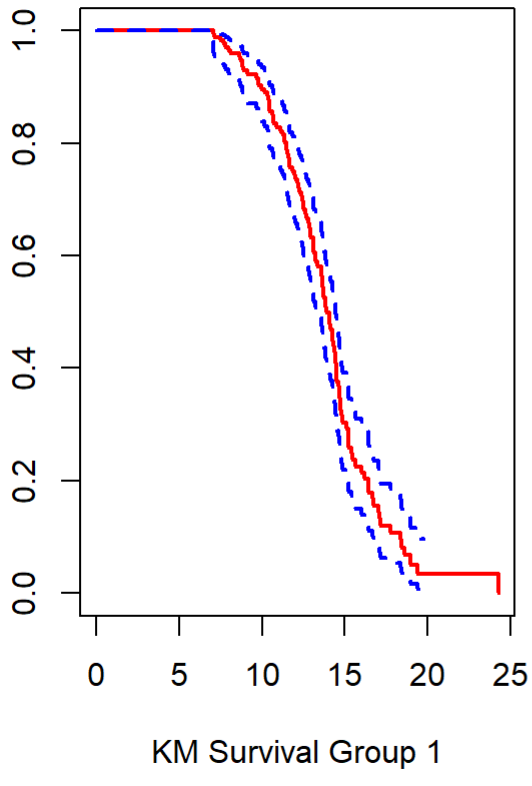
Biostatistics - Nonparametric Survival

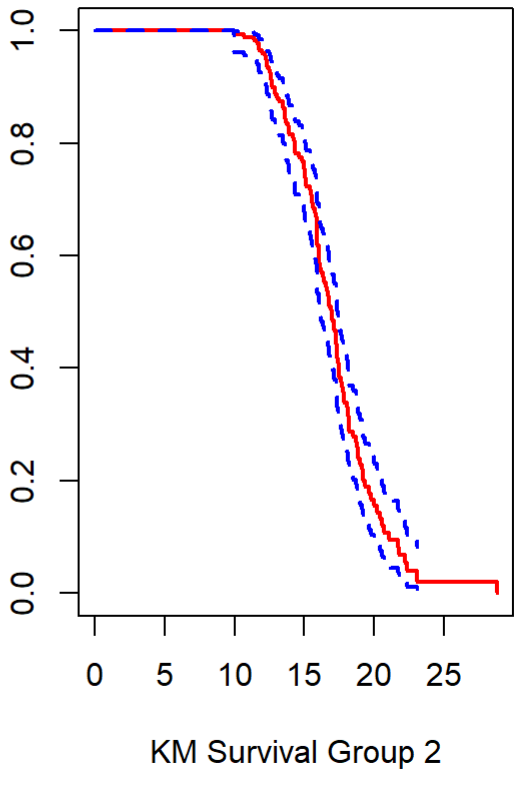
Rob Leonard (robleonard@tamu.edu
(mailto:robleonard@tamu.edu))

1) Kaplan-Meier

1a) Compute separate Kaplan-Meier survival curves for each of the two treatment groups. Make Plots.

```
data1 = read.csv("surv_times_data.csv", header=TRUE, sep = ",")   # read in original data
library(survival)
par(mfrow=c(1,2))  # set plots as side-by-side
data1$SurvObjGrp1 <- with(data1, Surv(Y1, event=Delta1))   # Delta1==1 means event occurred.  0
 is right censored
kmGrp1 <- survfit(SurvObjGrp1 ~ 1, data = data1, conf.type = "log-log")  # create survival curve
s for group 1
plot(kmGrp1, col=c("red", "blue", "blue"), lwd=2, xlab = "KM Survival Group 1")  # change colors
data1$SurvObjGrp2 <- with(data1, Surv(Y2, event=Delta2))   # Delta1==1 means event occurred.  0
 is right censored
kmGrp2 <- survfit(SurvObjGrp2 ~ 1, data = data1, conf.type = "log-log")  # create survival curve
s for group 1
plot(kmGrp2, col=c("red", "blue", "blue"), lwd=2, xlab = "KM Survival Group 2")  # change colors
```





Group 2 has the better survival prognosis. The survival probability is 1 through time period 13 or 14 for group 2 but only through time period 8 or 9 for group 1. The survival curve for group 2 is shifted right compared to group 1. The survival probability doesn't drop to 0 until after time period 25 for group 2, but dips to 0 before time period 25 in group 1.

1b) Estimate mean survival time and a 95% CI using KM for each group.

```
print(kmGrp1, print.rmean=TRUE)
```

```
## Call: survfit(formula = SurvObjGrp1 ~ 1, data = data1, conf.type = "log-log")
##
##          n     events     *rmean *se(rmean)     median    0.95LCL    0.95UCL
##    183.000    110.000     13.985      0.306     13.873     13.331     14.455
##     * restricted mean with upper limit =  24.3
```

```
print(kmGrp2, print.rmean=TRUE)
```

```
## Call: survfit(formula = SurvObjGrp2 ~ 1, data = data1, conf.type = "log-log")
##
##          n     events     *rmean *se(rmean)     median    0.95LCL    0.95UCL
##    183.000    119.000     17.053      0.295     16.950     16.080     17.441
##     * restricted mean with upper limit =  28.8
```

```
(ciGrp1 = c(13.985-1.96*0.306,13.985+1.96*0.306) )
```

```
## [1] 13.38524 14.58476
```

```
(ciGrp2 = c(17.053-1.96*0.295,17.053+1.96*0.295) )
```

```
## [1] 16.4748 17.6312
```

The estimated mean for group 1 is 13.985. The estimated 95% CI for the group 1 mean is (13.38524, 14.58476). The estimated mean for group 2 is 17.053. The estimated 95% CI for group 2 mean is (16.4748, 17.6312).

1c) Obtain the estimate and 95% CI for the 1st, 2nd, and 3rd quartiles of the survival times for each group, based on the Kaplan-Meier survival curves. For group 1:

```
quantile(kmGrp1, prob=c(0.25, 0.5, 0.75), conf.int=TRUE)
```

```
## $quantile
##       25       50       75
## 11.98686 13.87340 15.38351
##
## $lower
##       25       50       75
## 11.25834 13.33085 14.73089
##
## $upper
##       25       50       75
## 12.63385 14.45479 16.66306
```

For group 2:

```
quantile(kmGrp2, prob=c(0.25, 0.5, 0.75), conf.int=TRUE)
```

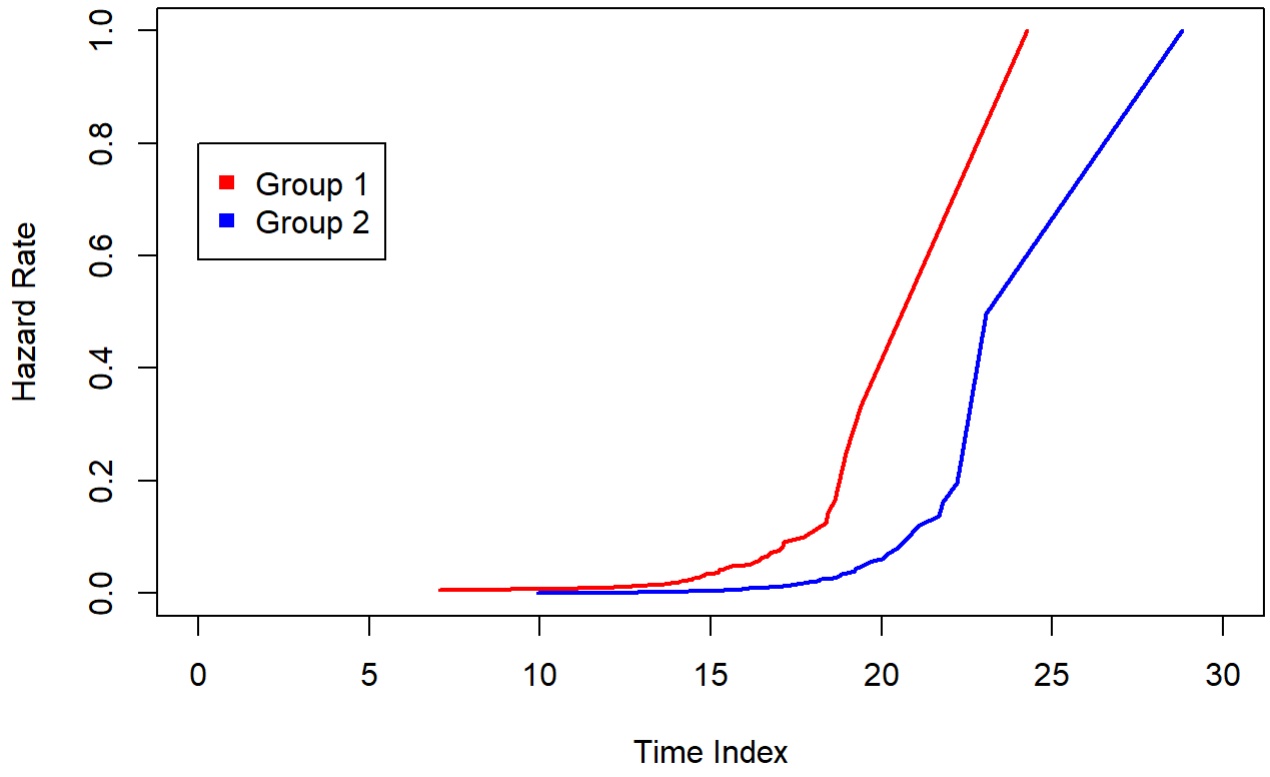
```
## $quantile
##       25       50       75
## 15.04833 16.95023 18.82140
##
## $lower
##       25       50       75
## 13.92103 16.07963 18.03539
##
## $upper
##       25       50       75
## 15.78946 17.44139 19.76284
```

	1st Quartile	2nd Quartile	3rd Quartile
Group1			
Estimate	11.98686	13.87340	15.38351
Confidence Interval	(11.25834,12.63385)	(13.33085 ,14.45479)	(14.73089 ,16.66306)
Group2			
Estimate	15.04833	16.95023	18.82140
Confidence Interval	(13.92103 ,15.78946)	(16.07963 ,17.44139)	(18.03539 ,19.76284)

1d) Plot hazard functions. Apply log rank test to the 2 survival curves.

```
data1dg1 <- data.frame(kmGrp1$n.censor, kmGrp1$n.event, kmGrp1$n.risk, kmGrp1$time)
hazGrp1 <- data1dg1$kmGrp1.n.event/data1dg1$kmGrp1.n.risk
hazard1 <- cbind(data1dg1$kmGrp1.time,hazGrp1)
hazard1 <- subset(hazard1, hazGrp1!=0)
plot(hazard1, col="red", lwd=2,  ylab = "Hazard Rate", xlab = "Time Index", type = "l", xlim=c(0
,30), ylim=c(0,1))

data1dg2 <- data.frame(kmGrp2$n.censor, kmGrp2$n.event, kmGrp2$n.risk, kmGrp2$time)
hazGrp2 <- data1dg2$kmGrp2.n.event/data1dg2$kmGrp2.n.risk
hazard2 <- cbind(data1dg2$kmGrp2.time,hazGrp2)
hazard2 <- subset(hazard2, hazGrp2!=0)
par(new=T);
plot(hazard2, col="blue", lwd=2,  ylab = "", xlab = "", type = "l", xlim=c(0,30), axes=FALSE)
legend(x=0,y=0.8,c("Group 1","Group 2"),col=c("red","blue"),pch=c(15,15))
```



The plots of the hazard rates look fairly similar in shape but group 2 is shifted right in time.

```
# combine data into a single dataset
dataA <- data.frame(data1[1],data1[2],data1[4])
dataA[4] <- 0
names(dataA)[1:4] = c("X","Y","Delta","Grp")
dataB <- data.frame(data1[1],data1[3],data1[5])
dataB[4] <- 1
names(dataB)[1:4] = c("X","Y","Delta","Grp")
data2 <- merge(dataA,dataB, all = TRUE)
# run log rank test
logrank1 <- survdiff(Surv(Y, event=Delta)~Grp, data = data2, rho = 0)
logrank1
```

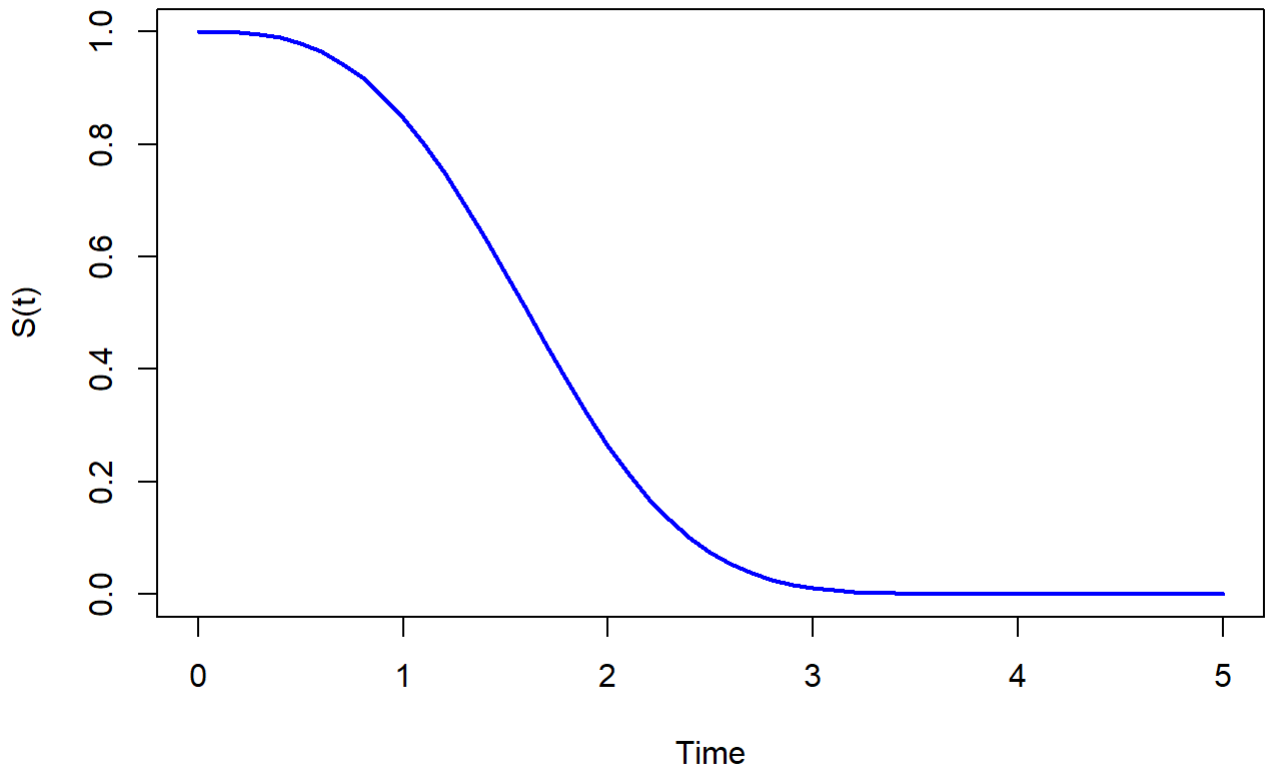
```
## Call:
## survdiff(formula = Surv(Y, event = Delta) ~ Grp, data = data2,
##     rho = 0)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## Grp=0 183      110     62.8      35.5      51.9
## Grp=1 183      119    166.2      13.4      51.9
##
##  Chisq= 51.9  on 1 degrees of freedom, p= 6e-13
```

We are testing $H_0 : \lambda_1(t) = \lambda_2(t)$ for all t, versus $H_a : \lambda_1(t) \neq \lambda_2(t)$ for at least one t. The Chi-Square test statistic is very high at 51.9 and the p-value is very low at $6*10^{-13}$. Thus, we reject the null hypothesis that the hazard rates are equal.

Problem 2

2a) Obtain the analytical form of the survival function, and plot it over time between 0 and 5.

```
time <- seq(0,5,0.1)
lambdaProb2 <- 0.5*(time^2)
fProb2 <- function(t) exp(-integrate( function(x) .5*(x^2), lower = 0, upper=t)$value)
survProb2 <- lapply(time, fProb2)
plot(time, survProb2, type = "l", lwd=2, col = "blue", xlim=c(0,5), ylab = "S(t)", xlab="Time")
```



2b) Calculate the mean of T.

The mean of T is the area under the survival curve. We have the exact form of the survival curve so the mean is the integral from 0 to infinity which is **1.62**.

```
meanProb2 <- integrate( function(x) exp(-(1/6)*x^3), lower = 0, upper=Inf)
meanProb2
```

```
## 1.622651 with absolute error < 1.6e-06
```

2c) Obtain the pth percentile of T. Also calculate values of 1st, 2nd and 3rd quartiles. The pth percentile is the smallest time at which the survival function is less than or equal to (1-p): $q_p = \inf[t : S(t) \leq (1-p)]$ => $inf[t : e^{-(1/6)t^3)} \leq (1-p)]$ => $inf[t : 1 - e^{-(1/6)t^3)} \geq p]$

```
(first <- (-6*log(.75))^(1/3))
```

```
## [1] 1.199558
```

```
(second <- (-6*log(.5))^(1/3))
```

```
## [1] 1.608146
```

```
(third <- (-6*log(.25))^(1/3))
```

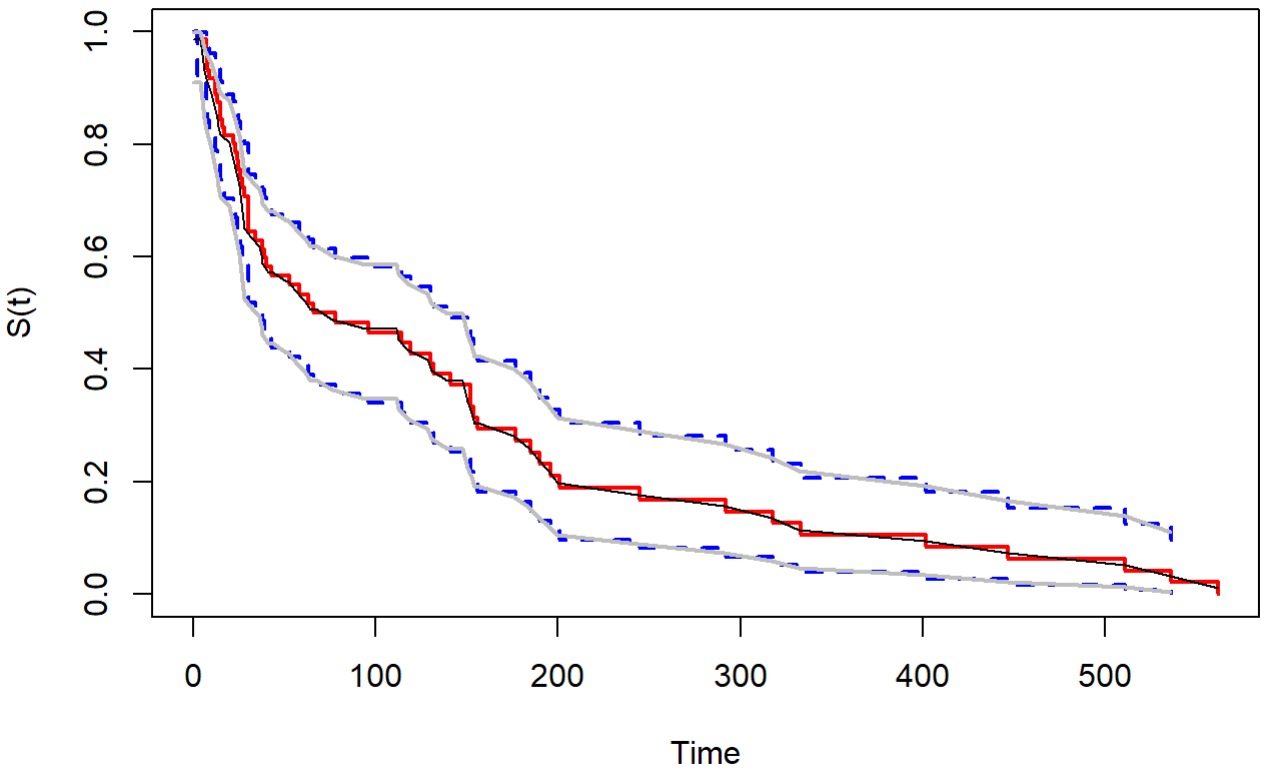
```
## [1] 2.026137
```

The first, second and third quartiles are **1.199558, 1.608146 and 2.026137**, respectively.

3) Kidney Data

3a) Obtain the estimate and the 95% CI's for the Survival function for the Kidney data using KM and NA methods. Plot them in same graph.

```
data(kidney)
kidney$SurvObj <- with(kidney, Surv(time, status))
kidneykm <- survfit(SurvObj ~ 1, data = kidney, conf.type = "log-log")
plot(kidneykm, col=c("red","blue","blue"),lwd=2, xlab="Time", ylab="S(t)")
# now plot Nelson Aalen
hazardNA3 <- kidneykm$n.event/kidneykm$n.risk
cum.hazard3 <- cumsum(hazardNA3)
var3 <- cumsum( kidneykm$n.event*(kidneykm$n.risk-kidneykm$n.event)/(((kidneykm$n.risk)^2)*(kidn
eykm$n.risk-1)))
sd3 <- sqrt(var3)
par(new=T);
plot(kidneykm$time, exp(-cum.hazard3), ylim=c(0, 1), ylab="", type="l", col="black", xlab="", ax
es=FALSE)
par(new=T);
plot(kidneykm$time, exp(-exp(log(cum.hazard3)-1.96*sd3/cum.hazard3)), col="gray",
     lwd=2, ylim=c(0, 1), type="l", ylab="", xlab="", axes=FALSE)
par(new=T);
plot(kidneykm$time, exp(-exp(log(cum.hazard3)+1.96*sd3/cum.hazard3)),  col="gray",
     lwd=2, ylim=c(0, 1), type="l", ylab="", xlab="", axes=FALSE)
```



3b) Obtain the estimate and 95% CI for the mean using the both estimators of S(t). You may use the bootstrap technique to construct the confidence interval for the mean.

#Kaplan-Meier Method:

```
print(kidneykm, print.rmean = TRUE)
```

```
## Call: survfit(formula = SurvObj ~ 1, data = kidney, conf.type = "log-log")
##
##          n     events      *rmean *se(rmean)     median    0.95LCL    0.95UCL
##       76.0       58.0      137.0       19.8       78.0       38.0      141.0
##     * restricted mean with upper limit =  562
```

```
# now find the CI, use for loop and sample rather than boot function
vectkm = rep(0,10000)
vectna = rep(0,10000)
for (i in 1:10000) {
  sampledata = kidney[sample(nrow(kidney),76, replace=TRUE),]
  maxtime = max(sampledata$time)
  samplekm <- survfit(SurvObj ~ 1, data = sampledata, conf.type = "log-log")
  vectkm[i] = survival:::survmean(samplekm, rmean=maxtime) [[1]]["*rmean"]
  # for Nelson-Aalen estimates
  hazardloop <- samplekm$n.event/samplekm$n.risk
  cum.hazardloop <- cumsum(hazardloop)
  Survivalloop <- exp(-cum.hazardloop)
  b = length(Survivalloop)
  Timediffloop = rep(0,b)
  Arealoop = rep(0,b)
  NAdataloop = data.frame(cbind(Survivalloop,samplekm$time,Timediffloop,Arealoop))
  colnames(NAdataloop) <- c("Survival", "Time","Timediff","Area")
  NAdataloop[2:b,3] = NAdataloop[2:b,2]-NAdataloop[1:(b-1),2]
  NAdataloop[1,3] = NAdataloop[1,2]
  NAdataloop$Area = NAdataloop$Survival*NAdataloop$Timediff
  vectna[i] = apply(NAdataloop[4],2,sum)
  }
vectkm = sort(vectkm)
CIlbKM = vectkm[250]
CIubKM = vectkm[9750]
```

Using Kaplan-Meier, **the estimated mean is 137.0 and the 95% CI is (100.0036278,177.6159904).**

#Nelson-Aalen Method:

```
NASurvival = exp(-cum.hazard3)
Timediff = rep(0,60)
Area = rep(0,60)
NAdata = data.frame(cbind(NASurvival,kidneykm$time,Timediff,Area))
colnames(NAdata) <- c("Survival", "Time","Timediff","Area")
for (i in 2:60) {
  NAdata[i,3] = NAdata[i,2]-NAdata[i-1,2]
}
NAdata[1,3] = NAdata[1,2]
NAdata$Area = NAdata$Survival*NAdata$Timediff
apply(NAdata[4],2,sum)
```

```
##     Area
## 131.1594
```

```
# 95% CI generation - mostly done above with KM estimates
vectna = sort(vectna)
CIlbNA = vectna[250]
CIubNA = vectna[9750]
```

Using Nelson-Aalen, **the estimated mean is 131.16 and the 95% CI is (94.8428074,167.5362624).**