Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

Question 1

$$\mu_1 = Average\ rating\ of\ exercisers$$

$$\mu_2 = Average\ rating\ of\ non-exercisers$$

$$H_0: \mu_1 - \mu_2 \leq D_0$$

$$H_1: \mu_1 - \mu_2 > D_0$$

Part a.

```
question_1 <- read.csv('Question 1.csv')
exercisers <- question_1[which(question_1$Exerciser == 'Yes'), ]
non_exercisers <- question_1[which(question_1$Exerciser == 'No'), ]

var.test(exercisers$Rating, non_exercisers$Rating, ratio = 1, alternative = 'two.sided')

##
##  F test to compare two variances
##
## data:  exercisers$Rating and non_exercisers$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
```

Variances are equal, so run Equal Variance t-test

```
t.test(exercisers$Rating, non_exercisers$Rating, alternative = "greater", mu = 0, paired
= FALSE, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  exercisers$Rating and non_exercisers$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
```

$p-value(0.009711)\ is\ less\ than\ \alpha(0.05),\ therefore\ reject\ null\ hypothesis.$

The data given does company's hypothesis that exercisers outperform non-exercisers.

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

Part b.

Company cannot say that exercisers outperform non-exercisers because the samples taken are independent of each other.

Question 2

$$\mu_1 = Average\ appraised\ value$$

$$\mu_2 = Average\ selling\ price$$

$$H_0: \mu_1 - \mu_2 = D_0$$

$$H_1: \mu_1 - \mu_2 \neq D_0$$

$$D_0 = 0$$

```
question_2 <- read.csv('Question 2.csv')
```

$$\alpha = 0.05$$

```
var.test(question_2$Value, question_2$Price, ratio = 1, alternative = 'two.sided')

##
##  F test to compare two variances
##
## data:  question_2$Value and question_2$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
```

Variances are not equal, so run Unequal Variance t-test

```
t.test(question_2$Value, question_2$Price, alternative = 'two.sided', mu = 0, paired = TR
UE, var.equal = FALSE)

##
##  Paired t-test
##
## data:  question_2$Value and question_2$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
```

$p - value(0.7236)\ is\ greater\ than\ \alpha(0.05),\ therefore\ do\ not\ reject\ null\ hypothesis$

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

$$\alpha = 0.01$$

```r
var.test(question_2$Value, question_2$Price, ratio = 1, alternative = 'two.sided', conf.l
evel = 0.99)
```

```
##
##  F test to compare two variances
##
## data:  question_2$Value and question_2$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
## 99 percent confidence interval:
##  0.3412558 1.1453329
## sample estimates:
## ratio of variances
##          0.6251812
```

Variances are not equal, so run Unequal Variance t-test

```r
t.test(question_2$Value, question_2$Price, alternative = 'two.sided', mu = 0, paired = TR
UE, var.equal = FALSE, conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  question_2$Value and question_2$Price
## t = -0.35493, df = 74, p-value = 0.7236
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -3.18021  2.42741
## sample estimates:
## mean of the differences
##                 -0.3764
```

$p - value(0.7236)\ is\ greater\ than\ \alpha(0.01),\ therefore\ do\ not\ reject\ null\ hypothesis$

$$\alpha = 0.1$$

```r
var.test(question_2$Value, question_2$Price, ratio = 1, alternative = 'two.sided', conf.l
evel = 0.9)
```

```
##
##  F test to compare two variances
##
## data:  question_2$Value and question_2$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
## 90 percent confidence interval:
##  0.4254522 0.9186731
## sample estimates:
```

```
## ratio of variances
##          0.6251812
```

Variances are equal, so run equal Variance t-test

```
t.test(question_2$Value, question_2$Price, alternative = 'two.sided', mu = 0, paired = TR
UE, var.equal = TRUE, conf.level = 0.9)

##
##  Paired t-test
##
## data:  question_2$Value and question_2$Price
## t = -0.35493, df = 74, p-value = 0.7236
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -2.142845  1.390045
## sample estimates:
## mean of the differences
##                 -0.3764
```

$p - value(0.7236)\ is\ greater\ than\ \alpha(0.10),\ therefore\ do\ not\ reject\ null\ hypothesis$

Using these sample data, we can say that there is no a statistically significant mean difference between the appraised values and selling prices of the houses sold in this suburban community, for the levels of significance of 0.1, 0.05 and 0.01 with a p-value of 0.7236 for all the 3 levels of significance.

Question 3

$$\sigma_1 = Variance\ in\ service\ time\ for\ Teller\ 1$$

$$\sigma_2 = Variance\ in\ service\ time\ for\ Teller\ 2$$

$$H0: \sigma_1/\sigma_2 = 1$$

$$H1: \sigma_1/\sigma_2 \neq 1$$

```
question_3 <- read.csv('Question 3.csv')

var.test(question_3$Teller1, question_3$Teller2, ratio = 1, alternative = 'two.sided', co
nf.level = 0.9)

##
##  F test to compare two variances
##
## data:  question_3$Teller1 and question_3$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
```

$p - value\ is\ 1.045e^{-08},\ which\ means\ it\ is\ statistically\ significant\ so\ reject\ H_0$
$Therefore\ the\ variance\ in\ service\ times\ differs\ between\ the\ 2\ tellers$

The data allows us to infer at the 10% significance level that the variance in service times differs between the two tellers.

Question 4

$$p_1 = Proportion\ of\ people\ who\ took\ Vioxx\ and\ developed\ heart\ problems$$

$$p_2 = Proportion\ of\ people\ who\ took\ placebos\ and\ developed\ heart\ problems$$

$$H_0: p_1 - p_2 \leq 0$$

$$H_1: p_1 - p_2 > 0$$

```
prop.test(c(45, 25), c(1287, 1299), alternative = 'greater', correct = FALSE)

##
##  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
## 95 percent confidence interval:
##  0.005219614 1.000000000
## sample estimates:
##     prop 1     prop 2
## 0.03496503 0.01924557
```

p-value is 0.006891, which is lesser than alpha of 0.05, hence we reject null hypothesis.

Therefore, we can conclude that Vioxx caused a *statistically significant* increase in the risk of developing serious heart problems.

From the point of view of patients, as a Vioxx user, these results would not cause me significant worry because

1. 25 subjects who took placebos also developed heart problems.
2. It does alleviate my pain.

```
(pop_vioxx = ceiling(2000000*(0.03496503)))

## [1] 69931
```

69,931 people from a population of 2 million would develop heart problems if all 2 million took Vioxx.

```
(pop_not_vioxx = ceiling(2000000*(0.01924557)))
```

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

```
## [1] 38492
```

38,492 people from a population of 2 million would develop heart problems if all 2 million did not take Vioxx.

Based on this, the results are practically significant to the company.

The company might get sued for millions of dollars, and will also lose reputation resulting in a lesser number of people buying the drugs made by them, which would cause them severe losses, maybe even resulting in bankruptcy.

Question 5

```r
question_5 <- read.csv('Question 5.csv')
table(question_5)
```

```
##           Retention
## Benefit      0   1
##   Health    18 107
##   Vacation  31 109
```

Part a.

The confounding effects in this comparison are as follows:

1. The type of people taken into consideration might be different.
2. The cost of healthcare is different in different states and so people in different states have different criteria for the states.
3. There might be internal transfers and certain people might have received benefits on both the coasts.

Part b.

$$p_1 = Proportion\ of\ retention\ after\ getting\ health\ benefits$$

$$p_2 = Proportion\ of\ retention\ after\ getting\ vacation\ benefits$$

$$H_0: p_1 - p_2 \geq 0.05$$

$$H_1: p_1 - p_2 < 0.05$$

```r
p_health_hat = 107/125
p_vacation_hat = 109/140
denom = sqrt(((p_health_hat)*(1 - p_health_hat)/125) + ((p_vacation_hat)*(1 - p_vacation_
hat)/140))

z_5_b = ((p_health_hat - p_vacation_hat) - 0.05)/ denom
pnorm(z_5_b, lower.tail = FALSE)
```

```
## [1] 0.2801272
```

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

p-value is 0.28 which is greater than alpha(0.05), therefore statistically not significant

Hence failure to reject the null hypothesis and accepting that giving health benefits increases retention rate while being effective.

Part c.

$$p_1 = Proportion\ of\ retention\ after\ getting\ health\ benefits$$

$$p_2 = Proportion\ of\ retention\ after\ getting\ vacation\ benefits$$

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0.05$$

```
prop.test(c(107, 109), c(125, 140), alternative = 'two.sided', correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(107, 109) out of c(125, 140)
## 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 is 0.1051 which is greater than alpha(0.05), therefore statistically not significant

Therefore we fail to reject the null hypothesis, and accept that statistically there is no significant difference in retention rates between the benefit plans.

Question 6

```
question_6 <- read.csv('Question 6.csv')
```

Part a.

$$\mu_1 = Average\ RINCOME\ in\ year\ 2000$$

$$\mu_2 = Average\ RINCOME\ in\ year\ 2008$$

$$H_0: \mu_1 - \mu_2 \geq 0\ \ i.e.\ income\ does\ not\ increase$$

$$H_1: \mu 1 - \mu 2 < 0\ \ i.e.\ income\ does\ increase$$

```
var.test(question_6$RINCOME_2000, question_6$RINCOME_2008, ratio = 1, alternative = 'two.
sided')
```

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

```
##
##  F test to compare two variances
##
## data:  question_6$RINCOME_2000 and question_6$RINCOME_2008
## F = 0.47789, num df = 1817, denom df = 1188, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4306743 0.5296807
## sample estimates:
## ratio of variances
##          0.4778902
```

Variances are not equal, so run unequal Variance t-test

```
t.test(question_6$RINCOME_2000, question_6$RINCOME_2008, alternative = 'less', var.equal
= FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  question_6$RINCOME_2000 and question_6$RINCOME_2008
## t = -8.1934, df = 1923.8, p-value = 2.29e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -7880.687
## sample estimates:
## mean of x mean of y
##  31230.75  41092.09
```

p-value ($2.29e^{\wedge -16}$) is statistically significant, and we reject the null hypothesis

Income has increased between 2000 and 2008.

Part b.

$$\mu_1 = Average\ RINCOME\ in\ year\ 2008$$

$$\mu_2 = Average\ RINCOME\ in\ year\ 2014$$

$$H_0: \mu_1 - \mu_2 \geq 0 \ \ i.e.\ income\ does\ not\ increase$$

$$H_1: \mu 1 - \mu 2 < 0 \ \ i.e.\ income\ does\ increase$$

```
var.test(question_6$RINCOME_2008, question_6$RINCOME_2014, ratio = 1, alternative = 'two.
sided')
```

```
##
##  F test to compare two variances
##
## data:  question_6$RINCOME_2008 and question_6$RINCOME_2014
## F = 0.8249, num df = 1188, denom df = 1522, p-value = 0.0004715
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7411993 0.9187172
```

```
## sample estimates:
## ratio of variances
##          0.8249016
```

Variances are not equal, so run unequal Variance t-test

```
t.test(question_6$RINCOME_2008, question_6$RINCOME_2014, alternative = 'less', var.equal
= FALSE, paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  question_6$RINCOME_2008 and question_6$RINCOME_2014
## t = -2.8351, df = 2648.9, p-value = 0.002308
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1743.641
## sample estimates:
## mean of x mean of y
##  41092.09  45247.37
```

p-value(0.002308) is statistically significant, so we reject the null hypothesis

Income has increased between 2008 and 2014.

Part c.

```
cpi_data = readxl::read_xlsx('U.S. CPI Annual.xlsx')

cpi_2000 = cpi_data[which(cpi_data$Year == '2000'), 2 ]
cpi_2008 = cpi_data[which(cpi_data$Year == '2008'), 2 ]
cpi_2014 = cpi_data[which(cpi_data$Year == '2014'), 2 ]
```

$$\mu_1 = Average\ adjusted\ RINCOME\ in\ 2008$$

$$\mu_2 = Average\ RINCOME\ in\ 2008$$

$$H_0: \mu_1 - \mu_2 \geq 0 \ \ i.e.\ income\ did\ not\ increase$$

$$H_1: \mu_1 - \mu_2 < 0 \ \ i.e.\ income\ increased$$

```
adj_rincome_2008_2000 <- question_6$RINCOME_2000*(215.25500/172.1917)

var.test(adj_rincome_2008_2000, question_6$RINCOME_2008, ratio = 1, alternative = 'two.si
ded')

##
##  F test to compare two variances
##
## data:  adj_rincome_2008_2000 and question_6$RINCOME_2008
## F = 0.74681, num df = 1817, denom df = 1188, p-value = 2.367e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6730247 0.8277444
## sample estimates:
```

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

```
## ratio of variances
##          0.7468101
```

Variances are not equal, so run unequal Variance t-test

```
t.test(adj_rincome_2008_2000, question_6$RINCOME_2008, alternative = 'less', var.equal =
FALSE, paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  adj_rincome_2008_2000 and question_6$RINCOME_2008
## t = -1.6001, df = 2276.8, p-value = 0.05485
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 58.16234
## sample estimates:
## mean of x mean of y
##  39041.22  41092.09
```

P-VALUE(0.05485) is not statistically significant, so we cannot reject the null hypothesis

Income has not increased between 2000 and 2008 when inflation is adjusted.

Part d.

$$\mu_1 = Average\ adjusted\ RINCOME\ in\ 2014$$

$$\mu_2 = Average\ RINCOME\ in\ 2014$$

$$H_0: \mu_1 - \mu_2 \geq 0\ \ i.e.\ income\ did\ not\ increase$$

$$H_1: \mu_1 - \mu_2 < 0\ \ i.e.\ income\ increased$$

```
adj_rincome_2014_2008 <- question_6$RINCOME_2008*(236.715/215.25500)

var.test(adj_rincome_2014_2008, question_6$RINCOME_2014, ratio = 1, alternative = 'two.si
ded')

##
##  F test to compare two variances
##
## data:  adj_rincome_2014_2008 and question_6$RINCOME_2014
## F = 0.99758, num df = 1188, denom df = 1522, p-value = 0.9665
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8963551 1.1110329
## sample estimates:
## ratio of variances
##          0.9975787
```

Variances are equal, so run equal Variance t-test

Statistical Exploration and Reasoning Assignment 3

Mitesh Ranmal Jain | 917883640

```
t.test(adj_rincome_2014_2008, question_6$RINCOME_2014, alternative = 'less', var.equal =
TRUE, paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  adj_rincome_2014_2008 and question_6$RINCOME_2014
## t = -0.037969, df = 2710, p-value = 0.4849
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2479.795
## sample estimates:
## mean of x mean of y
##  45188.80  45247.37
```

P-VALUE(0.4849) is not statistically significant, so we cannot reject the null hypothesis

Income has NOT increased between 2008 and 2014 when inflation is adjusted.

Part e.

1. Increase in income does not mean increase in spending capacity.
2. When comparing values, without taking into account inflation, it will not paint the true picture.