Stats Exploration Reasoning HW_3

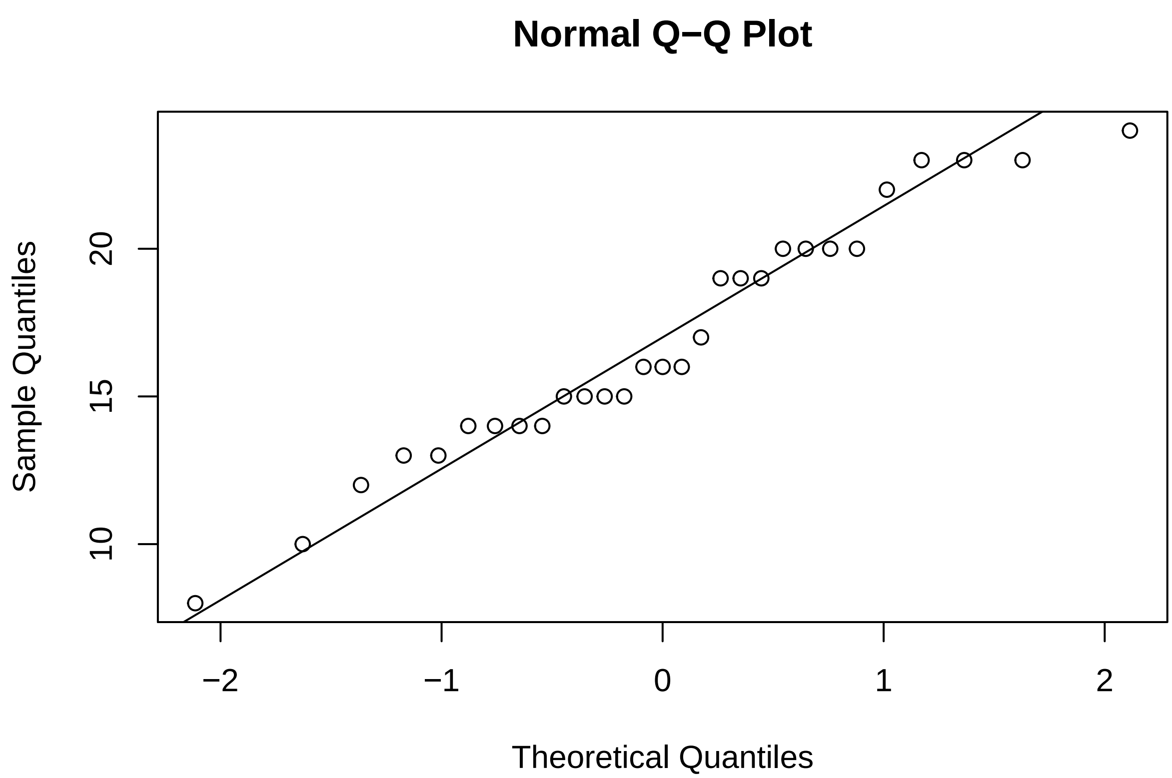
Shalini Mishra

10/20/2019

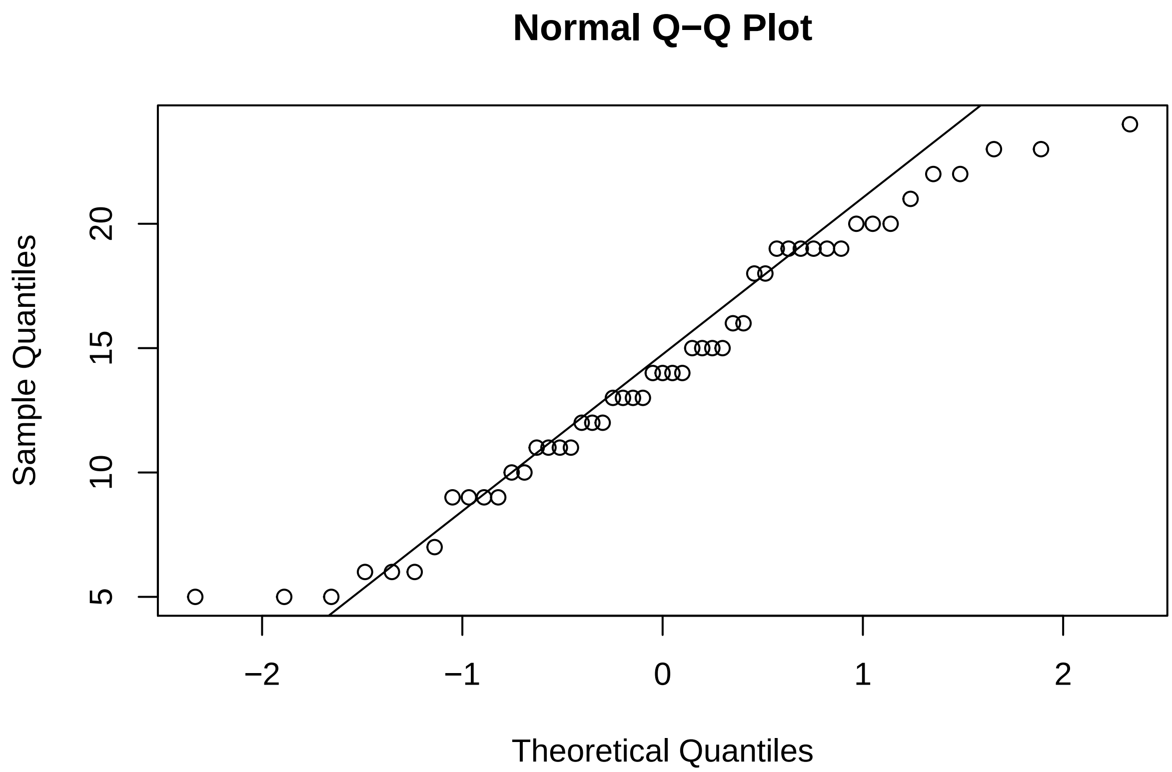
Question 1
Null Hypothesis is that there is no change in productivity between exercisers and non-exercisers
Alternate Hypothesis is that exercisers are more productive than non-exercisers
Let, μ_{ex} be mean productivity for exercisers and μ_{noex} for non-exercisers

- $H_0 : \mu_{ex} - \mu_{noex} = 0$
- $H_a : \mu_{ex} - \mu_{noex} > 0$

```
exercise1 <- read.csv('Question 1.csv')
#Assessing normality through informal technique
ex1_1 <- na.omit(dplyr::filter(exercise1,Exerciser=='Yes'))
qqnorm(ex1_1$Rating)
qqline(ex1_1$Rating)
```




```
ex1_2 <- na.omit(dplyr::filter(exercise1,Exerciser=='No'))
qqnorm(ex1_2$Rating)
qqline(ex1_2$Rating)
```



Because the points are approximately linear, We can conclude that the sample data come from a population that has a normal distribution.

a

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
yes_mean <- exercise1%>%filter(Exerciser=='Yes')%>%
    summarise(num=n(),mean_y=mean(Rating),sd_y=sd(Rating))
no_mean <- exercise1%>%filter(Exerciser=='No')%>%
    summarise(num=n(),mean_n=mean(Rating), sd_n=sd(Rating))
#equal variance test using F test
var.test(ex1_1$Rating,ex1_2$Rating,ratio=1,alternative='two.sided')
```

```
##
##  F test to compare two variances
##
## data:  ex1_1$Rating and ex1_2$Rating
## F = 0.5979, num df = 28, denom df = 50, p-value = 0.1454
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3171869 1.2001930
## sample estimates:
## ratio of variances
##          0.5979037
```

```
#As p value >0.05=> failed to reject null hypothesis: var.equal=TRUE
t.test(ex1_1$Rating,ex1_2$Rating,mu=0, alternative='g',var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ex1_1$Rating and ex1_2$Rating
## t = 2.3867, df = 78, p-value = 0.009711
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.8243938       Inf
## sample estimates:
## mean of x mean of y
##  16.86207  14.13725
```

As p-value<0.05 => Reject null hypothesis,H_0 At 5% significance level, there is sufficient evidence to support the claim that exercisers outperform non-exercisers on an average i.e. average rating of exercisers are higher than the non-exercisers'.

b

No, we can't infer that any difference between the two groups is due to exercise. We could have had it been a paired test sample i.e. employees' average rating pre-exercising and post exercising. In the given case, the difference can be attributed to other factors since we are looking at a different groups.

Question 2

Null Hypothesis is that there is no difference between appraisal value and selling prices
Alternate Hypothesis is that there is a difference
Let, μ_{ap} be mean appraisal value and μ_{sp} be mean selling price

- $H_0 : \mu_{ap} - \mu_{sp} = 0$
- $H_a : \mu_{ap} - \mu_{sp} \neq 0$

```
exercise2 <- read.csv('Question 2.csv')
head(exercise2)
```

```
##   House  Value  Price
## 1     1 119.37 121.87
## 2     2 148.93 150.25
## 3     3 130.39 122.78
## 4     4 135.70 144.35
## 5     5 126.30 116.20
## 6     6 137.08 139.49
```

```r
#F test to compare variances
var.test(exercise2$Value, exercise2$Price, ratio = 1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  exercise2$Value and exercise2$Price
## F = 0.62518, num df = 74, denom df = 74, p-value = 0.04503
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3949780 0.9895526
## sample estimates:
## ratio of variances
##          0.6251812
```

Since, p_value<0.05
At 5% significance level,
We have sufficient evidence to support the claim that the two variances are not equal:var.equal=FALSE

```r
#As we are considering same 75 cars for both, it's a paired sample
t.test(exercise2$Value, exercise2$Price, alternative='two.sided',
       mu=0,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  exercise2$Value and exercise2$Price
## t = -0.35493, df = 74, p-value = 0.7236
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.489448  1.736648
## sample estimates:
## mean of the differences
##                 -0.3764
```

Our p-value for the given test is 0.7236 >> 0.05 At 5% significance level, we don't have sufficient evidence to support our claim that difference exists between appraisal value and selling price
Due to the p-value being very high, For all the below significant values,
$\alpha = 0.01 << 0.7236 =>$ failed to reject H_0
$\alpha = 0.05 << 0.7236 =>$ failed to reject H_0
$\alpha = 0.10 << 0.7236 =>$ failed to reject H_0
It is appropriate to conclude that no difference exists between appraisal values and selling prices

Question 3 Null Hypothesis is that there is no difference between variances in service times of teller 1 and 2
Alternate Hypothesis is that there is a difference
Let, σ_1 be variance for teller 1 and σ_2 be variance for teller 2

- $H_0 : \sigma_1/\sigma_2 = 1$
- $H_a : \sigma_1/\sigma_2 \neq 1$

```r
exercise3 <- read.csv('Question 3.csv')
#head(exercise3)
var.test(exercise3$Teller1,exercise3$Teller2,ratio=1,
         alternative='two.sided', conf.level=0.90)
```

```
##
##  F test to compare two variances
##
## data:  exercise3$Teller1 and exercise3$Teller2
## F = 0.30561, num df = 99, denom df = 99, p-value = 1.045e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 90 percent confidence interval:
##  0.2192197 0.4260330
## sample estimates:
## ratio of variances
##          0.3056056
```

As, p-value<<0.10 => Reject the null hypothesis
At 10% significance level, we have sufficient evidence to support the claim that variance in service time differ for two tellers. Yes,The data allow us to infer at the 10% significance level that the variance in service times differs between the two tellers.

Question 4
Null Hypothesis is that prop. of patients who developed serious heart problems using Vioxx and using Placebo are equal Or more in placebo group than Vioxx
Alternate Hypothesis is higher prop. of serious heart patients in Vioxx group than placebo
Let, p_v be prop. of patients who developed serious heart problems using *Vioxx* for 18 months
and p_p be prop. of patients who developed serious heart problems using *placebo* for 18 months

- $H_0 : p_v - p_p \leq 0$
- $H_a : p_v - p_p > 0$ For quality assurance and healthcare, we consider significance value of 0.01 (confidence level of 99%).

```r
prop.test(c(45, 25), c(1287, 1299), alternative = "greater", correct = FALSE,
          conf.level=0.99)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(45, 25) out of c(1287, 1299)
## X-squared = 6.0657, df = 1, p-value = 0.006891
## alternative hypothesis: greater
## 99 percent confidence interval:
##  0.0008693274 1.0000000000
## sample estimates:
##     prop 1     prop 2
## 0.03496503 0.01924557
```

p-value $= 0.006891 < 0.01$
At 1% significance level, we have sufficient evidence to support the claim that the proportion of serious heart patients is higher for those who use Vioxx than placebo.

```r
(45/1287)*100
```

```
## [1] 3.496503
```

```r
(25/1299)*100
```

```
## [1] 1.924557
```

From Vioxx user's point of view,
Merely 3% of the total Vioxx users and ~2% of the total placebo users suffered from serious heart problems
Due to very minimal difference between both groups, I as a patient would have attributed this more to chance than something which is caused by Vioxx consumption

```r
print('Rough/manual Extrapolation for 2M americans')
```

```
## [1] "Rough/manual Extrapolation for 2M americans"
```

```r
#serious heart patients for Vioxx
(45/1287) * 2000000
```

```
## [1] 69930.07
```

```r
#serious heart patients for non-Vioxx(placebo)
(25/1299) * 2000000
```

```
## [1] 38491.15
```

```r
print('Using confidence intervals')
```

```
## [1] "Using confidence intervals"
```

```r
### For 99% confidence level, find the chi-sq (lower) and chi-sq (upper)
#z score for 95%
z99 <- qnorm(0.995,0,1,TRUE)
#proportion of heart patients for Vioxx
p_v <- 45/1287
#upper confidence limit
ucl_v <- p_v+ z99*sqrt((p_v*(1-p_v))/1287)
#lower confidence limit
lcl_v <- p_v- z99*sqrt((p_v*(1-p_v))/1287)
cat("99 percent Confidence Interval for Vioxx Heart patients proportion in true population:\n", round(l
```

```
## 99 percent Confidence Interval for Vioxx Heart patients proportion in true population:
##  0.022 0.048
```

```r
round(lcl_v,3)*2000000
```

```
## [1] 44000
```

```r
round(ucl_v,3)*2000000
```

```
## [1] 96000
```

```r
#proportion of heart patients for Placebo
p_p <- 25/1299
#upper confidence limit
ucl_p <- p_p+ z99*sqrt((p_p*(1-p_p))/1299)
#lower confidence limit
lcl_p <- p_p- z99*sqrt((p_p*(1-p_p))/1299)
cat("99 percent Confidence Interval for placebo Heart patients proportion in true population:\n", round
```

```
## 99 percent Confidence Interval for placebo Heart patients proportion in true population:
##  0.009 0.029
```

```r
round(lcl_p,3)*2000000
```

```
## [1] 18000
```

```r
round(ucl_p,3)*2000000
```

```
## [1] 58000
```

Analyzing simple manual extrapolation: 69K Vioxx users with serious health problems compared to 38K non vioxx users is a result which isn't something Merck would like to pursue and moreover they're inclined to avoid it. Analyzing extrapolation using CI:
Roughly between 44K and 96K people can contain the number of true population heart patients.it still is staggeringly high compared to the estimates for Placebo. That's a risk any pharma company would like to avoid
The results are both statistically significant(test result) and practically significant for Merck as a company to close its production and pull it out from the market
Merck can experience following losses:
Hamper their reputation which have both long term and short term repurcussions
Can instigate legal battle due to striking difference in numbers -expensive
Brach of trust and hampering of human lives

Question 5

a.
125 employees from West Coast - increased health benefits
240 employees from East Coast - increased vacation days
This is a case of independant samples. There can be many confounding factors
As per my understanding, confounding effect is something which hasn't been accounted for and can certainly hamper the final result of an experiment. In our experiment, we are considering samples from two different places. There can be confounding due to:
– Difference in location (east/west coast) – Distribution of employess as per Gender – Age Group - Irregular distribution of age groups in these samples can give inconsistent results – Departments under the company considered for each sample in this analysis- nature of retention might be different across different departments – Difference in sample size

b
Null Hypothesis is that difference in prop. of employees retained between groups with increased health

benefits and increased vacation days is only 0.05 or less
Alternate Hypothesis is higher prop. of employees retained from health benefits group by 0.05 or more
Let, p_h be prop. of employees retained from the group with increased health benefits
and p_v be prop. of employees reatined from the group with increased vacation days

- $H_0 : p_h - p_v \geq 0.05$
- $H_a : p_h - p_v < 0.05$

```
exercise5 <- read.csv('Question 5.csv')
#head(exercise5)
#nrow(exercise5)
#prop. of employees retained from the group with increased health benefits
ret_h <- length(which(exercise5$Benefit == 'Health' & exercise5$Retention == 1))
n_h <- 125
prop_h <- ret_h/n_h
#prop. of employees retained from the group with increased vacation days
ret_v <- length(which(exercise5$Benefit == 'Vacation' & exercise5$Retention == 1))
n_v <- 140
prop_v <- ret_v/n_v
#appropriate test statistic
z_score <- ((prop_h-prop_v) - 0.05)/sqrt((prop_h*(1-prop_h)/n_h)+(prop_v*(1-prop_v)/n_v))

p_value <- pnorm(z_score)
p_value
```

```
## [1] 0.7198728
```

Since, p_Value $> \alpha = 0.01$ => Failed to reject H_0
p_Value $> \alpha = 0.05$ => Failed to reject H_0
p_Value $> \alpha = 0.10$ => Failed to reject H_0
At all the significance levels mentioned above, there is not sufficient evidence to warrant rejection of the the claim that the difference in retention proportion will be higher than or equal to 0.05.
Yes,we have statistically significant evidence to support the fact we have higher retention to compensate for switching to health benefits.

c
Null Hypothesis is that there is no difference in prop. of employees retained between groups with increased health benefits and increased vacation days
Alternate Hypothesis is there is difference between the retention proportion of the two groups
Let, p_h be prop. of employees retained from the group with increased health benefits
and p_v be prop. of employees reatined from the group with increased vacation days

- $H_0 : p_h - p_v = 0$
- $H_a : p_h - p_v \neq 0$

```
prop.test(c(ret_h, ret_v),c(n_h,n_v), alternative = "two.sided",
          correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
```

```
## data:  c(ret_h, ret_v) out of c(n_h, n_v)
## X-squared = 2.6269, df = 1, p-value = 0.1051
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01486734  0.16972448
## sample estimates:
##    prop 1    prop 2
## 0.8560000 0.7785714
```

p_value=0.1051 > 0.05 => Failed to reject H_0
p_value=0.1051 > 0.10 => Failed to reject H_0
At 5% significance level, there is not sufficient evidence to support the claim that there is a difference in retention proportions for the two groups.
Same is the case for 10% significance level
No, there is no statistically significant difference in retention rates between the benefit plans based on the evidence provided to us.

Question 6

```r
#Reported incomes for 2000,2008,2014
exercise6 <- read.csv('Question 6.csv')
#CPI inndices for each year
annual_data <- readxl::read_excel('U.S. CPI Annual.xlsx')

head(exercise6)
```

```
##   RINCOME_2000 RINCOME_2008 RINCOME_2014
## 1         9000           NA        82500
## 2         9000        82500        82500
## 3           NA        45000           NA
## 4           NA        16250        13750
## 5           NA        32500           NA
## 6        21250           NA       175000
```

```r
nrow(exercise6)
```

```
## [1] 2817
```

```r
head(annual_data)
```

```
## # A tibble: 6 x 2
##    Year `All Urban Consumers - (CPI-U): U.S. city average: All items: 1982~
##   <dbl>                                                               <dbl>
## 1  1913                                                                9.88
## 2  1914                                                               10.0
## 3  1915                                                               10.1
## 4  1916                                                               10.9
## 5  1917                                                               12.8
## 6  1918                                                               15.0
```

```r
colnames(annual_data) <- c('Year','CPI')
```

The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. Indexes are available for the U.S. and various geographic areas. Average price data for select utility, automotive fuel, and food items are also available. (Source-https://www.bls.gov/cpi/)

a

Let, μ_{2008} be mean income in 2008
and μ_{2000} be mean income in 2000

- $H_0 : \mu_{2008} - \mu_{2000} \leq 0$
- $H_a : \mu_{2008} - \mu_{2000} > 0$ i.e.$\mu_{2008} > \mu_{2004}$

Please note we are not adjusting for inflation here

```r
income_2000 <- na.omit(dplyr::select(exercise6,RINCOME_2000))
income_2008 <- na.omit(dplyr::select(exercise6,RINCOME_2008))
#Conducting a test to check if population variances are equal using F test
var.test(income_2008$RINCOME_2008,income_2000$RINCOME_2000,ratio=1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  income_2008$RINCOME_2008 and income_2000$RINCOME_2000
## F = 2.0925, num df = 1188, denom df = 1817, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.88793 2.32194
## sample estimates:
## ratio of variances
##           2.092531
```

```r
#As p value <0.05, we reject the null hypothesis i.e. var.equal=FALSE

#As these are not same set of people, these are independant samples(paired=FALSE)
t.test(income_2008,income_2000,alternative='greater',mu=0,
       paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  income_2008 and income_2000
## t = 8.1934, df = 1923.8, p-value = 2.29e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  7880.687      Inf
## sample estimates:
## mean of x mean of y
##  41092.09  31230.75
```

As, p-value << 0.05 => Reject the null hypothesis
At 5% significance level,We have sufficient evidence to support the claim that income rose between 2000 and 2008

b
Let, μ_{2014} be mean income in 2014 and μ_{2008} be mean income in 2008

- $H_0 : \mu_{2014} - \mu_{2008} \leq 0$
- $H_a : \mu_{2014} - \mu_{2008} > 0$

Please note we are not adjusting for inflation here

```
income_2008 <- na.omit(dplyr::select(exercise6,RINCOME_2008))
income_2014 <- na.omit(dplyr::select(exercise6,RINCOME_2014))
#Conducting a test to check if population variances are equal using F test
var.test(income_2014$RINCOME_2014,income_2008$RINCOME_2008,ratio=1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  income_2014$RINCOME_2014 and income_2008$RINCOME_2008
## F = 1.2123, num df = 1522, denom df = 1188, p-value = 0.0004715
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.088474 1.349165
## sample estimates:
## ratio of variances
##           1.212266
```

```
#As p value <0.05, we reject the null hypothesis i.e. var.equal=FALSE

#As these are not same set of people, these are independant samples(paired=FALSE)
t.test(income_2014,income_2008,alternative='greater',mu=0,
       paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  income_2014 and income_2008
## t = 2.8351, df = 2648.9, p-value = 0.002308
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1743.641      Inf
## sample estimates:
## mean of x mean of y
##  45247.37  41092.09
```

As, p-value < 0.05 => Reject the null hypothesis
At 5% significance level,We have sufficient evidence to support the claim that income rose between 2008 and 2014

c
Let, μ_{adj} be inflation adjusted mean income in 2008 (w.r.t 2000) and μ_{2000} be mean reported income in 2000

- $H_0 : \mu_{adj} - \mu_{2000} \leq 0$
- $H_a : \mu_{adj} - \mu_{2000} > 0$

```
library(dplyr)
CPI_08 <- annual_data%>% select(CPI)%>% filter(annual_data$Year==2008)
CPI_00 <- annual_data%>% select(CPI)%>% filter(annual_data$Year==2000)
inf_rate <- (CPI_08-CPI_00)/CPI_00
#inflation rate in 2008 considering 2000 as the base year
inf_rate
```

```
##         CPI
## 1 0.2500895
```

```
class(income_2008)
```

```
## [1] "data.frame"
```

```
adj_factor <- rep((1-inf_rate),length(income_2008))
income_2008 <- cbind(income_2008,adj_factor)
#head(income_2008)
income_2008$adjusted_income <- income_2008$RINCOME_2008*income_2008$CPI
#Conducting a test to check if population variances are equal using F test
var.test(income_2008$adjusted_income,income_2000$RINCOME_2000,ratio=1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  income_2008$adjusted_income and income_2000$RINCOME_2000
## F = 1.1768, num df = 1188, denom df = 1817, p-value = 0.00191
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.061707 1.305780
## sample estimates:
## ratio of variances
##           1.176768
```

```
#Reject null hypothesis i.e. var.equal=FALSE

t.test(income_2008$adjusted_income,income_2000,alternative='greater',mu=0,
       paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  income_2008$adjusted_income and income_2000
## t = -0.42269, df = 2392.3, p-value = 0.6637
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2032.296       Inf
## sample estimates:
## mean of x mean of y
##  30815.39  31230.75
```

As p-value>> 0.05 => Failed to reject the null hypothesis,
At 5% significance level, We don't have sufficient evidence to support the claim that income rose between 2000 to 2008 i.e. average income in 2008 was higher than the average income in 2000

d

Let, μ_{adj} be inflation adjusted mean income in 2014 (w.r.t 2008) and μ_{2008} be mean reported income in 2008

- $H_0 : \mu_{adj} - \mu_{2008} \leq 0$
- $H_a : \mu_{adj} - \mu_{2008} > 0$

```
library(dplyr)
#obtaining CPI index for 2008 and 2014
CPI_08 <- annual_data%>% select(CPI)%>% filter(annual_data$Year==2008)
CPI_14 <- annual_data%>% select(CPI)%>% filter(annual_data$Year==2014)
inf_rate <- (CPI_14-CPI_08)/CPI_08
#inflation rate in 2014 considering 2008 as the base year
inf_rate
```

```
##          CPI
## 1 0.09969571
```

```
#Real income in 2014 (compared to 2008)=(Reported Income in 2014)*(1- Inflation rate)
adj_factor <- rep((1-inf_rate),length(income_2014))
income_2014 <- cbind(income_2014,adj_factor)
head(income_2014)
```

```
##    RINCOME_2014       CPI
## 1         82500 0.9003043
## 2         82500 0.9003043
## 4         13750 0.9003043
## 6        175000 0.9003043
## 8         32500 0.9003043
## 10        45000 0.9003043
```

```
income_2014$adjusted_income <- income_2014$RINCOME_2014*income_2014$CPI

#Conducting a test to check if population variances are equal using F test
var.test(income_2014$adjusted_income,income_2008$RINCOME_2008,ratio=1,
         alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  income_2014$adjusted_income and income_2008$RINCOME_2008
## F = 0.9826, num df = 1522, denom df = 1188, p-value = 0.7468
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8822604 1.0935625
## sample estimates:
## ratio of variances
##          0.9825994
```

```r
#failed to reject the null hypothesis, var.equal=TRUE
t.test(income_2014$adjusted_income,income_2008$RINCOME_2008,alternative='greater',mu=0,
       paired=FALSE,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  income_2014$adjusted_income and income_2008$RINCOME_2008
## t = -0.25497, df = 2710, p-value = 0.6006
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2651.066       Inf
## sample estimates:
## mean of x mean of y
##  40736.40  41092.09
```

As p-value>> 0.05 => Failed to reject the null hypothesis,
At 5% significance level, We don't have sufficient evidence to support the claim that income rose between 2008 to 2014 i.e. average income in 2014 was higher than the average income in 2000

e

I have learned that just comparing raw incomes of two different years is an incorrect way. We have to treat for inflation to have a fair omparison between the years. Otherwise there is high possibility of us making biased decisions which in turn will affect the business decisions we make. we alsways have to make sure when we are comapring, the base should be same