STA3030F Project 2

Kiara Beilinsohn

14 March 2019

Introduction

The objective for this assignment is to analyse the data from two sample datasets. The datasets C1 and D2 were utilised The first segment of data for this project was obtained from 30 different people who filled in a form before and after it was changed (paired t test), the times were recorded. The data was collected by the internal revenue service as they were looking for ways to improve the wording and format of its tax return forms.The aim for this study is to test if there is significant differences in the completion time. The second segment of this data was obtained from a study perfomed by the Columbia University with regards to two different departments and the amount of times lectures said “uh” and “ah” in lectures to fill the gap between words. 50 observations were taken from each department. This assignment focuses on testing whether there are significant differences between each of the variences among the groups in the different data sets. It will focus on the differences in means and variences, using a variety of different methods; namely Bootstrap resampling and Standard Normal theory. The results will also provide us with a confidence interval for the difference between the two means.

The data provided is as follows:
form1 = 23, 59, 68, 122, 74, 90, 70, 87, 155, 120, 124, 103, 54, 90, 124, 80, 69, 123, 76, 71, 94, 167, 69, 105, 98, 73, 79, 61, 121, 56
form2 = 88, 114, 81, 41, 108, 92, 52, 54, 103, 50, 135, 76, 143, 124, 151, 96, 76, 128, 60, 127, 109, 122, 88, 109, 90, 56, 105, 64, 127, 104

Question 1

Question 1, involves a Bootstrapping method. I created 5000 Bootstrapped statistics by replicating the original sampling process. I resample by placing all observations from the original sample into a hat, then drawing the wanted observations from the hat with replacement to form a Bootstrap sample. A mean is then calculated for each Bootstrap sample, the first 5 of these values can be seen below. I then subtract these means from the two Bootstrap samples of the different original samples to obtain a specific Bootstrap statistic, which I store in an array. This is done 5000 times.

The hypotheses for this test are::
$H_0: \mu_n = \mu_p$ (There is no difference in the population means)
$H_1: \mu_n \neq \mu_p$ (The form1 population mean is different to the form2 population mean)

I let $\overline{x_b}$ denote the Bootstrap statistic and $\overline{x_d}$ denote the difference between the means of the original samples. Therefore the sampling error is $\overline{x_b} - \overline{x_d}$.Then the observed

discrepencies are calculated which is $\overline{x_d} - 0$ under the null hypothesis. The p-value is then calculated using $p = Pr(\overline{x_b} - \overline{x_d} \geq \overline{x_d} - 0) = Pr(\overline{x_b} \geq 2\overline{x_d} - 0)$ and this is done under the Bootstrap assumption : where $(\overline{x_b} - \overline{x} \approx \overline{x} - \mu)$. The p-value through Bootstrapping is calculated via

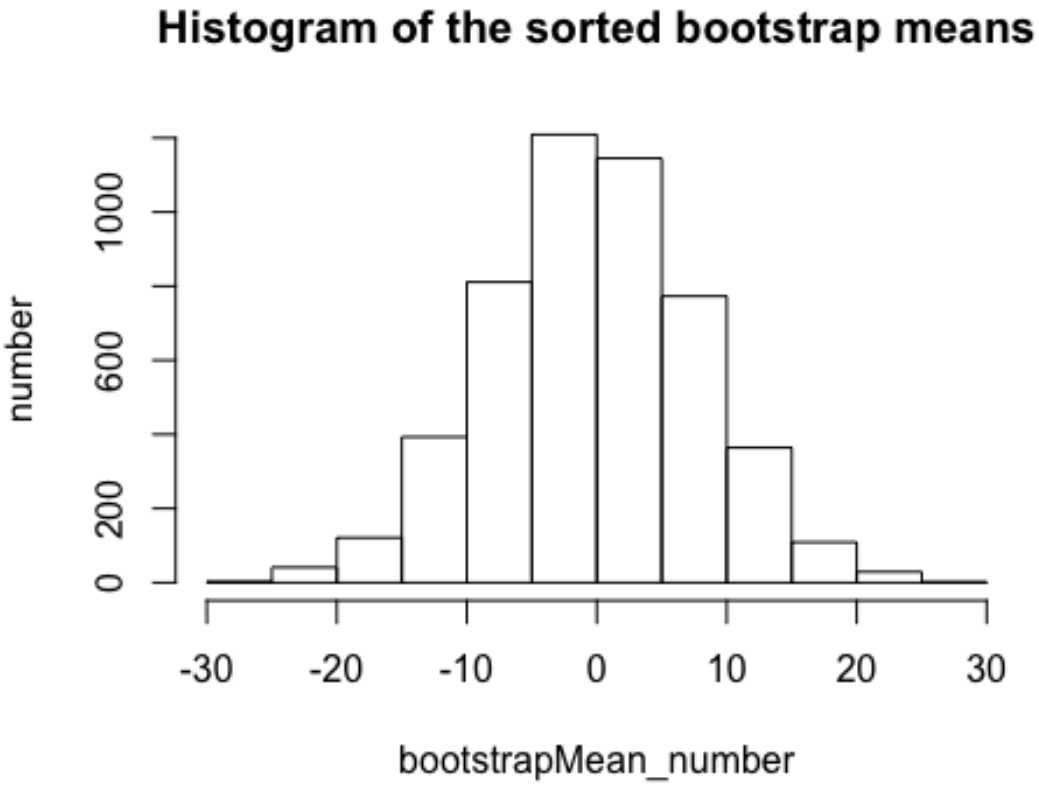
$$pValue = \frac{sum}{sizeOfSample}$$

and in this case it is equal to 0.161.

The p-value is defined as the probability of observing a similar or more extreme result, assuming the null hypothesis is true. With the calculated pvalue, it can be concluded that there is strong evidence to fail to reject the null hypothesis and conclude that there is no difference in the performance on the two forms and that the population means are equal.

```
## [1] "First five values:"

## [1] 1.133333
## [1] 0.9
## [1] 3.433333
## [1] 4.233333
## [1] -8.466667

## [1] "The pvalue is:"

## [1] 0.161
```

In addition the analysis of the Bootstrap mean through a histogram displays a symetrical distribution around the mean is approximately 0.0252 which looks approximately like the Normal distribution, this can be seen below:



Question 2

The following assumptions can be made: the Bootstrap principle where $(\overline{x_b} - \overline{x} \approx \overline{x} - \mu)$. There's an assumption that the information is unbiased as it is made up of independant measurements that is representative of the sample.

Bootstrapping was performed by resampling. The resampling was repeated 5000 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. It was done by sorting the array of means ana obtaining the lower 0.025 limit and the upper 0.975 limit for the Bootstrap means.The 95% confidence interval for the difference in means was calculated via the Bootstrap principle: $(\overline{x_b} - \overline{x} \approx \overline{x} - \mu)$). It is calculated as follows:

```
## [1] "The 95% confidence intervals for the means is"

## [-26.567 ; 4.933]
```

Question 3

The data used for Question 3 is as follows:

dept1 = 4, 9, 8, 7, 4, 4, 8, 7, 9, 0, 7, 4, 7, 8, 4, 4, 7, 5, 0, 3, 3, 1, 3, 3, 4, 7, 10, 8, 7, 5, 1, 5, 3, 4, 5, 10, 5, 8, 4, 4, 4, 9, 5, 9, 5, 7, 3, 8, 10, 4
dept2 = 5, 4, 9, 5, 5, 5, 6, 4, 5, 5, 4, 0, 6, 7, 4, 3, 3, 6, 7, 4, 5, 1, 8, 9, 5, 9, 4, 7, 0, 9, 5, 6, 4, 5, 5, 6, 5, 5, 3, 4, 5, 6, 4, 3, 8, 3, 5, 7, 8, 8

Using Bootstrapping:

Question 3, involves a Bootstrapping method. I created 5000 Bootstrapped statistics by replicating the original sampling process. I resample by placing all observations from the original sample into a hat, then drawing the wanted observations from the hat with replacement to form a Bootstrap sample. The varience is then calculated for each Bootstrap sample, the first 5 of these values can be seen below. I then create a ratio of the sample1 varience over sample2 varience to obtain the F statistic (which can be seen below). The pvalue is then calculated from the F statistic. This was repeated 5000 times and an F distribution was observed.

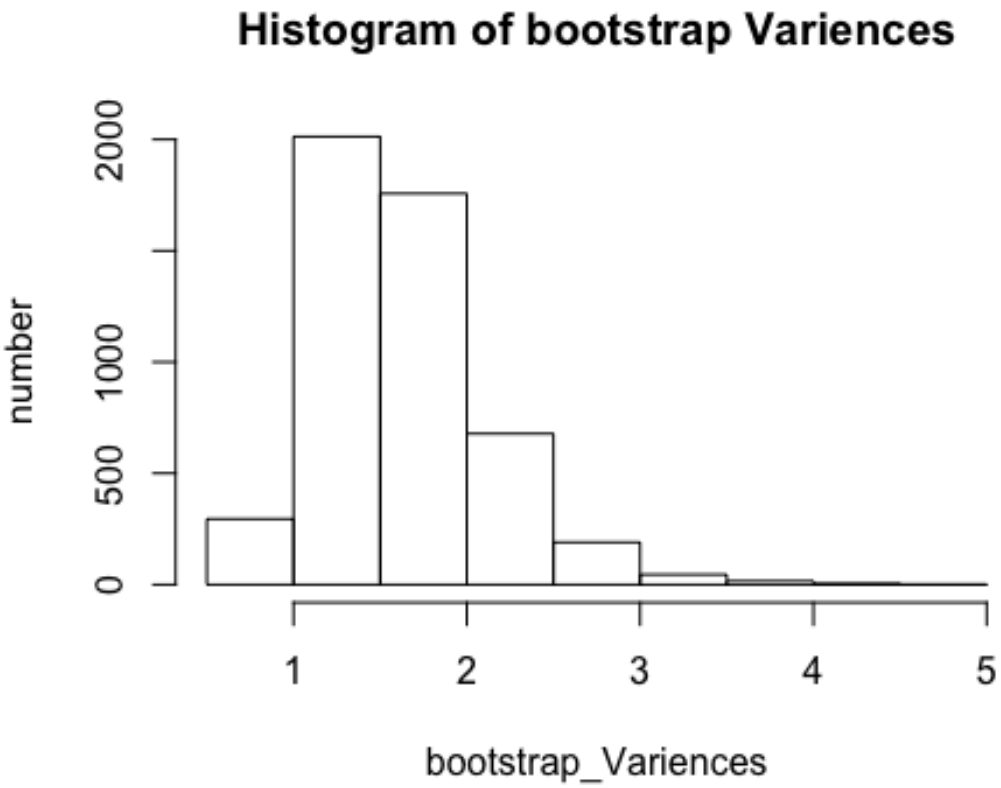
The hypotheses for this test are:
$H_0: \theta_1^2 = \theta_2^2$ (there is no difference in population variences)
$H_1: \theta_1^2 \neq \theta_2^2$ (The population variences are not equal)

```
## [1] "The ratio of variences is:"

## [1] 1.543412

## [1] "The pvalue is:"

## [1] 1
```

The first five values of the Bootstrap samples generated are:

```
## [1] "First five values:"

## [1] 1.222809
## [1] 1.150668
## [1] 1.664381
## [1] 1.524102
## [1] 0.9223945
```

An anyalysis of the Bootstrap variences has been plotted as a histogram, this demonstrates a relatively F distributed distribution, with a mean of approximately 1.7 as seen below:



Question 4

Using Normal theory

The Normal theory used to calculate the ratio of variences is done by calculating the F statsitic, this involves calculating the F statistic which is, dept1/ dep2 (varience1/varience2) and measuring that against the critical value which is F ($F_{n-1,m-1}$) with n-1 = 49 and m-1 =49 degrees of freedom.

```
## [1] "The ratio of variences is "

## [1] 1.543412

## [1] 0.9338914
```

Question 5

An interpretation of results can be seen as follows:

A comparision of the results from Q3 Bootstrapping method and Q4 Normal theory method shows: the Bootstrapping method and the Normal theory method vary in terms of the pvalues, however still in both cases we fail to reject H_0 because both of the pvalues are >

0.05, as the pvalue obtained from Bootstrapping is 1 and the pvalue obtained from Normal theory is 0.9338914. These differences occur due to random sampling in the Bootstrap method. Therefore, we can conclude at the 5% significance level that $\theta_1^2 = \theta_2^2$.

Determining whether or not the variences are equal based on their pvalues is important in determining which test statistic to compute with two sample problems, using the behrn Fisher test or the pooled varience test for the means.

Conclusion

Two sample problems can make use of various methods such as Bootstrapping and Normal theory to perform hypothesis tests on their means, in addition these methods can be used to test for equality of variences.

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

```
# putting the data in arrays
# dependant data
form1 =
c(23,59,68,122,74,90,70,87,155,120,124,103,54,90,124,80,69,123,76,71,94,167,6
9,105,98,73,79,61,121,56   )
form2 =
c(88,114,81,41,108,92,52,54,103,50,135,76,143,124,151,96,76,128,60,127,109,12
2,88,109,90,56,105,64,127,104)

size1 = length(form1)
size2 = length(form2)

xbar = mean(form1) - mean(form2)

# Question 1
# testing weather there is a significant differnece in the means
B = 5000
bootMeans = numeric(B)

for(i in 1:B)
{
  samp1 <- sample(form1, replace= TRUE, size = size1)
  samp2 <- sample(form1, replace= TRUE, size = size2)
  bootMeans[i] <- mean(samp1) - mean(samp2)
```

```r
}

# printing out the first 5 bootstrapMean values:
print("First five values:")
for(i in 1:5)
{
  print(bootMeans[i])
}
# getting the pvalue
pvalue =sum(abs(bootMeans)>=abs(2*xbar-0))/B
print("The pvalue is:")
pvalue


# Question 2


bootMeans=sort(bootMeans)
c1 <- round(bootMeans[125], digits = 3) # lower bound thats 2.5%
c2 <- round(bootMeans[4875], digits = 3) # upper bound thats 97.5%

#Extra
hist(bootMeans, main = "Histogram of the sorted bootstrap means", xlab =
"bootstrapMean_number", ylab = "number")

# Basic Bootstrap Interval
bbll <- round((xbar + (xbar - c2)), digits = 3)
bbul <- round((xbar + (xbar - c1)), digits = 3)
bbci <- c(bbll, bbul)
print("The 95% confidence intervals for the means is")
bbci


# independant data
dept1 <-
c(4,9,8,7,4,4,8,7,9,0,7,4,7,8,4,4,7,5,0,3,3,1,3,3,4,7,10,8,7,5,1,5,3,4,5,10,5
,8,4,4,4,9,5,9,5,7,3,8,10,4)
dept2 <-
c(5,4,9,5,5,5,6,4,5,5,4,0,6,7,4,3,3,6,7,4,5,1,8,9,5,9,4,7,0,9,5,6,4,5,5,6,5,5
,3,4,5,6,4,3,8,3,5,7,8,8)


# Question 3: using bootstraps to test the variences for equality

#Extra
hist(boot, main ="Histogram of bootstrap Variences", xlab =
"bootstrap_Variences", ylab= "number")

# Question 3: using bootstraps to test the variences for equality
```

```r
BB = 5000
boot = numeric(BB)

for(i in 1:BB)
{
  number = sample(1:50, replace = T, size = 50)
  sample1 = var(dept1[number])
  sample2 = var(dept2[number])
  boot[i] = sample1/sample2
}

ratio <- var(dept1)/var(dept2)
print("The ratio of variences is:")
print(ratio)
pVal <- sum(abs(boot>= boot- 2*ratio))/BB
print("The pvalue is:")
print(pVal)


# Question 4: Using normal theory
p.normFtest = var.test(dept2,dept1, alternative = "two.sided" )
print(p.normFtest)
print(p.normFtest$p.value)
```