Censoring and Truncation - Highlighting the Differences [*]

Micha Mandel

The Hebrew University of Jerusalem, Jerusalem, Israel, 91905

July 9, 2007

[*] Micha Mandel is a Lecturer, Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel (email: msmic@huji.ac.il)

Abstract

Survival data are very common in the medical science, actuarial science, astronomy, demographic, and many other scientific areas. The most typical characteristic of survival data is incompleteness, where by far the most common models are those of censoring and truncation. Although quite different in nature, the left truncation and the right censoring models result in non-parametric estimates which are very similar in form. However, in other models of truncation and censoring this similarity breaks and estimation should be based on the special properties of each type of incompleteness. In this note, two simple models, one of censoring with a known censoring distribution and the other of truncation with a known truncation distribution, are contrasted and estimation is discussed. The goal is to highlight the differences between the two types of incomplete data in a way that can help analyzing and interpreting more complicated survival data.

KEYWORDS: survival analysis, selection bias, censored data, truncated data

1 INTRODUCTION

Censoring and truncation are common features of survival data, both are taught in most survival analysis courses. Although very different in nature, many statisticians tend to confuse between them, probably due to the very similar form of their non-parametric maximum likelihood estimates (NPMLEs). In this note, the mechanisms that create censored and truncated data are explored and their differences are highlighted.

We focus on the random censoring model with a known distribution of censoring and on the random truncation model with a known distribution of the truncation variable and show that the NPMLEs have different forms. These simple models illustrate nicely the different natures of truncation and censoring and provide guidelines for analyzing truncated and censored data that are non-standard.

Sections 2 presents the censoring and truncation models and derives their NPMLEs. Section 3 remarks briefly on identification problems and discusses estimation in the fully non-parametric cases. Section 4 demonstrates estimation and identification using a simple example from a study on critically ill patients. Section 5 concludes the paper with a short discussion.

2 THE MODELS

It is easy to understand censoring and truncation as deriving from the following model, which for simplicity is presented for the continuous case. Let $X \sim F$ and $Y \sim G$ be independent positive random variables with densities f and g, respectively. We will look at the representation (U, V, Δ) of (X, Y), where $U = \min(X, Y)$, $V = \max(X, Y)$, $\Delta = I\{X \leq Y\}$ and $I\{A\}$ is the indicator function of the event A takes on the value 1 on A and 0 on its complement. The density/probablity function of (U, V, Δ) is given by

$$h(u, v, \delta) = \begin{cases} f(u)g(v) & u \leq v \ , \ \delta = 1 \\ f(v)g(u) & u < v \ , \ \delta = 0 \ , \end{cases} \tag{1}$$

and will be referred to as the complete data density.

2.1 Censoring

Using (1), the *random right censoring model* is described by observing only the data (U, Δ), and the less common *random left censoring model* is defined by the data (V, Δ). Thus, it is clearly seen that random censoring is a missing data model. We next consider right censored data, (U, Δ), and describe non-parametric estimation of F. Estimation of G and estimation under the left censoring model is analogous.

Let $\bar{F}(x) := P(X > x)$ denote the survival function of X, and let $\bar{F}(x-) = P(X \geq x)$. The hazard function of X is defined by $\lambda_X(x) := f(x)/\bar{F}(x-)$ and plays an important role in survival analysis. The hazard function λ_X identifies the distribution F by the relation $\bar{F}(x-) = \exp\{-\Lambda_X(x)\}$, where $\Lambda_X(x) := \int_0^x \lambda_X(u)du$ is the cumulative hazard function. The NPMLE of F under the random censoring model is obtained by first estimating the hazard and then transforming to F; the details follow.

Let $(u_1, \delta_1), \ldots, (u_n, \delta_n)$ be n independent realizations of (U, Δ). The naive empirical estimate of $h(u, 1) = \int_{v \geq u} h(u, v, 1)dv = f(u)\bar{G}(u-)$ is just

$$\hat{h}(u, 1) = \frac{1}{n} \sum_{i=1}^{n} I\{u_i = u, \delta_i = 1\}, \tag{2}$$

where $\bar{G}(u-) := P(Y \geq u)$. Next, note that

$$\bar{K}(u-) := P(U \geq u) = P(X \geq u)P(Y \geq u) = \bar{F}(u-)\bar{G}(u-), \tag{3}$$

because of independence. The empirical estimate of $\bar{K}(u-)$ is

$$\hat{\bar{K}}(u-) = \frac{1}{n} \sum_{i=1}^{n} I\{u_i \geq u\}. \tag{4}$$

Using the relation $\lambda_X(u) = h(u, 1)/\bar{K}(u-)$ and (2) and (4), a natural estimate of $\Lambda_X(u)$ is

$$\hat{\Lambda}_X(u) = \sum_{k=1}^{m} \frac{\hat{h}(t_k, 1)}{\hat{\bar{K}}(t_k-)} I\{t_k \leq u\},$$

where $t_1 < t_2 < \cdots < t_m$ are the distinct values of $u_1, \ldots, u_n$. The resulting Nelson-Aalen estimate of $\bar{F}(u-)$ is $\exp\{-\hat{\Lambda}_X(u)\}$.

In the discrete case, the hazard function is defined in a similar way as $\lambda_X(t) = P(X = t)/P(X \geq t)$. The relation between the hazard and the survival function in the discrete case is $\bar{F}(x-) = \prod_{k|t_k<x}\{1 - \lambda_X(t_k)\}$, where $t_1, t_2, \ldots$ are the support points of F. Letting

$\hat{\lambda}_X(t_k) = \hat{h}(t_k, 1)/\hat{\bar{K}}(t_k-)$, the NPMLE of $\bar{F}$, known as the Kaplan-Meier or the product-limit estimate is given by

$$\hat{\bar{F}}_{KM}(x-) = \prod_{k|t_k<x} \{1 - \hat{\lambda}_X(t_k)\}. \tag{5}$$

The Kaplan-Meier and the Nelson-Aalen estimators are asymptotically equivalent and usually have very close values when data are continuous. However, they may differ considerably when data are discrete and contain many ties. Note that when deriving the Kaplan-Meier and the Nelson-Aalen estimates, neither G nor the fact that G is known was used.

Finally, the notion of the risk group should be mentioned. The risk group at time u contains all subjects who have not failed nor censored before time u. It is easily seen from (4) that $n\hat{\bar{K}}(u-)$ is just the number of subjects in the risk group at time u, i.e., the size of the risk group. Since in the survival analysis terminology the hazard function at u stands for the instantly probability of failing at u conditionally on being alive up to time u, the estimate of the hazard can be intuitively understood as the proportion of failures among those who are at risk to fail (see (2) and (4)).

2.2 Truncation

Using (1), the *random right truncation model* of F (or left truncation of G) is described by observing the data (U, V) only if $\Delta = 1$. The way the model is presented shows explicitly its nature and the difference between truncation and censoring. Whereas censoring is a model of missing observations (missing V values), truncation is a model of selection bias (selecting only those data points whose Δ values are 1). Therefore, estimation using truncated data is naturally based on methods for selection bias models (e.g., Vardi, 1985).

The likelihood of an observation is the density of $(U, V)|\Delta = 1$ and it equals $h(u, v, 1)/P(\Delta = 1)$; it can be written as

$$\frac{f(u)g(v)I\{u \leq v\}}{\int f(z)\bar{G}(z-)dz} = \frac{f(u)\bar{G}(u-)}{\int f(z)\bar{G}(z-)dz} \times \frac{g(v)}{\bar{G}(u-)} I\{u \leq v\}. \tag{6}$$

The right hand side of (6) factors the density of $(U, V)|\Delta = 1$ to the marginal density of $U|\Delta = 1$ (first term) and the conditional density of $V|U, \Delta = 1$ (second term). The marginal density of U in the sample, $f^*(u) = f(u)\bar{G}(u-)/\int f(z)\bar{G}(z-)dz$, is a biased (or weighted) density with a weight $\bar{G}(z-)$ at z. Thus, as smaller u is, as larger the weight $\bar{G}(u-)$ is; or phrasing it differently, small values of X have a better chance to be smaller than an independent value of Y and hence have a better chance to be included in the sample.

Since in our model G is known, the conditional density of $V|U, \Delta = 1$ can be omitted from the likelihood providing a simple algorithm for estimating F: First estimate the biased density using the empirical law,

$$\hat{f}^*(u) = \frac{\widehat{f(u)\bar{G}(u-)}}{\int f(z)\bar{G}(z-)dz} = \frac{1}{n}\sum_{i=1}^{n} I\{u_i = u\}, \tag{7}$$

and then use the inverse transformation $f(u) \propto f^*(u)/\bar{G}(u-)$ to obtain an inverse weighting estimate for F

$$\hat{\bar{F}}_{IW}(x-) = \frac{\sum_{i=1}^{n} I\{u_i \geq x\}\left[\bar{G}(u_i-)\right]^{-1}}{\sum_{i=1}^{n}\left[\bar{G}(u_i-)\right]^{-1}}. \tag{8}$$

Note the important role of G in this estimate and the importance of the assumption that it is known. Note also that data on V, which in this model are data on Y, are not used at all by (7) and (8).

In summary, both $\hat{\bar{F}}_{KM}$ and $\hat{\bar{F}}_{IW}$ use only the data (U, Δ), but for the latter model only points with $\Delta = 1$ are observed and can be used. This shows that truncation is a more severe incomplete data scenario and in general one can expect $\hat{\bar{F}}_{KM}$ to perform better than $\hat{\bar{F}}_{IW}$ if the complete data are the same. The example in section 4 demonstrates this point further.

3 FURTHER TOPICS

Identification. The problem of identifiability, that is, whether F is estimable from the data, has not been mentioned, although exists in both models. Here again an important difference between truncation and censoring appears. For the right censoring model where the available data are realizations of (U, Δ), if there is an interval $\mathrm{I} = (y_{\max}, x_{\max})$ on which $G = 1$ and $F < 1$, then F is not identifiable on I. In other words, X values that are larger than the possible maximum value of Y $(y_{\max})$ cannot be seen and hence F cannot be estimated there. Nonetheless, F is estimable at all points smaller than $y_{\max}$. For the right truncation model, the identification problem is different. In fact, F is either identifiable everywhere or not identifiable at all, since the only estimable function is $F/F(y_{\max})$. If $F(y_{\max}) = 1$, then F is identifiable, otherwise one cannot estimate F and must replace his/her parameter of interest to the conditional distribution presented above. It is important to distinguish between these two identification problems, and especially to notice that many truncation models force one to estimate parameters which are different than those one initially targeted. The example in the next section illustrates this point further.

Fully non-parametric models. The estimate of F in the right censoring model does not involve G and can still be calculated when G is unknown. This is not the case in the right truncation model. When G is not known, it should be estimated by the data and then plugged in to provide the estimate of F. Using (6), the likelihood of n observations is given by

$$\prod_{i=1}^{n} \frac{f(u_i)\bar{G}(u_i-)}{\int f(z)\bar{G}(z-)dz} \times \prod_{i=1}^{n} \frac{g(v_i)}{\bar{G}(u_i-)}.$$

Equation (7) shows that for any G, the maximum value of the first term of the likelihood is $\prod_{j=1}^{n} n^{-1} \sum_{i=1}^{n} I\{u_i = u_j\}$, which is independent of G. Thus, the algorithm of obtaining the NPMLE is to first estimate G from the conditional likelihood of $V|U$, namely $\prod_{i=1}^{n} g(v_i)/\bar{G}(u_i-)$, and then plug in the estimate in (8). Shen (2003) shows that this procedure provides the product-limit estimate of F using the conditional likelihood of $U|V$. Thus, although in the fully non-parametric left truncation model the classical product-limit estimate of F is the NPMLE, it is so only as a coincidence of the product-limit being algebraically equal to an inverse weighting estimate. This point is insightful and should be mentioned much more frequently than it does.

4 EXAMPLE

During eight consecutive days in February 2003, professional teams screened all patients in five hospitals in Israel and determined, based on pre-defined criteria, all those who were critically ill (CI) during the last 24 hours. Those selected were followed-up until death or until improvement. Table 1 presents data on the first day and length of CI status for those found outside intensive care units. For example, of the 30 patients deteriorated on the third day, nine were CI for three days and of the 31 subjects entered the study on the seventh day, three were CI for three days. More details can be found in Simchen et.al. (2007).

Table 1: Length of being critically ill by day of enrollment

day entered	Length of being CI (days)								
the survey	1	2	3	4	5	6	7	8	total
1	17	11	6	1	3	2	0	1	41
2	8	13	10	1	1	1	0	0	34
3	11	7	9	2	0	0	1	0	30
4	11	12	0	0	0	2	0	1	26
5	11	9	3	3	1	2	0	0	29
6	13	5	3	2	1	1	1	0	26
7	14	4	3	5	2	2	0	1	31
8	12	9	4	0	1	0	2	0	28
total	97	70	38	14	9	10	4	3	245

To demonstrate the influence of censoring and truncation assume that subjects were followed only for 5 days. Thus, X, the number of days a patient is CI, and $Y \equiv 5$, the follow-up time, are obviously independent. Under the truncation model, no information is given on columns 6-8. Under the censoring model, the number of observations in columns 6-8 is known, but the exact lengths of CI status of these 17 observations are not known. The difference between truncation and censoring in this scenario is in identification. From the censored data, one can estimate $\bar{F}(6-) = P(X \geq 6)$ by the proportion of patients still deteriorated after day 5 and get $\hat{\bar{F}}(6-) = 17/245 \cong .069$, but F is not identifiable on values greater than 5. From the truncated data, $P(X \geq 6)$ is not identifiable and hence inference can be made only on $F/F(5)$. It is easy to check that in this simple case the estimates of $F/F(5)$ in the truncation and censoring scenarios match. This is not always the case as shown in the next example.

Suppose now that follow-up ends on the eighth day; thus, those who enter on the k'th day are followed for $8 - k + 1$ days. Here Y is the length of follow-up and its distribution depends on the probability of entering the study on each day of the survey. It is reasonable to assume that the distribution of Y is uniform on $\{1, 2, \ldots, 8\}$ and that it is independent of X, the length of CI status. The data available are described by the stepped line in

Table 2: Estimation of F using censored, truncated and complete data - a comparison.

day		1	2	3	4	5	6	7	8
Censored Data	# failed	97	61	31	7	4	3	0	1
	# censored	16	13	5	3	3	1	0	0
	# survived	132	58	22	12	5	1	1	0
	Hazard λ_X	.396	.462	.534	.318	.333	.600	0	1
	$\hat{\bar{F}}_{KM}$	1	.604	.325	.151	.103	.069	.028	.028
Truncated Data	# failed	97	61	31	7	4	3	0	1
	Weight ($\bar{G}$)	1	.875	.750	.625	.500	.375	.250	.125
	$\hat{f}^*$	.475	.299	.152	.034	.020	.015	0	.005
	$\hat{f} \propto \hat{f}^*[\bar{G}]^{-1}$	.399	.287	.170	.046	.033	.033	0	.033
	$\hat{\bar{F}}_{IW}$	1	.601	.315	.145	.099	.066	.033	.033
Complete Data	# failed	97	70	38	14	9	10	4	3
	$\hat{\bar{F}}_{COM}$	1	.604	.318	.163	.106	.069	.029	.012

Table 1, where right truncation means that nothing is observed to the right of this line and censoring means that at each row only the number of observations to the right of the line is known, but their distribution there is not. Note that there is no problem here of identifiability and from both hypothetical data F is estimable.

Table 2 compares estimates of F assuming truncated and censored data. For comparison, the last block of the table presents an estimate based on the complete data, $\bar{F}_{COM}(x-) = n^{-1} \sum_i I\{x_i \geq x\}$. The first block describes estimation under the censoring scenario, where the first three lines summarizes the available data. The hazard λ_X is computed by dividing the number of failures with the total number of failures, censored and survivors, which comprise the risk set. The estimate of $\bar{F}$ follows from (5). The second block presents calculation under the truncation scenario. The data needed are the number of failures at each time and they are identical to that in the censoring case. Estimation of $\bar{F}$ follows equations (7) and (8) with G being the uniform distribution.

The three estimates agree well, with $\hat{\bar{F}}_{KM}$ closer to $\hat{\bar{F}}_{COM}$ than $\hat{\bar{F}}_{IW}$. This is expected since censored data contain information on the complete data that truncated data do not.

It is also seen that both $\hat{\hat{F}}_{KM}$ and $\hat{\hat{F}}_{IW}$ assign mass only to times where failures observed.

Although the example is somewhat synthetic and estimation could be based on the complete data, it illustrates scenarios that are very common in practice. The interested reader is referred to the complete CI study of Simchen et.al. (2007), and to the huge survival analysis literature.

5 CONCLUDING REMARKS

Censoring is defined by observing a random set to which an observation belongs while truncation means observing the exact lifetime value, but only if it belongs to a certain random set. This difference should be kept in mind when analyzing and interpreting survival data, and should be emphasized in survival analysis courses. Although in the fully non-parametric models of left truncation and right censoring the NPMLEs take similar forms both use the risk sets principle, there are many situations in which the NPMLEs under truncation and censoring differ. The simple model presented here is only one such example.

References

[1] Shen, P.S., (2003), "The Product-Limit Estimate as an Inverse-Probability-Weighted Average", *Communication in Statistics - Theory and Methods*, 32, 1119-1133.

[2] Simchen, E., Sprung, C. L., Galai, N., Zitser-Gurevich, Y., Bar-Lavi, Y., Levi, L., Zveibil, F., Mandel, M., Mnatzaganian, G., Goldschmidt, N., Ekka-Zohar, A., Weiss-Salz, I., (2007). "Survival of Critically Ill Patients Hospitalized In and Out of Intensive care", *Critical Care Medicine*, 35, 449-457.

[3] Vardi, Y. (1985), "Empirical Distributions in Selection Bias Models", *The Annals of Statistics*, 13, 178-203.