Homework Lab B

Kevin Ayala

October 18, 2018

Question 1 Part A

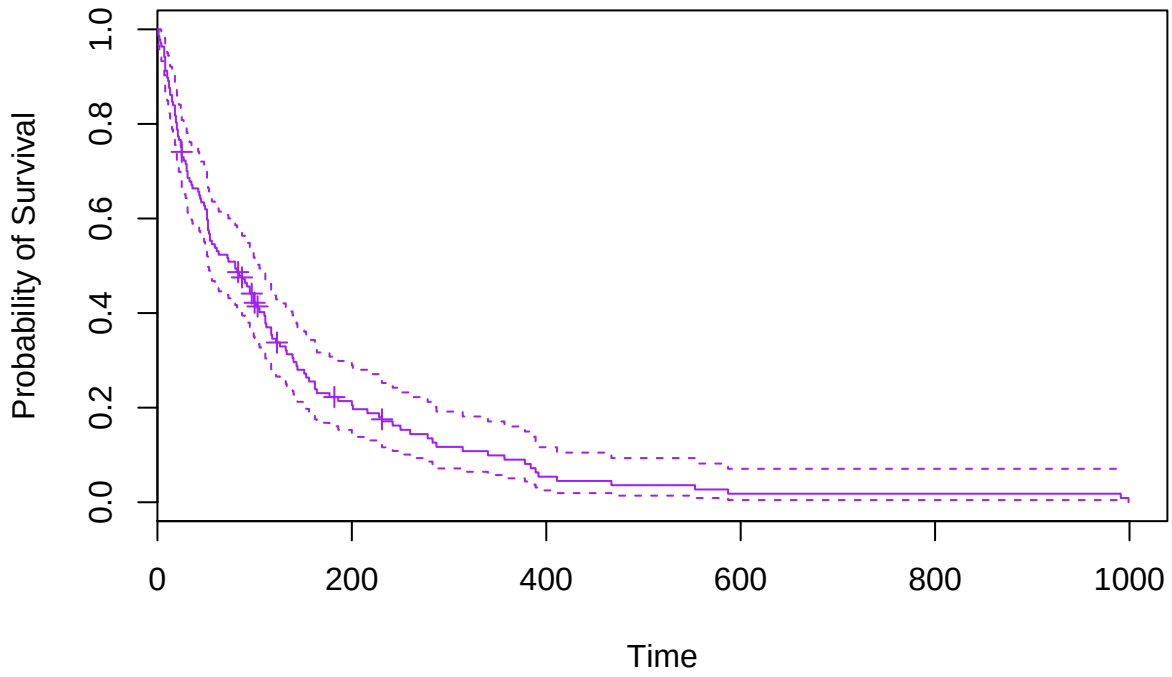
```
getwd()
```

```
## [1] "/Users/kevinlorenzoayala/Downloads"
```

```
library(survival)
vets.data<-read.table("/Users/kevinlorenzoayala/Downloads/vets.txt")
head(vets.data)
```

```
##    V1 V2
## 1  72  1
## 2 411  1
## 3 228  1
## 4 126  1
## 5 118  1
## 6  10  1
```

```
vets.km<-survfit(Surv(vets.data$V1,vets.data$V2)~1)
plot(vets.km, xlab= "Time", ylab="Probability of Survival", conf.int=TRUE,
     mark.time=TRUE, col = "Purple")
```



Question 1 Part B

```
quantile(vets.km, probs = c(.25,.5,.75))[1]
```

```
## $quantile
##  25  50  75
##  25  80 162
```

```
#Grabbed the first row to grab the times associated with the quantiles
```

Question 2 Part A

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

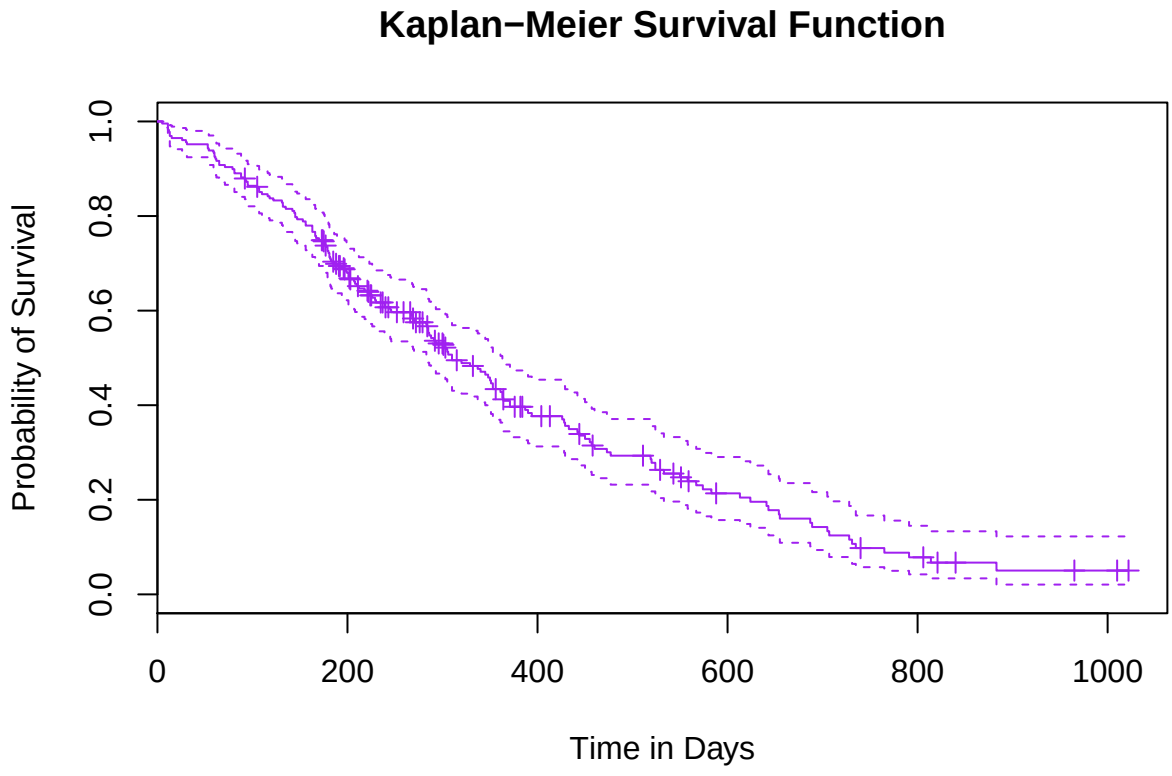
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
lung.data<-read.table("/Users/kevinlorenzoayala/Downloads/lung.txt")
head(lung.data)
```

```
##   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
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

```
lungs.km<-survfit(Surv(lung.data$time,lung.data$status)~1)
plot(lungs.km,xlab="Time in Days",ylab="Probability of Survival", conf.int=TRUE,
     mark.time=TRUE, main="Kaplan-Meier Survival Function", col = "purple")
```



Question 2 Part B

```
max(lungs.km$time[lungs.km$time<150]) #Getting closest observed time to 150, which is 147
```

```
## [1] 147
```

```
summary(lungs.km,times = 147)
```

```
## Call: survfit(formula = Surv(lung.data$time, lung.data$status) ~ 1)
##
##  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
```

Survival rate at this time is approx. .793, 95% CI is (.742, .848)

Question 2 Part C

```
quantile(lungs.km, .5, conf.int = TRUE)  #the median survival time is 310 with a 95% CI of (285, 363)
```

```
## $quantile
##  50
## 310
##
## $lower
##  50
## 285
##
## $upper
##  50
## 363
```

```
summary(lungs.km, time=310)
```

```
## Call: survfit(formula = Surv(lung.data$time, lung.data$status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   310     85     107    0.495  0.0352        0.431        0.569
```

Question 2 Part D

```
#Filtering out desired gender into subsets
lung.f <- filter(lung.data, sex == 2)
head(lung.f) #subsetted females
```

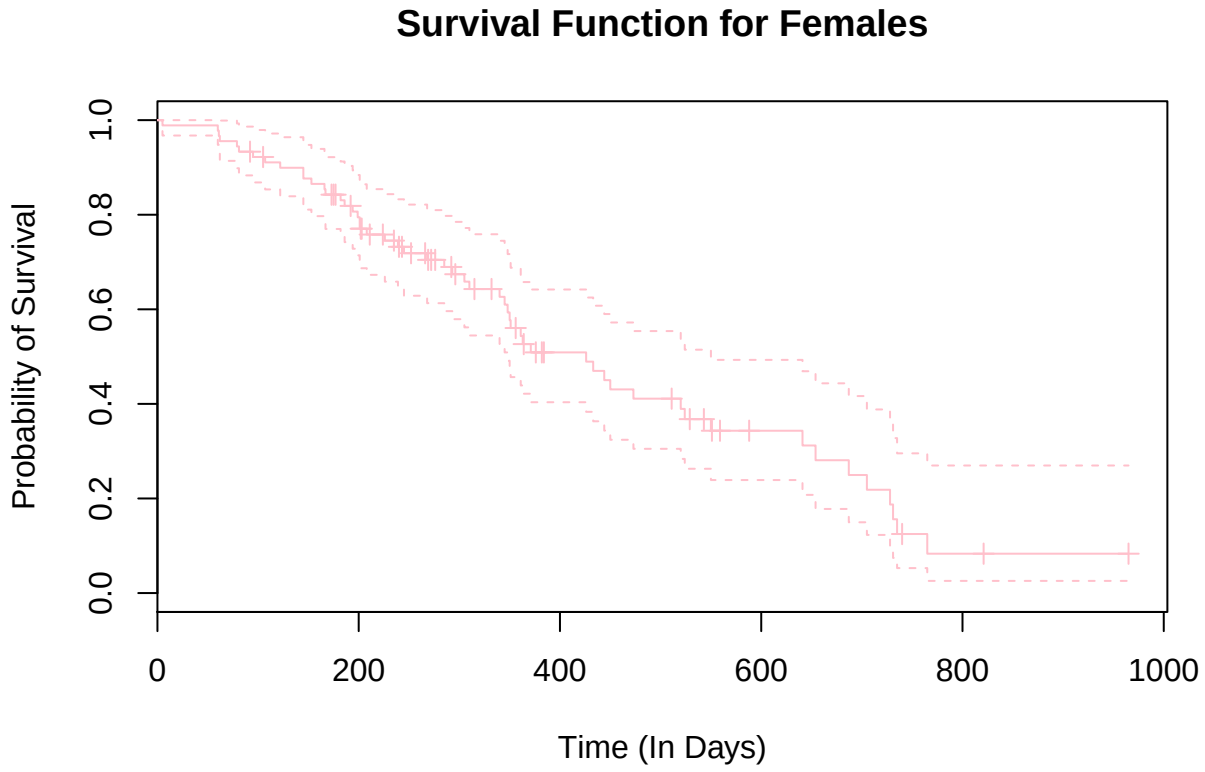
```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    7  310      2  68   2       2       70        60      384      10
## 2   11  361      2  71   2       2       60        80      538       1
## 3   16  654      2  68   2       2       70        70       NA      23
## 4   11  728      2  68   2       1       90        90       NA       5
## 5    1   61      2  56   2       2       60        60      238      10
## 6    6   81      2  49   2       0      100        70     1175      -8
```

```
lung.m <- filter(lung.data, sex == 1)
head(lung.m) #subsetted for the male case.
```

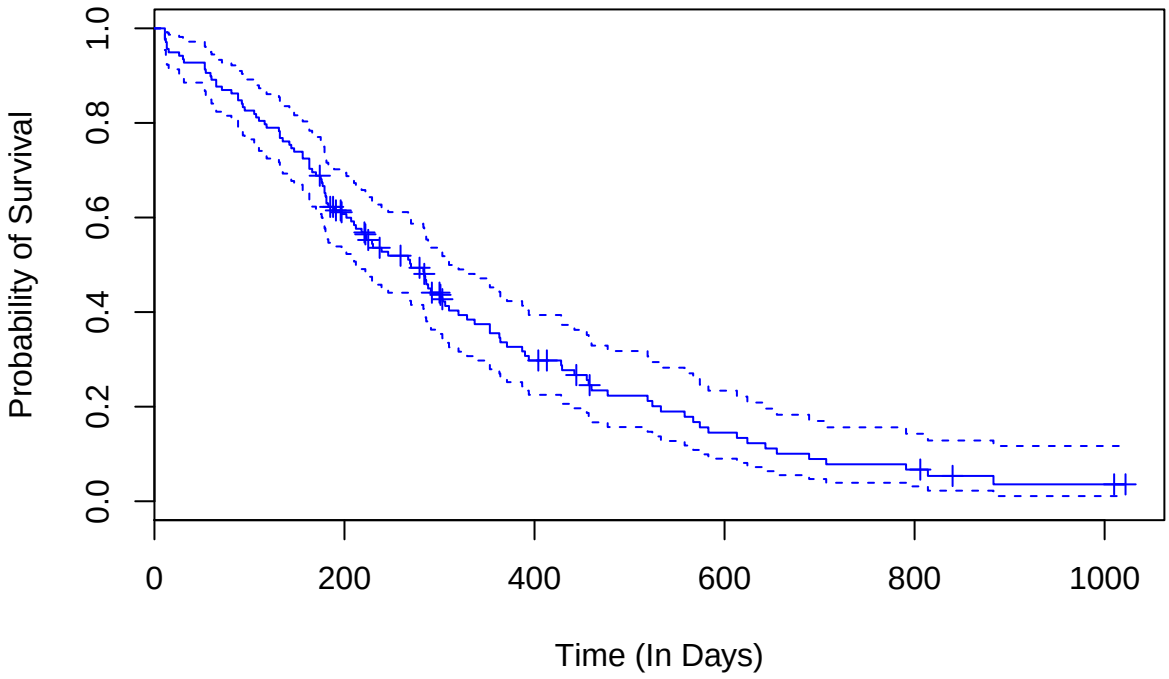
```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
```

```
## 5    1  883      2  60    1       0      100        90       NA       0
## 6   12 1022      1  74    1       1       50        80      513       0
```

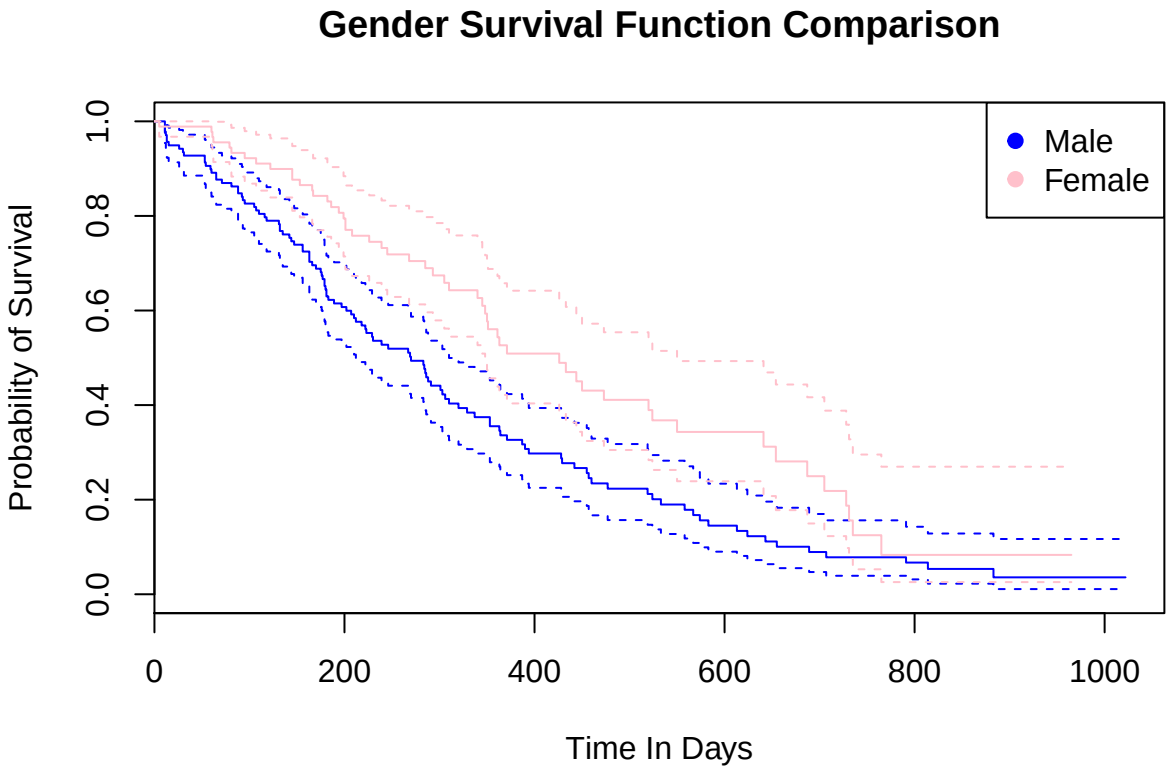
```
lungs.f.km<-survfit(Surv(lung.f$time,lung.f$status)~1) #female case
lungs.m.km<-survfit(Surv(lung.m$time,lung.m$status)~1) #male case

plot(lungs.f.km,xlab="Time (In Days)",ylab="Probability of Survival",
     main="Survival Function for Females",conf.int=TRUE, mark.time=TRUE, col = "pink")
```




```
plot(lungs.m.km,xlab="Time (In Days)",ylab="Probability of Survival",
     main ="Survival Function for Males",conf.int=TRUE, mark.time=TRUE, col = "blue")
```




```
#plot of both gender survival functions on one graph for easy comparison
lungs.genders<- survfit(Surv(time, status)~sex, data=lung.data)
plot(lungs.genders, xlab="Time In Days", ylab="Probability of Survival",
     main="Gender Survival Function Comparison", col=c("blue","pink"),
     conf.int = TRUE)
legend("topright",legend=c("Male","Female"), col=c("blue","pink"), pch=rep(19,2))
```



It seems that in general by the "Gender Survival Function Comparison" graph above, that females tend to have higher survival rates than men. This may be due to the stereotype of men having more addictive personality than females. Men and Women survival rate seem to be about the same at around 780 days since the 95% confidence intervals intersect, indicating possible same survival rate at that time and beyond.

Question 2 Part E

```
quantile(lungs.f.km, probs=.5)
```

```
## $quantile
##  50
## 426
##
## $lower
##  50
## 348
##
## $upper
##  50
## 550
```

```
summary(lungs.f.km, times<-426)
```

```
## Call: survfit(formula = Surv(lung.f$time, lung.f$status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   426     26      38    0.489   0.061        0.383        0.625
```

```
quantile(lungs.m.km, probs = .5)
```

```
## $quantile
##  50
```

```
## 270
##
## $lower
##  50
## 212
##
## $upper
##  50
## 310
```

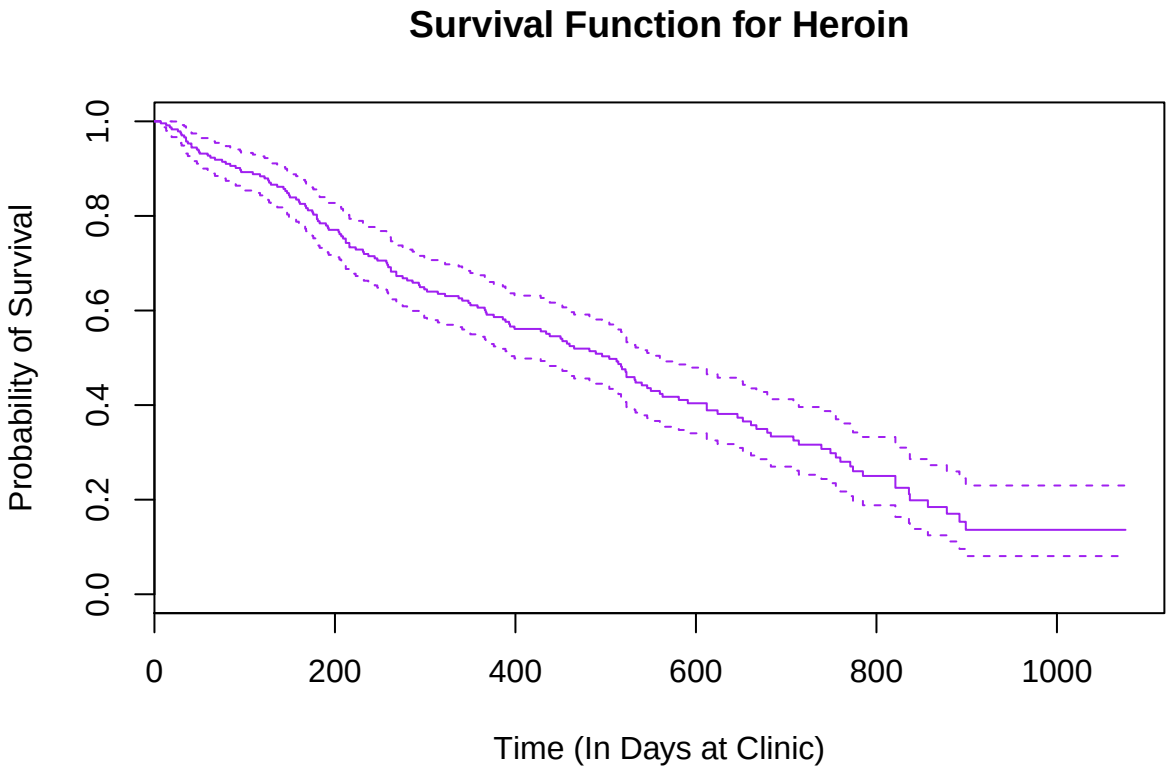
```
summary(lungs.m.km, times<- 270)
```

```
## Call: survfit(formula = Surv(lung.m$time, lung.m$status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   270     59      68    0.494  0.0436        0.415        0.587
```

The median of time for females is 426 with a confidence interval of (348,550), and the median of time for males is 270 with a confidence interval of (212, 310). The survival rates for females and males are .489 with a CI of (.415, .587) and the equivalent for males at the median, survival rate is .494 with a CI of (.415, .587). Since these similar survival rates occur at around 14 months (426 days) for women and at around 8 months (270 days) for men, then it definitely seems that women indeed have stronger survival rates than men. However, does not tell the full story because men and women have possible similar survival rates (confidence interval for survival rates intersect) at a later time which in this case is around 780 days. (780 days is an eyeball estimate)

Question 3 Part A

```
load("heroin.Rdt")
heroin.km <- survfit(Surv(heroin$Time, heroin$Status)~1)
plot(heroin.km, xlab = "Time (In Days at Clinic)", ylab="Probability of Survival",
     conf.int = TRUE, col= "purple", main = "Survival Function for Heroin")
```

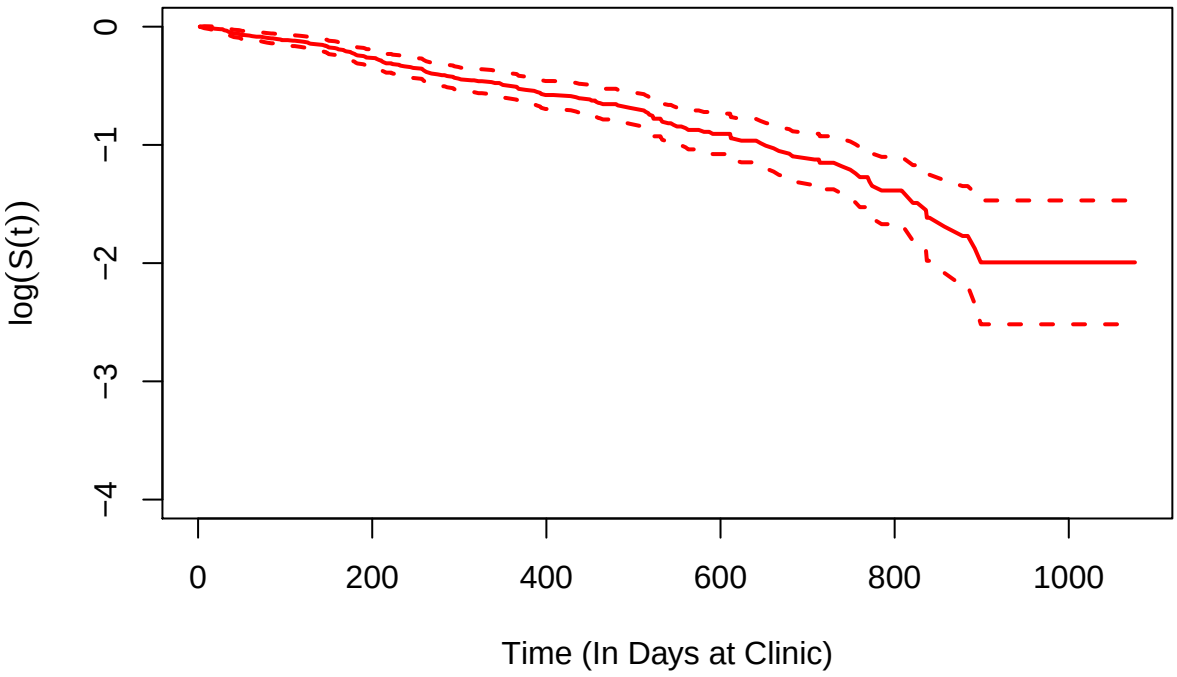


Question 3 Part B

```
m_j=heroin.km$n.event
n_j=heroin.km$n.risk
V_j=(m_j/(n_j*(n_j-m_j)))
CV_j = cumsum(V_j)
head(CV_j) #for the variance
```

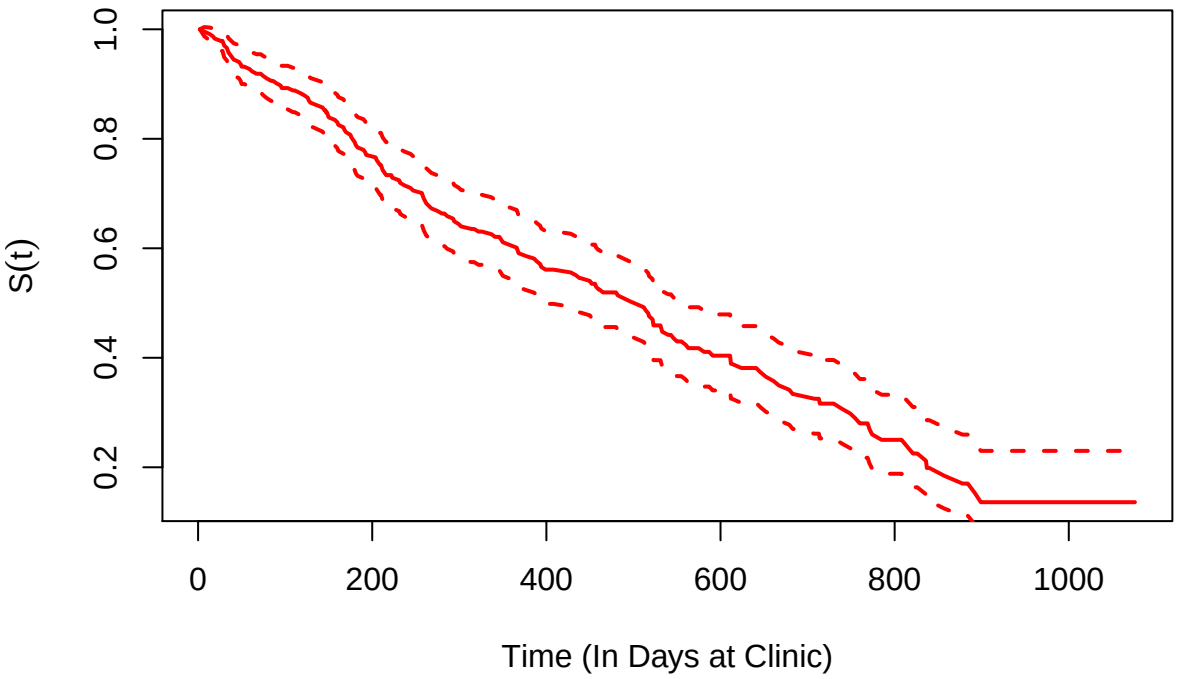
```
## [1] 0.000000e+00 1.803101e-05 3.621614e-05 5.455736e-05 7.305669e-05
## [6] 9.171619e-05
```

```
lowerboundlimit = log(heroin.km$surv) - 1.96*sqrt(CV_j)
upperboundlimit = log(heroin.km$surv) + 1.96*sqrt(CV_j)

plot(heroin.km$time,log(heroin.km$surv),lwd=2,type="l",ylim=c(-4,0),
xlab="Time (In Days at Clinic)",col = "red", ylab=expression(log(hat(S)(t))))

lines(heroin.km$time,lowerboundlimit,lty=2,col=2,lwd=2) #lowerbound Confidence Interval Graph
lines(heroin.km$time,upperboundlimit,lty=2,col=2,lwd=2) #upperbound Confidence Interval Graph
```



Question 3 Part C

```
plot(heroin.km$time,heroin.km$surv,lwd=2,type="l", col="red",
     xlab="Time (In Days at Clinic)",ylab=expression(hat(S)(t)))
lines(heroin.km$time,exp(lowerboundlimit),lty=2,col=2,lwd=2)
lines(heroin.km$time,exp(upperboundlimit),lty=2,col=2,lwd=2)
```



Question 3 Part D

The test statistic we will use: $T = \frac{log(\hat{s}(t) - log(p_0))}{sqrt(V_t)} = Z_{score}$

```r
max(heroin.km$time[heroin.km$time<365])
```

```
## [1] 358
```

```r
at.one.year <- summary(heroin.km, times = 358)
shat_358 <- at.one.year$surv
shat_358
```

```
## [1] 0.6060647
```

```r
at.one.year
```

```
## Call: survfit(formula = Surv(heroin$Time, heroin$Status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   358    124      87    0.606  0.0331        0.545        0.675
```

```r
heroin.km$std.err[heroin.km$time == 358] # the denominator in the test statistic above
```

```
## [1] 0.05458958
```

```r
z <- (log(shat_358)-log(.5))/heroin.km$std.err[heroin.km$time == 358]
z
```

```
## [1] 3.524093
```

```r
pnorm(-abs(z))
```

```
## [1] 0.0002124677
```

For a one tail test, H_0:s(t) = .5, H_a: s(t) < .5 at year one. Our p-value is .0002124677 in a one tailed test. We reject the null and conclude based on the evidence that at least 50% of patients are discharged within one year.

Quetion 3 Part E

```r
quantile(heroin.km, probs = .7)
```

```
## $quantile
##  70
## 749
##
## $lower
##  70
## 661
##
## $upper
##  70
## 836
```

```r
summary(heroin.km, times=749)
```

```
## Call: survfit(formula = Surv(heroin$Time, heroin$Status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   749     34     137    0.298  0.0363        0.235        0.379
```

```r
quantile(heroin.km, probs = .8)
```

```
## $quantile
##  80
```

```
## 837
##
## $lower
##  80
## 774
##
## $upper
## 80
## NA
```

```
summary(heroin.km, times = 837)
```

```
## Call: survfit(formula = Surv(heroin$Time, heroin$Status) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   837     16     146    0.199  0.0369        0.138        0.286
```

Our confidence interval at the the 70th percentile, at time 749 is (.235, .379). In the 80th percentile, we have that the upperbound is NA or does not exist, this is becuase there is no upperbound confidence limit that ever falls below 20% at the 80th percentile time of 837.