereas a longitudinal study is necessary to collect data on residual lifetimes.

Likelihood function based on n observations of the residual lifetimes $(y_1, \ldots, y_n)$ takes the following form:

$$\prod_{i=1}^{n}\frac{1-F_Z(y_i)}{\int_0^{\infty}uf_Z(u)du}. \tag{10}$$

Maximum likelihood estimation of parametric models is straightforward using this equation. For nonparametric models, the following iterative algorithm proposed by Denby and Vardi (1986) is useful.

Denby-Vardi (DV) Algorithm. Start with arbitrary positive numbers $p^{(0)}=(p_1^{(0)}, \ldots, p_n^{(0)})$ such that $\sum_{i=1}^{n}p_i^{(0)}=1$. Update the values of $p^{(m)}=(p_1^{(m)}, \ldots, p_n^{(m)})$ by the following:

$$p_j^{(m+1)}=\left(\sum_{i=1}^{n}\frac{r_i^{(m)}}{y_{(i)}}\right)^{-1}\frac{r_j^{(m)}}{y_{(j)}} \qquad j=1,\ldots,n,\ m=0,1,\ldots,$$

where

$$r_j^{(m)}=p_j^{(m)}\sum_{i=1}^{j}\frac{1}{\sum_{k=i}^{n}p_k^{(m)}} \qquad j=1,\ldots,n,$$

and $y_{(i)}$ denotes the ith order statistics of the observations of residual lifetimes. The algorithm provides a solution $\hat{p}=(\hat{p}_1, \ldots, \hat{p}_n)$ to the problem of maximizing $\prod_{i=1}^{n}\frac{\sum_{j=i}^{n}p_j}{\sum_{j=1}^{n}y_{(j)}p_j}$, which corresponds to the nonparametric version of the likelihood function given by (10). The nonparametric maximum likelihood estimate (NPMLE) of the population lifetime distribution F_Z is then obtained by a simple transformation $\hat{F}_Z\left(y_{(i)}\right)=\sum_{j=1}^{i-1}\hat{p}_j$. Denby and Vardi's algorithm was designed to find the NPMLE under a decreasing density constraint, which for this case, correspond to the residual lifetime density. For a detailed discussion on the validity of the use of the DV algorithm to obtain estimates of the population lifetime distribution, see Momma (1996).

3.3 Simulation Studies

To explore the properties of the NPMLE in finite samples, two simulations with 50 runs each were carried out. In the first experiment, n=200, and in the second, n=500. The underlying population distribution in both cases is exponential with mean 50, and the estimates were obtained using the DV algorithm. In the course of the study, it was discovered that larger number of iterations of the algorithm did not necessarily produce better estimates. In fact, estimates of the population lifetime distribution $\hat{F}_Z\left(y_{(i)}\right)=\sum_{j=1}^{i-1}\hat{p}_j$ after 50 iterations showed an irregular and unnatural pattern compared to the estimates after 10 iterations. Furthermore, estimates of mean lifetimes started to deviate from the true value at larger number of iterations, while the value of the nonparametric likelihood function continued to increase. A possible explanation is that since a discrete function is being fitted to a continuous function, too much fine tuning to a particular set of data produces undesirable results. Simulation results suggest that it is best to stop the iteration when the increase rate of the nonparametric likelihood function begins to slow down. For this reason, the algorithm was stopped after 10 iterations in this study.

It was also found that the DV method has a tendency to under-estimate the population lifetimes, and the bias was quite severe in some cases. **Figures 1** and **2** depict histograms of the 50 mean lifetimes calculated from the estimated lifetime distribution $\hat{F}_Z$ corresponding to n=200 and n=500, respectively. The true mean value is 50. The figures clearly indicate a substantial downward bias, with no clear indication of improvement in larger sample size. Even with 500 observations, some estimates took values close to zero. This is caused by the local bias at smaller values of the residual lifetime y. The residual lifetime density $\hat{f}_{Y^*}\left(y_{(i)}\right)=\frac{\sum_{j=i}^{n}\hat{p}_j}{\sum_{j=1}^{n}y_{(j)}\hat{p}_j}$ is often over-estimated for values of y near the origin. This translates to the downward bias of the lifetime estimate, since the mean of the population lifetime is the reciprocal of the residual

Figure 1. Estimated Mean Lifetimes based on 200 Observations

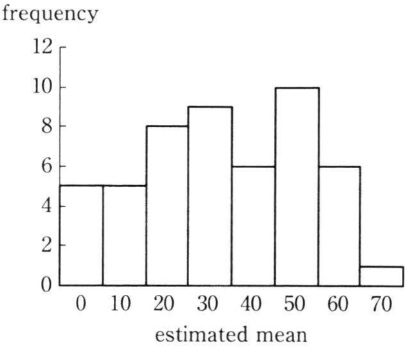
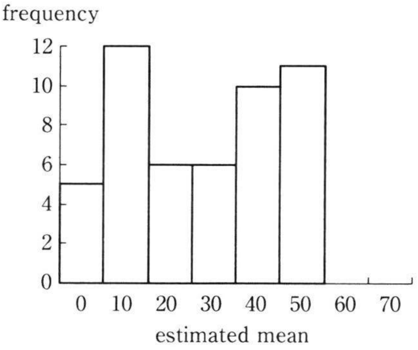


Figure 2. Estimated Mean Lifetimes based on 500 Observations



lifetime density at zero, as seen from (9).

Denby and Vardi proposed to correct the bias by flattening the peak of the density $\hat{f}_{Y^*}$ near the origin. More specifically, they propose to substitute $\hat{f}_{Y^*}$ by the following:

$$\hat{\hat{f}}_{Y^*}(y) = c\hat{f}_{Y^*}\left(\hat{F}_{Y^*}(\alpha)\right) \quad 0 < y \le \hat{F}_{Y^*}^{-1}(\alpha) \tag{11}$$

$$= c\hat{f}_{Y^*}(y) \qquad \hat{F}_{Y^*}^{-1}(\alpha) < y, \tag{12}$$

where c is a normalizing constant chosen so that $\hat{\hat{f}}_{Y^*}$ integrates to 1. As for the value of α, they suggest a small fraction, for example, 0.1. The bias-corrected lifetime distribution $\hat{\hat{F}}_Z$ is then obtained by transforming $\hat{\hat{f}}_{Y^*}$. Clearly, after adjusting for the bias, $\hat{\hat{F}}_Z(y)=0$ for all values of y such that $0<y\le\hat{F}_{Y^*}^{-1}(\alpha)$. Consequently, the bias-corrected lifetime distribution $\hat{\hat{F}}_Z$ takes a rather peculiar form. In order to prevent $\hat{\hat{F}}_Z$ from taking zero values, Momma (1996) suggested linearly increasing the values $\hat{\hat{f}}_{Y^*}$ by small amounts as y approaches zero.

Another limitation of the DV bias correction method is that it is designed only to reduce the values of the estimated density of the residual lifetimes near the origin. As a result, when the residual lifetime density in the vicinity of the origin is under-estimated before bias correction, the DV correction exacerbates the downward bias, and consequently over-estimates the lifetime distribution, although the amount is usually not substantial.

In view of the above, a different method is proposed here as a means to correcting the bias. In this approach, the peak value of the empirical histogram is employed to estimate the density of the residual lifetime at the smallest observed value $y_{(1)}$. Then, for a pre-determined small value of α, values of the bias-corrected density estimate $\hat{\hat{f}}_{Y^*}$ for $0<y\le\hat{F}_{Y^*}^{-1}(\alpha)$ are calculated by linearly connecting the estimated value of the density at $\hat{F}_{Y^*}^{-1}(\alpha)$ and the value of the empirical histogram at $y_{(1)}$. In case the uncorrected estimate of the residual lifetime density at $\hat{F}_{Y^*}^{-1}(\alpha)$ exceeds the value of the peak of the empirical histogram, take the first value of y such that $\hat{f}_{Y^*}(y)$ is smaller than the peak value, and proceed as above. The corrected value of $\hat{\hat{f}}_{Y^*}$ in turn is transformed and scale adjusted to obtain the values of $\hat{\hat{p}}=(\hat{\hat{p}}_1,\ldots,\hat{\hat{p}}_n)$ with the condition that $\sum_{i=1}^{n}\hat{\hat{p}}_i=1$. Finally, estimates of the lifetime distribution is obtained from the relation $\hat{\hat{F}}_Z\left(y_{(i)}\right)=\sum_{j=1}^{i-1}\hat{\hat{p}}_j$.

Histograms of the estimated mean lifetimes based on 200 observations, using the Denby-Vardi bias correction method and the empirical histogram method are shown in **Figures 3** and **4**, respectively. The value of α was set to 0.1 for both methods. In a number of cases where the residual lifetime density near the origin was substantially over-estimated (the lifetime distribution under-estimated), $\hat{F}_{Y^*}^{-1}(y_{(1)})>\alpha$ so that condition (11) did not hold for any observed values of y. But these are clearly the cases that most needed the bias-correction. Accordingly, in such cases, the value of the estimated density corresponding to the smallest observation $y_{(1)}$ was substituted by the value of the estimated density at the second smallest observation $y_{(2)}$, and the entire density normalized. As can be seen from **Figures 1**, **3** and **4**, both bias correc-

Figure 3. Estimated Mean Lifetimes Denyb-Vardi Bias Correction Method

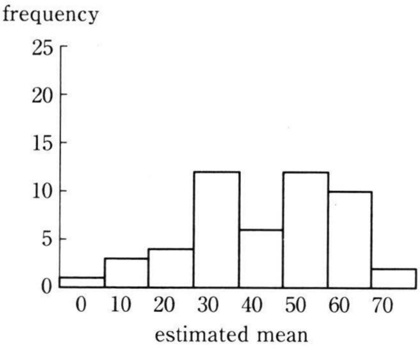
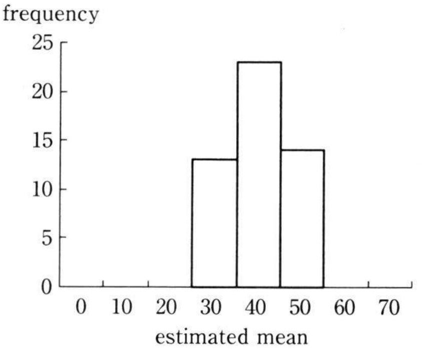


Figure 4. Estimated Mean Lifetimes Bias Correction by Empirical Histogram



tion methods produced better estimates of the mean lifetimes than the uncorrected case, but the empirical histogram method proved to be more effective.

Figures 5 and **6** illustrate the estimated lifetime distribution functions without bias correction (uncorrected), bias-corrected using the Denby-Vardi method (DV) and bias-corrected using the empirical histogram (histogram), along with the theoretical distributions. **Figure 5** is an example of the case where, without bias correction, the lifetime distribution is under-estimated, while **Figure 6** corresponds to the case of over-estimation. The effectiveness of the Denby and Vardi's correction method depends heavily on a few values of the estimated density of Y^* in the vicinity of $\hat{F}_{Y^*}^{-1}(\alpha)$, and therefore its performance is rather unstable. In some cases where the uncorrected $\hat{f}_{Y^*}$ substantially over-estimated the residual lifetime density near the origin, the method reduced the bias drastically, while in other seemingly similar cases, it had minimal effect. In contrast, the method using the empirical histogram produced consistently good estimates.

The empirical histogram method is not without faults. A major difficulty of this method is in finding the "adequate" empirical histogram. Since the objective is to estimate the value of the density, it is not enough to detect its shape. In order to obtain an estimate of the density at a

Figure 5. Estimated Lifetime Distribution under-estimating lifetimes

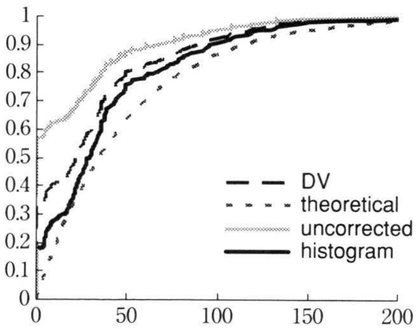
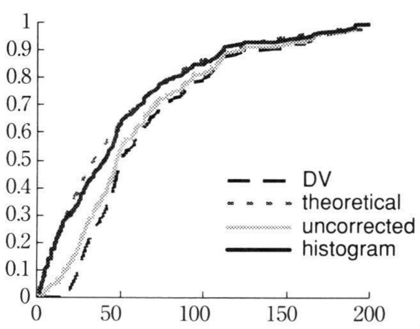


Figure 6. Estimated Lifetime Distribution over-estimating lifetimes



single point, class width of the empirical histogram must be made as small as possible (and therefore the number of classes made as large as possible) while retaining the distributional form. Typically, histograms with about 600 bars gave good estimates of the value of the density at its peak. It is advisable to start with an empirical histogram with a moderate number of classes, and make the class width smaller while keeping the shape of the histogram intact. It takes a trained eye to pick out the right histogram for estimation. Depending on the histogram chosen, the results of the bias-correction may vary considerably. Therefore, the method should be used with caution.

Empirical histogram is also useful to determine whether the unadjusted NPMLE of the lifetime distribution is biased downward for a particular case. As has been noted already, since the mean lifetime of the whole population corresponds to the reciprocal of the density of the residual lifetime at zero, the value can be estimated from the empirical histogram of the residual lifetimes. Results of the simulation studies indicate that when the unadjusted NPMLE contains noticeable downward bias, estimated mean lifetime using the empirical histogram always exceeded the NPMLE by a substantial amount. This being the case, mean estimate from the empirical histogram can be used to test the existence of the downward bias of the NPMLE. When the NPMLE of the mean is significantly smaller than the estimate based on the empirical histogram, bias correction should be implemented.

3.4 Censored Observations

It is safe to assume that data on residual lifetimes will almost always include observations censored from above. When an observation of the residual lifetimes is censored after a period τ, the density of the residual lifetime of the intercepted item becomes

$$f_{Y^*}(y) = \begin{cases} \dfrac{1-F_Z(y)}{\int_0^\infty uf_Z(u)du} & if \quad y \le \tau \\ \\ \dfrac{\int_\tau^\infty [1-F_Z(x)]dx}{\int_0^\infty uf_Z(u)du} & if \quad y > \tau \end{cases} . \tag{13}$$

The likelihood function including right censoring will then be of the form

$$L = \prod_{i=1}^{n} \left(\frac{1-F_Z(y)}{\int_0^\infty uf_Z(u)du} \right)^{I_i} \left(\frac{\int_\tau^\infty [1-F_Z(x)dx]}{\int_0^\infty uf_Z(u)du} \right)^{1-I_i} , \tag{14}$$

where $I_i=1$ if the residual lifetime of item i is observed and $I_i=0$ if it is censored. Parametric estimation based on (14) is straightforward. For nonparametric cases, since the DV algorithm was originally proposed for cases including random right censoring, the method is still applicable with only minor changes.

4. DISCRETE OBSERVATIONS AND GENERAL WEIGHTED DISTRIBUTIONS

A couple of other problems concerning intercepted sampling methods are discussed briefly in this Section.

4.1 Discrete Observations

Intercepted sampling is commonly used in epidemiological studies, where the object of the study is to investigate the survival time of a certain disease. For clinical trials of a prevalent disease, the initiation time of the disease is most likely unobservable. In addition, patients will likely start a treatment program when diagnosed with a disease, so the residual lifetime is also unobservable. As such, the only observable variable is the stage, or the condition of the disease at the sampling time point. Assume that the observable discrete variable W^* indicating the condition of the disease takes the value i when $x_{(i)} \leq X^* < x_{(i+1)}$, where X^* is as before, the duration time of the disease at interception. Then,

$$\begin{aligned} P_{W^*}(i) &= F_{X^*}(x_{i+1}) - F_{X^*}(x_i) \\ &= \frac{1}{\int_0^\infty u f_Z(u)du} \left\{ \int_0^{x_{i+1}} [1 - F_Z(u)]du - \int_0^{x_i} [1 - F_Z(u)]du \right\} \\ &= \frac{1}{\int_0^\infty u f_Z(u)du} \left\{ \int_{x_i}^{x_{i+1}} (1 + u f_Z(u))du - (x_{i+1} F_Z(x_{i+1}) - x_i F_Z(x_i)) \right\} \end{aligned} \qquad (15)$$

where $P_{W^*}(\mathrm{i}) = P(W^* = i)$.

Example Exponential distribution

When the underlying lifetime distribution is exponential so that $F_Z(z) = 1 - e^{-\lambda z}$, the above relation (15) yields

$$p_{W^*}(i) = x_{i+1}\left(\lambda e^{-\lambda x_{i+1}} - e^{-x_{i+1}}\right) - x_i\left(\lambda e^{-\lambda x_i} - e^{-x_i}\right) - \left(e^{-x_{i+1}} - e^{-x_1}\right).$$

In reality, the condition of the disease is not determined solely by the length of its duration time. It depends on the patient's various physical conditions, among other things. A more elaborate model, such as regression type models, should be formulated to study such data.

4.2 General Weighted Distributions

As stated in Section 2, the distributional forms of the lifetime distribution and the residual life time distribution of the items in the intercepted population were obtained under the assumption of a stationary Poisson birth process of the whole population. In some cases, this assumption does not hold. Consider, for example, the case of the time spent on the internet. People are on and off the internet throughout the day. This means that for each person, a stochastic process is formed where there is an on time (time spent on-line) and an off time (time spent off the net). In other words, every person's internet usage forms a renewal process with two alternating states. It is natural to assume that the length of the on time and the off time are related to each other, and therefore a standard Poisson assumption is not likely to hold for this case.

One way to generalize the distributional forms for cases such as this is to consider a more general weighted density, where observations from the weighted density follows

$$f_{Z^*}(z) = \frac{\omega(z) f_Z(z)}{\int_0^\infty \omega(u) f_Z(u) du}.$$

Here, $\omega(z)$ is a monotone weighting function to be estimated from the data. When $\omega(z)=z$, it becomes the standard length-biased density, whereas $\omega(z)=z^2$ produces size biased density, often used for spatial data. For a semi-parametric estimation of this type of model, see Sun and Wang (2006).

5. CONCLUDING REMARKS

Intercepted sampling is a convenient way to obtain survival data. When observations on the residual lifetimes of the intercepted items are used to estimate the population lifetime distribution, downward bias is often detected. Simulation studies suggest that with a proper use of the empirical histogram, this bias of the NPMLE can be controlled. Comparison of the properties of the moment based estimates and the NPMLE is an object of further study.

REFERENCES

1. Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *J. Am. Stat. Assoc.*, 97, 201-209.
2. Asgharian, M., and Wolfson, D. B. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *Ann. Statist.*, 33, 2109-2131.
3. Blumenthal, S. (1967). Proportional sampling in life length studies. *Tehcnometrics*, 9, 205-218.
4. Cox, D.R. (1962). *Renewal Theory*, John Wiley and Sons, New York.
5. Cox, D.R. (1969). Some sampling problems in technology, In *New Developments in Survey Sampling*. Johnson, N. L. and Smith, H., Jr. eds. 506-527. Wiley-Interscience.
6. Denby, L. and Vardi, Y. (1986). The survival curve with decreasing density. *Technometrics*, 28, 359-367.
7. Momma, M. (1991). Estimation of lifetime models using intercepted sampling methods, *Ph. D. Dissertation*, Princeton University.
8. Momma, M. (1996). Nonparametric estimation of a lifetime distribution based on observations from a recurrence time density. 東洋大学経済研究所『経済研究年報』第 21 号, 111-136.
9. Sun, J. and Wang B. (2006). Sieve estimates for biased survival data. *IMS Lecture Notes-Monograph Series*, Recent Developments in Nonparametric Inference and Probability, 50, 127-143.
10. Vardi, Y. (1982). Nonparametric estimation in the presence of length bias, *Ann. Statist.*, 10, 616-620.
11. Vardi, Y. (1988). Statistical models for intercepted data. *J. Am. Stat. Assoc.*, 83, 183-97.