PSTAT 175 LAB B

Yanjie Qi

2019/10/20

library(survival)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.1 ✔ dplyr 0.8.3  
## ✔ tidyr 1.0.0 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

Import the datasets:

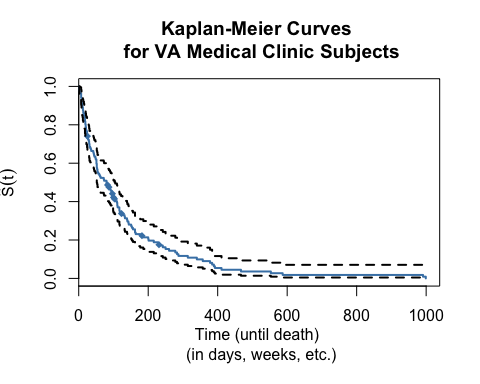
vets <- read.table("~/Desktop/Fall 2019/PSTAT 175/LAB B/vets.txt", quote="\"", comment.char="")  
heroin <- read.table("~/Desktop/Fall 2019/PSTAT 175/LAB B/heroin.txt", quote="\"", comment.char="")

## 1

# create vectors  
vet.time <- vets$V1  
vet.cns <- vets$V2  
vet.Surv <- Surv(vet.time, vet.cns)  
vet.fit <- survfit(vet.Surv ~ 1, conf.int=0.95)

## a)

# plot the kaplan-meier estimator with 95% CI  
plot(vet.fit,main="Kaplan-Meier Curves \n for VA Medical Clinic Subjects",  
 xlab="Time (until death) \n (in days, weeks, etc.)",  
 ylab=expression(hat(S)(t)),col=c("steelblue","black","black"),lwd=2,  
mark.time = TRUE,mark=18)



## b)

## calculate estimate of the 75th, 50th, and 25th percentiles  
x<-min(vet.fit$time[vet.fit$surv < .75])  
y<-min(vet.fit$time[vet.fit$surv < .50])  
z<-min(vet.fit$time[vet.fit$surv < .25])  
print(c("75th percentile =", x, "50th percentile =", y, "25th percentile =", z))

## [1] "75th percentile =" "25" "50th percentile ="  
## [4] "80" "25th percentile =" "162"

## 2

load the dataset

data(lung, package = "survival")  
attach(lung)

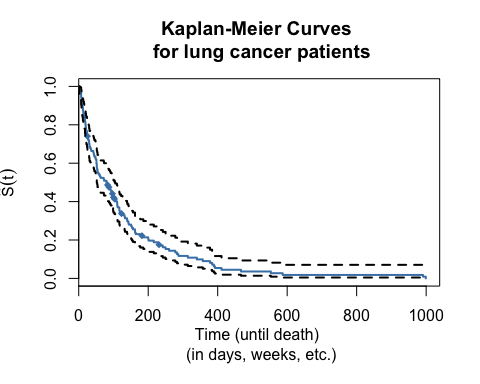
## a)

create the vectors:

lung.time <- lung$time  
lung.cts <- lung$status  
lung.surv <- Surv(lung.time, lung.cts)  
lung.fit <- survfit(lung.surv ~ 1, conf.int = 0.95)

Plot the Kaplan-Meier estimator:

plot(vet.fit,main="Kaplan-Meier Curves \n for lung cancer patients",  
 xlab="Time (until death) \n (in days, weeks, etc.)",  
 ylab=expression(hat(S)(t)),col=c("steelblue","black","black"),lwd=2,  
mark.time = TRUE,mark=18)



## b)

find the time near 150 days:

t<-max(lung.fit$time[lung.fit$time<150])  
t

## [1] 147

Get the estimate and confidence interval:

# by using summary function  
summary(lung.fit, times=t)

## Call: survfit(formula = lung.surv ~ 1, conf.int = 0.95)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 147 180 47 0.793 0.0269 0.742 0.848

## c)

Find the median survival time:

tm <- median(lung.fit$time)  
tm

## [1] 274

Get the estimate and confidence interval:

summary(lung.fit, times=tm)

## Call: survfit(formula = lung.surv ~ 1, conf.int = 0.95)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 274 106 93 0.575 0.0338 0.513 0.645

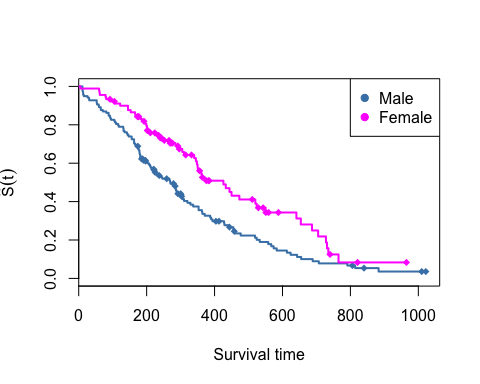
## d)

# Set Up the Vector  
lu.s = survfit(lung.surv~sex,data=lung)  
lu.s

## Call: survfit(formula = lung.surv ~ sex, data = lung)  
##   
## n events median 0.95LCL 0.95UCL  
## sex=1 138 112 270 212 310  
## sex=2 90 53 426 348 550

Plot the seperated estimators:

plot(lu.s,xlab="Survival time",ylab = expression(hat(S)(t)),lwd=2,  
 col=c("steelblue","magenta"), mark.time = TRUE,mark=18)  
legend("topright",legend=c("Male","Female"),  
 col=c("steelblue","magenta"),pch=rep(19,2))

 Generally Speaking, women have better survival rates. It is consistent except in the time around 750, the survival rate is almost the same.

## e)

Get the estimate and confidence interval:

summary(lu.s, times = tm)

## Call: survfit(formula = lung.surv ~ sex, data = lung)  
##   
## sex=1   
## time n.risk n.event survival std.err   
## 274.0000 58.0000 68.0000 0.4937 0.0436   
## lower 95% CI upper 95% CI   
## 0.4152 0.5870   
##   
## sex=2   
## time n.risk n.event survival std.err   
## 274.0000 48.0000 25.0000 0.7045 0.0501   
## lower 95% CI upper 95% CI   
## 0.6129 0.8098

From these intervals, we could conclude that female survival rate is still higher than male’s, since, in the end, female have 0.4937 survival rate with CI (0.4152,0.5870), but as for male, they have 0.7045 survival rate with CI (0.6129,0.8098). Accordingly, female sample indeed have higher survival rate. It might not be the whole story, since the plot in part(d) tells us there is a certain time, male and female samples have the identical survival rate, so the female survival rate is not always higher than male survival rate.

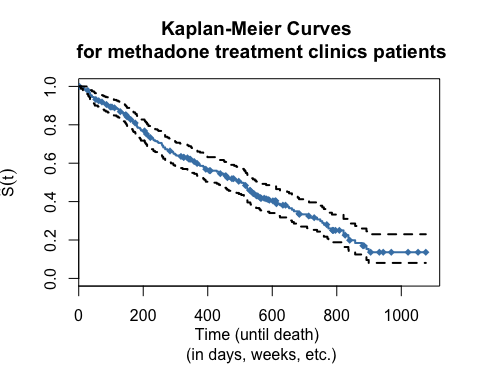
## 3

## a)

# create the vector needed  
her.surv <- Surv(heroin$Time, heroin$Status)  
her.fit <- survfit(her.surv ~ 1, conf.int = 0.95)

Plot the Kaplan-Meier Estimate with 95% CI:

plot(her.fit,main="Kaplan-Meier Curves \n for methadone treatment clinics patients",  
 xlab="Time (until death) \n (in days, weeks, etc.)",  
 ylab=expression(hat(S)(t)),col=c("steelblue","black","black"),lwd=2,  
mark.time = TRUE,mark=18)



## b)

extract mj and nj from the survfit function output:

mj = her.fit$n.event  
nj = her.fit$n.risk

calculate mj/(nj∗(nj−mj)) and the cumulative sum:

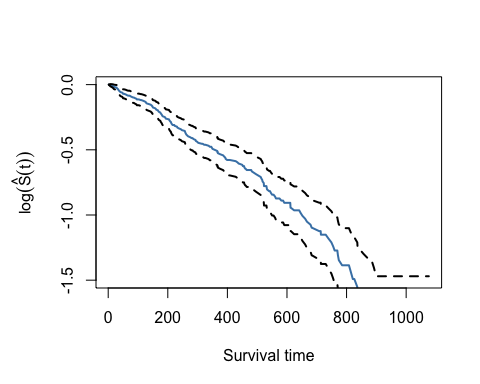
Vj = mj/nj/(nj-mj)  
cVj = cumsum(Vj)

Then calculate the 95% CI:

lowerCI = log(her.fit$surv) - 1.96\*sqrt(cVj)  
upperCI = log(her.fit$surv) + 1.96\*sqrt(cVj)

Plot the estimate of S(t):

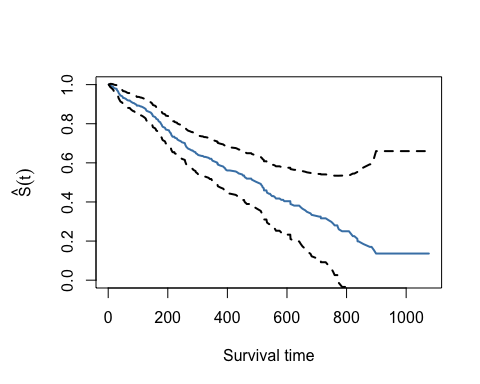
par(mar=c(5,5,4,2))  
plot(her.fit$time,log(her.fit$surv),lwd=2,type="l",ylim=c(-1.5,0),  
xlab="Survival time",ylab=expression(log(hat(S)(t))), col="steelblue")  
lines(her.fit$time,lowerCI,lty=2,col=1,lwd=2)  
lines(her.fit$time,upperCI,lty=2,col=1,lwd=2)



## c)

Transform the plot from part b:

lowerCI = her.fit$surv - 1.96\*sqrt(cVj)  
upperCI = her.fit$surv + 1.96\*sqrt(cVj)  
par(mar=c(5,5,4,2))  
plot(her.fit$time,her.fit$surv,lwd=2,type="l",ylim=c(0,1),  
xlab="Survival time",ylab=expression(hat(S)(t)), col="steelblue")  
lines(her.fit$time,lowerCI,lty=2,col=1,lwd=2)  
lines(her.fit$time,upperCI,lty=2,col=1,lwd=2)

 We could see plot in part c has no differences with the plot in part b.

## d)

we could get for one year what percentage of patients are in the clinic from summary:

summary(her.fit, times=365)

## Call: survfit(formula = her.surv ~ 1, conf.int = 0.95)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 365 122 87 0.606 0.0331 0.545 0.675

we should use two-sided hypothesis. Hypothesis Test: H0: S >= 50% vs. Ha: S < 50%

t = max(her.fit$time[her.fit$time<365])  
her.fit$surv[her.fit$time==t]

## [1] 0.6060647

test statistics：

z = (log(her.fit$surv[her.fit$time==t])-log(0.5))/her.fit$std.err[her.fit$time==t]  
z

## [1] 3.524093

As shown above, the test statistic is greater than the critical value 1.96, so we reject H0 and conclude that less than 50% are being discharged in one year.

Then, we coud compute the p-value:

pnorm(-z,mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

## [1] 0.0002124677

So, the p-value is 0.0002124677. ## e)

q1 <- min(her.fit$time[her.fit$lower<0.3])   
q2 <- max(her.fit$time[her.fit$upper>0.3])   
c(q1,q2)

## [1] 661 826

The time interval for 70th percentile is 661 826.

summary(her.fit, times = c(q1,q2))

## Call: survfit(formula = her.surv ~ 1, conf.int = 0.95)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 661 46 130 0.357 0.0359 0.294 0.435  
## 826 18 14 0.225 0.0368 0.163 0.310

Then, for 80th percentile:

n1 <- min(her.fit$time[her.fit$lower<0.2])   
n2 <- max(her.fit$time[her.fit$upper>0.2])   
summary(her.fit,times = c(n1,n2))

## Call: survfit(formula = her.surv ~ 1, conf.int = 0.95)  
##   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 774 27 141 0.260 0.0364 0.1978 0.342  
## 1076 1 9 0.136 0.0364 0.0807 0.230

c(n1,n2)

## [1] 774 1076

We know time=1076 is the end of dataset and it is censored. Therefore, we could not use this way to find the 80th percentile. Instead, we use the estimate for smaller s(t).