Assignment1

Kiara Beilinsohn

26 February 2019

## Introduction

The objective of this study is to analyse the data from the Income and Expenditure Survey conducted by Statistcs South Africa for a given year. The aim is to investigate confidence intervals for the age part of the data, perform hypothesis on the age data and construct a 95% confidence intervel for the upper quartile of the household income data. In addition the anyalysis of histograms is performed. To solve this problem we will use both the bootstrap sampling method and standard normal theory.

## The data

The sample of the data was taken from the Income and Expenditure Survey by Statistics South Africa. The whole dataset is seen first, then the age data, and lastly the hhinc data (created with my index 143). A section of this data can all be seen below:

## [1] "All the data"

## PROVINCE GENDER RACE AGE AREATYPE HHSIZE HHEXPENDITURE HHINCOME  
## 1 EC Female BLACK 22 1 2 18675.37 17089.80  
## 2 WC Male COLOURED 63 1 2 19615.78 20365.39  
## 3 EC Male BLACK 35 1 3 24653.62 24751.26  
## 4 EC Female COLOURED 49 2 7 20402.55 22780.69  
## 5 EC Male BLACK 21 1 3 20410.74 18716.14  
## 6 WC Female COLOURED 42 1 4 22524.18 22264.18

## [1] "Age data"

## [1] 33 24 34 44 50 24 23 39 46 35 33 40 35 26 42 63 51 28 41 29 64 42 38  
## [24] 36 61 29 34 57 34 51 38 55 55 28 52 31 54 48 62 50 37 27 51 46 41 49  
## [47] 60 29 57 31 50 56 31 35 29 55 33 32 62 40 48 37 26 39 50 41 60 26 39  
## [70] 36 59 57 24 54 41 35 30 31 22 40 24 35 51 41 59 48 49 60 35 57 37 33  
## [93] 65 64 45 40 49 41 39 34

## [1] "hhinc data"

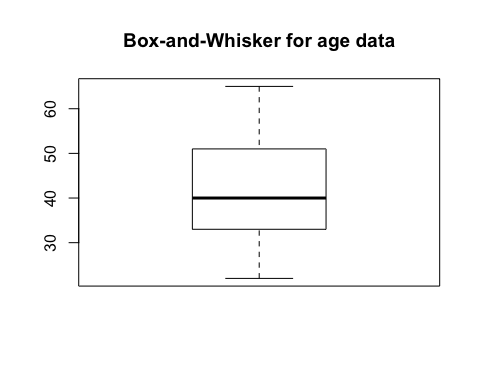
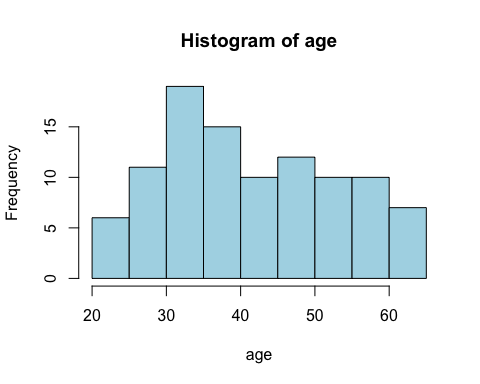
## [1] 22853.87 19588.37 23302.44 23189.00 24660.93 24542.43 20685.32  
## [8] 23957.60 20033.32 23633.64 19722.42 26316.74 22017.55 22847.04  
## [15] 23682.11 24782.54 27864.24 16748.71 24172.17 24134.13 21460.69  
## [22] 23333.42 21064.30 18453.05 21233.03 21695.57 21918.49 18530.07  
## [29] 24799.02 22107.97

## Question 1: Explanitary data analysis

Histograms, boxplots and summaries are shown below with regards to the age and hhinc data. This is done by running the hist(), boxplot(), summary() and IQR() functions.

For Age data: The histogram produced shows that the data does not fit ant particular distrobution, it is therefore a non-parametric distrobution as it doesnt fit for example generic distributions such as the Normal distribution.

The boxplot shows that the distribution is slightly skewed to the right and the mean for the data is 42.00, the lower quartile is 33.00 and the upper quartile is 51.00 (as seen by the summary below), thus the Inter quatile range is 18.

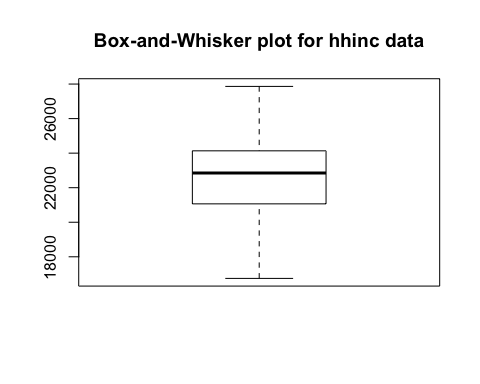
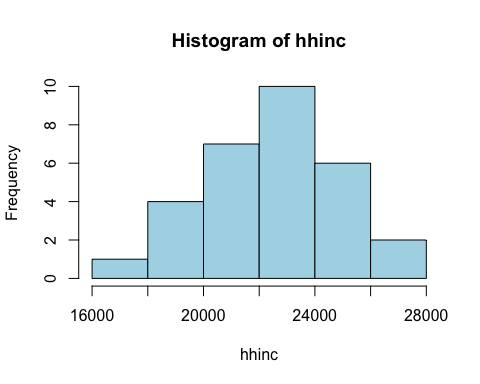


## [1] "summary of age data"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 22.00 33.00 40.00 42.11 51.00 65.00

## [1] 18

For the hhinc data: The histogram produced shows that the data is relativly symetrically distributed around the median and displays a Normal distrubution. The boxplot also conveys a symetrical distribution with a mean of 22444, the lower quatile being 21106 and the upper quartile being 2409, thus the Inter Quatile range is 2983.515. This can be seen below:



## [1] "summary of hhinc data"

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16749 21106 22850 22444 24090 27864

## [1] 2983.515

## Question 2: 95% confidence interval for the population mean age

## a) Bootstrapping approach

We make the assumption of the bootstrap principle where . Theres an assuption that the information is unbiased as it is made up of independant mesurements that is representative of the sample.

Bootstrapping was performed by resampling. The resampling was repeated 4000 times with replacement. The mean for each bootstrap sample was placed into an array. This information (in the array) was then used to calculate the 95% confidence interval for the mean.

The confidence interval is as calculated below using Bootstrapping:

## ( 39.89 , 44.32 )

## b) Using normal theory

This is based on the central limit theorem and therefore for large enough n we can assume

.

In order to calculate the confidence interval manually the following formula was used:

The confidence interval is as calculated below using normal theory:

## (39.8 , 44.42)

## Question 3: Using the bootstrap resampling theory to arrive at a conclusion about the following hypothesis test.,

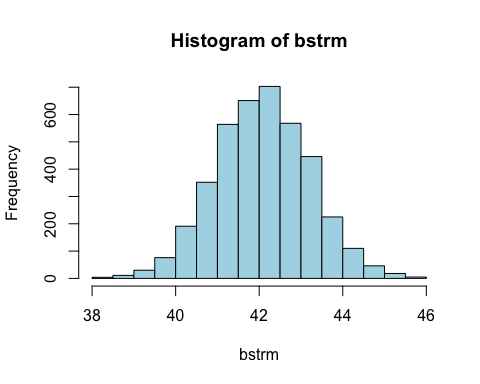
Using boostrap statistics, we first find the sampling errors which are . Then see the observed discrepency where . Now we sum the bootstrap sample errors that are less than the discrepency divided by the total number of bootstrap sample. In order to calculate the pValue the number of means over 43 plus the error portion was was calculated, then to get the pvlaue the number of those above 43 plus the error was divided by the size of the full sample, as seen:

It can be determened that we fail to reject as the pvalue is significanly larger then 0.05, as it is pValue and we can conclude at the 5% significance level that 43.

The pvalue can be seen below:

## [1] 0.5015

## Question 4: histogram of the distribution of the sample mean for age



The histogram reveals a symetrical shape (symetrical around the mean 42.11 ), which appears to have equal tails and be approximatly normally distributed.

## Question 5: 95 % confidence interval for the household income upper quartile

The confidence interval was worked out via the formula:

Since the confidence interval is for , we use a correction factor of 0.5, and a probability p equal to 0.75, and n equal to the length of hhinc (in this case it is 30). in this case r and s are the corresponding postitions in the sorted hhinc data (ie the data is now in order from highest to lowest) to obtain the confidence interval. As shown below:

## ( 23302.44 , 24799.02 )

## Conclusion

The data from the Income and Expenditure Survey conducted by Statistcs South Africa for a given year was analysed through histograms (which was found to be normal for question 4), boxplots and summaries - for question one. Various methods such as the bootstrap resampling method and normal theory method was used to construct a 95% confidence intervals and conduct hypothesis, this was done for question two. It was found that the confidence interval for the bootstrapping method was: (39.89 , 44.32 ) and the normal theory confidence intervak was ( 39.7950124 , 44.4249876 ). For question 3 the pValu was found to be pValue thus failing to reject at the 5% significance level. For question 5 the 95 % confidence interval for the household income upper quartile was found to be ( 2.33024410^{4}, r upperBound ).

## Plagarism Decleration

I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own. My work will not be copied by another student with the intention of passing it off as his own work.

Signed:

## R code

## R studio code:  
# Assignment 1 Kiara Beilinsohn BLNKIA001  
  
# loading in data and declaring variables  
mydata <-read.table("incexp.txt",header=TRUE)  
age = mydata$AGE[143:(143+99)]  
hhinc =mydata$HHINCOME[143:(143+29)]  
  
# checkcing the size of each vector  
length(age)  
length(hhinc)  
  
### QUESTION 1 ### exploritary data analysis:  
hist(age, col = "lightblue")  
boxplot(age,main="Box-and-Whisker")  
summary(age)  
  
hist(hhinc, col = "lightblue")  
boxplot(hhinc,main="Box-and-Whisker")  
summary(hhinc)  
  
  
# QUESTION 2 Bootstrapping for age approach a)  
  
bstr <- matrix(0,ncol=100,nrow=4000)  
for( i in 1:4000)  
{  
 samp=sample(age,size=100,replace=TRUE)  
 bstr[i,]=samp  
}  
  
#sorting the means:  
bstrm=apply(bstr,1,mean)  
  
#print(bstrm)  
bootstrapMean <- mean(bstrm)  
  
sortmean=sort(bstrm)  
# 95 % CI   
print("Bootstrapping method")  
lowBound <- sortmean[0.025\*4000]  
upBound <- sortmean[0.975\*4000]  
cat("(",lowBound, ",",upBound,")")  
  
# standard normal theory approach b)  
#x-u/sd/sqrtn - formula to use  
sampleMean <- mean(age)  
standardDev <- sd(age)  
n <- length(age)  
sqrtn <- sqrt(n)  
  
  
#getting the crit value for conf interval, using a t value as the population varience is unknown:  
tval<-qt(0.975,99)  
lowerBound<-sampleMean-tval\*(standardDev/sqrtn)  
upperBound<-sampleMean+tval\*(standardDev/sqrtn)  
print("Theory method")  
cat("(",round(lowerBound,2)," , ",round(upperBound,2),")",sep="")  
  
##########################################################################  
# QUESTION 3 hypothesis test H0: u<= 43 H1: U>43  
#meanGta43  
  
# bootstrap mean - sample mean   
# count bootstrap means >   
bEr <- bootstrapMean - 43  
meansOver43 <- (bstrm > 43 + bEr )  
sum <- sum(meansOver43)  
  
pValue <- sum/4000 #the number of whole sample  
pValue  
##############################################################################  
  
# QUESTION 4   
hist(bstrm, col = "lightblue")  
# it is a symetrical distribution which appears to have equal tals and be approximatly normally distributed  
  
  
# QUESTION 5   
cf <- 0.5  
p <- 0.75  
n <- length(hhinc)  
z <- qnorm(0.975)  
sortHhinc <- sort(hhinc)  
# getting r the lower limit for the confidence interval  
  
r <- round(cf +n\*p+(-z)\*sqrt(n\*p\*(1-p)))  
lowerBound <- sortHhinc[r]  
  
# getting s the upper limit for the confidence interval  
s <- round(cf +n\*p+(z)\*sqrt(n\*p\*(1-p)))  
upperBound <- sortHhinc[s]  
  
# conf interval  
cat("(", lowerBound, " , ", upperBound, ")")