KAPLAN-MEIER ESTIMATE

Dr. Mutua Kilai

Department of Pure and Applied Sciences

Jan-April 2024

Kaplan-Meier Estimator of Survival

- Let t_i denote an ordered observed value. The **emprical survivor function(esf)** denoted by $S_n(t)$ is defined by:

$$S_n(t) = \frac{Numberof \;\; observations \;\; > t}{n} = \frac{\{t_i > t\}}{n} \qquad (1)$$

- Example: Calculate the Emprical Survivor function for the following data:

9	13	13	18	23	28	31	34	45	48	161

Solution

t	0	9	13	18	23	28	31	34	45	48	161
$S_n(t)$	$\frac{11}{11}$	$\frac{10}{11}$	$\frac{8}{11}$	$\frac{7}{11}$	$\frac{6}{11}$	$\frac{5}{11}$	$\frac{4}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	0

```
library(survival)

aml <- c(9,13,13,18,23,28,31,34,45,48,16)

status <- rep(1,11)

esf.fit <- survfit(Surv(aml, status)~1)

plot(esf.fit,conf.int=F,xlab="time until relapse (in weeks)"
ylab="proportion without relapse",lab=c(10,10,7))
```

- The **esf** is a consistent estimator of the true survivor function $S(t)$.
- The exact distribution of $nS_n(t)$ for each fixed t is binomial (n, p) where $n =$ number of observations and $p = P(T > t)$

Kaplan-Meier Estimate

- The K-M adjusts the esf in order to reflect the presence of right-censored observations.
- Let $y_{(i)}$ denote the i^{th} distinct ordered censored or uncensored observation and is the right interval of the interval I_i
- Let $R(t)$ denote the **risk set just before time t** and let:
  - n_i denote the number in $R(y_{(i)})$ and the number alive and not censored just before $y_{(i)}$
  - d_i the number died at time $y_{(i)}$
  - p_i P(surviving through I_i|alive at beginning I_i)
  - $q_i = 1 - p_i =$ P(die in I_i|alive at beginning I_i)
- Recall the general multiplication of joint events A_1 and A_2

$$P(A_1 \cap A_2) = P(A_2|A_1)P(A_1)$$

Cont'd

- From this repeated rule the survivor function can be expressed as:

$$S(t) = P(T > t) = \prod_{y_{(i)} \leq t} p_i$$

- The estimates of p_i and q_i are:

$$\hat{q}_i = \frac{d_i}{n_i}$$

- and

$$\hat{p}_i = 1 - \hat{q}_i = 1 - \frac{d_i}{n_i}$$

- Formally

$$\hat{S(t)} = \prod_{y_{(i)} \leq t} \hat{p}_i = \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

where

- n_i the number of subjects at risk at time t_i

- d_i is the number of individuals who fail at time t_i

Advantages of K-M Estimate

- It is simple and straightforward to use and interpret
- it is a nonparametric estimator, so it constructs a survival curve from the data and no assumptions is made about the shape of the underlying distribution
- it gives a graphical representation of the survival function(s), useful for illustrative purposes

Example

Consider the following data and calculate the K-M estimate

subject	time	event
1	3	0
2	5	1
3	7	1
4	2	1
5	18	0
6	16	1
7	2	1
8	9	1
9	16	1
10	5	0

where

- subject: is the individuals identifier
- time: is the time to event (in years)
- event: is the event status ($0 =$ censored, $1 =$ event happened)

Solution

- We first need to count the number of distinct event times, ignoring censored observations we have 5 distinct event times.
- We make a table and fill as follows;
  - $y_{(j)}$ gives the ordered distinct event times
  - d_j gives the number of observations for each distinct event time
  - R_j gives the remaining number of individuals at risk. For this, the distribution of time (censored and not censored) is useful.

y_j	d_j	R_j	$1 - \frac{d_j}{R_j}$	$S_{KM}(t)$
2	2	10	0.800	0.800
5	1	7	0.857	0.686
7	1	5	0.800	0.548
9	1	4	0.750	0.411
16	2	3	0.333	0.317

- We can compute the KM estimator in R using the following.

We enter the data in R

```r
# create dataset
dat <- data.frame(
  time = c(3, 5, 7, 2, 18, 16, 2, 9, 16, 5),
  event = c(0, 1, 1, 1, 0, 1, 1, 1, 1, 0))
```

We then run the K-M estimator using the survfit() and Surv() functions as follows:

```r
# KM
library(survival)

km <- survfit(Surv(time, event) ~ 1,
  data = dat
)
```

- The Surv() function accepts two arguments:
  - the time variable
  - the event variable
- The $\sim$ in the survfit() function indicates that we estimate the Kaplan-Meier without any grouping.
- We can plot the K-M as follows

```
library(survminer)

# plot
ggsurvplot(km,
  conf.int = FALSE,
  legend = "none"
)
```

- The crosses on the survival curve denote the censored observations.
- The advantage with the ggsurvplot() function is that it is easy to draw the median survival directly on the plot:

```
ggsurvplot(km,
  conf.int = FALSE,
  surv.median.line = "hv",
  legend = "none"
)
```

Confidence Interval

- To obtain the confidence limits for the product limit estimator, we first use the delta method in order to obtain the variance of $\log(S(\hat{t}))$
- The delta method allows one to approximate the variance of a continuous variable $g(.)$ of a random variable.
- If a random variable X has mean μ and variance σ^2 then $g(X)$ will have approximate mean $g(\mu)$ and variance $\sigma^2 \times [g'(\mu)]^2$ for a sufficiently large sample size.
- Applying the delta's formula we get the variance of $\log S(\hat{t})$ as

$$var\left(\log \hat{S}(t_k)\right) = \sum_{t_i \leq t} var \log(1 - \hat{q}_i) \approx \sum_{t_i \leq t} \frac{d_j}{n_j(n_j - d_j)} \quad (2)$$

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

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

- Unfortunately, confidence intervals computed based on this variance may extend above one or below zero.
- A more satisfying approach is to find the confidence intervals for the complementary log-log transformation of $\hat{S}(t)$ as follows:

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

- Theory tells us that for each fixed value t

$$\hat{S}(t) \sim N(S(t), \hat{var}\left(\hat{S}(t)\right))$$

- Thus at time t an approximate $(1-\alpha)\times 100\%$ confidence interval for the probability of survival $S(t) = P(T > t)$ is given by:
$$\hat{S}(t) \pm z_{\frac{\alpha}{2}} s.e(\hat{S}(t))$$

Exercise

Find the K-M estimator for the following data (n = 21) and obtain the 95% CI for $S(t)$ when $t = 21$

$$6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+$$

Thank You!