STAT645 - Homework 7

Salih Kilicli

10/1/2019

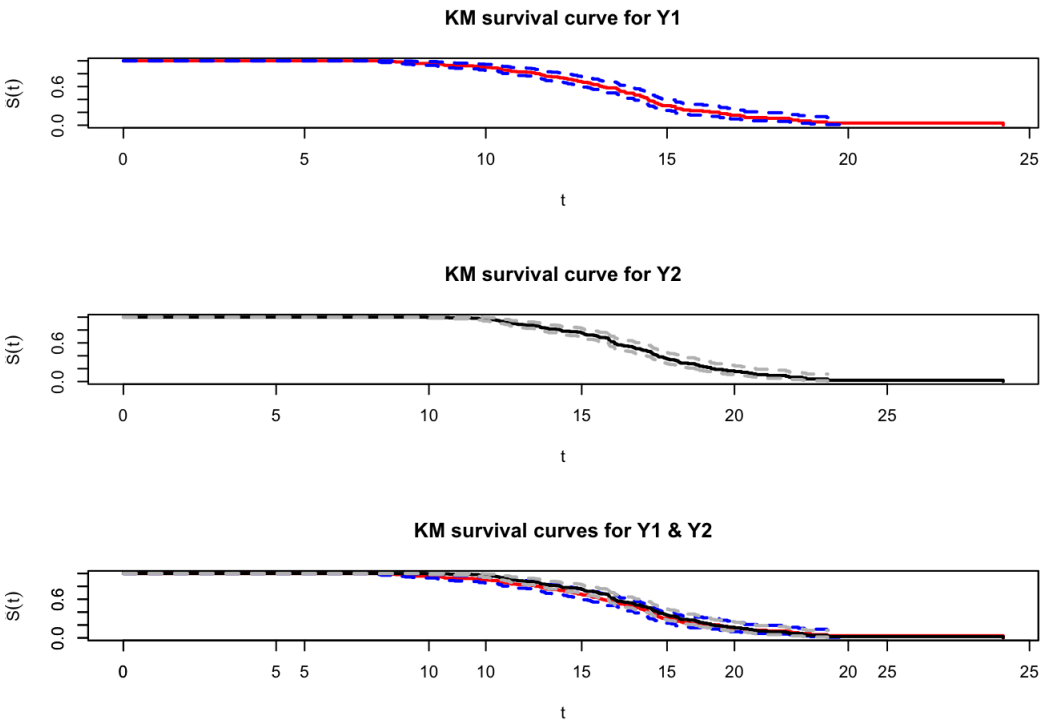
```
rm(list=ls())
knitr::opts_chunk$set(echo = TRUE)
```

**Problem 1: For the “surv times data.txt” data given in the data folder under the Course Content tab of eCampus:*

```
setwd("/Users/younique/Desktop/STAT645/Data")
survtimes= read.csv("surv_times_data.csv", header=TRUE)
```

(a) Compute separate Kaplan-Meier survival curves for each of the two treatment groups. Make two plots, one for each of the two treatment groups, showing the estimated survival curves, and 95% pointwise confidence interval and comment on the differences between the curves for the two treatment groups. Which group has the better survival prognosis?

```
library(survival)
survtimes$SurvObj1 = with(survtimes, Surv(Y1, Delta1==1))
survtimes$SurvObj2 = with(survtimes, Surv(Y2, Delta2==1))
one = survfit(SurvObj1 ~ 1, data=survtimes)
two = survfit(SurvObj2 ~ 1, data=survtimes)
par(mfrow=c(3,1))
plot(one,col=c("red","blue","blue"),xlab='t',ylab=expression(hat(S)(t)),main="KM survival curve for Y1",lwd=2)
plot(two,col=c("black","grey","grey"),xlab='t',ylab=expression(hat(S)(t)),main="KM survival curve for Y2",lwd=2)
plot(one, col=c("red","blue","blue"), xlab='t', ylab=expression(hat(S)(t)), main="KM survival curves for Y1 & Y2"
  , lwd=2)
par(new=T)
plot(two, col=c("black","grey","grey"), lwd=2)
```



Y2 data has a better survival prognosis since survival time probability reduces at later times and in general the plot of survival curve for Y2 takes higher values than plot of survival curve for Y1, i.e., subjects survive for longer time in Y2 data.

(b) Obtain the estimate and 95% CI for the mean survival time for each group, based on the Kaplan-Meier survival curves.

```
print("The estimate and 95 percent CI for the mean survival time for group 1 given below")
```

```
## [1] "The estimate and 95 percent CI for the mean survival time for group 1 given below"
```

```
group1=print(one, print.rmean=TRUE) #prints rmean and other percentiles
```

```
## Call: survfit(formula = SurvObj1 ~ 1, data = survtimes)
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##    183.000    110.000     13.985      0.306     13.873     13.600
##    0.95UCL
##     14.482
##     * restricted mean with upper limit =  24.3
```

```
rmean1=13.985
cat("The mean survival time estimate for group 1 is =", rmean1,"\n")
```

```
## The mean survival time estimate for group 1 is = 13.985
```

```
se_rmean1=0.306
CI1 = rmean1+c(-1,1)*(1.96)*se_rmean1
names(CI1)=c("Lower","Upper")
CI1
```

```
##    Lower    Upper
## 13.38524 14.58476
```

```
print("The estimate and 95 percent CI for the mean survival time for group 2 given below")
```

```
## [1] "The estimate and 95 percent CI for the mean survival time for group 2 given below"
```

```
group2=print(two, print.rmean=TRUE) #prints rmean and other percentiles
```

```
## Call: survfit(formula = SurvObj2 ~ 1, data = survtimes)
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##    183.000    119.000     17.053      0.295     16.950     16.168
##    0.95UCL
##     17.481
##     * restricted mean with upper limit =  28.8
```

```
rmean2=17.053
cat("The mean survival time estimate for group 2 is =", rmean2,'\n')
```

```
## The mean survival time estimate for group 2 is = 17.053
```

```
se_rmean2=0.295
CI2 = rmean2+c(-1,1)*(1.96)*se_rmean2
names(CI2)=c("Lower","Upper")
CI2
```

```
##   Lower   Upper
## 16.4748 17.6312
```

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

```
print("Quartile estimates of survival times of group 1 and corresponding 95 percent CI's are given below")
```

```
## [1] "Quartile estimates of survival times of group 1 and corresponding 95 percent CI's are given below"
```

```
quantile(one, 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
## 11.34816 13.60025 14.76136
##
## $upper
##       25       50       75
## 12.67745 14.48226 16.74518
```

```
print("Quartile estimates of survival times of group 2 and corresponding 95 percent CI's are given below")
```

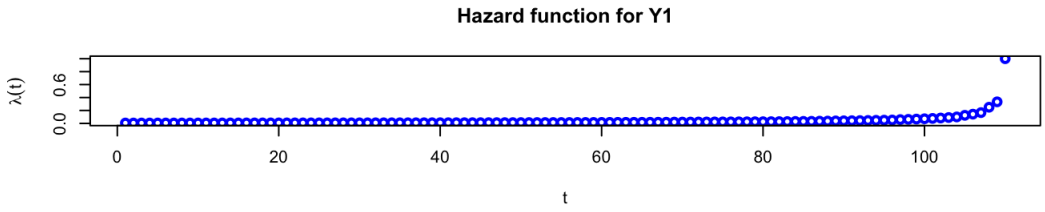
```
## [1] "Quartile estimates of survival times of group 2 and corresponding 95 percent CI's are given below"
```

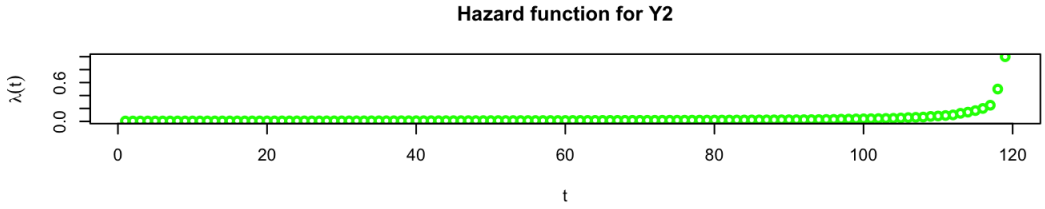
```
quantile(two, prob=c(0.25, 0.5, 0.75), conf.int=TRUE)
```

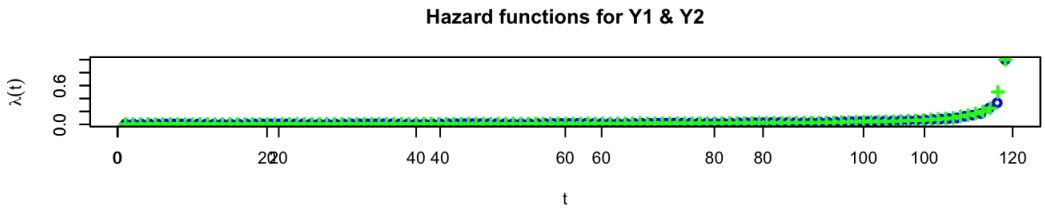
```
## $quantile
##       25       50       75
## 15.04833 16.95023 18.82140
##
## $lower
##       25       50       75
## 14.26997 16.16761 18.10029
##
## $upper
##       25       50       75
## 15.86906 17.48058 20.00283
```

(d) Plot the hazard functions for the two groups. Then use the log rank test to formally test for equality of the two survival curves. Report the chi-square statistic and the p-value. Comment on the results.

```
par(mfrow=c(3,1))
hone=summary(survfit(Surv(Y1,Delta1)~1, data=survtimes))
hazard1 = (hone$n.event)/(hone$n.risk)
plot(hazard1, xlab='t', ylab=expression(lambda(t)), main="Hazard function for Y1", col="blue", lwd=2)
htwo=summary(survfit(Surv(Y2,Delta2)~1, data=survtimes))
hazard2 = (htwo$n.event)/(htwo$n.risk)
plot(hazard2, xlab='t', ylab=expression(lambda(t)), main="Hazard function for Y2", col="green", lwd=2)
plot(hazard1, xlab='t', ylab=expression(lambda(t)), pch=1, main="Hazard functions for Y1 & Y2", col="blue", lwd=2
  )
par(new=T)
plot(hazard2, xlab='', ylab='', pch=3, col="green", lwd=2)
```








```
Y=c(survtimes$Y1, survtimes$Y2)
Delta=c(survtimes$Delta1, survtimes$Delta2)
TX=rep(0:1, each=183)
survdiff(Surv(Y,Delta)~TX)
```

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

```
cat("Chi-square statistic for the log-rank test is given by: \n Chi-squared =", 51.9)
```

```
## Chi-square statistic for the log-rank test is given by:
##  Chi-squared = 51.9
```

```
cat("p-value for the log-rank test is given by: \n p =", 6e-13)
```

```
## p-value for the log-rank test is given by:
##  p = 6e-13
```

Hypotheses for log-rank test are given by: $H_0 : \lambda_1(t) = \lambda_2(t)$ for $t \leq \tau$, whereas $H_a : \lambda_1(t) \neq \lambda_2(t)$ for at least one t.

Since the $p - value$ found is almost 0, we don't have enough evidence to reject the null hypothesis. Therefore, the Null hypothesis holds and $\lambda_1(t) = \lambda_2(t)$ for $t \leq \tau$. Notice here, log-rank test doesn't detect a difference when survival curves cross like in our case. In Problem 1a we have seen that survival curves are actually slightly different.

Problem 2: Suppose that the hazard function of T is $\lambda(t) = 0.5t^2$.

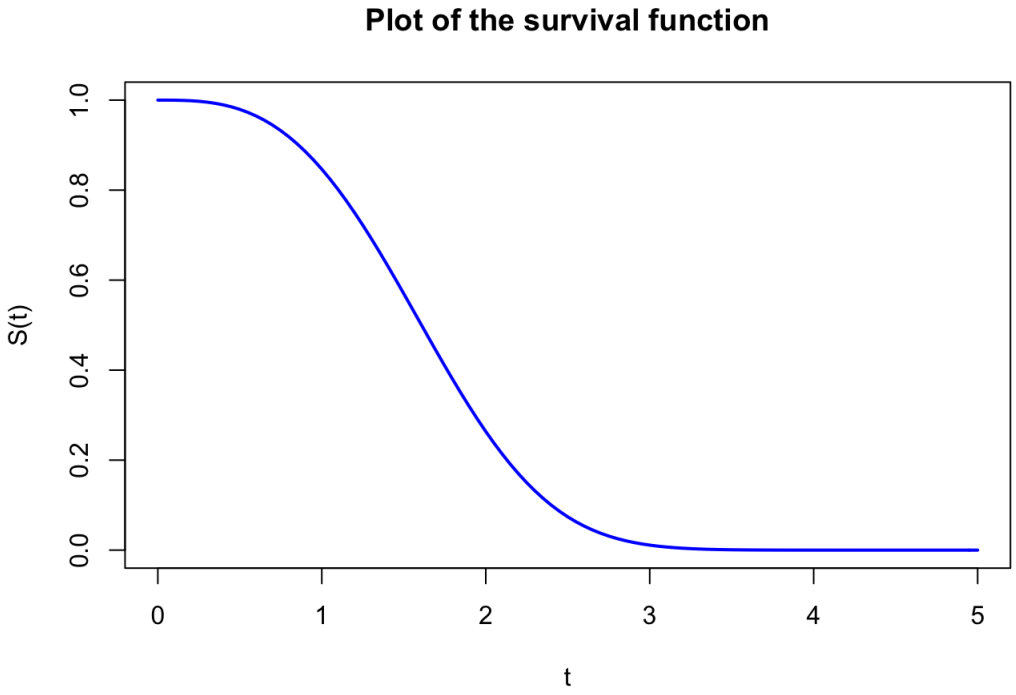
(a) Obtain the analytical form of the survival function, and plot it over time between 0 and 5.

Since hazard function is given by: $\lambda(t) = 0.5t^2$, the cumulative hazard function can be written as:

$$\Lambda(t) = \int_0^t 0.5u^2\,du = [\frac{u^3}{6}]\Big|_0^t = \frac{t^3}{6}$$

The, the survival function is obtained by: $S(t) = e^{-\Lambda(t)} = e^{-\frac{t^3}{6}}$.

```
par(mfrow=c(1,1))
S = function(t) {exp(-(t^3)/6)}
plot(S,0,5,xlab="t",ylab="S(t)",col="blue",main="Plot of the survival function", lwd=2)
```



(b) What is the mean of T?

Since we have the actual $S(t)$ (survival function), rather then it's estimate, the mean of T can be found by integrating $S(t)$ over 0 to ∞.

```
Mu=integrate(S, lower=0, upper=Inf)
print("The mean of the T is given by")
```

```
## [1] "The mean of the T is given by"
```

```
Mu
```

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

(c) Obtain the pth percentile of T. Also, provide the value of $Q1$, $Q2$ and $Q3$.

p^{th} percentile of T is given by:

$inf(t : \hat{S}(t) \leq (1-p))$, however, since we have the actual value of $S(t)$ rather than the estimate it can be found by solving the equality $S(t) = (1-p)$. Then, solving $e^{\frac{-t^3}{6}} = (1-p)$ for t yields:

$$t = (-6ln(1-p))^{\frac{1}{3}} = (ln(\frac{1}{(1-p)^6}))^{\frac{1}{3}}$$

.

Then, the values of $Q1$, $Q2$ and $Q3$ can be given by setting $p = 0.25, 0.5, 0.75$ respectively.

$$Q1 = (-6ln(1-0.25))^{\frac{1}{3}}, Q2 = (-6ln(1-0.5))^{\frac{1}{3}}, Q3 = (-6ln(1-0.75))^{\frac{1}{3}}$$

as below.

```
Q1 = (-6*log(1-0.25))^{1/3}
Q2 = (-6*log(1-0.5))^{1/3}
Q3 = (-6*log(1-0.75))^{1/3}
percentiles=c(Q1,Q2,Q3)
names(percentiles)=c("Q1", "Q2", "Q3")
print("The percentile values for Q1, Q2, and Q3 are given respectively by")
```

```
## [1] "The percentile values for Q1, Q2, and Q3 are given respectively by"
```

```
percentiles
```

```
##       Q1       Q2       Q3
## 1.199558 1.608146 2.026137
```

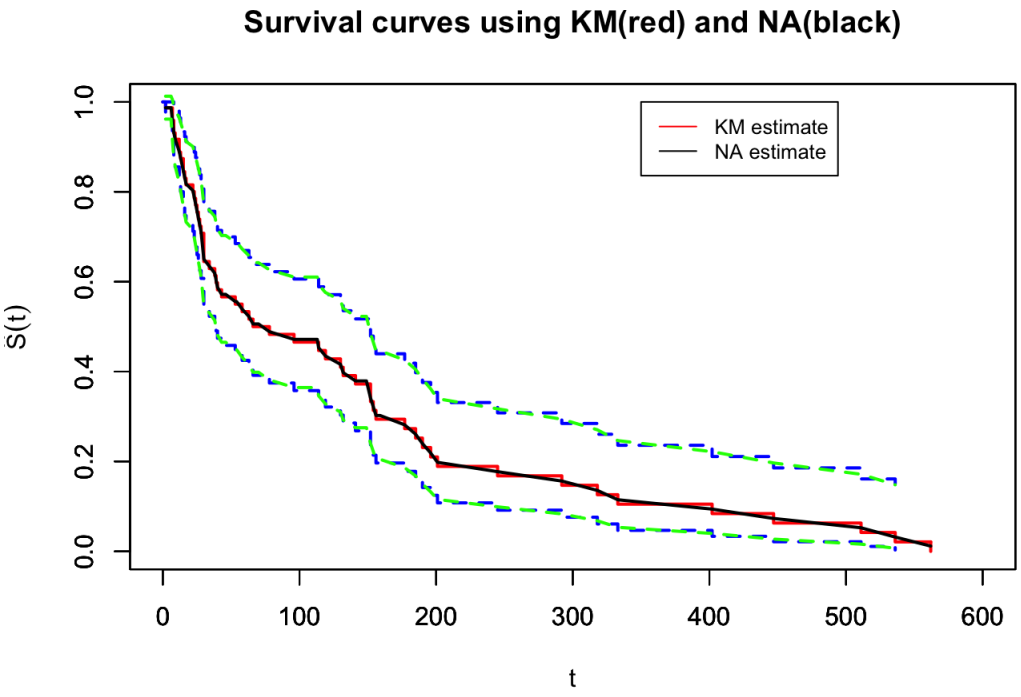
Problem 3: Consider the kidney data in the survival library of R.

(a) Obtain the estimate and 95% CI for the survival function $S(t)$ using the Kaplan-Meier and Nelson-Aalen approach, and plot them in the same figure.

```
library(survival)
par(mfrow=c(1,1))
data(kidney)
head(kidney)
```

id	time	status	age	sex	disease	frail
1	8	1	28	1	Other	2.3
1	16	1	28	1	Other	2.3
2	23	1	48	2	GN	1.9
2	13	0	48	2	GN	1.9
3	22	1	32	1	Other	1.2
3	28	1	32	1	Other	1.2

```
kidney$SurvObj = with(kidney, Surv(time, status==1))
km = survfit(SurvObj ~ 1, data=kidney)
hazard=km$n.event/km$n.risk
cumhaz=cumsum(hazard)
nsquared=(km$n.risk)^2                          # var estimate = cumsum(\dfrac{di(ni-di)}{ni^2(ni-1)})
sd=sqrt( cumsum(km$n.event*(km$n.risk-km$n.event)/(nsquared*(km$n.risk-1))) )  #sd = sqrt(var estimate)
t=km$time
plot(km, col=c("red","blue","blue"), ylim=c(0,1), xlim=c(0,600), xlab='t', ylab=expression(hat(S)(t)), main="Surv
  ival curves using KM(red) and NA(black)", lwd=2)
par(new=T)
plot(t, exp(-cumhaz), col="black", ylim=c(0,1), xlim=c(0,600), xlab='', ylab='', type='l', lwd=2)
par(new=T)
plot(t, exp(-cumhaz+1*1.96*sd), col="green", ylim=c(0,1), xlim=c(0,600), xlab='', ylab='',type='l', lty=2, lwd=2)
par(new=T)
plot(t, exp(-cumhaz-1*1.96*sd), col="green", ylim=c(0,1), xlim=c(0,600), xlab='', ylab='',type='l', lty=2, lwd=2)
legend(350, 1, legend=c("KM estimate", "NA estimate"),col=c("red", "black"), lty=1:1, cex=0.8)
```



The plots of survival curves using Kaplan-Meier and Nelson-Aalen approaches and their 95% CIs given by the plot above.

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

```
km = summary(survfit(Surv(kidney$time,kidney$status) ~ 1))
my.hazard=km$n.event/km$n.risk
cum.hazard=cumsum(my.hazard)
myvar=cumsum(km$n.event*(km$n.risk-km$n.event)/(km$n.risk^2*(km$n.risk-1)) )
mysd=sqrt(myvar)
surv_hat_1 = c(1, km$surv)
time_hat = c(0, km$time)
mu_hat_1 = 0
for(i in 2:length(surv_hat_1)) {
mu_hat_1 = mu_hat_1 + surv_hat_1[i - 1] * (time_hat[i] - time_hat[i - 1])
}
print(paste("Estimate of the mean survival time using K-M approach is ",round(mu_hat_1,3)))
```

```
## [1] "Estimate of the mean survival time using K-M approach is  137.02"
```

```
#Another approach to find mean and 95% CI together for K- M estimate:

km.fit = survfit(Surv(kidney$time,kidney$status) ~ 1)
print(km.fit,print.rmean=TRUE)
```

```
## Call: survfit(formula = Surv(kidney$time, kidney$status) ~ 1)
##
##          n     events      *rmean *se(rmean)     median    0.95LCL
##       76.0       58.0      137.0       19.8       78.0       39.0
##    0.95UCL
##      152.0
##     * restricted mean with upper limit =  562
```

```
lb=137.0-1.96*19.8; ub=137.0+1.96*19.8
print(paste("95% CI : (",round(lb,3),round(ub,3),")"))
```

```
## [1] "95% CI : ( 98.192 175.808 )"
```

```
# Mean survival time estimates for the Nelson-Aalen estimator of S(t).

surv_hat_2 = c(1, exp(-cum.hazard))
time_hat = c(0, km$time)
mu_hat_2 = 0
for(i in 2:length(surv_hat_2)) {
mu_hat_2 = mu_hat_2 + surv_hat_2[i - 1] * (time_hat[i] - time_hat[i - 1])
}
print(paste("Estimate of the mean survival time using Nelson-Alan approach is ",round(mu_hat_2,3)))
```

```
## [1] "Estimate of the mean survival time using Nelson-Alan approach is  141.761"
```

```
# Using bootstrap

B=1000
n=nrow(kidney)
b.mu1=integer()
b.mu2=integer()
for(b in 1:B){
ind=sample(1:n,n,replace = TRUE)
b.dat=kidney[ind,]
b.kmfit = survfit(Surv(b.dat$time,b.dat$status) ~ 1)
b.haz=b.kmfit$n.event/b.kmfit$n.risk
b.cumhaz=cumsum(b.haz)
b.sv1=c(1,b.kmfit$surv)
b.sv2=c(1,exp(-b.cumhaz))
b.time=c(0,b.kmfit$time)
mu_hat_1 = 0
for(i in 2:length(b.sv1)){
mu_hat_1 = mu_hat_1 + b.sv1[i - 1] * (b.time[i] - b.time[i - 1])
}
b.mu1[b]=mu_hat_1
mu_hat_2 = 0
for(i in 2:length(b.sv1)){
mu_hat_2 = mu_hat_2 + b.sv2[i - 1] * (b.time[i] - b.time[i - 1])
}
b.mu2[b]=mu_hat_2
}
mean(b.mu1)
```

```
## [1] 136.9765
```

```
mean(b.mu2)
```

```
## [1] 144.9894
```

```
quantile(b.mu1, c(0.025,0.975))
```

```
##     2.5%    97.5%
## 100.5836 178.1091
```

```
quantile(b.mu2, c(0.025,0.975))
```

```
##     2.5%    97.5%
## 108.8427 185.7913
```