University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad •

SPES BONA

Student name: Shamiso Chikamhi

Student number: CHKSHA007

STA3030F: Project 1

13 March 2020

1 Table of contents.

2 Introduction

2.1 Objective.

The objective of this project is to use the bootstrap technique to explore the behavior of different data sets, in order to give an estimate of what the population parameters might be given the samples. This method is based on the bootstrap assumption that the behavior of the bootstrap statistics around the original sample statistic is reflects the behavior of the sample statistic around the parameter value of the population.

2.2 Question 1

Question one is based on an Income and Expenditure Survey conducted in South Africa. From this the age of the heard of the household as well as the household income is explored. Summary statics as well as boxplot ad histograms are used to understand the data. A 95% confidence interval of the mean age if constructed and the mean age is tested on the null hypothesis that it is less than or equal to 43. The bootstrap mean age is also plotted on a histogram to get a picture of the distribution. Lastly, the 90% confidence interval of household income median is constructed.

2.3 Question 2

This question is based on the data collected for crops yielded after being grown using different fertilisers, A and B. the objective of the experiment was to find if the two fertilisers would yield different amounts of crops. The first step was to test whether they mean amount of crop yielded for each fertiliser was different. In addition, the confidence interval of the mean difference was computed. Lastly the variance was tested for equality using the bootstrap approach as well as the normal theory.

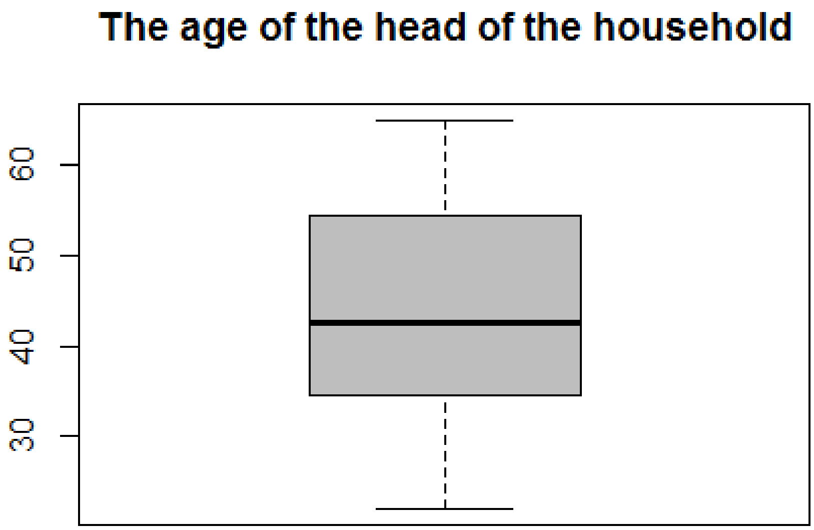
2.4 Question 3

Question 3 is based on data collected from 60 stores of how much novelty items are sold given the price ranges of R9, R10 and R11. This question will conduct an analysis of variance test using the bootstrap technique testing whether the mean sales at the 3 prices are equal or not.

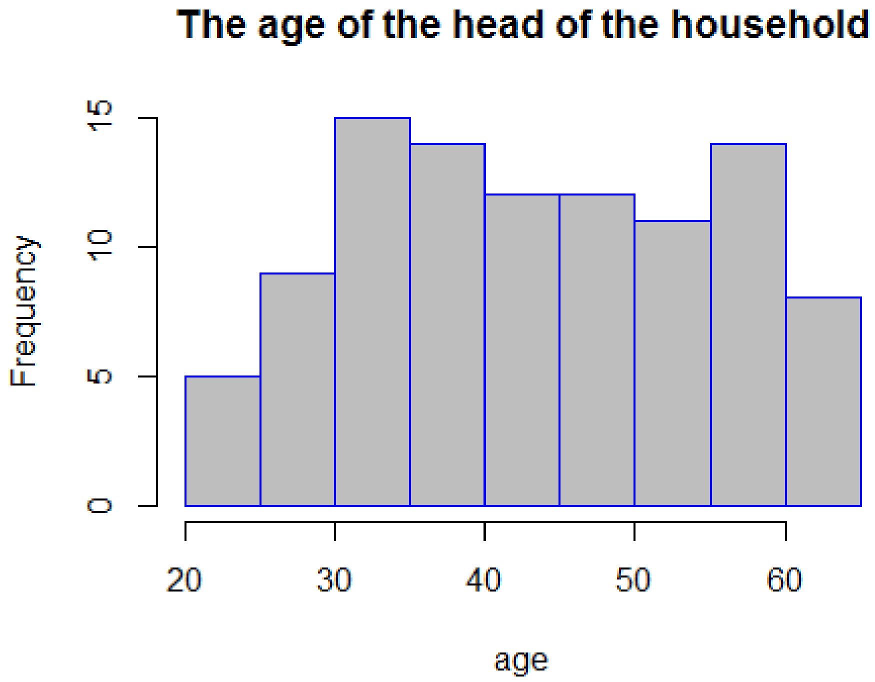
3 Question 1

3.1 Part A: Exploratory data analysis.

Below is the boxplot and the histogram representing the distributions of the mean age variable.



3.2

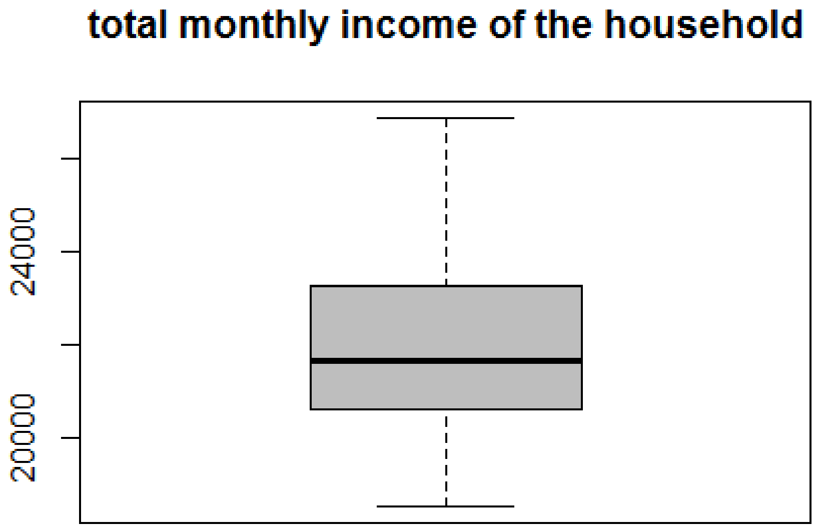


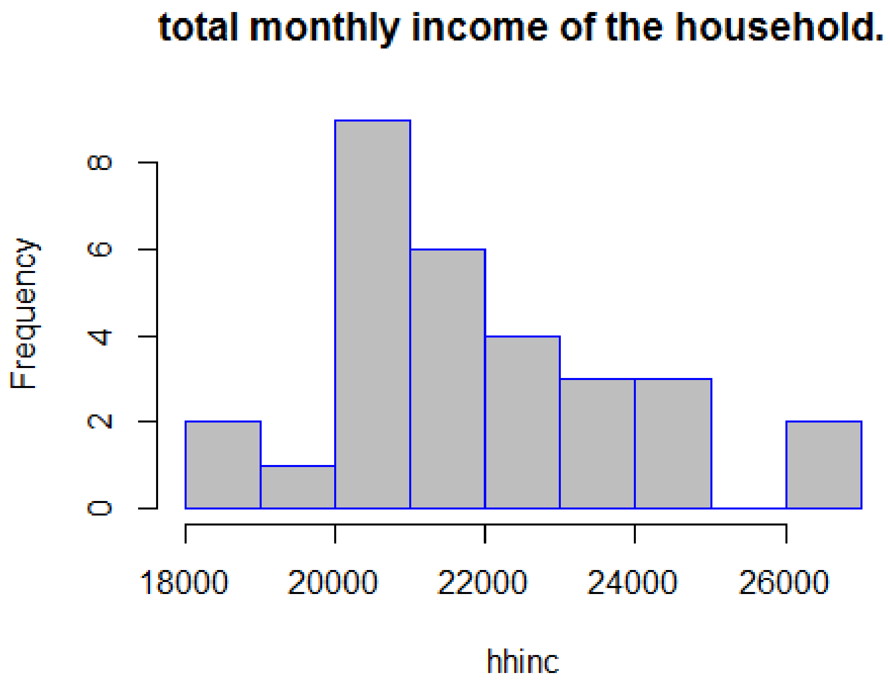
```
## The summary statistics of the age variable are given as:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   22.00   34.75   42.50   43.82   54.25   65.00
```

Confrimed by the summary statistics above, the age of the heard of household ranges from 22 years to 65 years old. It has a mean of 42.5 years old. This shows that there is quite a wide variety in the age of the household heards in South Africa. We also see from the histogram that te sample does not provide a normal distribution in the age of households heards. The hieghets number of househld heard are 30 to 40 years old.

The histogram and the box plot of the house hold income is shown below.





```
## The summary statistics of the household income variable are given as:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18530   20608   21662   21940   23151   26873
```

The minimum household income here is R13 530 while the maximum is R26 873. This is quite a big difference. the majority of the households in the data earn between R20 000 and R22 000. The mean household income is R21 662.

3.3 Part B: 95% confidence interval.

To obtain the confidence interval using the bootstrap technique;

- I performed the bootstrap process ranked the bootstrap mean age from smallest to largest.
- This allowed me to select the element on at the position 2.5% and 97.5% which gave me the bootstrap bounds of 41.57 and 46.1.
- I calculated the error bound by $\bar{X}b - \bar{X}$ to get -2.25 and 2.28
- Finally, the lower and upper bounds where calculated as $\bar{X} - error$ under the bootstrap assumption stated earlier. The bound where 41.54 and 46.07.

```
## The bootstrap 95% confidence interval is (41.57,46.1)

## The lower bound error is -2.25 and the upper bound error is 2.28
## The 95% confidence interval for the population mean age is (41.54,46.07)
```

The confidence interval (41.54, 46.07) means that, there is a 0.95 probability that these bounds are correct. We are 95% confident that we have the correct bounds stated are correct. The confidence interval has a fairly small range. This is a good range because we are able to accurately point out the mean age.

3.4 Part C Hypothesis test.

H_0: $\mu_{age} \leq 43$

H_1: $\mu_{age} > 43$

- Assume that H_0 is true, then $\mu_{age} \leq 43$.
- The sampling error given by $\bar{X} - \mu$ is 0.82 and $\bar{X} - \mu \geq 0.82$
- By the bootstrap assumption, $\bar{X}b - \bar{X} \geq 0.82$ hence $\bar{X}b \geq 44.64$
- After performing 4000 bootstraps to get bootstrap sample mean, the number of the bootstrap means greater than 44.64 was 970 which gave a p-value of 0.2424 =$\frac{970}{4000}$

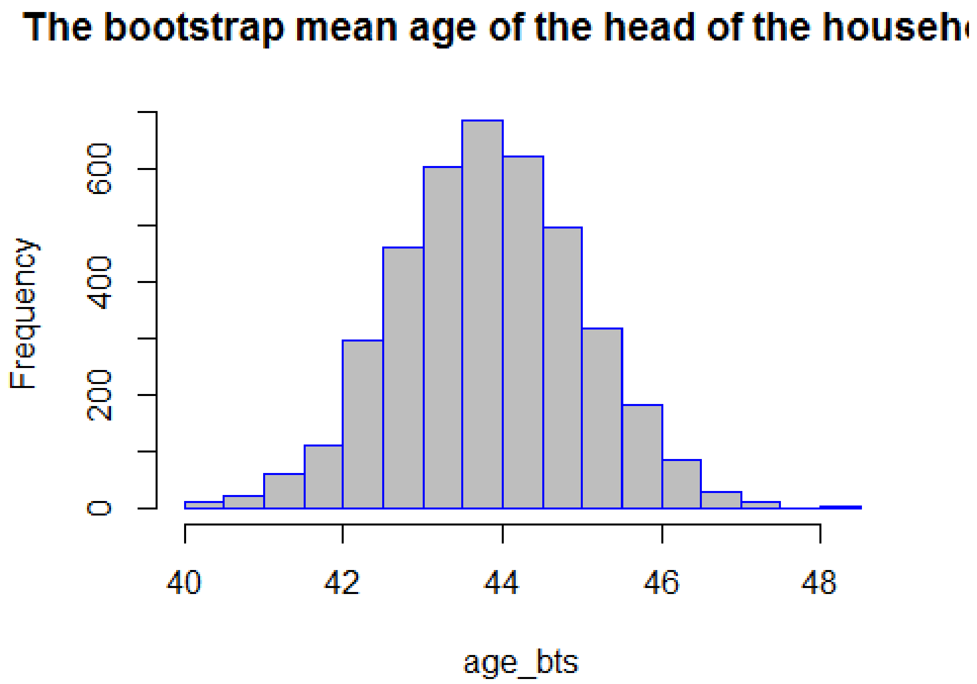
```
## The sample mean is 43.82 and the sampling error is 0.82

## The number bootstrap means greater than 44.64 is:

## [1] 970

## Hence the p-value is:0.2425

## The p-value is 0.2425 which is significantly large hence we fail to reject
null hypothesis and conclude that the mean age is less than 43.
```

The p value above is quite small and at a 5% significance level we can reject H_0 and conclude that the mean age for household heads is less than or equal to 43 years old.

This result almost tallies with our sample mean of 43.83 years old. This may suggest that the sample represents the population well.

3.5 Part D distribution of the sample mean age.

Below is a histogram showing the bootstrap mean age of the heads of households. It shows a normal distribution around 43.5 years old. This can be taken as the mean of the distribution and hence the bootstrap mean of the age is 43.5. this is not far off the sample mean of 43.8. and it also close to mean age of 43 that was tested by hypothesis in parct c above.



3.6

3.7 Part E: 90 % confidence interval of the median household income.

To obtain the confidence interval using the bootstrap technique;

- I performed the bootstrap process ranked the bootstrap median household income from smallest to largest.
- This allowed me to select the element on at the position 5% and 95% which gave me the bootstrap bounds of 20 803.5 and 22 107.97
- I calculated the error bound by $Mb - M$ to get -1 016.82 and 675.04
- Finally, the lower and upper bounds where calculated as $M - error$ under the bootstrap assumption stated earlier. The bound where 22 679 and 20 987.14.

We are 90% confident that these bounds are correct and there is 0.95 probability in repeated sampling of stating the correct bounds.

```
## The sample median is 21662.18

## Bootstrap 90% confidence interval is (20803.5,22107.97)

## The lower bound error is -1016.82 and the upper bound error is 675.04

## The 90% confidence interval for the household income median is

(22679,20987.14)
```

4 Question 2

The summaries of the different crop yields are given below. The mean yield from fertilizer A is 551.5 while that from fertilizer B is 576.8. The minimum for fertilizer A is 466 which is less than the minimum for fertilizer B of 486. The maximum yields however are almost similar with A having 669 and B having 665.

```
## The data summary of yield from fertiliser A is:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   466.0   511.0   546.5   551.5   586.0   669.0

## The data summary of yield from fertiliser B is:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   486.0   538.5   581.5   576.8   603.5   665.0
```

4.1 Part A Mean difference hypothesis

H_0: The mean yield from the two fertilizers are not different, $\mu b - \alpha a = 0$

H_1: The mean yields are different, $\mu b - \mu a \neq 0$

If the null hypothesis true, error = original sample mean difference, 25.25.

After performing 5000 bootstraps to get bootstrap mean differences, the differences greater than or less than 25.25 were 627. This gave a p-value of 0.1254.

```
## The original sample mean difference 25.25

## The number of bootstrap differences execeding or are less than the sample
mean difference:

## [1] 627

## The p-value is 0.1254

## The p-value is 0.1254 which is significantly large hence we fail to reject
null hypothesis and conclude that the mean crop yields from the two fertilise
rs are equal.
```

The p- value is significantly large and hence we cannot reject the null hypothesis and conclude that the mean difference is different. This result agrees with the sample means that have a difference of 25.25. We cannot however conclude that which of the two fertilizers gave the better mean crop yield as our tests was two sided.

4.2 Part B 95% confidence interval for the difference in means

To obtain the confidence interval using the bootstrap technique;

- Using the bootstrap mean difference in part a above, they can be ranked in ascending order.

- This allowed me to select the element on at the position 2.5% and 97.5% which gave me the bootstrap bounds of -32.15 and 32.7. Under the null hypothesis of no difference these are equal to error bound.

- Finally, the lower and upper bounds where calculated as $(\bar{X}b - \bar{X}a) - \ error$ under the bootstrap assumption stated earlier. The bound where -7.45 and 57.4.

```
## The bootstrap difference 95% confidence interval is (-32.15, 32.7)

## The lower bound error is -32.15 and the upper bound error is 32.7

## The 95% confidence interval for the mean yield difference is (-7.45, 57.4)
```

We are 95% confident these bound are correct and there is 0.95 probability in repeated sampling of stating the correct bounds. The range of the differences in this case is quite wide.

4.3 Part C Test for the variances for equality.

H_0: The variance in crop yield of fertilizer A is equal to that of fertilizer B

H_1: The variances are different

After performing 5000 bootstraps to get bootstraps f-rations, the following was found.

- The original F- ratio $(\frac{Var(A)}{var(B)})$ was 1.038

- In the bootstrap ratios, ratios greater than this f ratio were counted and are 2241.

- This gave a p-value of $\frac{2241}{5000}$ = 0.4482.

This p-value is too large and hence we fail to reject the null hypothesis and conclude that the variances of the crop yields from the two fertilizers are equal. Our original sample ratio is almost equal to 1 and it supports this evidence.

```
## The variance of yield from fetilier A is 2741.947 while that of fertilizer
B is 2641.145. The original sample F-ratio is  1.038166

## The number of bootstrap F-ratio greater than the original F-ratio:

## [1] 2241

## The p-value is 0.4482

## The p-value is 0.4482 which is significantly large hence we fail to reject
null hypothesis and conclude that the variane of the crop yields from the two
fertilizers are equal.
```

4.4 Part D Comparison of the results from bootstrapping and normal theory

The theoretical p-value is obtained from an f-distribution with 19,19 degrees of freedom and is given as 0.5321. This value is also large, and we fail reject null hypothesis and conclude that the variances are different.

```
## An F-test with 19,19 Degrees of freedom is conducted and the p-value is 0.
5321058

## The p-value is 0.5321058 which is significantly large hence we fail to rej
ect null hypothesis and conclude that the variane of the crop yields from the
two fertilisers are equal.
```

The bootstrap p-value `0.4482` does not differ significantly from the theoretical p-value 0.5321. this may suggest that our samples are well representing the population, which is the case because the data is based on the whole experiment where one-acre lands where used for each fertilizer and the yields where recorded.

5 Question 3

5.1 Part A test whether there are significant differences in means between the groups.

H_0: The mean sales at the 3 prices is the same, μ1 = μ2 = μ3

H_1: At least on mean differs from the others

- This is an ANOVA test
- The grand mean given by $Y.. = \sum_{j=1}^{ni}\frac{Yij}{ni}$ = 146.116. the mean sales (Yi.) of the different prices are also given in the output below.
- The SSE $=\sum_{i=1}^{k}\sum_{j=1}^{ni}(Yij - Yi.)$ = 41897.55
- SST $\sum_{i=1}^{k}(Yi. - Y..)$ = 5010.633
- Hence the original F-ratio is $\frac{SST/(k-1)}{SSE/(N-k)}$ = 3.408

```
## The grand mean is 146.1167

## The mean for price1 is 153.6, for price2 is 151.5 and for price3 is 133.25
The SSE is 41897.55 and the SST is 5010.633

## The original F-ratio is 3.408387
```

4000 bootstraps where conducted using the same procedure as above to get the bootstrap F-ratios. The results are discussed in part c.

5.2 Part B print-out of the first 3 bootstrap samples generated

The following where some of the sample generated in the bootstrap procedure.

```
## Samples for price 1:

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## [1,]  139  126  105  210   96   91  158  148  190   180   175   170   137
## [2,]  129  128  141  134  174  186  176  117  127   197   110   133   126
## [3,]  148  126  186  127  172  134  174   91  167   129   148   127   180
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## [1,]   186    96   140   106   152   141   110
## [2,]    91   126   158   126   210   146   170
## [3,]   171   171   186   141    99    91   167

## Samples for price 2:

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## [1,]  126  129  141  110  152  176  129  129  134   122   186   126   140
## [2,]  145  170  157  141  127  110  176  176  148   175   148    91   186
## [3,]  167  159  106  116  128  141  117  148  110   152   134   110   134
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## [1,]   117   152   144   122   176   172   117
## [2,]   103   133   171   168   129   146   145
## [3,]   186   180    96   137   172   129   128

## Samples for price 3:

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## [1,]  181  181  158  106  129  139  157  152  164   134   146   134   122
## [2,]  210  186  158  148  148  171  148  180  176    96   137   131   176
## [3,]  113  181  158  152   96  127  103  133  170    96   152   110   126
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20]
## [1,]   175   145   170   158   133   117   152
## [2,]   158   157   134   172   166   181   141
## [3,]   157   129   129   158   134   105   127
```

5.3

5.4 Part C p-value and conclusion based on the bootstrap results.

The p-value obtained by dividing the number of bootstrap F-ratios greater than the sample ratio by 4000 is 0.0874. this evidence suggests that at a 5% significance level we do no reject the null hypothesis and conclude that at least on mean differs.

The theoretical p -value is 0.0400 which suggests that the mean differences are equal.

```
## The bootstrap pvalue is 0.0874

## The theoretical pvalue is 0.03997518
## The p-value is 0.0874 which is significantly large hence we fail to reject
null hypothesis and conclude that there are no significant differences in mea
ns between the groups.
```

5.5 Part D Interpretation of the results.

Our theoretical p-value and the bootstrap p-value differ which may suggest that the sample was not well representing the data. However according to the result that the mean sales differ, I would advise the manager to explore further tests to see which of the sales means differ. The sample mean in this case shows that the mean sales with price3 R11 is 133.25 which is less than the mean sales for price1 and price2 which are 153.6 and 151.5 respectively. We can hence suggest to sale the novelty items at either R10 or R9.

6 Appendix: R Code

6.1 Question 1

```
# Defining data
mydata <- read.delim("F:/project 1/incexp.txt")
index = 167
age = mydata$AGE[index:(index+99)]
hhinc =mydata$HHINCOME[index:(index+29)]
```

6.1.1 Part A Exploratory data analysis.

```
# AGE variable

boxplot(age,main = "The age of the head of the household", col = "gray" )
```

```
cat("The summary statistics of the age variable are given as:")
summary(age)
hist(age, breaks = 15, main ="The age of the head of the household", col ="gray", bor
der="blue" )
```

```
# House hold income variable
boxplot(hhinc, main = "total monthly income of the household", col = "gray" )
```

```
cat("The summary statistics of the household income variable are given as:")
summary(hhinc)
hist(hhinc, main ="total monthly income of the household.", col ="gray", border="blue
" )
```

6.1.2 Part B 95% confidence interval for the population mean age.

```
sample_mean_age = mean(age)

# Bootstrap procedure for the mean
age_bts <- matrix(0,ncol=100,nrow=4000)
for( i in 1:4000)
{samp=sample(age,size=100,replace=TRUE)
age_bts[i,]=samp}
age_bts=apply(age_bts,1,mean)
age_sort_bts=sort(age_bts)

# Bootstrap confidence interval
cat("The bootstrap 95% confidence interval is (",age_sort_bts[0.025*4000],",",age_sor
t_bts[0.975*4000],")")

# Lower and upper bound error
lb_error = age_sort_bts[0.025*4000] - sample_mean_age
ub_error = age_sort_bts[0.975*4000] - sample_mean_age
cat("The lower bound error is", lb_error, "and the upper bound error is",ub_error)
```

```
#Confidence interval bounds
bound1 = sample_mean_age - lb_error
bound2 = sample_mean_age - ub_error

#95% confidence interval
if (bound1<bound2){
  cat("The 95% confidence interval for the population mean age is (", bound1, ",", bo
und2, ")")
}  else{
  cat("The 95% confidence interval for the population mean age is (", bound2, ",", bo
und1, ")")
}
```

6.1.3 Part C Hypothesis test that age is less than or equal to 43 H0, H1

```
age_null_hyp = 43
error = sample_mean_age - age_null_hyp
cat("The sample mean is", sample_mean_age, "and the sampling error is", error)

# vector of elements that are greater than smaple mean plus error
pv_elements <- age_bts > (sample_mean_age+error)
cat("The number bootstrap means greater than", sample_mean_age+error, "is:")
sum(pv_elements)
cat("Hence the p-value is:")
p_value = (sum(pv_elements))/4000

# P-value interpretation
if (p_value < 0.05){
  cat ("The p-value is", p_value, "which is significantly small hence we reject null
hypothesis and conclude that the mean age is greater than 43.")
} else {
cat ("The p-value is", p_value, "which is significantly large hence we fail to reject
null hypothesis and conclude that the mean age is less than 43.")
}
```

6.1.4 Part D histogram of the distribution of the sample mean age.

```
hist(age_bts, breaks = 20, main ="The bootstrap mean age of the head of the household
", col ="gray", border="blue")
```

6.1.5 Part E 90 % confidence interval of the median household income.

```
sample_median_hhinc = median(hhinc)
cat("The sample median is", sample_median_hhinc)

# Bootstrap procedeure fro the meadian of household income
hhinc_bts <- matrix(0,ncol=30,nrow=4000)
for( i in 1:4000)
{samp1=sample(hhinc,size=30,replace=TRUE)
hhinc_bts[i,]=samp1}
hhinc_bts=apply(hhinc_bts,1,median)
hhinc_sort_bts=sort(hhinc_bts)

# Bootstrap confidence interval
```

```
cat("Bootstrap 90% confidence interval is (",hhinc_sort_bts[0.05*4000],",",hhinc_sort
_bts[0.95*4000],")")

# Lower and upper bound error
lb1_error = hhinc_sort_bts[0.025*4000] - sample_median_hhinc
ub1_error = hhinc_sort_bts[0.975*4000] - sample_median_hhinc
cat("The lower bound error is", lb1_error, "and the upper bound error is", ub1_error)

# Confidence Interval bounds
bound3 = sample_median_hhinc - lb1_error
bound4 = sample_median_hhinc - ub1_error
# 90% confidence interval
if (bound1<bound2){
  cat("The 90% confidence interval for the house hould income median is (", bound3, "
,", bound4, ")")
}  else{
  cat("The 90% confidence interval for the house hould income median is  (", bound3,
",", bound4, ")")
}
```

6.2 Question 2

```
# Defining data
data2 <-  read.delim("F:/project 1/E1.txt")
fertA = data2$Fert_A
fertB = data2$Fert_B

# Data summary
cat("The data summary of yield from fertiliser A is:")
summary(fertA)
cat("The data summary of yield from fertiliser B is:")
summary(fertB)
```

6.2.1 Part A Test whether there are significant differences in means between the two groups. H0 and H1

```
# Original sample mean differences
samp_meanA = mean(fertA)
samp_meanB = mean(fertB)
samp_meanDiff = samp_meanB - samp_meanA
cat("The original sample mean difference", samp_meanDiff)

# Bootstrap for the mean differences between the 2 groups
all_A_B =c(fertA,fertB)
bstA=bstB=matrix(0,ncol=20,nrow=5000)
for(j in 1:5000)
{samp2=sample(all_A_B,size=40,replace=TRUE)
bstA[j,] =samp2[1:20]
bstB[j,]=samp2[21:40]
}

bstA_mean=apply(bstA,1,mean)
bstB_mean=apply(bstB,1,mean)
bst_mean_diff=bstB_mean-bstA_mean


# Vector of elements that are greater than smaple mean plus error
```

```r
pv_diff <- (bst_mean_diff > samp_meanDiff)|(bst_mean_diff < -samp_meanDiff)
cat("The number of bootstrap differences execeding or are less than the sample mean d
ifference:")
sum(pv_diff)
p_value1 = sum(pv_diff)/5000
cat("The p-value is", p_value1)


# Interpretaion of the p-value
if (p_value1 < 0.05){
  cat ("The p-value is", p_value1, "which is significantly small hence we reject null
hypothesis and conclude that the mean crop yields from the two fertilisers are not eq
ual.")
} else {
  cat ("The p-value is", p_value1, "which is significantly large hence we fail to rej
ect null hypothesis and conclude that the mean crop yields from the two fertilisers a
re equal.")
}
```

6.2.2 Part B 95% confidence interval for the difference in means

```r
# Bootstrap confidence interval
sort_bst_mean_diff = sort(bst_mean_diff)
cat("The bootstrap difference 95% confidence interval is (", sort_bst_mean_diff[0.025
*5000],",",sort_bst_mean_diff[0.975*5000],")")

#Lower and upper bound error
lb2_error = sort_bst_mean_diff[0.025*4000]
ub2_error = sort_bst_mean_diff[0.975*4000]

cat("The lower bound error is", lb2_error, "and the upper bound error is", ub2_error)

# Confidence interval bounds
bound5 = samp_meanDiff - lb2_error
bound6 = samp_meanDiff - ub2_error
# Confidence interval
if (bound5<bound6){
  cat("The 95% confidence interval for the mean yield difference is (", bound5, ",",
bound6, ")")
}  else{
  cat("The 95% confidence interval for the mean yield difference is (", bound6, ",",
bound5, ")")
}
```

6.2.3 Part C Test for the variances for equality.

```r
# Original sample F-ratio
varA = var(fertA)
varB = var(fertB)
samp_F = varA/varB
cat("The variance of yield from fetilier A is", varA,"while that of fertiliser B is",
varB,". The original sample F-ratio is ", samp_F)

# Bootstrap F-ratio
bstA_var =apply(bstA,1,var)
bstB_var =apply(bstB,1,var)
bst_F = bstA_var/bstB_var
```

```
# Vector of elements that are greater than the sample F ratio
pv_F <- (bst_F > samp_F)
cat("The number of bootstrap F-ratio greater than the original F-ratio:")
sum(pv_F)
p_value2 = sum(pv_F)/5000
cat("The p-value is", p_value2)


# Interpretaion of the p-value
if (p_value1 < 0.05){
  cat ("The p-value is", p_value2, "which is significantly small hence we reject null
hypothesis and conclude that the variance of the crop yields from the two fertilisers
are different.")
} else {
  cat ("The p-value is", p_value2, "which is significantly large hence we fail to rej
ect null hypothesis and conclude that the variane of the crop yields from the two fer
tilisers are equal.")
}
```

6.2.4 Part D Comparison of the results from bootstrapping and normal theory

```
# Calculating the p-value
df1 = length(fertA)-1
df2 = length(fertB)-1
p_value3 = pf(samp_F, df1,df2)
cat("An F-test with", df1,",",df2, "Degrees of freedom is conducted and the p-value i
s", p_value3)

# Interpreting the p-value
if (p_value1 < 0.05){
  cat ("The p-value is", p_value3, "which is significantly small hence we reject null
hypothesis and conclude that the variance of the crop yields from the two fertilisers
are different.")
} else {
  cat ("The p-value is", p_value3, "which is significantly large hence we fail to rej
ect null hypothesis and conclude that the variane of the crop yields from the two fer
tilisers are equal.")
}
```

6.3 Question 3

```
# Defining data
data3 <- read.delim("F:/project 1/aov1.txt")
data3 <- na.omit(data3)
p1= data3$Price_9
p2 = data3$Price_10
p3 = data3$Price_11

price = list("numeric",length=3)
price[[1]]=c(p1)
price[[2]]=c(p2)
price[[3]]=c(p3)

n=c(length(price[[1]]),length(price[[2]]),length(price[[3]]))
```

```
N=sum(n)
k=length(price)
```

6.3.1 Part A test whether there are significant differences in means between the groups H0 H1

```
#ANOVA WITH ORIGINAL DATA
RATIOS = Yi.=vector("numeric")
Y.. = mean(unlist(price))
cat("The grand mean is", Y..)
SSE = SST = 0


for(j in 1:k)
{Yi.[j] = mean(price[[j]])
SSE = SSE+sum((price[[j]]-Yi.[j])^2)
SST = SST+n[j]*(Yi.[j]-Y..)^2
}

cat("The mean for price1 is",Yi.[1],"for price 2 is", Yi.[2], "and for price 3 it is"
, Yi.[3],"The SSE is",SSE, "and the SST is",SST)
F_sample = (SST/(k-1))/(SSE/(N-k))
cat("The original F-ratio is", F_sample)

# ANOVA with bootstrap samples
B3 = 5000
ALLprice = unlist(price)

bst1=bst2=bst3 = matrix(0, nrow = 5000, ncol = 20 )
for(h in 1:B3)
{SSE = SST = 0
Yi.=vector("numeric")
price = list("numeric",length=3)
bst_anova = sample(ALLprice, size=60, replace=TRUE)
bst1[h,] = bst_anova[1:20]
bst2[h,] = bst_anova[21:40]
bst3[h,] = bst_anova[41:60]

price[[1]]=bst_anova[1:20]
price[[2]]=bst_anova[21:40]
price[[3]]=bst_anova[41:60]

for(j in 1:k)
{Yi.[j] = mean(price[[j]])
SSE = SSE+sum((price[[j]]-Yi.[j])^2)
SST = SST+n[j]*(Yi.[j]-Y..)^2
}
RATIOS[h]=(SST/(k-1))/(SSE/(N-k))
}
```

6.3.2 Part B print-out of the first 3 bootstrap samples generated

```
# First 3 bootstrap samples
cat("Samples for price 1:")
bst1[1:3,]
cat("Samples for price 2:")
```

```
bst2[1:3,]
cat("Samples for price 3:")
bst3[1:3,]
```

6.3.3 Part C p-value and conclusion based on the bootstrap results.

```
p_value4 = length(RATIOS[RATIOS>F_sample])/B3

theory_pvalue = pf(F_sample, k-1, N-k, lower.tail = FALSE)
cat("The bootstrap pvalue is", p_value4)
cat("The theoretical pvalue is", theory_pvalue)


# Interpretation of the p-value
if (p_value1 < 0.05){
  cat ("The p-value is", p_value4, "which is significantly small hence we reject null
hypothesis and conclude that there are significant differences in means between the g
roups.")
} else {
  cat ("The p-value is", p_value4, "which is significantly large hence we fail to rej
ect null hypothesis and conclude that there are no significant differences in means b
etween the groups.")
}
```

Department of Statistical Sciences Plagiarism Declaration form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will NOT be marked.

COURSE CODE: **STA** 3030F

COURSE NAME: Statistical inference and modelling

STUDENT NAME: Shamiso Chikmhi

STUDENT NUMBER: CHKSHA007

TUTORS NAME: ____________ TUT. GROUP #: 1

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
3. This tutorial/report/project is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else’s assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University’s plagiarism rules (http://www.uct.ac.za/uct/policies/plagiarism_students.pdf).

Signature: slchikamhi Date: 13-03-2020