

## Chapter 4. Nonparametric Estimation of Survival Func. $S(t)$

In Chapter 3, we assumed the distribution of a failure time can be specified. Then, we estimated the corresponding parameters through MLE. In fact, we never know the true distribution comes from which distribution family. In this chapter, we focus on estimating the survival function of the failure time without any specified distribution assumption.

Suppose the  $n$  failure time  $T_1, \dots, T_n$  can be observed exactly. Then the empirical distribution function of the survival function

$$S^e(t) = \frac{1}{n} \sum_{i=1}^n I(T_i \geq t)$$

is an unbiased and consistent estimator of the survival function  $S(t) = \Pr(T \geq t)$ . However, in most application, the data is observed incompletely due to the censoring mechanism. Later, we will discuss the right censoring cases first, and then extend the idea to include left truncation.

We start by the life table skill in Section 4.1 and extend the idea to provide the product limit estimator (or Kaplan-Meier (K-M) estimator) in Section 4.2. There are several points of view to K-M estimator other than a modification of the life table. We will show them In Section 4.3. In Section 4.4, we extend the K-M estimator to include left-truncation data.

### 4.1 Life table:

The life table estimate of the survival function is obtained by first dividing the period of observation into a series of time intervals. Let  $0 = t_0^* < t_1^* < \dots < t_{m-1}^* < t_m^* = \infty$  be the cut points of the intervals. Define

- $d_j$  : the number of death in the  $j$ th interval  $[t_{j-1}, t_j)$
- $c_j$  : the number of censoring in the  $j$ th interval
- $n_j$  : the number of individuals who are alive (at risk of death) at the beginning of the  $j$ th interval
- $h_j$  : the conditional probability of death in the  $j$ th interval given that the individual alive at the beginning of the  $j$ th interval

Without censoring, that is  $c_j = 0$ ,  $h_j$  can be estimated by  $d_j/n_j$ . With censoring, we assume the censoring process such that the censoring time occur uniformly throughout the  $j$ th interval. Thus, the average number of individuals who are at risk during this interval is  $n_j - c_j \leq n'_j = n_j - \frac{c_j}{2} \leq n_j$ . Therefore,  $\hat{h}_j = d_j/n'_j$ .

This implies  $S^*(t) = \prod_{j=1}^k (1 - d_j/n_j')$  for  $t_k^* < t \leq t_{k+1}^*$ .

## 4.2 Kaplan-Meier estimator (1958, or name product limit estimator)

Let  $0 = t_{(0)} < t_{(1)} < \dots < t_{(r-1)} < t_{(r)} < t_{(r+1)} = \infty$  be ordered distinct death times. Define

- $d_j$  : the number of death at  $t_{(j)}$
- $n_j$  : the number of individual who are alive (at risk of death) at time  $t_{(j)}$
- $h_j$  : the conditional probability of death at time  $t_{(j)}$  given that the individual still alive at  $t_{(j)}$

The Kaplan-Meier estimate of survival function becomes

$$\hat{S}(t) = \prod_{j=1}^k (1 - h_j) = \prod_{j=1}^k (1 - d_j/n_j) \text{ for } t_{(k)} < t \leq t_{(k+1)}.$$

Note:

- (1) If the largest observation is censored, the  $\hat{S}(t)$  is defined up to the largest observation.
- (2) If censoring time ties with death time, always treat the censoring time a little larger than death time.

Example:

	4	5	5+	6	7	7	8+	9	...
$d_j$	1	1	0	1	1	1	0	1	
$c_j$	0	0	1	0	0	0	1	0	
$n_j$	17	16	15	14	13	12	11	10	

Then,  $\hat{S}(4) = 1$ ,  $\hat{S}(4.5) = 16/17$ ,  $S(6.5) = \frac{16}{17} \times \frac{15}{16} \times \frac{13}{14} = 0.82$ , etc

### 4.2.1 Standard error of K-M estimator

Given  $n_j$ ,  $d_j \sim B(n_j, 1 - p_j)$  where  $p_j = \Pr(\text{survive at the end of interval} \mid \text{survive at the beginning of the interval})$ . Thus,  $Var(1 - d_j/n_j) = Var(d_j/n_j) = p_j(1 - p_j)/n_j$ . However,  $\hat{p}_j = 1 - d_j/n_j$ . This implies  $Var(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j = (\frac{d_j}{n_j})(1 - \frac{d_j}{n_j})/n_j$ . Recall that  $\hat{S}(t) = \prod_{j=1}^k (1 - d_j/n_j)$ . To get  $Var(\hat{S}(t))$ , we first assume  $\hat{p}_j$  ( $j = 1, \dots, k$ ) are independent (in fact, there are not) and calculate  $Var(\ln \hat{S}(t))$ . Then apply delta method to get  $Var(\hat{S}(t))$ .

(1) By delta method, we have  $Var(\ln \hat{p}_j) \simeq (\frac{1}{p_j})^2 Var(\hat{p}_j)$ . Thus

$$\begin{aligned} Var(H(t)) &= Var(-\ln \hat{S}(t)) \simeq \sum_{j=1}^k (\frac{1}{p_j})^2 Var(\hat{p}_j) \\ &= \sum_{j=1}^k \frac{(1 - p_j)}{p_j n_j}. \end{aligned}$$

(2) By delta method again,

$$\begin{aligned} Var(\hat{S}(t)) &= Var(\exp\{H(t)\}) \simeq [\exp\{H(t)\}]^2 \sum_{j=1}^k \frac{(1 - p_j)}{p_j n_j} \\ &= [S(t)]^2 \sum_{j=1}^k \frac{(1 - p_j)}{p_j n_j}. \end{aligned}$$

This implies

$$\hat{Var}(\hat{S}(t)) = [S(t)]^2 \sum_{j=1}^k \frac{(1 - \hat{p}_j)}{\hat{p}_j n_j} = [\hat{S}(t)]^2 \sum_{j=1}^k \frac{d_j}{(n_j - d_j) n_j}.$$

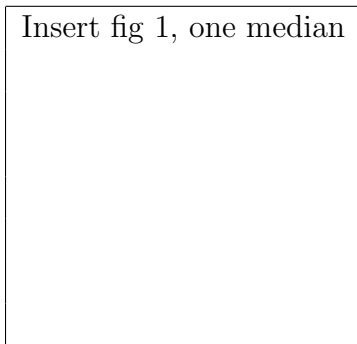
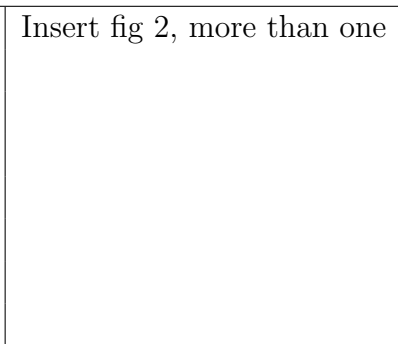
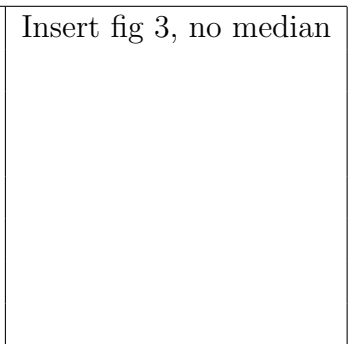
The above estimation names Greenwood formula (1972). The corresponding standard error is

$$se(\hat{S}(t)) = \hat{S}(t) \sqrt{\sum_{j=1}^k \frac{d_j}{(n_j - d_j) n_j}}.$$

Then a 95% C.I. for  $S(t)$  becomes  $\hat{S}(t) \pm 1.96 \times se(\hat{S}(t))$ . Again, this may give a C.I. that outside 0 and 1. We are better create a 95% C.I. of  $\hat{S}(t)$  through  $\ln - \ln \hat{S}(t)$  which takes value from  $-\infty$  to  $\infty$ .

#### 4.2.2 Median $m$ estimation

Recall that  $m$  is the median  $T$  if  $S(m+) \leq 0.5$  and  $S(m-) \geq 0.5$ . There are three possibilities when we apply K-M plot to search  $\hat{m}$ .

Insert fig 1, one median	Insert fig 2, more than one	Insert fig 3, no median
		

Since  $Var(\hat{S}(\hat{m})) \simeq [\frac{dS(m)}{dm}]^2 Var(\hat{m})$ , we have  $\hat{Var}(\hat{m}) \simeq Var(\hat{S}(\hat{m}))/[\hat{f}(\hat{m})]^2$ .

### 4.2.3 Mean $\mu$ estimation

Recall that  $\mu = \int_0^\infty xf(x)dx = \int_0^\infty S(x)dx$ . Thus  $\mu$  can be estimated by the total area under the K-M curve. It is important to note that when the largest observation is right censored, we often consider the truncated mean. That is the area under the K-M curve up to the largest observation.

### 4.2.4 Cumulative hazard function estimator

$$(1) \quad \tilde{H}(t) = -\ln \hat{S}(t) = -\sum_{j=1}^k \ln(1 - d_j/n_j)$$

$$(2) \quad \hat{H}(t) = \sum_{j=1}^k h_j = \sum_{j=1}^k d_j/n_j$$

Note: As  $d_j/n_j$  is small,  $\tilde{H}(t) \simeq \hat{H}(t)$ . People prefers to use  $\hat{H}(t)$ .

Similarly, we have  $V\hat{a}r(\hat{H}(t)) = \sum_{j=1}^k \frac{d_j}{n_j(n_j-d_j)}$ .

## 4.3 Three others points of view to K-M estimator

### 4.3.1 Sketch of nonparametric maximum likelihood estimator

Suppose the  $n$  observations are  $(y_i, \delta_i)$  ( $i = 1, \dots, n$ ) with  $r$  distinct failure times  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$  where  $r \leq n$ . Assume the survival function  $S(t)$  is a step function with jumps at  $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ . Then, the corresponding likelihood function is

$$L(h_1, \dots, h_r) = \prod_{i=1}^n h(y_i)^{\delta_i} S(y_i)$$

where  $h(t) = \Pr(T = t | T \geq t)$  and  $S(t) = \Pr(T \geq t)$ . Let  $h_i = \Pr(T = t_{(i)} | T \geq t_{(i)})$ , the log likelihood function becomes

$$\begin{aligned} l(h_1, \dots, h_r) &= \sum_{i=1}^n [\delta_i \ln h(y_i) + \ln S(y_i)] \\ &= \sum_{j=1}^r d_j \ln h_j + \sum_{i=1}^n \ln \left[ \prod_{t_{(j)} < y_i} (1 - h_j) \right] \\ &= \sum_{j=1}^r d_j \ln h_j + \sum_{i=1}^n \sum_{t_{(j)} < y_i} \ln(1 - h_j) \\ &= \sum_{j=1}^r d_j \ln h_j + \sum_{j=1}^r (n_j - d_j) \ln(1 - h_j) \end{aligned}$$

where  $d_j = \sum_{i=1}^n \delta_i I(y_i = t_{(j)})$  and  $n_j = \sum_{i=1}^n I(y_i \geq t_{(j)})$ . Put  $\partial l / \partial h_i = 0$ , we have  $\hat{h}_i = d_i/n_i$  ( $i = 1, \dots, r$ ). This generalize nonparametric MLE is identical to the K-M estimator. (i.e  $\hat{S}(t) = \prod_{t_{(j)} < t} (1 - \hat{h}_j)$ ).

To see the covariance structure of  $\hat{S}(t)$  and  $\hat{S}(s)$ , we first consider the asymptotic properties of  $\tilde{h}$ . Since

$$\frac{\partial^2 l}{\partial h_i \partial h_j} = 0, \forall i \neq j$$

and

$$\left[ \frac{-\partial^2 l}{\partial h_i \partial h_i} \right]^{-1} = \frac{\hat{h}_i(1 - \hat{h}_i)}{n_i},$$

this implies  $(\hat{h}_1, \dots, \hat{h}_r)$  converges to  $N((h_1, \dots, h_r), \Sigma)$  in distribution where  $\Sigma = \text{diag}\{h_i(1 - h_i)/n_i\}$ . By delta method, we then have  $\text{Var}(\ln(1 - \hat{h}_i)) = h_i/[(1 - h_i)n_i]$ .

Suppose  $s < t$ . Let  $A = \sum_{t_{(i)} < s} \ln(1 - \hat{h}_i)$  and  $B = \sum_{s \leq t_{(i)} < t} \ln(1 - \hat{h}_i)$ . It is trivial that  $\text{cov}(A, B) = 0$ . Thus

$$\begin{aligned} \text{cov}(\ln \hat{S}(t), \ln \hat{S}(s)) &= \text{cov}(A + B, A) = \text{cov}(A, A) \\ &= \text{Var}(\ln \hat{S}(s)) = \sum_{t_{(i)} < s} \frac{d_i}{n_i(n_i - d_i)}. \end{aligned}$$

By delta method again, for any  $t, s$ , we have

$$\text{cov}(\hat{S}(t), \hat{S}(s)) = \hat{S}(t)\hat{S}(s) \sum_{t_{(i)} < s} \frac{d_i}{n_i(n_i - d_i)}.$$

PS: In fact,  $h_1$  will effect  $h_2$ , therefore, we need to say when  $n$  is large, then  $h_1, h_2$  are almost uncorrelated. That is,  $h_1, h_2$  are asymptotically uncorrelated.

#### 4.3.2 Redistribute the mass to the right algorithm.

- (1) Put the  $n$  observations in ascending order
- (2) Allocate mass  $1/n$  to each of them
- (3) Starting with the smallest observation, look for censored observations. Each time you encounter a censored observation, redistribute its mass eventually to all larger observations (censored and uncensored).

This algorithm will assign "mass" equals to the K-M estimator at each failure time.

Example:

$Y_{(i)}$	9	13	13+	18	23	28+	31	34	45+	48	161+
at start	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$	$\frac{1}{11}$
step 1	$\frac{1}{11}$	$\frac{1}{11}$	0	$\frac{1}{11} + \frac{1}{11*8}$	$\frac{9}{88}$	$\frac{9}{88}$	$\frac{9}{88}$	$\frac{9}{88}$	$\frac{9}{88}$	$\frac{9}{88}$	$\frac{9}{88}$
step 2	$\frac{1}{11}$	$\frac{1}{11}$	0	$\frac{9}{88}$	$\frac{9}{88}$	0	$\frac{9}{88} + \frac{9}{88*5}$	$\frac{27}{220}$	$\frac{27}{220}$	$\frac{27}{220}$	$\frac{27}{220}$
step 3	$\frac{1}{11}$	$\frac{1}{11}$	0	$\frac{9}{88}$	$\frac{9}{88}$	0	$\frac{27}{220}$	$\frac{27}{220}$	0	$\frac{81}{440}$	$\frac{81}{440}$
$\hat{S}(t)$	0.91	0.82		0.72	0.61		0.49	0.37		0.18	

### 4.3.3 Self-consistency

If we can observe the exact failure times  $T_i$ , then the expectation of the empirical distribution estimator

$$E[S^e(t)] = \frac{1}{n} \sum_{i=1}^n E[I(T_i \geq t)] = S(t)$$

Now, we observe only  $(Y_i, \delta_i)$ , the self consistent estimator given by Efron (1967) is defined by setting

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^n E[I(T_i \geq t) | Y_i, \delta_i].$$

Since

$$E[I(T_i \geq t) | Y_i, \delta_i = 1] = I(Y_i \geq t)$$

and

$$E[I(T_i \geq t) | Y_i, \delta_i = 0] = \begin{cases} \frac{S(t)}{S(Y_i)} & \text{if } t > Y_i \\ 1 & \text{if } t \leq Y_i < T_i \end{cases},$$

we have

$$\begin{aligned} \hat{S}(t) &= \frac{1}{n} \sum_{i=1}^n [\delta_i I(Y_i \geq t) + (1 - \delta_i) \frac{S(t)}{S(Y_i)} I(t < Y_i) + (1 - \delta_i) I(t \leq Y_i)] \\ &= \frac{1}{n} \sum_{i=1}^n [I(Y_i \geq t) + (1 - \delta_i) \frac{S(t)}{S(Y_i)} I(t < Y_i)]. \end{aligned}$$

Set

$$S^{(n)}(t) = \frac{1}{n} \sum_{i=1}^n [I(Y_i \geq t) + (1 - \delta_i) \frac{S^{(n-1)}(t)}{S^{(n-1)}(Y_i)} I(t < Y_i)]$$

where  $S^{(0)} = \frac{1}{n} \sum_{i=1}^n I(Y_i \geq t)$ .

Remark: K-M estimator (P-L estimator), NMLE, RDMTR and SC estimator provide the same result.

### 4.5 Left truncation

Suppose we observe  $(t_i, k_i)$  with  $t_i \geq k_i$  ( $i = 1, \dots, n$ ). If we just use  $t_i$  to estimate  $S(t)$ , it is biased. Similar as right-censored case, here we need only to redefine the number at risk at time  $t_{(i)}$  where  $t_{(1)} < t_{(2)} < \dots < t_{(r)}$  are  $r$  ordered distinct death time. The K-M estimator is still

$$\hat{S}(t) = \prod_{t_{(i)} < t} (1 - d_i/n_i),$$

however, the number at risk is being modified as

- $n_i$  = number at risk at time  $t_{(i)} = \sum_{j=1}^n I(k_j < t_{(i)} \leq t_j)$