Applied Survival Analysis Using R
Chapter 3: Nonparametric Survival Curve Estimation

Qi Guo

Department of Mathematical Sciences
The University of Texas at Dallas

April, 10 2019

Kaplan-Meier estimator

- In this chapter we will discuss non-parametric estimators of the survival function

Kaplan-Meier estimator

KM estimator is the product over the failure times of the conditional probabilities of surviving to the next failure time.

- Formally, it given by:

$$\hat{S}(t) = \prod_{t_i \leq t}(1 - \hat{q}_i) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i}) \tag{1}$$

- where n_i is the number of subjects at risk at time t_i, and d_i is the number of individuals who fail at that time, so $q_i = \frac{d_i}{n_i}$ is failure probability

Example

- The example data in `Table 1.1` may be used to illustrate the construction of the *KM estimate*, as shown in `Table 3.1`.

Table 3.1 Kaplan-Meier estimate

t_i	n_i	d_i	q_i	$1-q_i$	$S_i = \prod(1-q_i)$
2	6	1	0.167	0.833	0.833
4	5	1	0.200	0.800	0.666
6	3	1	0.333	0.667	0.444

Fig. 3.1 Right-continuous Kaplan-Meier survival function estimate

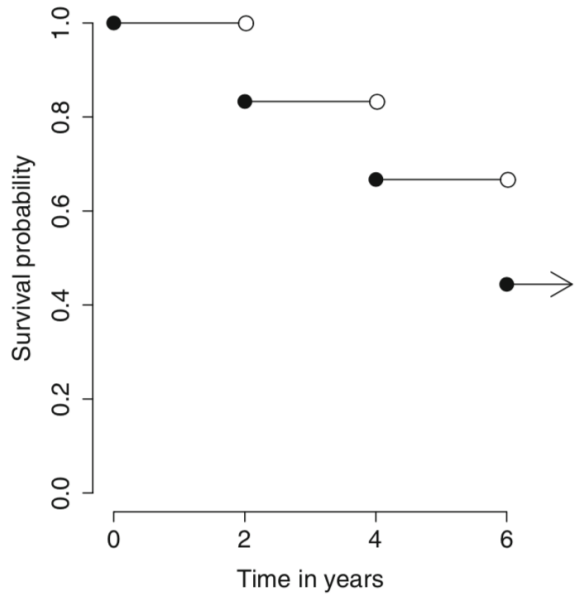


Table 1.1 Survival data

Patient	Survtime	Status
1	7	0
2	6	1
3	6	0
4	5	0
5	2	1
6	4	1

Get Variance By Delta Method

- To obtain confidence limits for the product-limit estimator, we first use what is known as the "**`delta method`**" to obtain the variance of $log(\hat{S}(t))$

$$var(\log(\hat{S}(t_k))) = \sum_{t_i \le t} var(\log(1 - \hat{q}_i)) \approx \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{2}$$

- To get the variance of $\hat{S}(t)$ itself, we use the delta method again to obtain:

$$var(\hat{S}(t_k)) \approx [\hat{S}(t)]^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \tag{3}$$

log-log transformation

- A more satisfying approach is to find confidence intervals for the complementary *log-log transformation* of $\hat{S}(t)$ as follows:

$$var(\log([-\log \hat{S}(t_k)]) \approx \frac{1}{[\log \hat{S}(t)]^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{4}$$

R

- To obtain estimates of the *Kaplan-Meier estimator* in `R` for the data in `Table 1.1`, we first load the "**`survival`**" library, and then enter the data

```
> library(survival)
> tt <- c(7,6,6,5,2,4)
> cens <- c(0,1,0,0,1,1)
> Surv(tt, cens)
[1] 7+  6   6+  5+  2   4
```

- For the estimation itself we use the "**`survfit`**" function

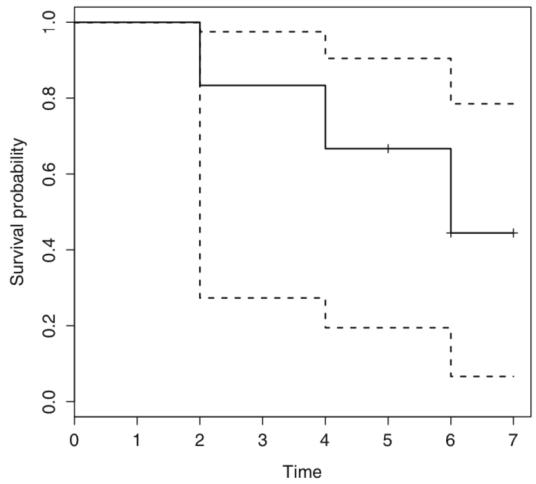
```
> result.km <- survfit(Surv(tt, cens) ~ 1, conf.type="log-log")
> result.km
[1] records   n.max   n.start   events   median   0.95LCL   0.95UCL
        6         6         6        3        6         2        NA
```

R

- To see the full *Kaplan-Meier estimate*, and plot it, we use the "**summary**" and "**plot**" functions:

```
> summary(result.km)
[1]
time    n.risk    n.event   survival    std.err    lower95%CI    upper95%CI
2          6          1       0.833        0.152        0.2731         0.975
4          5          1       0.667        0.192        0.1946         0.904
6          3          1       0.444        0.222        0.0662         0.785

> plot(result.km)
```



Nelson-Aalen Estimator

- Based on the relationship of $S(t)$ and $h(t)$.
- An estimate of the cumulative hazard function is the sum of the estimated hazards up to a time t_i:

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{5}$$

- and the *survival function* estimate is simply

$$S(t) = e^{-H(t)} \tag{6}$$

R

- *The Nelson-Aalen estimate* may be obtained using the "**survfit**" function with the option `type = "fh"`

```
> result.fh <- survfit(Surv(tt, cens) ~ 1, conf.type="log-log",
 type="fh")
> summary(result.fh)
[1]
time   n.risk   n.event   survival   std.err   lower95%CI   upper95%CI
  2        6         1      0.846     0.155       0.2401        0.981
  4        5         1      0.693     0.200       0.1799        0.925
  6        3         1      0.497     0.248       0.0585        0.841
```

Median and CI for Median

- Formally, the median survival time may be defined as: $\hat{t}_{med} = \inf\{t : \hat{S}(t) \leq 0.5\}$
- To find a $1 - \alpha$ confidence interval for the median:

$$-z_{\alpha/2} \leq \frac{g\{\hat{S}(t)\} - g(0.5)}{\sqrt{var[g\{\hat{S}(t)\}]}} \leq z_{\alpha/2} \tag{7}$$

- where $g(\mu) = \log[-\log(\mu)]$ and $var[g\{\hat{S}(t)\}]$ is given by equation (4)

Left Truncation

Left Truncation

Instead of examining the time from entry into the clinical trial until censoring or death, let us use as the time origin the time of diagnosis.

Table 3.3 Data from Table 1.1, with the addition of the time of diagnosis

Patient	Diagnosis	Survtime	Censor	SurvtimeDiag
1	−2	7	0	9
2	−5	6	1	11
3	−3	6	0	9
4	−3	5	0	8
5	−2	2	1	4
6	−5	4	1	9
X	−4	−2	1	

The time units are still the same, with time 0 indicating the time of entry into the trial and the time "Diagnosis" indicating the prior time of diagnosis. The new variable "SurvtimeDiag" denotes the time from diagnosis until censoring or death. The variables "Survtime" and "Censor" are as they were in Table 1.1. The new "Patient X" is a hypothetical patient with a short time from diagnosis until death. Practically speaking, such a patient is never observed; even if we somehow had a record of his diagnosis and early death, we could not possibly know for certain if that person would have entered the trial had he lived long enough. Such patients with short survival times are less likely to be enrolled in the trial than other patients, resulting in length-biased sampling

Figure

Fig. 3.8 Data from Table 1.1, now with diagnosis times

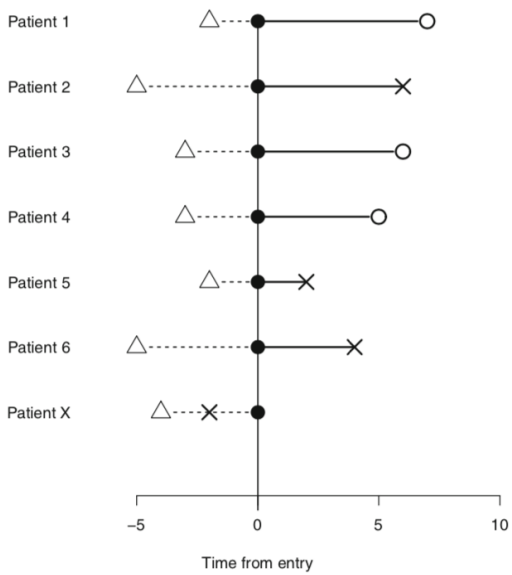
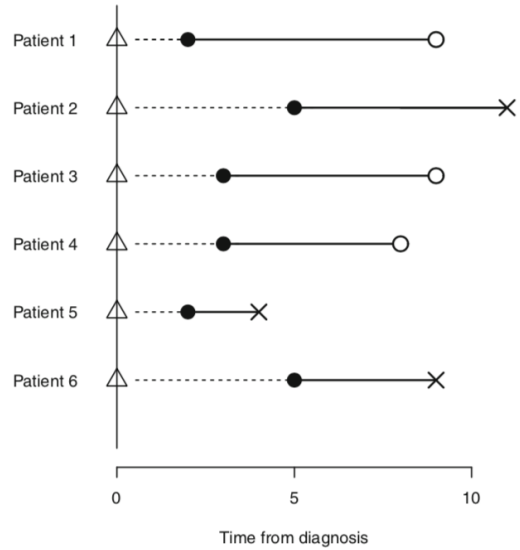


Fig. 3.9 Time from diagnosis to death. Entry into the clinical trial is denoted by *solid circles*. The *dashed lines* are "left truncation" times. Had the event occurred during these intervals, the patient would not have been observed



- We have used the terms "`tm.enter`" and "`tm.exit`" for the *left truncation* and *survival times*, respectively.

R

```
> tt <- c(7, 6, 6, 5, 2, 4)
> status <- c(0, 1, 0, 0, 1, 1)
> backTime <- c(-2, -5, -3, -3, -2, -5)
> tm.enter <- -backTime
> tm.exit <- tt - backTime
> result.left.trunc.km <- survfit(Surv(tm.enter, tm.exit, status,
type="counting") ~ 1,  conf.type = "none")
> summary(result.left.trunc.km)
[1]
time   n.risk   n.event  entered   censored   survival   std.err
 4         4         1        0         0        0.750       0.217
 9         4         1        0         2        0.562       0.230
11         1         1        0         0        0.000         NAN
> result.left.trunc.naa <- survfit(Surv(tm.enter, tm.exit, status,
type="counting") ~ 1, type="fleming-harrington", conf.
type="none")
> summary(result.left.trunc.naa)
[1]
time   n.risk   n.event  entered   censored   survival   std.err
4         4         1        0         0        0.779       0.225
9         4         1        0         2        0.607       0.248
11        1         1        0         0        0.223         Inf
```

Data

- A serious problem arises with *left-truncated data* if the risk set becomes *empty at an early survival time*.
- Consider for example the *Channing House data*, "`ChanningHouse`".
- This data is subject to *left truncation* because subjects who die at older ages are more likely to have enrolled in the center than patients who died at younger ages.

```
> head(ChanningHouse)
[1] sex     entry    exit     time    cens
1 Male       782      909      127       1
2 Male      1020     1128      108       1
3 Male       856      969      113       1
4 Male       915      957       42       1
5 Male       863      983      120       1
6 Male       906     1012      106       1
```

Transform data and using KM and NAA estimator

- Transform the data

```
>ChanningHouse <- within(ChanningHouse,
entryYears <- entry/12exitYears <- exit/12)
>ChanningMales <- ChanningHouse[ChanningHouse$sex == "Male",]
```

- KM estimator

```
>result.km <- survfit(Surv(entryYears, exitYears, cens,
type="counting") ~ 1, data=ChanningMales)
>plot(result.km, xlim=c(64, 101), xlab="Age",
ylab="Survival probability", conf.int=F)
```

- NAA estimator

```
>result.naa <- survfit(Surv(entryYears, exitYears, cens,
type="counting") ~ 1, type="fleming-harrington",
data=ChanningMales)
>lines(result.naa, col="blue", conf.int=F)
```

KM.68 and Plot

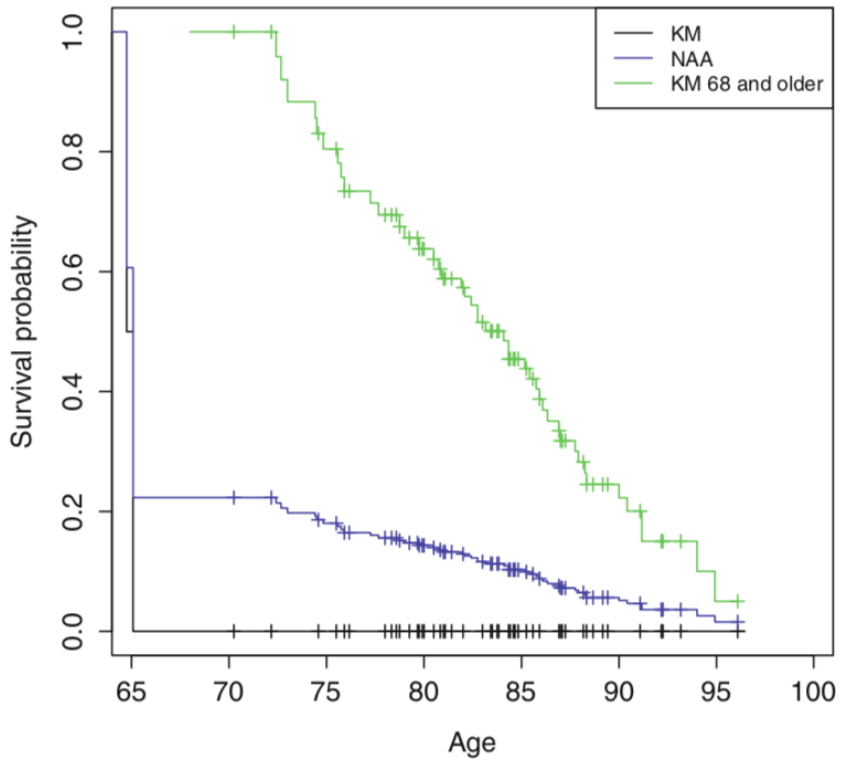
- Men reach the age of 68, using the "`start.time`" option:

```
>result.km.68 <- survfit(Surv(entryYears, exitYears, cens,
type="counting") ~ 1, start.time=68, data=ChanningMales)
```

- Plot

```
> lines(result.km.68, col="green", conf.int=F)
> legend("topright", legend=c("KM", "NAA", "KM 68 and older"),
lty=1, col=c("black", "blue", "green"))
```

Plot



- Apparently, *the KM 68 and older* is much better behaved.