Survival Models: Week 4

Kaplan-Meier Estimation of Survival Function

How might we estimate the survival function if we have censored data? Possible options include:

1. Remove the censored observations.
2. Treat censored observations as complete observations, that is, ignore the censoring.
3. Use Kaplan-Meier Estimation.

Approaches 1 and 2 are not ideal!

Why approach 1 & 2 are not ideal.

observe 100 people over 10 years.

10 deaths 10 left at time 5.

At time 10, you know the number of deaths is between 10 ~ 20, out of 100.

How to estimate $S(10)$?

① Remove the censored data.

loss of information.

10 censored lives to 5 at least (only out of 95)

② Ignore the censoring

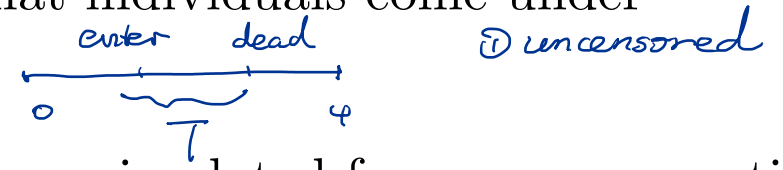
(assuming censored are all dead).

Then $S(10) = 0.8$, underestimated

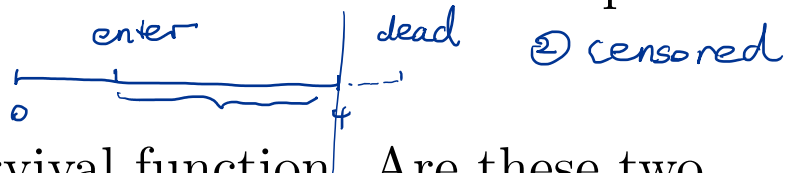
(or overestimated $F(10)$)

R Example

- Consider an example with 50 observations, some of which are censored.
- Investigation lasts for 4 years. The times that individuals come under observation are simulated from $U(0, 4)$.



- The future life times for these observations are simulated from an exponential distribution with mean 5.



- Use approaches 1 and 2 to estimate the survival function. Are these two methods appropriate?

R Example

```
#problems with (1) and (2) if censored observations
#investigation last for 4 years
set.seed(1)
enter<-runif(50,0,4)
uncensored<-rexp(50,0.2)
censored<-ifelse(uncensored+enter>4,4-enter,uncensored)
iscensored<-ifelse(uncensored+enter>4,1,0)
round(rbind(uncensored,censored,iscensored)[,1:5],1)
           [,1] [,2] [,3] [,4] [,5]
uncensored  5.0  7.2  0.2  1.6  6.6
censored    2.9  2.5  0.2  0.4  3.2
#using complete data
plot(ecdf(uncensored),verticals=T,main=NULL)
```

```
#treating censored as actual values
lines(ecdf(censored),verticals=T,col="red")
#removing censored values
lines(ecdf(censored[-which(iscensored==1)]),verticals=T,col="blue")
#actual exp(0.2) cdf
lines(ecdf(rexp(1000,0.2)),verticals=T,col="green")
```

R Example

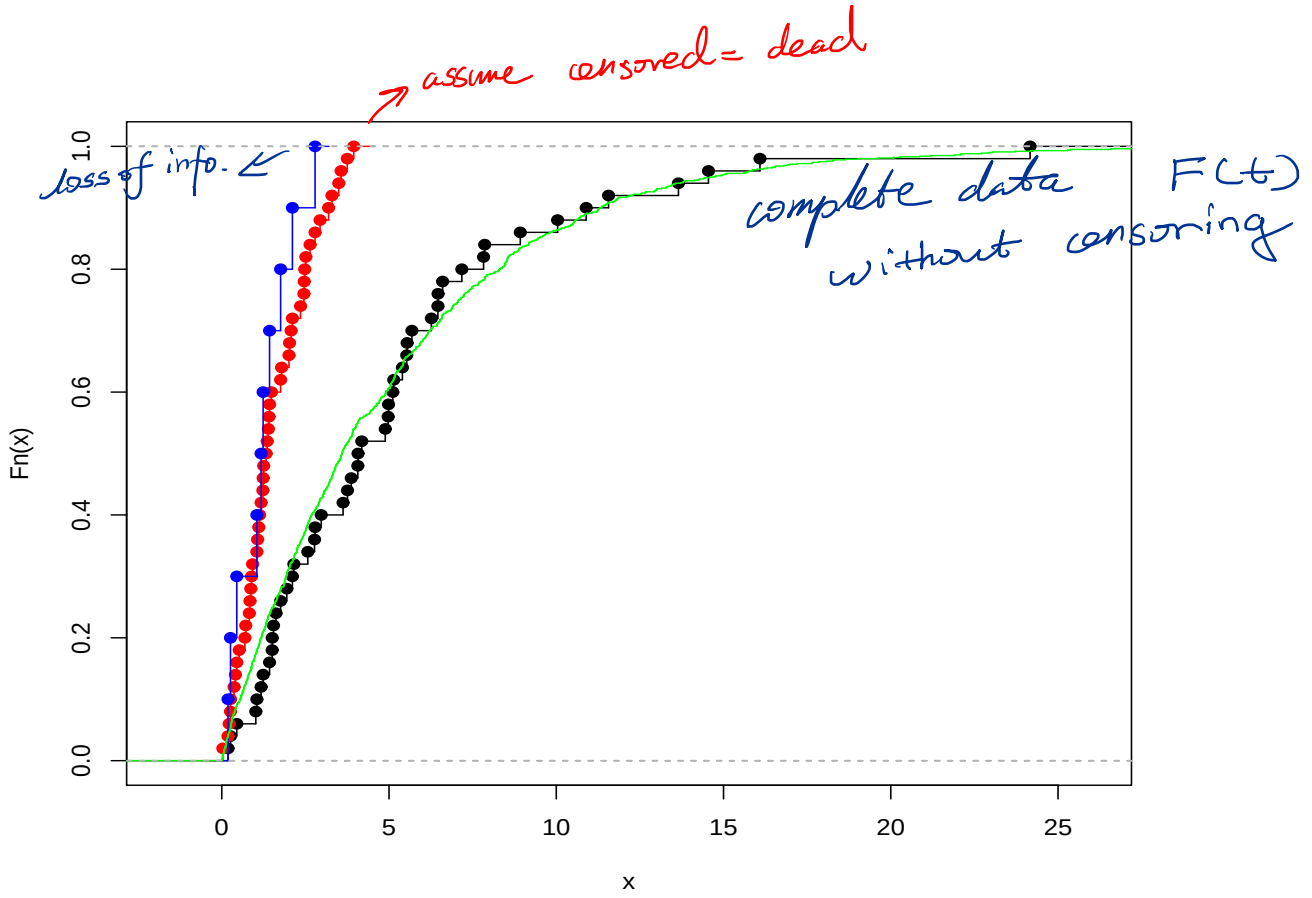


Figure 1: Different methods of estimating $F(t)$.

Kaplan-Meier Estimation of Survival Function

Some notation will be needed in order to define the Kaplan-Meier (KM) estimator:

- N: number of individuals under investigation.
- m: the total number of observed deaths among the N individuals.
- $t_1 < t_2 < ... < t_k$ are the k distinct time of deaths. $[k = m$ if no tied deaths$]$.
- d_j: number of deaths at time t_j. $[d_1 + d_2 + ... + d_k = m]$.
- r_j: number of individuals alive at time t_j. (under investigation)

Kaplan-Meier - Heuristic Derivation

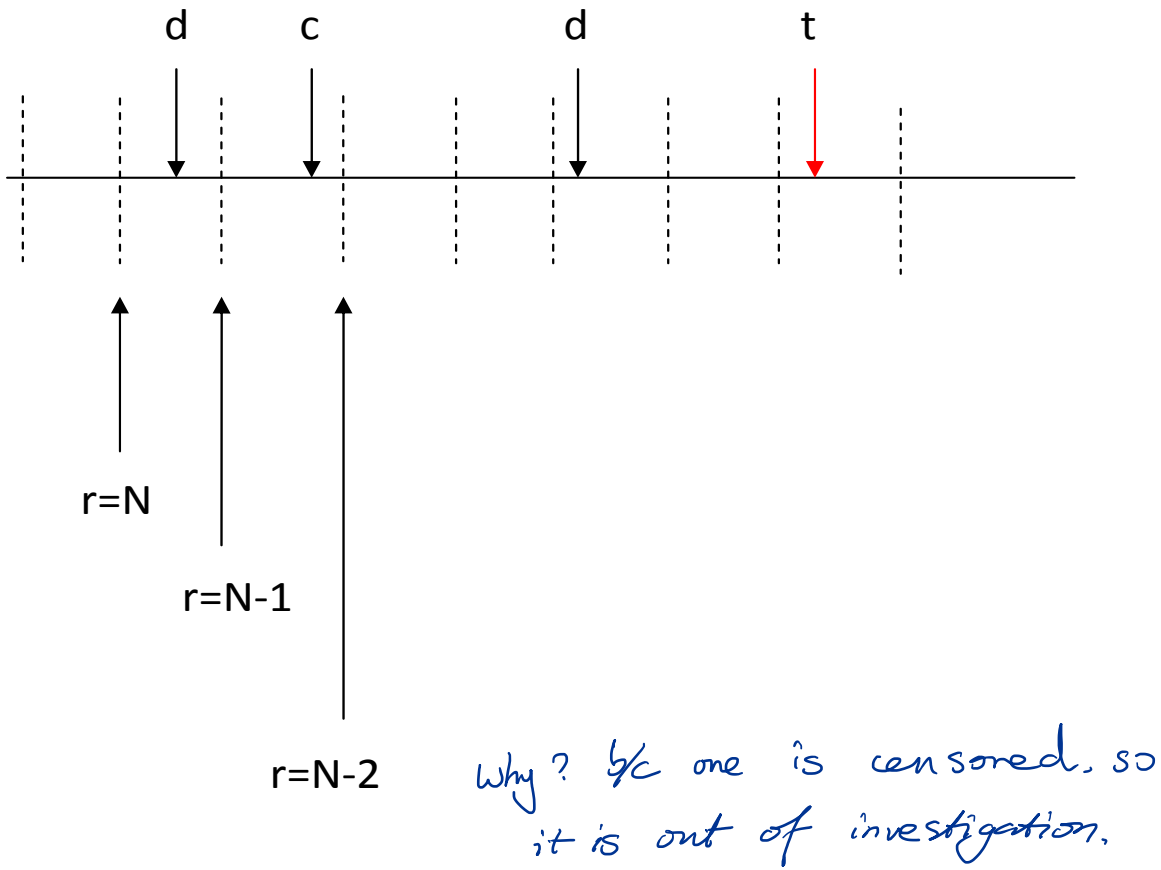


Figure 2: Schematic of observed deaths (d) and censored observations (c).

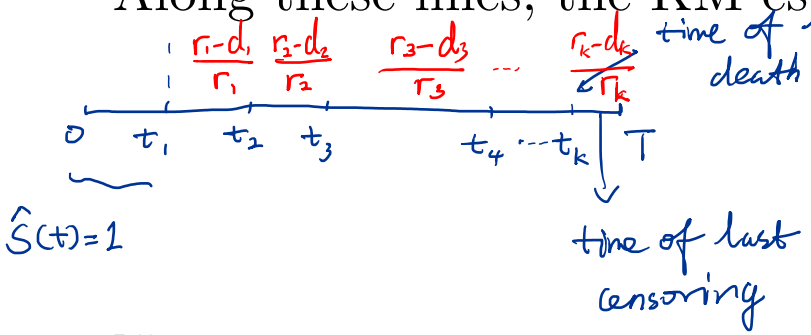
Kaplan-Meier - Heuristic Derivation

- Split interval into "small" segments of time - only one event can occur in each segment. An individual can be dead, alive or censored.
- Estimate the probability of surviving each segment of time as 1 if no death occurs and $\frac{r-1}{r}$ if a death occurs. one death, so remained # r minus 1.
- For example we could estimate the probability of surviving to time t, in the previous slide as $1 \times \frac{N-1}{N} \times 1 \times 1 \times 1 \times \frac{N-3}{N-2}$.
- The probability of surviving a given sequence of segments that contain w distinct times of death can then be estimated as:

$$\prod_{i=1}^{w} \frac{r_i - 1}{r_i}.$$

Kaplan-Meier - Heuristic Derivation

Along these lines, the KM estimate of the survival function is given by:



$$\hat{S}(t) = \prod_{t_j \leq t} \frac{r_j - d_j}{r_j}.$$

Notes:

- In the above we have d_j and not 1 because in practice we may observed tied deaths.
- Estimator is 1 before first observed time of death, that is, $\hat{S}(t) = 1$ for $t < t_1$.
- If largest time under observation is not a death $\hat{S}(t)$ will not "fall" to zero, that is, $\hat{S}(t) > 0$ for $t > t_k$. Usually set $\hat{S}(t) = 0$ for $t > t_k$ or at the time of the last censoring.

Kaplan-Meier - Example

if you have a death & a censored at the same time, we assume death comes first & include the censored one.

The observed survival times (or times of censoring) of $N = 7$ individuals after a particular operation (in months) were: $8, 12, 12*, 17, 22, 27*, 30$. The "$*$" correspond to censored observations. The KM estimate of the survival function is:

t_j	r_j	$\frac{r_j - d_j}{r_j}$	$\hat{S}(t) = \prod_{t_j \le t} \frac{r_j - d_j}{r_j}$
8	7	$\frac{7-1}{7}$	$\frac{7-1}{7}$
12	6	$\frac{6-1}{6}$	$\frac{7-1}{7} \times \frac{6-1}{6}$
17	4	$\frac{4-1}{4}$	$\frac{7-1}{7} \times \frac{6-1}{6} \times \frac{4-1}{4}$
22	3	$\frac{3-1}{3}$	$\frac{7-1}{7} \times \frac{6-1}{6} \times \frac{4-1}{4} \times \frac{3-1}{3}$
30	1	$\frac{1-1}{1}$	$\frac{7-1}{7} \times \frac{6-1}{6} \times \frac{4-1}{4} \times \frac{3-1}{3} \times \frac{0}{1}$

the first time of death

$\hat{S}(t) = 1$, $t < 8$

$\hat{S}(t) = \frac{6}{7}$, $8 \le t < 12$

KM estimate only changes when there is a death.

$r_j = 6$ $\quad \hat{S}(t) = \frac{6-1}{6} \times \frac{6}{7} = \frac{5}{7}$, $12 \le t < 17$

$r_j = 4$ $\quad \hat{S}(t) = \frac{4-1}{4} \times \frac{5}{7} = \frac{15}{28}$, $17 \le t < 22$

$\hat{S}(t) = \frac{6}{7} \cdot \frac{5}{6} \cdot \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{0}{1} = 0$ $\quad 30 \le t < \infty$

- What if the last one is censored instead of death?

  then: $\hat{S}(t) = \cdots \cdot \frac{1}{1} > 0$ leave it as this

  But, conventionally, we set it as 0.

R - Example

The R package "survival" implements KM estimation (and Cox Regression). We will analyse the "lung" dataset available in this package. This dataset contains information on the survival time of patients with advanced lung cancer. In addition to survival there is information on age, sex, and other variables.

```
> library(survival)
> data(lung)
> lung[1:3,]
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
```

Note: status=1 corresponds to a censored observation and status=2 to death. sex=1 corresponds to Male.

R - Example

```
#one curve for both males and females
censored<-ifelse(lung$status==2,1,0)
km.est<-survfit(Surv(lung$time,censored)~1,conf.type="plain")
plot(km.est,ylab="survival function",xlab="time")
#different curves for males and females
km.est1<-survfit(Surv(lung$time,censored)~lung$sex)
plot(km.est1,ylab="survival function",xlab="time")
```

R Example

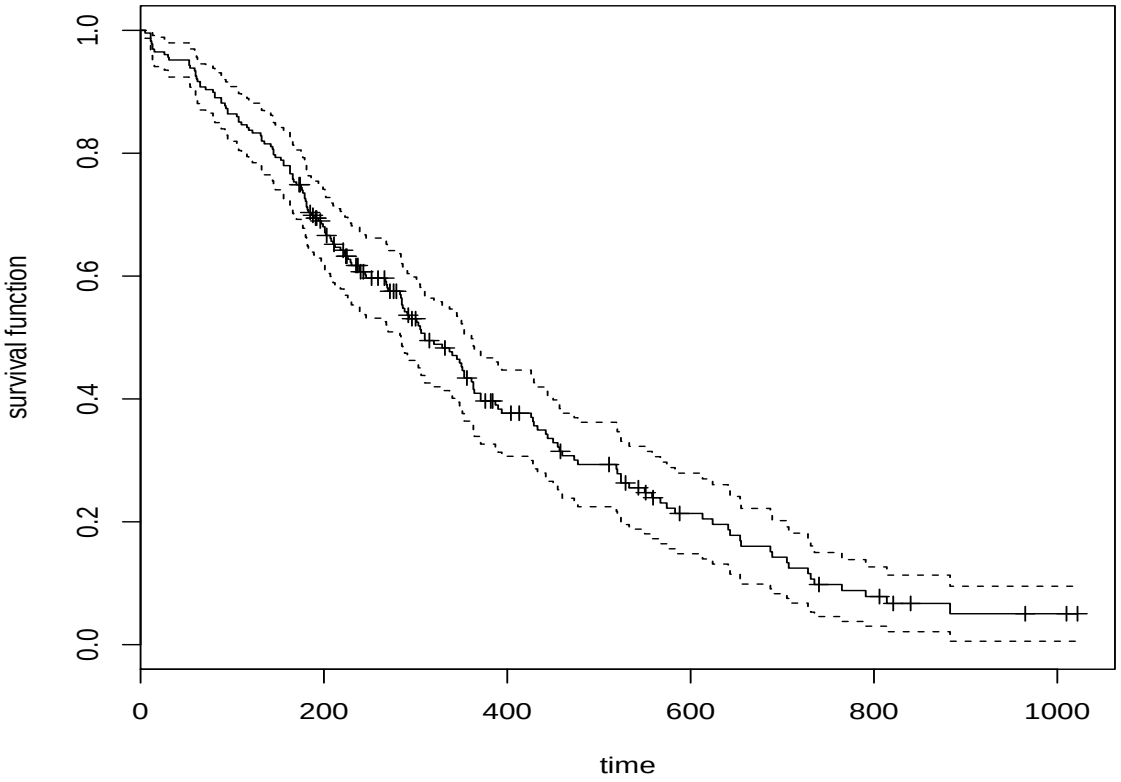


Figure 3: KM curve for lung data.

R Example

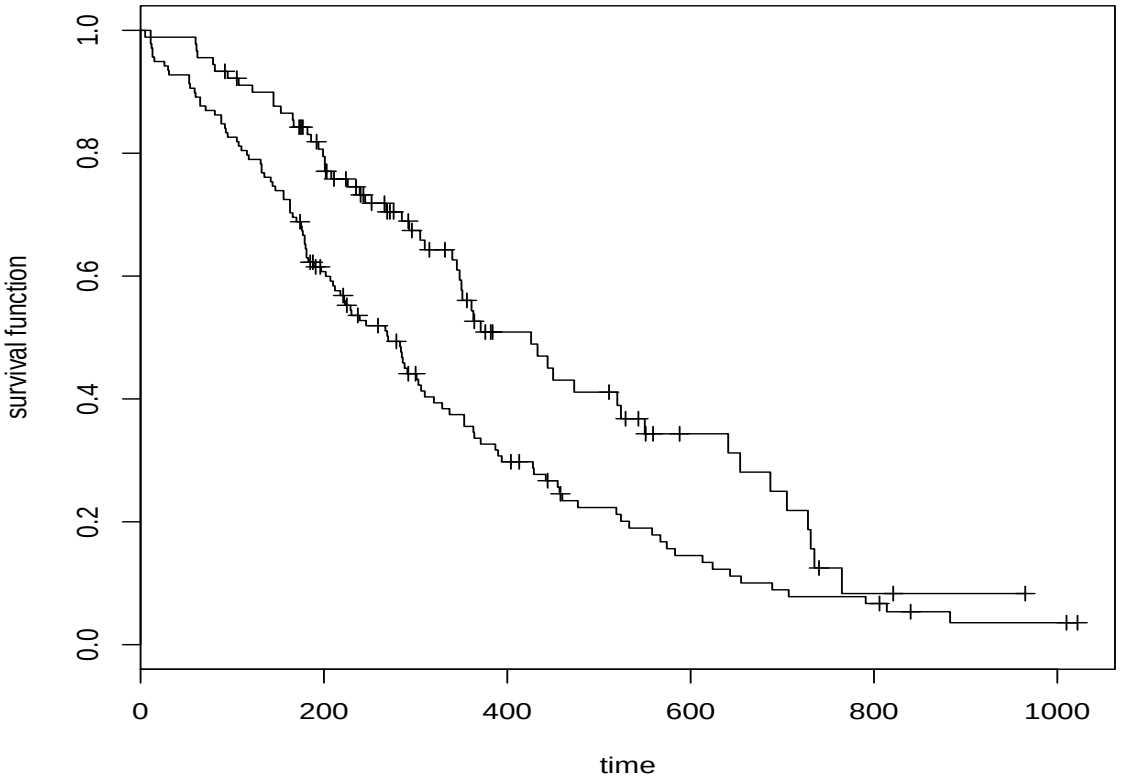


Figure 4: KM curve for lung data - by sex. Higher curve is for females.

R Example

```
> summary(km.est)
Call: survfit(formula = Surv(lung$time, censored) ~ 1,
conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    228       1   0.9956 0.00438      0.98704       1.0000
   11    227       3   0.9825 0.00869      0.96541       0.9995
   12    224       1   0.9781 0.00970      0.95906       0.9971
   13    223       2   0.9693 0.01142      0.94691       0.9917
.
.
.
```

R Example

```
#investigation last for 30 years
set.seed(1)
enter<-runif(50,0,30)
uncensored<-rexp(50,0.1)
censored<-ifelse(uncensored+enter>30,30-enter,uncensored)
iscensored<-ifelse(uncensored+enter>30,0,1)
#complete data
km.est<-survfit(Surv(uncensored)~1)
plot(km.est,ylab="Survival function",xlab="time")
#censored data using KM
km.est1<-survfit(Surv(censored,iscensored)~1)
lines(km.est1,col="red")
#treading censored values as actual values
km.est2<-survfit(Surv(censored)~1)
lines(km.est2,col="blue")
```

R Example

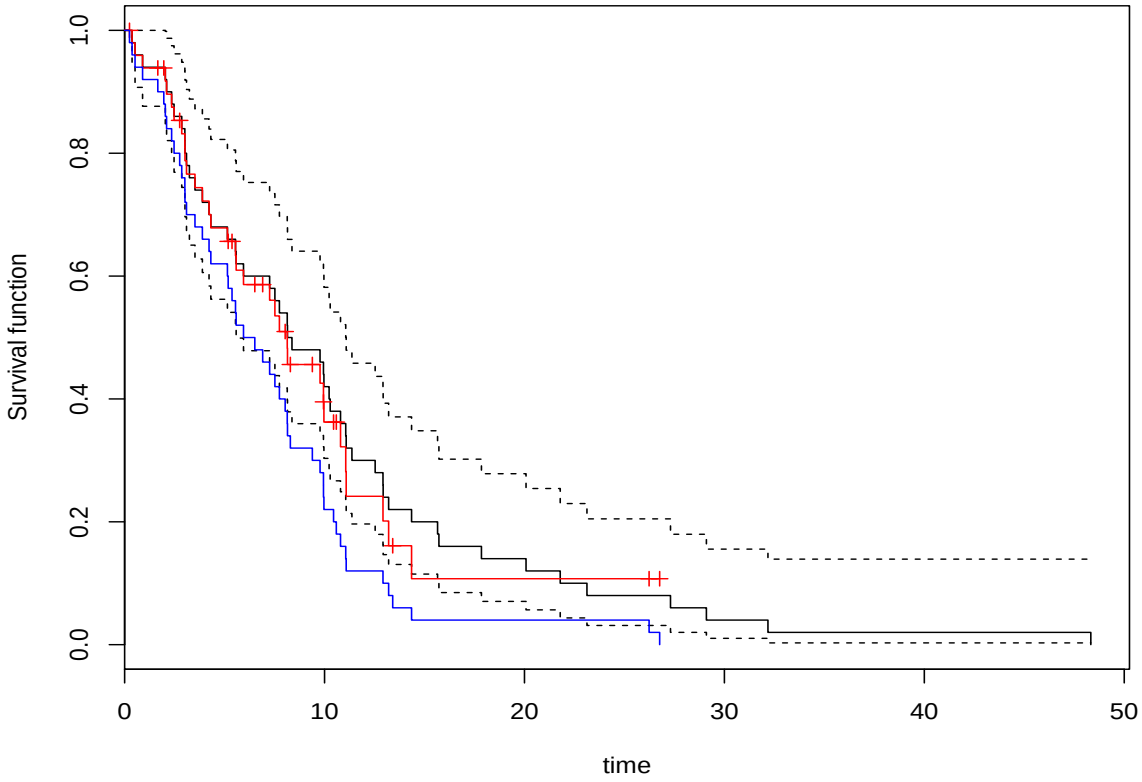


Figure 5: Showing that KM curve "works" for censored data

The Delta Method

The delta method provides a means of approximating the variance of a function of a random variable X, denoted $g(X)$. The random variable X has mean μ and variance σ^2. To use the delta method we approximate $g(X)$ by the first two terms of a Taylor series expansion about μ:

$$g(X) \approx g(\mu) + (X - \mu)g'(\mu)$$

Using the above we have:

$$var(g(X)) = var(X - \mu)(g'(\mu))^2 = \sigma^2(g'(\mu))^2$$

The Delta Method

An example, consider the function of the random variable X, $exp(X)$. Using the delta method the variance of $exp(X)$ can be approximated as:

$$var(exp(X)) \approx \sigma^2 exp(2\mu)$$

Notes:

- To use delta method the function $g(X)$ must be "smooth" - can be approximated by a Taylor series.
- The delta method assumes that the function is approximately linear over the expected range of X.

According to Delta Method

$$E[g(x)] \approx g(\mu)$$

if a question asks you to find expectation using Delta Method.

Variance of KM

based on the Delta method.

We will first consider the variance of:

Why to take log?
① multiplication ⇒ summation ⇒ var(sums) = sum(var)

$$log[\hat{S}(t)] = \sum_{t_j \le t} log(\frac{r_j - d_j}{r_j}) = \sum_{t_j \le t} log(\hat{p}_j).$$

We will *assume* that the number of survivors among the r_j at risk is $\text{Bin}(r_j, p_j)$, where p_j is estimated by $\hat{p}_j$. This means that $\hat{p}_j$ has variance $\frac{1}{r_j}\hat{p}_j(1-\hat{p}_j)$. Using this fact and the delta method we can estimate the variance of $log(\hat{p}_j)$ by:

$$var[log(\hat{p}_j)] = \frac{1}{\hat{p}_j^2}\frac{1}{r_j}\hat{p}_j(1-\hat{p}_j) = \frac{d_j}{r_j(r_j - d_j)}$$

Variance of KM

Now assuming that what happens for each group of at risk individuals (the r_j) are independent we have the following:

$$var\{log[\hat{S}(t)]\} = \sum_{t_j \leq t} var[log(\hat{p}_j)] = \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}$$

Using the above result, we can obtain an estimate of the variance of KM by noting that the KM is equal to "exp()" of the quantity above. This implies that:

$$var[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}$$

This is known as Greenwood's formula.

Greenwood's formula.

$$\log(\hat{S}(t)) = \sum_{t_j \le t} \log \frac{r_j - d_j}{r_j} = \sum_{t_j \le t} \log(\hat{p}_j)$$

$$\hat{p}_j = \frac{r_j - d_j}{r_j} \qquad d_j \sim \text{Bin}(r_j, 1-p_j)$$

d_j: # of deaths; $1-p_j$: prob of surviving; r_j: # of trials

$$E(\hat{p}_j) = 1 - E\left(\frac{d_j}{r_j}\right) = 1 - (1-p_j) = p_j \qquad \text{unbiased estimator}$$

$$\text{Var}(\hat{p}_j) = \text{Var}\left(\frac{d_j}{r_j}\right) = \frac{1}{r_j^2} r_j (1-p_j) \cdot p_j$$

$$\approx \frac{1}{r_j} \hat{p}_j (1-\hat{p}_j)$$

By delta method

$$\text{Var}[\log(\hat{p}_j)] = \frac{1}{\hat{p}_j^2} \cdot \frac{1}{r_j} \hat{p}_j (1-\hat{p}_j) = \frac{d_j}{r_j (r_j - d_j)} \qquad \text{(Delta method)}$$

square of deriv of $\log(\hat{p}_j)$ → $\frac{1}{\hat{p}_j^2}$

Var of $\hat{p}_j$ → $\frac{1}{r_j} \hat{p}_j (1-\hat{p}_j)$

$$\text{Var}[\log(\hat{S}(t))] = \text{Var}\left[\sum_{t_j \le t} \log(\hat{p}_j)\right] = \sum_{t_j \le t} \frac{d_j}{r_j (r_j - d_j)}$$

Using above results

$$\text{Var}(\hat{S}(t)) = \text{Var}\left(e^{\log(\hat{S}(t))}\right)$$

$$= (\hat{S}(t))^2 \cdot \text{Var}[\log \hat{S}(t)]$$

$$= (\hat{S}(t))^2 \cdot \sum_{t_j \le t} \frac{d_j}{n_j(r_j - d_j)}$$

Constructing CI for $S(t)$

① $\hat{S}(t) \pm 1.96$ s.d.$(\hat{S}(t))$

or

2. based on assumption that $\hat{S}(t) \sim$ Normal → not always. (though not that appropriate)

only holds when s.d. is very small.

② we are interested in ~~the~~ $\log(\hat{S}(t))$ (better way to construct CI)

$\log(\hat{S}(t)) \pm 1.96$ s.d.$(\log \hat{S}(t))$

→ $\log \hat{S}(t) \sim$ Normal

CI for $\log S(t)$

$\log(\hat{S}(t)) = \sum_{t_j \le t} \log(\hat{p}_j)$ → $\sim$ Normal

if t_j is ↑, i.i.d. apply CLT

⇒ normal distribution.

Back transform CI for $\log S(t)$

$\Rightarrow$ CI for $S(t)$

(e^{min}, e^{max})

↓

can guarantee > 0

but cannot guarantee ≤ 1

sometime we find $e^{max} > 1$,

we manually set it to 1.

Aside: The Bootstrap

Example: Based on the following sample of size $n = 7$ from a particular population $X_1 = 9, X_2 = 10, X_3 = 22, X_4 = 7, X_5 = 13, X_6 = 13, X_7 = 7$ we wish to estimate the quantity $exp(\mu/10)$, where μ is the unknown population mean. We would also like to assign a standard error to our estimate and/or produce a 95% confidence interval.

- Our estimate would be $exp(\bar{x}/10)$.
- But, how do we assign a standard error to our estimate $exp(\bar{x}/10)$?
- The delta-method is one approach but may not help us to construct a 95% confidence interval.
- Another approach is to use the Bootstrap.

The Bootstrap works by taking repeated samples (with replacement) from the observed $n = 7$ observations. Some possible Bootstrap samples are:

1. $X_3, X_3, X_7, X_4, X_3, X_3, X_1$
2. $X_1, X_2, X_4, X_7, X_6, X_6, X_5$.

Typically, we will take $B = 1000$ or more Bootstrap samples. For each of these samples we then recompute our original statistic, in this case $exp(\bar{x}/10)$. For example for the first sample above we would exponentiate the mean of the values $X_3, X_3, X_7, X_4, X_3, X_3, X_1$. This value is denoted $S(X)_1^*$, the first Bootstrap replicate. There will be B Bootstrap replicates.

Finally, the Bootstrap estimate of the variance is simply the sample variance of the B Bootstrap replicates. A 95% Bootstrap confidence interval can be obtained by taking the 2.5% and 97.5% percentiles of the B Bootstrap replicates.

```
> set.seed(123)
> x<-c(9,10,22,7,13,13,7)
> exp(mean(x)/10) # estimate
[1] 3.180832
>
```

```
> #using Boostrap
> res<-rep(0,1000) #B=1000
>
> for(i in 1:1000) {
+
+ xsamp<-sample(x,length(x),replace=TRUE)
+ res[i]<-exp(mean(xsamp)/10)
+
+ }
>
> #bootstrap estimate of standard error
> sqrt(var(res)) # estimate of standard error
[1] 0.6356967
>
> #bootstrap 95% CI
> quantile(res,c(0.025,0.975))
2.5%    97.5%
2.390323 4.746951
```

A histogram of the Bootstrap replicates can also provide information about the sampling distribution.

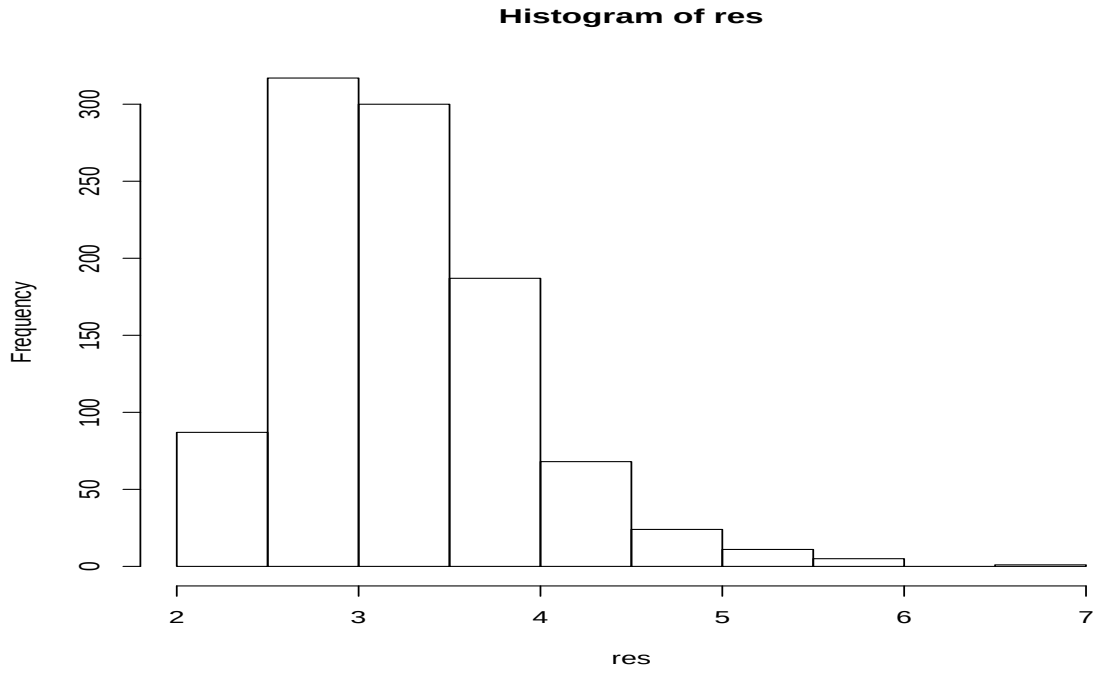


Figure 6: Bootstrap replicates

Appears slightly skewed. The "standard" est $\pm$ approach does not allow for this!

The Bootstrap is easy to apply no matter how complicated the statistic.