UC Davis

Katie Truong & Gordon Hung

5/25/16

Project 2
STA 104

jail.csv

I. Introduction:

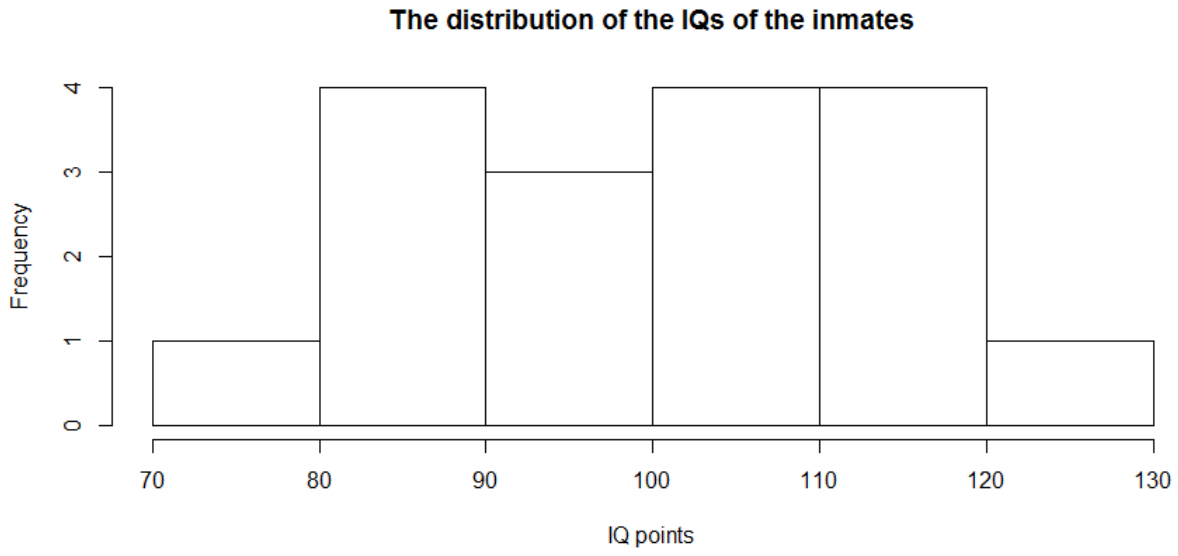
The data consists of the records of IQs and genders of a sample of inmates. Based on the given information, we are trying to answer two questions: Are the median IQ of those in prison is lower than the overall median IQ (question 1), and if the median differs by gender (question 2). These questions are important because they challenge the fairness of the justice system (have they ever wrongly convicted someone who has extremely low IQ) or to develop educational programs inside the prison that are suitable for both genders and their level of intelligence.

II. Summary of the data:

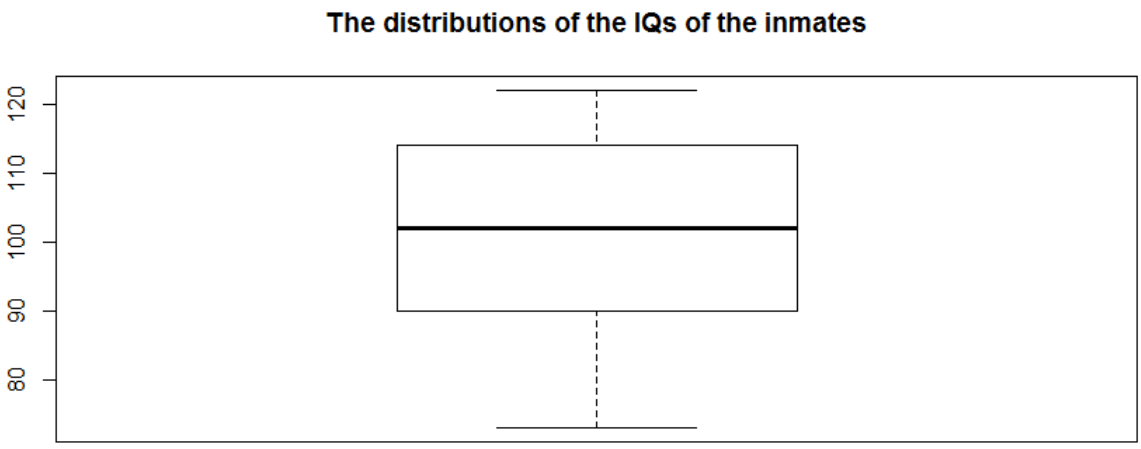
The data consists of the records of IQs and genders of 17 inmates. The numbers of observations in the sample are unevenly distributed between two genders: there are 8 observations of male and 9 observations of female.

The summary of the data set is given below:

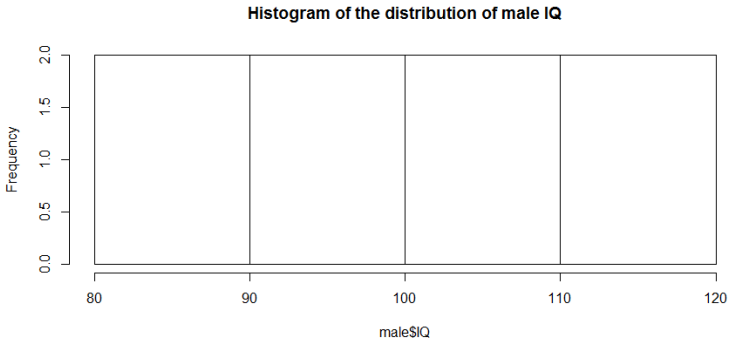
	Min	1st quarter	Median	Mean	3rd quarter	Max	SD
All	73.0	90.0	102.0	101.4	114.4	122.0	14.69
Male	82	90.25	100.0	100.4	111.0	120.0	13.71
Female	73.0	90.0	104.0	102.3	116.0	122.0	16.28

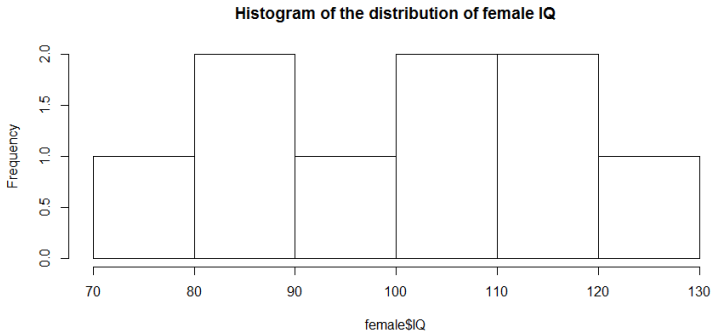


The histogram of the overall distribution shows that it's roughly symmetric, with no inherent skewness.



The boxplot shows that the IQ points has no outliers.





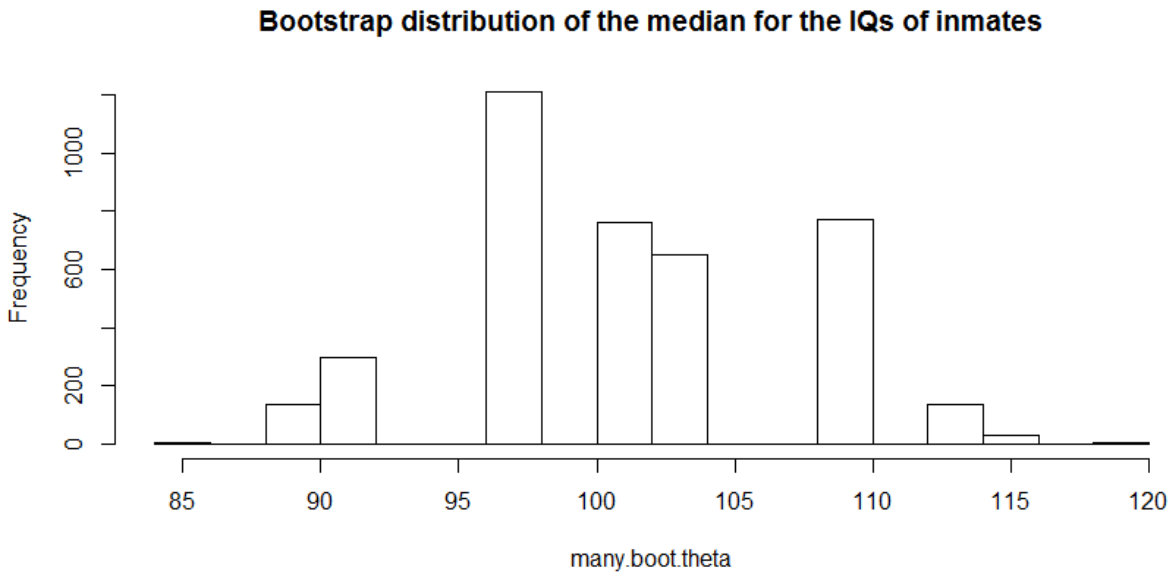
The histograms of the distribution of IQs between for each group also shows that they are roughly symmetric, with no skewness.

However, since the sample size is very small (n = 17) and unevenly distributed between two groups, and we can't conclude anything about the whole population of inmates based on such small sample, we still have to use bootstrapping to add more variances and variabilities to the sample.

At this point, I would divide the Analysis and the Interpretation of the report into two groups to tackle the two questions separately.

Question 1:

IIIa. Analysis:



The distribution of the median of the IQs of inmates for 4000 bootstraps shows skewness and unevenness in many part of the distribution. Using the Shapiro test function with the null hypothesis that the distribution is normal, we have a p-value of 2.2×10^{-16}, which rejects the null hypothesis as well as the normality assumption for the distribution.

Since the bootstrap distribution is not normal and showing skewness, the BCA method would work best in this situation. Using the boot package, we are able to extract the bootstrap 95% confidence interval of the median IQ to be (89, 110).

The confidence interval has a width of 21 and a center of 99.5, which is close but smaller to the median of the observation.

IVa. Interpretation

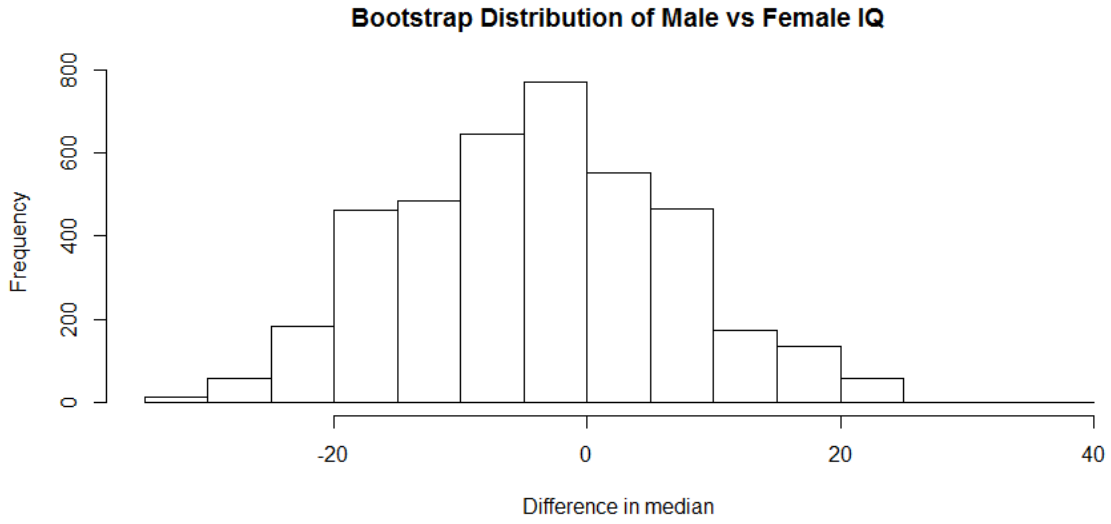
If we randomly sample the recorded IQ of 17 inmates for 4000 times and calculate the median, we can be 95% confident that the true median of the recorded IQs would be between (89, 110).

Since the BCA 95% confidence interval is not strictly below 100, we can't say that the median for those in the prison is lower than the overall median.

Question 2:

IIIb. Analysis:

I perform 4000 bootstraps for the difference in median of the recorded IQs between the male and female inmates. Since the square estimate of the difference in median is unknown, we resort to use simple resampling method.



The bootstrap distribution of the difference in median of male and female recorded IQs appears to be roughly symmetric. However, it doesn't pass the Shapiro test of normality with a p-value of

The 95% percentile confidence interval of the difference in median IQ between male and female inmates are (-25, 16).

To double check, we also look at the 95% percentile confidence interval, which is (-24, 17).

So we can see that the two confidence intervals aren't not much different, which confirms the validity of our BCA confidence interval.

The center of the BCA confidence interval is -4.5, while the width of the confidence interval is 20.5.

IVb. Interpretation:

If we randomly resample the two groups of inmates for 4000 times and record the difference in median IQ between male and female inmates, we can be 95% confident that the true difference of the median IQs between two genders are between -25 and 16, with the negative value corresponds to the occasion that the male IQ is lower.

Since the 95% confidence interval covers zero, we can conclude there is no significant difference in the median of the IQs of the two groups.

V. Conclusion:

By using the nonparametric bootstrap methods and BCA confidence intervals for the sample of 17 inmates, we are able to draw two conclusions from the sample. Firstly, the median IQ of the inmates is not different than the overall IQ of the population (which is known to be equal to 100). Secondly, there is no significant difference between the median IQ of the male inmates and of the female inmates. In other word, IQ is an indiscriminate statistics, which hardly depends on one's background, occupation, or gender. The legal system and educational system structure should be built based on this conclusion.

```
library(boot)
setwd("C:/R/STA 104/Project 2")
jail = read.csv("jail.csv")

summary(jail$IQ)
sd(jail$IQ)
hist(jail$IQ, main = "The distribution of the IQs of the inmates", xlab = "IQ
points")
boxplot(jail$IQ, main = "The distributions of the IQs of the inmates")
shapiro.test(jail$IQ)

male = subset(jail, gender == "M")
summary(male$IQ)
sd(male$IQ)
hist(male$IQ, main = "Histogram of the distribution of male IQ")
shapiro.test(male$IQ)

female = subset(jail, gender == "F")
summary(female$IQ)
sd(female$IQ)
hist(female$IQ, main = "Histogram of the distribution of female IQ")
shapiro.test(female$IQ)

###
###QUESTION 1###

##Estimate theta
R = 4000
boot.median= sapply(1:B,function(i){
  boot.sample = sample(jail$IQ,length(jail$IQ), replace = TRUE)
  boot.median = median(boot.sample)
  boot.t                    =                    (boot.median                    -
median(jail$IQ))/(sd(boot.sample)/sqrt(length(jail$IQ)))
  return(c(boot.median,boot.t))
})
jail.1 = jail$IQ
median.fun = function(jail.1,random){
  boot.data = jail$IQ[random]
  return(median(boot.data))
}
median.obs = median(jail.1)
median.boot = median(boot.median[1,])
B = 4000
many.boot.theta = sapply(1:B, function(i){
  boot.sample = sample(jail$IQ, length(jail$IQ),replace  = TRUE)
  theta.i = median(boot.sample)
  return(theta.i)})
hist(many.boot.theta,main="Bootstrap distribution of the median for the IQs of
inmates")
shapiro.test(many.boot.theta)
##Not normal

library(boot)
```

```
theta.fun = function(jail, random.indices){
  boot.data = jail$IQ[random.indices]
  theta.i = median(boot.data)
  return(theta.i)
}
r.bootstraps = boot(jail,theta.fun, R = 4000)
r.bootstraps
alpha = 0.05
r.ci = boot.ci(r.bootstraps,type = c("perc","bca"),conf = 1-alpha)
r.ci

##QUESTION 2
B = 4000
boot.two = sapply(1:B,function(i){
  group1 = male$IQ
  group2 = female$IQ
  boot.group1 = sample(group1, length(group1),replace = TRUE)
  boot.group2 = sample(group2, length(group2),replace = TRUE)
  theta.i = median(boot.group1) - median(boot.group2)
  return(theta.i)
})
hist(boot.two,main = "Bootstrap Distribution of Male vs Female IQ",xlab =
"Difference in median")

#Percentile method
alpha = 0.05
ci.percentile = as.numeric(quantile(boot.two, c(alpha/2, 1-alpha/2)))

##Preparing for the boot library
theta.fun.2 = function(jail, random.indices){
  boot.data = jail$IQ[random.indices]
  theta.i = median(boot.two)
  return(theta.i)
}

##Bootstrap
r.bootstraps.2 = boot(jail, theta.fun.2, R = 4000)
r.bootstraps.2

##Doesn't work somehow

##Source:  http://stats.stackexchange.com/questions/22945/confidence-interval-
for-the-difference-of-two-means-using-boot-package-in-r
medianDiff = function(jail, indexVector) {
  m1 = median(subset(jail[indexVector, 1], jail[indexVector, 2] == "M"))
  m2 = median(subset(jail[indexVector, 1], jail[indexVector, 2] == "F"))
  m = m1 - m2
  return(m)
}

totalBoot = boot(jail, medianDiff, R = 4000)
totalBootCI = boot.ci(totalBoot)
```