HW 4

2022-10-09

Contents

Question 1

a)

We will begin the test by stating the null and alternative hypothesis:

Null hypothesis: $H_0 : \mu_{VitC} = \mu_{Placebo}$

Alternative hypothesis: $H_A : \mu_{VitC} < \mu_{Placebo}$

We will set the alpha value to 0.05.

```
vit_c = c(6, 7, 7, 7, 8, 7, 7, 8, 7, 8, 10, 6, 8, 5, 6)
vit_c.df = data.frame(vit_c)
placebo = c(10, 12, 8, 6, 9, 8, 11, 9, 11, 8, 12, 11, 9, 8, 10, 9)
placebo.df = data.frame(placebo)
q1_test = mean(vit_c) - mean(placebo)
```

We get a test statistic of -2.3042.

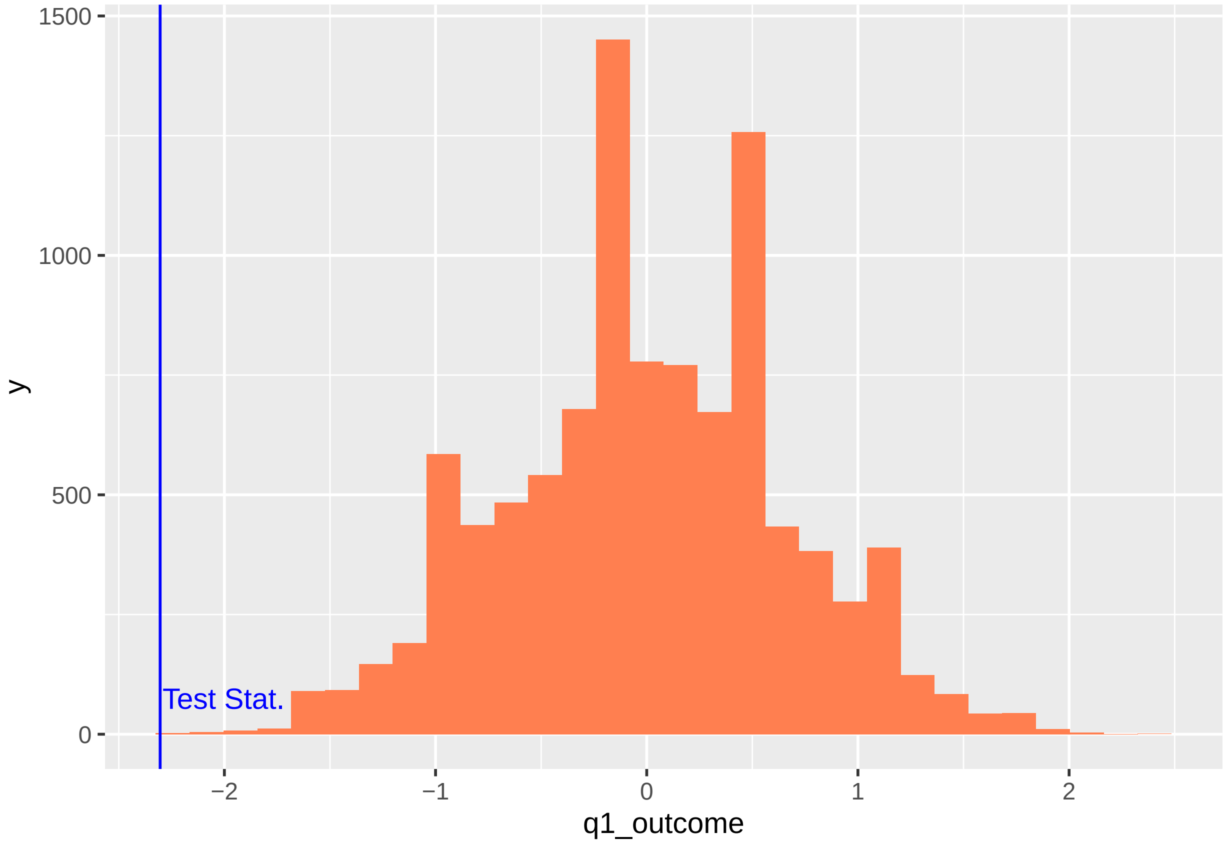
Since we are conducting a permutation test, we will pool the data together. We will then randomly sample 15 to be X_{VitC}, and the rest will be assigned as $X_{Placebo}$.

```
q1_vec = c(vit_c, placebo)
N = 10000
q1_outcome = numeric(N)
for(i in 1:N)
{ index = sample(31, 15, replace=FALSE)
  q1_outcome[i] = mean(q1_vec[index]) - mean(q1_vec[-index])
}
q1_outcome.df <- data.frame(q1_outcome)
```

Plotting our result:

```
ggplot(q1_outcome.df, aes(x = q1_outcome)) + geom_histogram(fill = "coral") + geom_vline(xintercept = q
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Calculating our p-value.

```
(sum(q1_outcome <= q1_test))/(N)
```

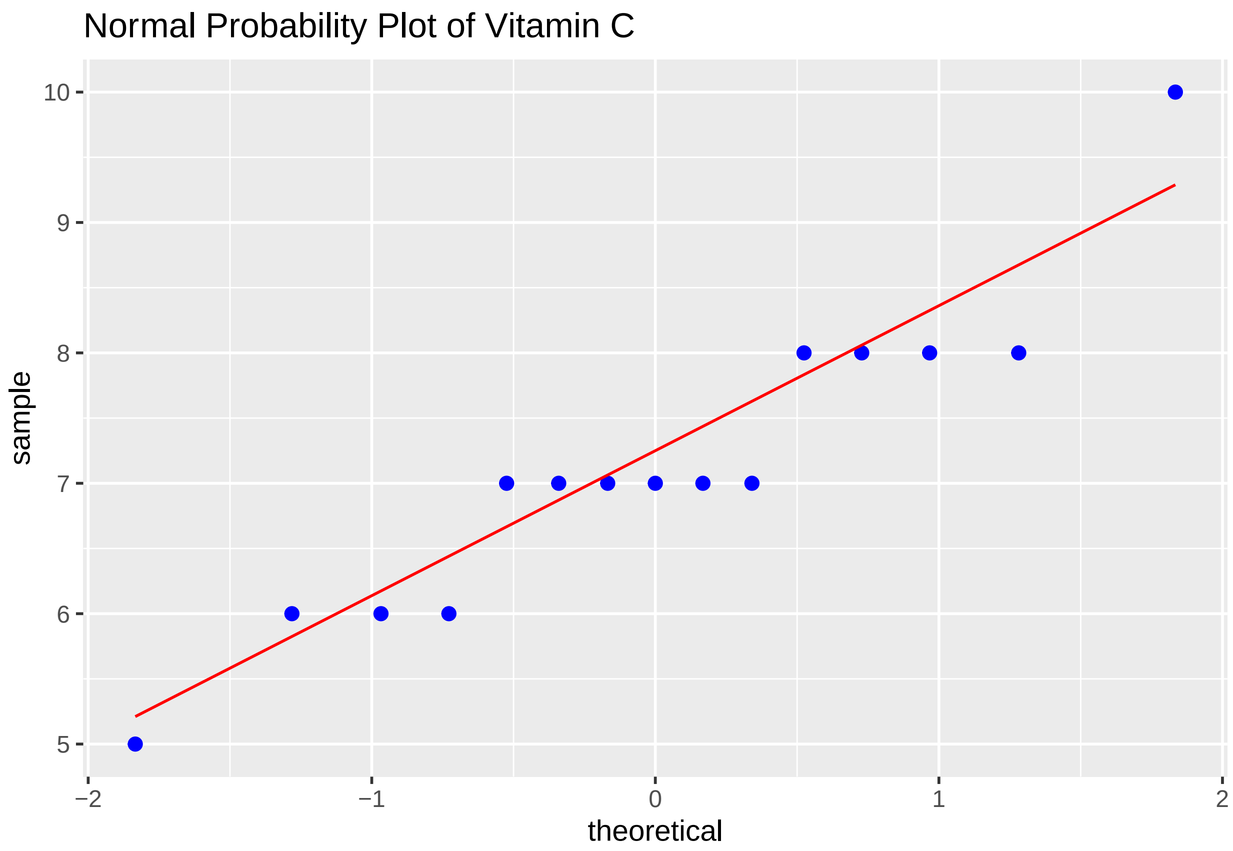
```
## [1] 0
```

Since our p-value is 0, we can reject the null hypothesis that the recovery time with Vitamin C is the same as the recovery time without Vitamin C. Therefore our conclusion is that the recovery time with Vitamin C is quicker than without it.

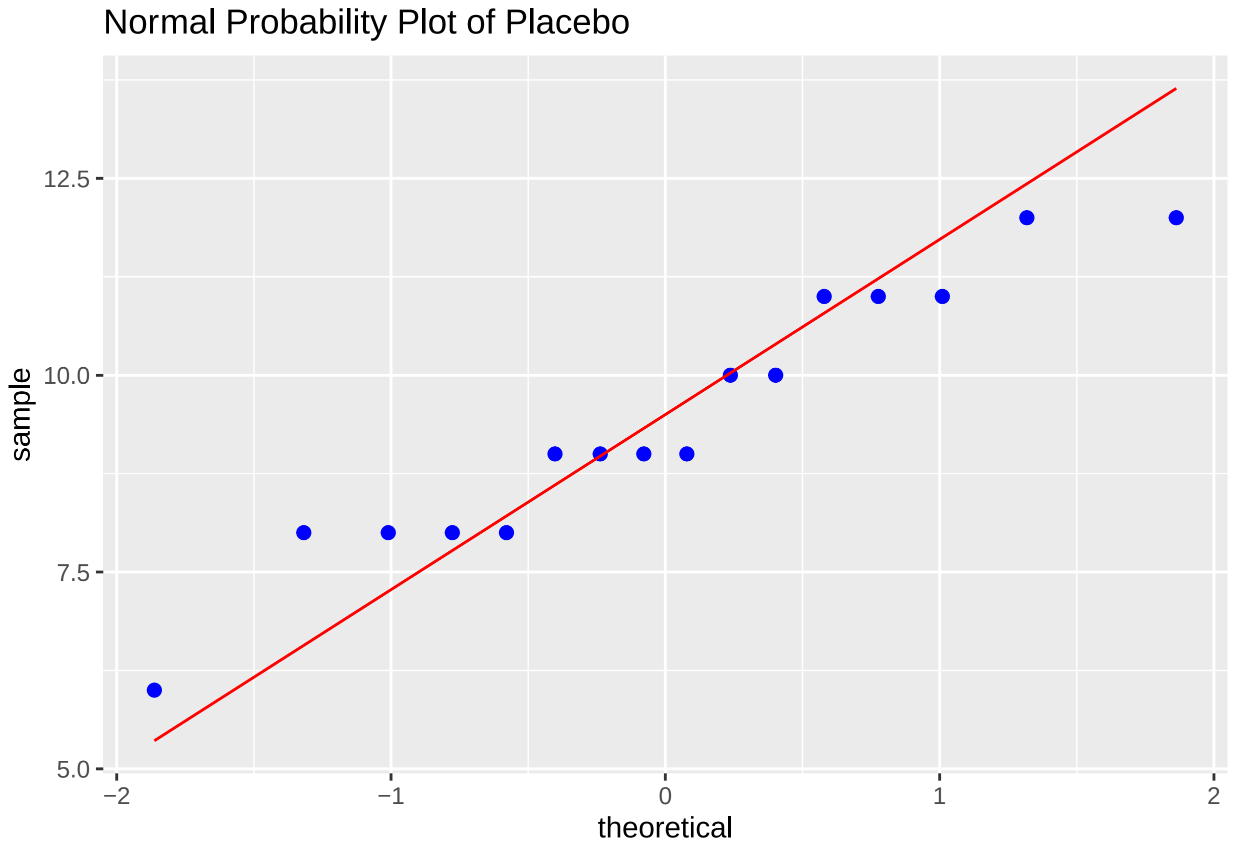
b)

To conduct a t-test, we must first check for normality of our data since both samples have less than 25 observations. We will do this by using stat_qq plot.

```
ggplot(data= vit_c.df, aes(sample = vit_c)) + stat_qq(size=2, col='blue') + stat_qqline(col='red') + gg
```



```
ggplot(data=placebo.df, aes(sample = placebo)) + stat_qq(size=2, col='blue') + stat_qqline(col='red')+ g
```



Looking at the normal probability plot for both samples, it looks like they both follow an approx. normal distribution which mean we can use t.test to test out hypothesis.

```
# Prep data in a dataframe
treatment = c(rep("vit_c", length(vit_c)), rep("placebo", length(placebo)))
recovery = c( vit_c, placebo)
q1b.df = data.frame(treatment, recovery)
# Conduct t.test
t.test(recovery ~ treatment, alternative="greater")
```

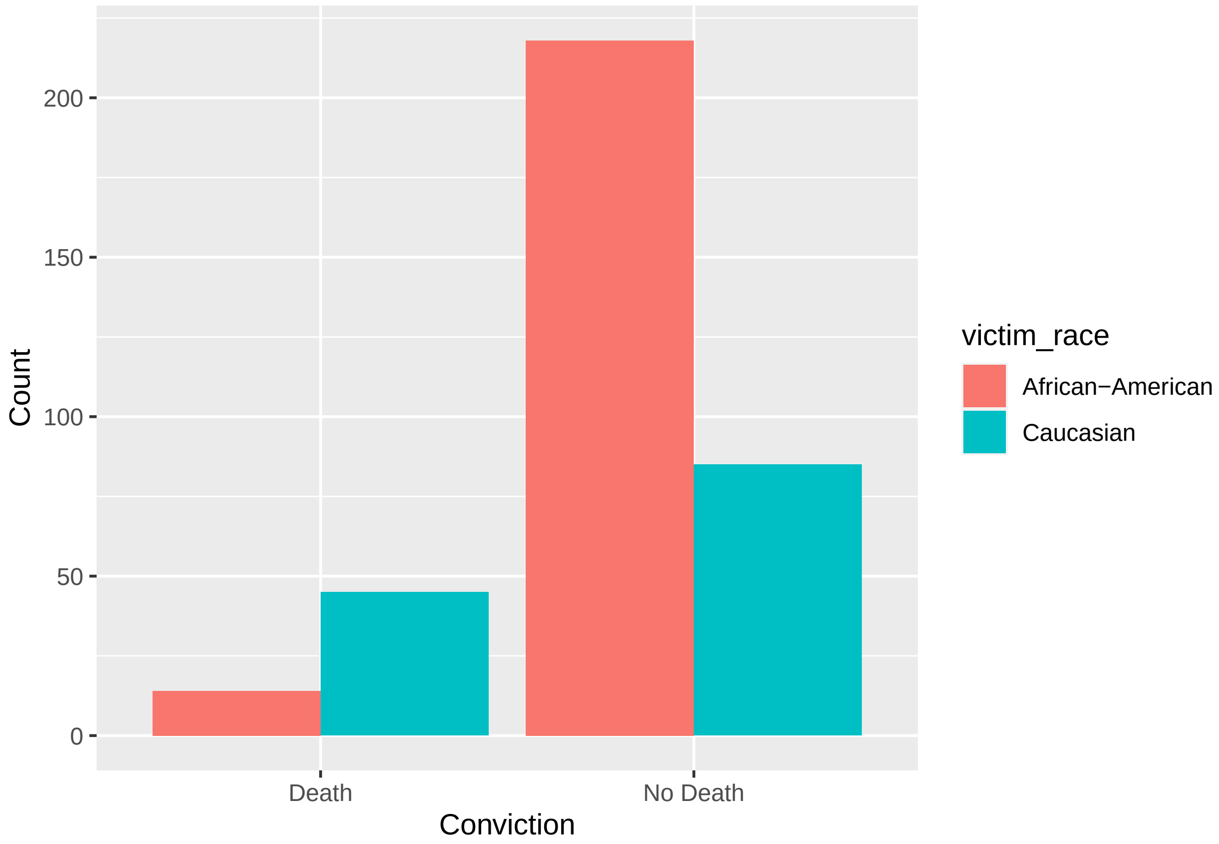
```
##
##  Welch Two Sample t-test
##
## data:  recovery by treatment
## t = 4.445, df = 27.08, p-value = 6.723e-05
## alternative hypothesis: true difference in means between group placebo and group vit_c is greater tha
## 95 percent confidence interval:
##  1.421325      Inf
## sample estimates:
## mean in group placebo   mean in group vit_c
##              9.437500              7.133333
```

From our t.test we get a p-value of 6.723e-05 which is smaller than our set alpha value of 0.05, so we can reject our null hypothesis that the mean recovery time between placebo and Vitamin C groups are equal. We can conclude that the mean recovery time of the Vitamin C group is less than the placebo group on average with a significance level of 0.05.

Question 2

a)

```r
count = c(45, 14, 85, 218)
death = c("Death", "Death", "No Death", "No Death")
victim_race = c("Caucasian", "African-American", "Caucasian", "African-American")
q2.df = data.frame(count, death, victim_race)
ggplot(q2.df, aes(x = death, y = count, fill = victim_race)) + geom_bar(stat="identity", position = "doc
```



b)

Based on our observation from our bar graph, we want to test whether face affects the conviction. Our null and alternative hypothesis are as follows:

Null hypothesis: $H_0 : P_{Caucasian} = P_{African-American}$

Alternative hypothesis: $H_A : P_{Caucasian} > P_{African-American}$

We will set the alpha value to 0.05.

Because both of our samples have more than 25 observations, we can assume that the data follows an approx. normal distribution.

To test the difference between 2 samples, we use prop.test in R. We will use the Agresti-Coull version of prop.test.

```
prop.test(c(45+1,14+1), c(130+2,232+2), alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c45 + 1 out of 130 + 214 + 1 out of 232 + 2
## X-squared = 49.141, df = 1, p-value = 1.191e-12
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2112574 1.0000000
## sample estimates:
##     prop 1     prop 2
## 0.34848485 0.06410256
```

From our prop.test, we get a p-value of 1.191e-12. Comparing this to our set alpha value of 0.05, it is definitely less. This means we can reject our null hypothesis that the proportion of death sentence conviction where the victim was Caucasian is not the same compared to the proportion of death sentence convictions where the victim was African-American. We can then conclude with a significance level of 0.05 that these proportions are different and that the race of the victim does affect whether an African-American murder is sentenced to death in Georgia and when the victim is Caucasian, the proportion of death sentence convictions is higher than when the victim is African-American.

Question 3

```
q3_data = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/CloudSeedingData.csv")
```

a)

We want to test whether injecting silver iodide **increased** rainfall. This means we want to conduct a right-sided test. We will set up our null and alternative hypothesis as follows:

Null hypothesis: $H_0 : \mu_{Seeded} = \mu_{Unseeded}$

Alternative hypothesis: $H_A : \mu_{Seeded} > \mu_{Unseeded}$

We will set the alpha value to 0.05.

```
nrow(filter(q3_data, TREATMENT=="UNSEEDED"))
```

```
## [1] 26
```

```
nrow(filter(q3_data, TREATMENT=="SEEDED"))
```

```
## [1] 26
```

Both samples contain 26 observations, so we can assume they follow an approx. normal distribution. Because of this, we can use t-test to test our claim.

```r
t.test(q3_data$RAINFALL ~ q3_data$TREATMENT, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  q3_data$RAINFALL by q3_data$TREATMENT
## t = 1.9982, df = 33.855, p-value = 0.02689
## alternative hypothesis: true difference in means between group SEEDED and group UNSEEDED is greater
## 95 percent confidence interval:
##  42.63408      Inf
## sample estimates:
##   mean in group SEEDED mean in group UNSEEDED
##               441.9846               164.5885
```

From our t-test, we get a p-value of 0.02689. Comparing this to our set alpha value of 0.05, we can reject our null hypothesis that the amount of precipitation from cumulus clouds injected with silver iodide is the same as cumulus clouds not injected with silver iodide. Therefore, we can conclude with a significance level of 0.05 that the cumulus clouds injected with silver iodide produce more rainfall compared to the clouds that were not injected with silver iodide.

Our 95% confidence interval from our t.test is **(42.63408, Inf)**. This means we can say with 95% confidence that the difference in rainfall between cumulus clouds injected with silver iodide and cumulus clouds not injected with silver iodide is at least 42.63408. The upper bound of the confidence interval is infinity because this is a right-sided test.

b)

We want to test whether the standard deviation of rainfall between cumulus clouds injected with silver iodide and cumulus clouds not injected with silver iodide is the same. If the standard deviations are the same, then the ratio between the two will be 1. Otherwise is will be $\neq 1$ Knowing this, we will set up our null and alternative hypothesis as follows:

Null hypothesis: $H_0 : \frac{\sigma_{Seeded}}{\sigma_{Unseeded}} = 1$

Alternative hypothesis: $H_A : \frac{\sigma_{Seeded}}{\sigma_{Unseeded}} \neq 1$

We will set the alpha value to 0.05.

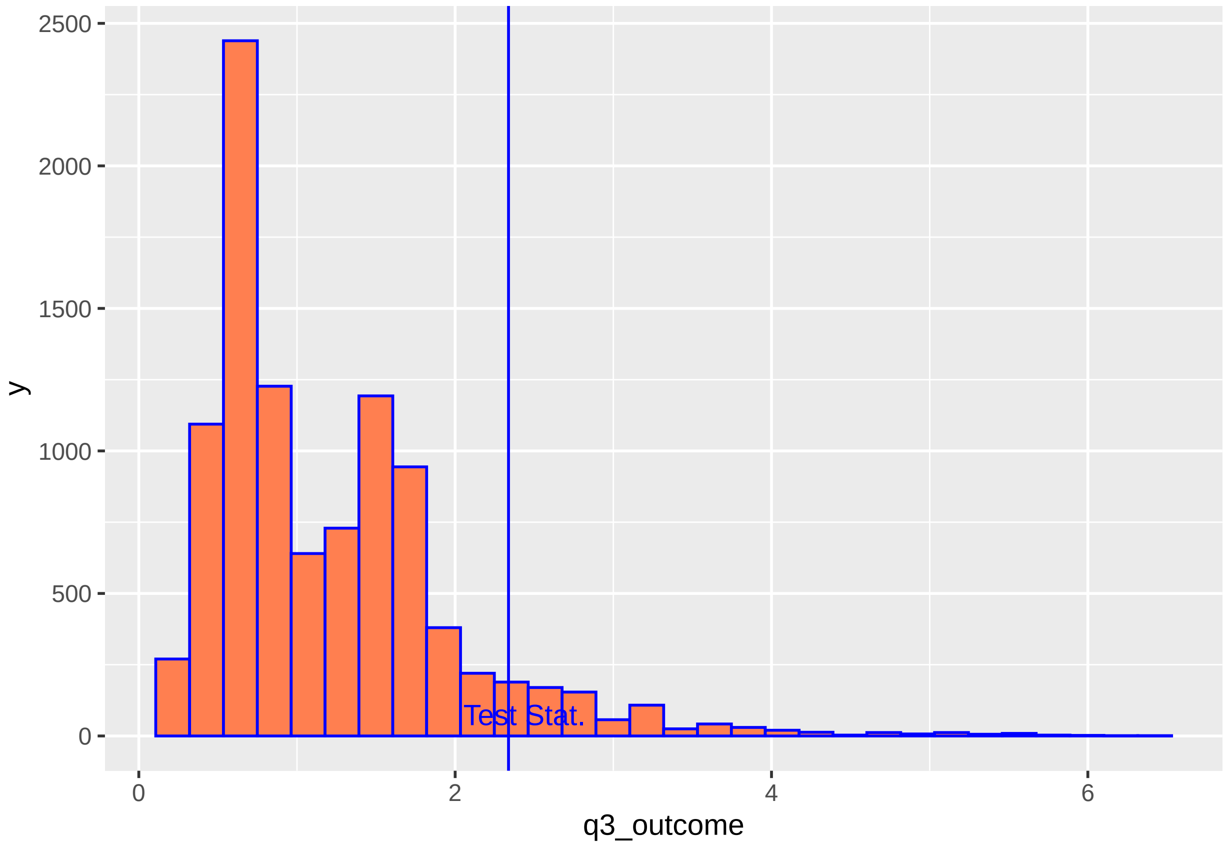
To conduct a permutation test, we will pool the data together and sample 26 observations randomly. These will be the seeded observations and the rest will be the unseeded observations. We will do 10,000 permutations.

```r
rainfall = q3_data$RAINFALL
q3_test_stat = (sd(~RAINFALL, data=filter(q3_data, TREATMENT=="SEEDED")) / sd(~RAINFALL, data=filter(q3
N = 10000
q3_outcome = numeric(N)
for(i in 1:N)
{ index = sample(52, 26, replace=FALSE)
  q3_outcome[i] = (sd(rainfall[index]) / sd(rainfall[-index]))
}
q3_outcome.df = data.frame(q3_outcome)
(sum(q3_outcome >= q3_test_stat) + sum(q3_outcome <= (-1*q3_test_stat)))/(N)
```

```
## [1] 0.0767
```

```
ggplot(q3_outcome.df, aes(x = q3_outcome)) + geom_histogram(fill = "coral", color = "blue") + geom_vlin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From our permutation test, we get a test statistic of 2.3374 and a p-value of 0.0718. This is greater than out set alpha value of 0.05, therefore we fail to reject our null hypothesis that the standard deviation of rainfall in cumulus clouds injected with silver iodide is the same as the standard deviation of cumulus clouds not injected with silver iodide.

c)

From Q3 in the Assignment 4 document:

To start we must take the log transformation of X_{Seeded} and $X_{Unseeded}$

```
seeded = filter(q3_data, TREATMENT=="SEEDED")
unseeded = filter(q3_data, TREATMENT=="UNSEEDED")
log_seeded = log(seeded$RAINFALL)
log_unseeded = log(unseeded$RAINFALL)
```

We know $ln(\bar{X}_1) - ln(\bar{X}_2)$ estimates $ln(\frac{\tilde{\mu_1}}{\tilde{\mu_2}})$. Therefore, we will take the mean of the log transformed data and then subtract the the estimate $ln(\frac{\tilde{\mu_1}}{\tilde{\mu_2}})$.

```
mean(log_seeded) - mean(log_unseeded)
```

```
## [1] 1.143781
```

We get $ln(\frac{\tilde{\mu_1}}{\tilde{\mu_2}}) = 1.143871$. To get rid of the natural log on the left hand side, we will make both sides the exponents of e.

```
exp(1.143781)
```

```
## [1] 3.138613
```

Finally we get $\frac{\tilde{\mu_1}}{\tilde{\mu_2}} = 3.138613$. This tells us that the median rainfall of seeded clouds are 3.138613 more times as large compared to the median rainfall of unseeded clouds.

Question 4

We want tot test whether the proportion of males who experienced new hair growth is greater in males who took minoxidil compared to males who took the placebo. To begin our test, we will state the null and alternative hypothesis:

Null hypothesis: $H_0 : P_{Minoxidil} = P_{Placebo}$

Alternative hypothesis: $H_A : P_{Minoxidil} > P_{Placebo}$

We will set the alpha value to 0.05.

Since both samples have more than 25 observations, we can assume that they follow approximately normal distributions. This means we can use prop.test to test for a difference in the two proportions.

```
prop.test(c(99,62), c(310,309), alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c99 out of 31062 out of 309
## X-squared = 11.331, df = 1, p-value = 0.0003811
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.06124968 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.3193548 0.2006472
```

From our prop.test, we get a p-value of 0.0003811. This is smaller than our set alpha value of 0.05, so we can reject our null hypothesis that the proportion of men than experienced new hair growth is the same between groups that used minoxidil and those who used a placebo. We can conclude with a significance level of 0.05 that the proportion of men that experienced new hair growth with minoxidil is larger than those who experienced new hair growth with a placebo.

From our prop.test, we get a 95% confidence interval of **(0.06124968, 1.0)**. This means that we can say with 95% confidence that the difference between the proportion of men that experienced new hair growth with minoxidil and the proportion of men that experience new hair growth with a placebo is at least 0.06125.

Question 5

a)

We want to test whether giving our chocolate bars had an effect on the answers to Q9. Since doing a two-sided test would not provide any useful information, we will do a right-sided test to test whether the use of chocolate bars had a positive effect:

Null hypothesis: $H_0 : \mu_{Chocolate} = \mu_{NoChocolate}$

Alternative hypothesis: $H_A : \mu_{Chocolate} > \mu_{NoChocolate}$

We will set the alpha value to 0.05.

To use a permutation test, we will pool our data together and randomly sample from the pool. We will do 10,000 permutations.

```
# Read data
q5_data = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/chocnochocratings.csv")
q5_test_stat = mean(~ Q9, data=filter(q5_data, GroupName=="Chocolate")) - mean(~ Q9, data=filter(q5_data
q5_test_stat
```

```
## [1] 0.2033333
```

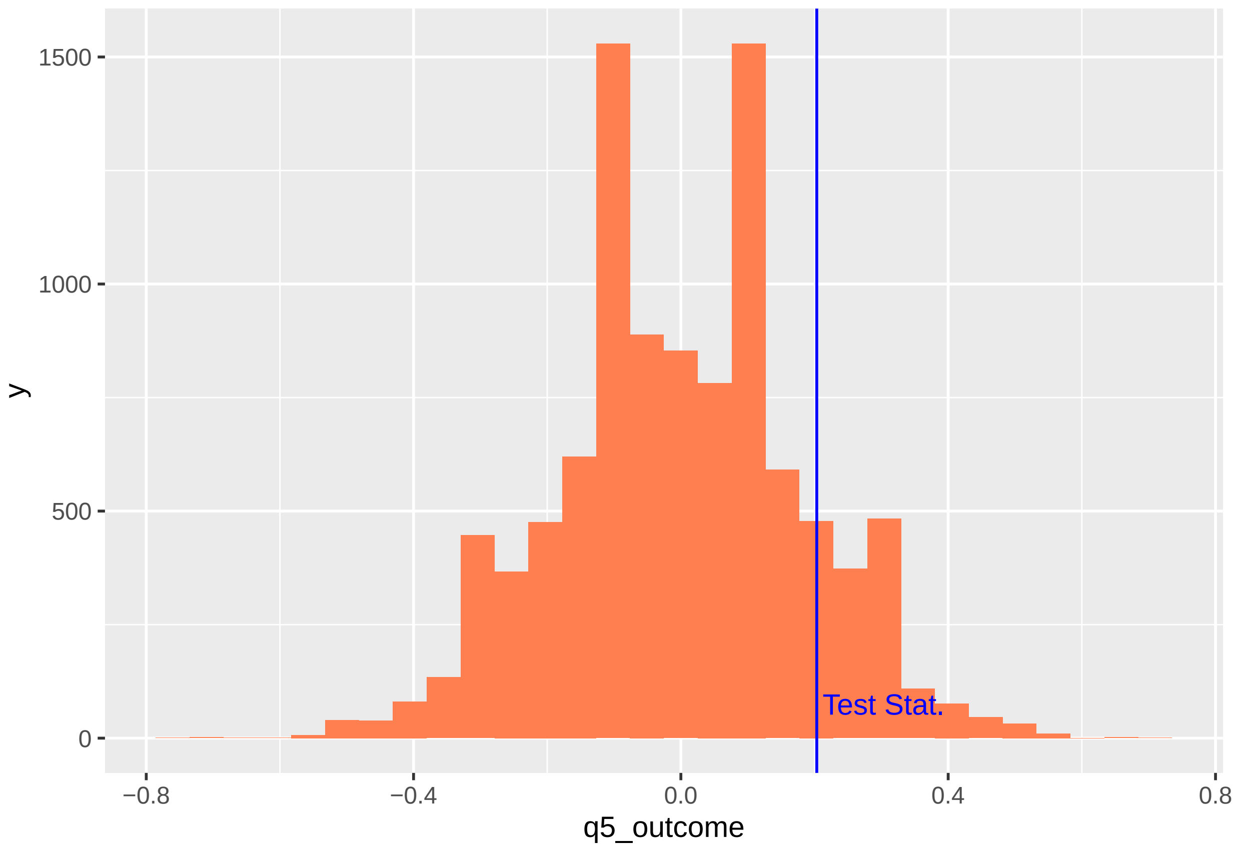
We get a test statistic of $\bar{X}_{Chocolate} - \bar{X}_{NoChocolate} = 0.2033$

```
Q9_result = q5_data$Q9
N = 10000
q5_outcome = numeric(N)
for(i in 1:N)
{ index = sample(98, 50, replace=FALSE) # Out of 98 total observations, randomly select 50 to be "Choco
  q5_outcome[i] = mean(Q9_result[index]) - mean(Q9_result[-index])
}
q5_outcome.df = data.frame(q5_outcome)
(sum(q5_outcome >= q5_test_stat))/(N)
```

```
## [1] 0.161
```

```
ggplot(q5_outcome.df, aes(x = q5_outcome)) + geom_histogram(fill = "coral") + geom_vline(xintercept = q5
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From our permutation test, we get a p-value of 0.1671. This is greater than our set alpha value of 0.05 and therefore we fail to reject our null hypothesis that the response for Q9 for students who were given chocolate is the same as the response for Q9 for students who were not given chocolate on average. We can conclude with a 0.05 significance level that giving students chocolate does not effect their opinion on the professor compared to other professors they have had.

b)

We want to test whether giving our chocolate bars had an effect on the **overall** responses. We will once again do a right-sided test to test for positive effect. We will begin by stating our null and alternative hypothesis:

Null hypothesis: $H_0 : \mu_{Chocolate} = \mu_{NoChocolate}$

Alternative hypothesis: $H_A : \mu_{Chocolate} > \mu_{NoChocolate}$

We will set the alpha value to 0.05.

Since both groups have over 25 observations ($n_{Chocolate} = 50, n_{NoChocolate} = 48$), we can assume that the data follows an approximately normal distribution. Therefore, we can use t.test to test our claim.

```
t.test(q5_data$Overall ~ q5_data$GroupName, alternative="greater") #perform t-test with the greater in
```

```
##
##  Welch Two Sample t-test
##
## data:  q5_data$Overall by q5_data$GroupName
## t = 1.6616, df = 95.93, p-value = 0.04993
## alternative hypothesis: true difference in means between group Chocolate and group NOChoc is greater
## 95 percent confidence interval:
##  9.400666e-05          Inf
## sample estimates:
## mean in group Chocolate    mean in group NOChoc
##                4.072000                3.849792
```

From our t.test, we get a p-value of 0.04993. Since this value is smaller than our set alpha value of 0.05, we can reject our null hypothesis that the overall response for students who were given chocolate is the same as the overall response for students who were not given chocolate on average. We can conclude with a 0.05 significance level that giving students chocolate bars does have a positive affect the overall rating of the professor.

The meaning of our p-value is the probability that we observe results that are at least as extreme as the ones we observed in our sample. This is our evidence against our null hypothesis. In order to reject our null hypothesis, the p-value must be lower than our set criteria of 0.05, meaning there would less than a 5% chance to observe more extreme results. Since our p-value was not less than 0.05, the evidence provided in our sample was not "convincing" enough to conclude something different from our null hypothesis.

c)

We should not use a t-test for the test in part a) because the variable being tested, Q9, is comparing the professor to other professors at the university. It is likely not the case that all students surveyed have taken courses from the exact same instructors. Since it is not an assumption under a t-test that the observations are exchangeable, and therefore assumed to be the same, it is not recommended to use t-test for this particular test in a).

Question 6

a)

To test the claim that the proportion of Albertans that support the adoption of a provincial sales tax has increased since March 2018, we begin by stating our null and alternative hypothesis:

Null hypothesis: $H_0 : P_{2020} = P_{2018}$

Alternative hypothesis: $H_A : P_{2020} > P_{2018}$

We will set the alpha value to 0.05.

To conduct a permutation test, we pool the data together. From this, we will randomly sample 900 observations which will be for 2020 and the rest of the observations will be for 2018.

```r
# Prep the data by combining and calculating the test statistic
q6_2020 = c(rep(1, 358), rep(0, 900-358))
q6_2018 = c(rep(1, 225), rep(0, 900-225))
q6_combined = c(q6_2020, q6_2018)
q6_test_stat = mean(q6_2020) - mean(q6_2018)
q6_test_stat
```

```
## [1] 0.1477778
```

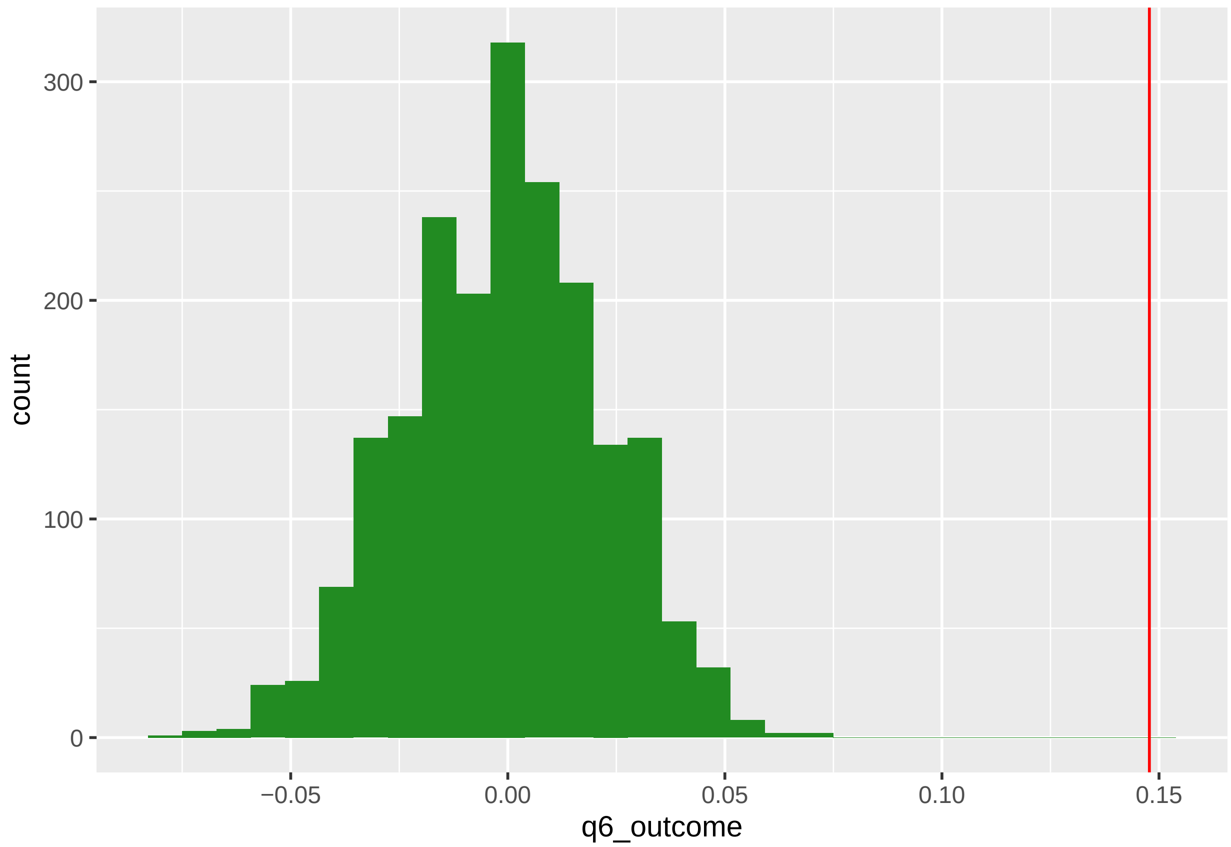
We get a test statistic of 0.1478.

```
# Conduct the permutation test
N = 2000
q6_outcome = numeric(N)
for(i in 1:N)
{ index = sample(1800, 900, replace=FALSE)
  q6_outcome[i] = mean(q6_combined[index]) - mean(q6_combined[-index])
}
q6_outcome.df = data.frame(q6_outcome)
(sum(q6_outcome >= q6_test_stat))/(N)
```

```
## [1] 0
```

```
ggplot(q6_outcome.df, aes(x=q6_outcome))+ geom_histogram(fill = "forestgreen") + geom_vline(xintercept
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



We get a p-value of 0. This means we can reject our null hypothesis that the proportion of Albertans that support the adoption of a provincial sales tax has stayed the same since March 2018. We can then conclude that the proportion of Albertans that support the adoption of a provincial sales tax has increased since 2019 with a significance of 0.05.

b)

Since both of our samples have more than 25 observations, we can assume that both follow an approx. normal distribution.

```
prop.test(c(358,225), c(900,900), alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c out of c358 out of 900225 out of 900
## X-squared = 44.876, df = 1, p-value = 1.049e-11
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1119479 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.3977778 0.2500000
```

From our prop.test, we get a p-value of 1.049e-11. To get our z-value, we take the square root of 44.876 . We will then take the positive value since the proportion of 2020 is greater than 2018 and we are subtracting 2018 from 2020. We get a z-score of 6.6900.

Comparing our p-value to our set alpha value of 0.05, we can reject the null hypothesis that that the proportion of Albertans that support the adoption of a provincial sales tax has stayed the same since March 2018 and conclude that the proportion of Albertans that support the adoption of a provincial sales tax has increased since 2019 with a significance of 0.05. This is the same conclusion we got in part a).

c)

The conclusion from part a) and b) was that $p_{2020} > p2018$.

From our prop.test in part b), we got a 95% confidence interval of $p_{2020} > p2018$ of **(0.1119, 1.0)**. This means we can say with 95% confidence that the proportion of Albertans that support a provincial sales tax in 2020 is at least 0.1119 greater than the proportion of Albertans that support a provincial sales tax in 2018.

Question 7

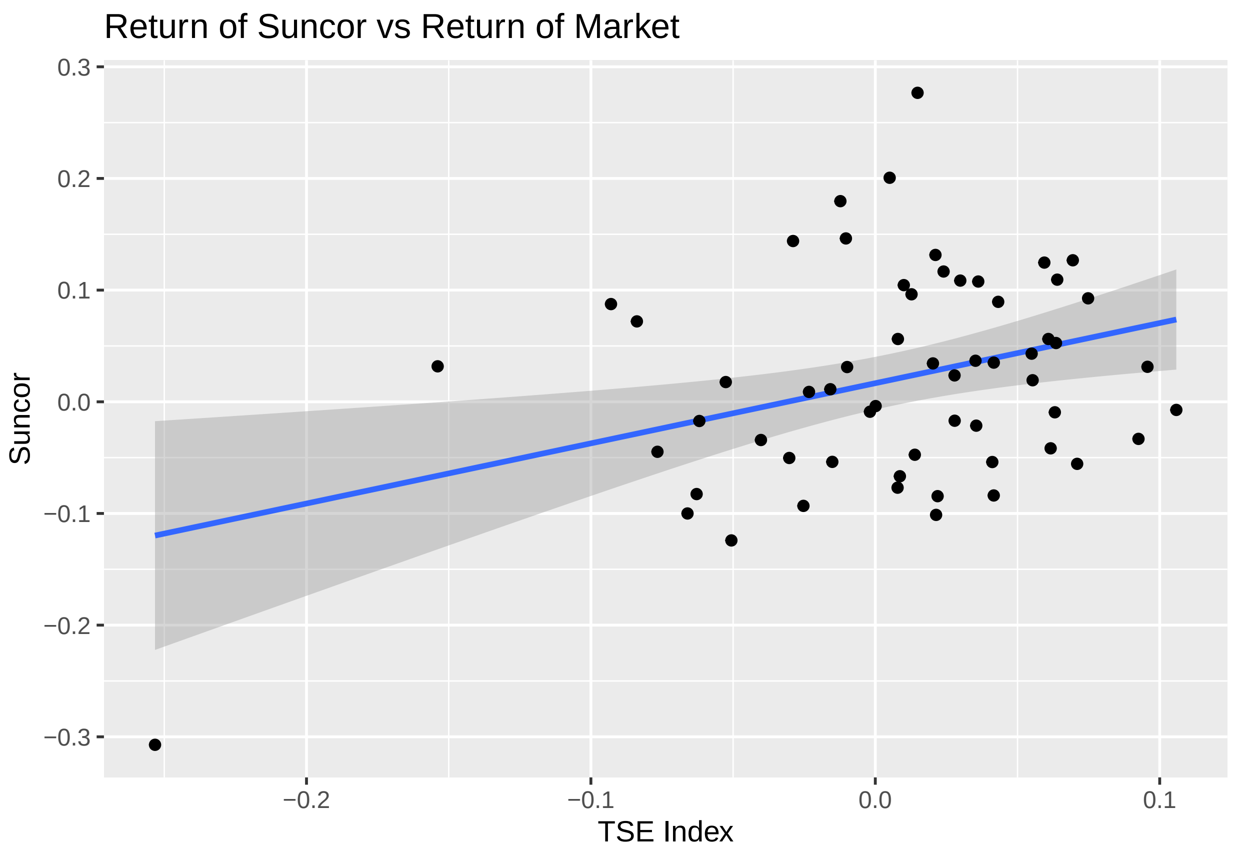
```
capmdata = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/capm.csv")
head(capmdata, 3)
```

```
##      Suncor TSE.Index
## 1  0.008850 -0.023314
## 2  0.035088  0.041661
## 3 -0.016949  0.027904
```

a)

```
ggplot(capmdata, aes(x = TSE.Index, y = Suncor)) + geom_smooth(method = "lm") + geom_point() + xlab("TS
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



b)

```r
predicted_return = lm(Suncor ~ TSE.Index, data=capmdata)
predicted_return$coef
```

```
## (Intercept)   TSE.Index
##  0.01664794  0.53869099
```

The linear regression equation representing the model is $R_{\hat{Stock_{A,i}}} = 0.01665 + 0.5387R_{Market,i}$.

c)

The intercept (β_0) is the expected return of Suncor when the return of the TSE Index is 0%. The TSE.Index (β_1) is the coefficient of the TSE Index in the linear regression equation. This means that for 1% return in the TSE Index, the Suncor will be 0.5387%.

d)

```r
0.01664794+0.53869099*(0.04)
```

```
## [1] 0.03819558
```

A 4% return in the TSE Index will result in a predicted average return of Suncor stock of 3.8200%.

e)

There are two conditions that must be met for our linear regression model to be valid.

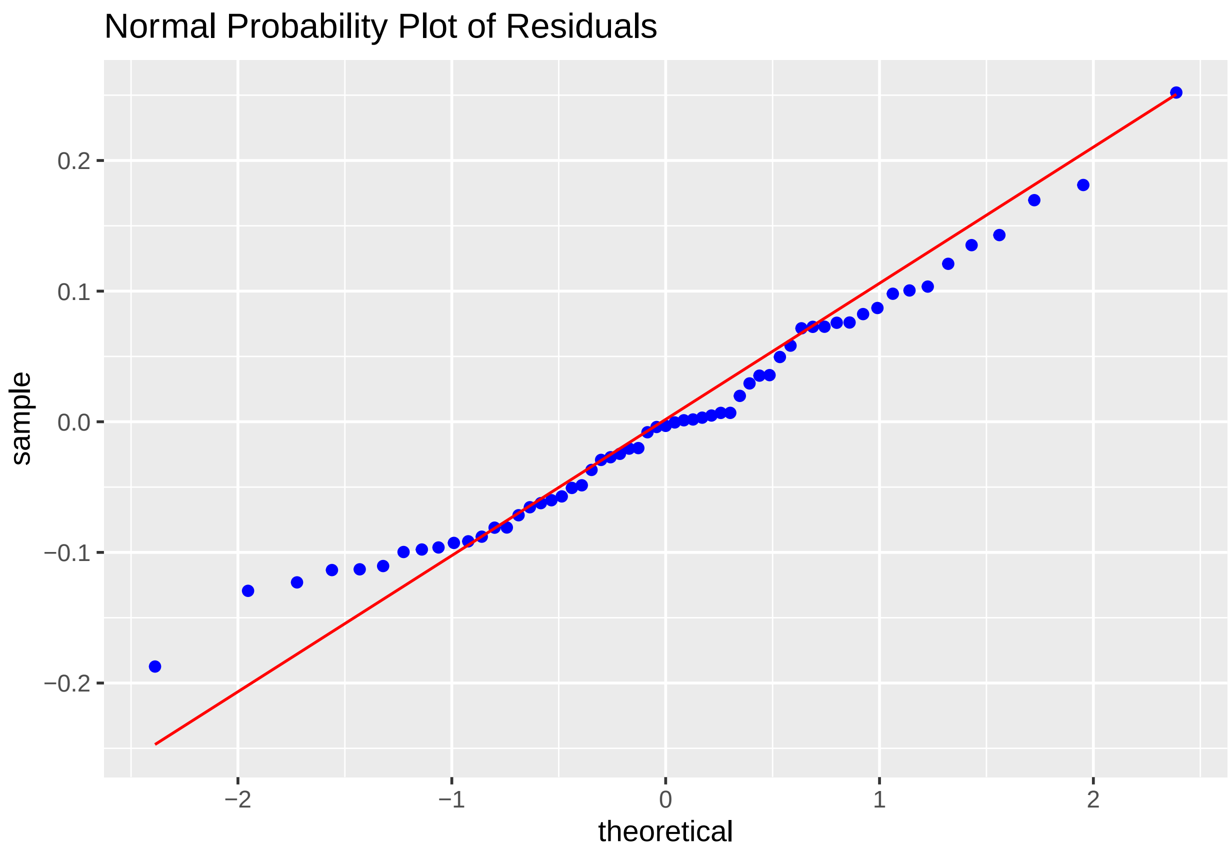
1. Normality of residuals: The dependent variable (Suncor stock returns) must be normally distributed with a mean of μ and standard deviations of σ. To check this we will plot a stat_qq plot of the residuals since $e_i = y_i - \hat{y}_i$, if y is normally distributed, so will the residuals.

2. Homoscedasticity: For each distinct value of the independant variable (TSE.Index return), the dependent variable (Suncor stock returns) has the same standard deviation σ. To check this, we will plot a scatterplot of the fitted values and the residuals.

```
# Get the and residuals fitted values
predicted.return.suncor = predicted_return$fitted.values
ei_suncor = predicted_return$residuals
q7_diagnostic.df = data.frame(predicted.return.suncor, ei_suncor)
```

Normality of Residuals Plot:

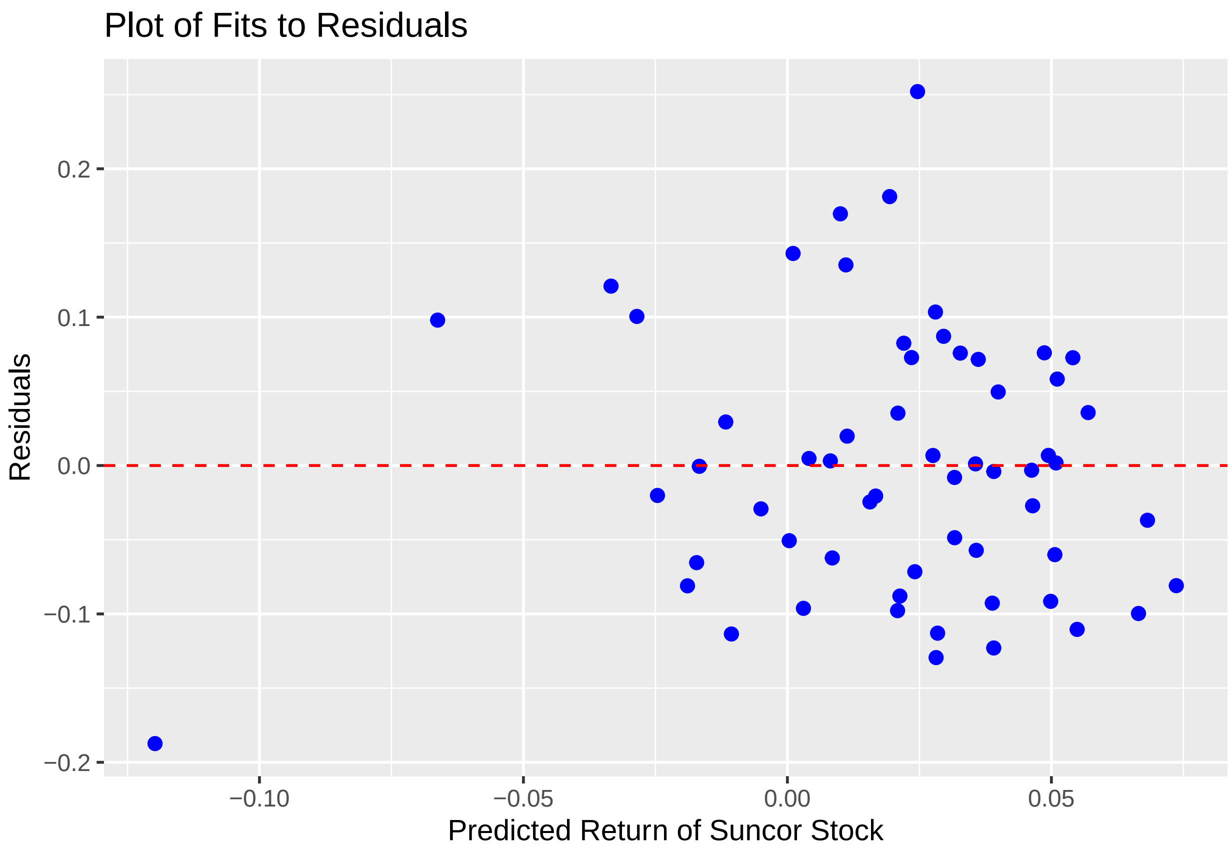
```
ggplot(q7_diagnostic.df, aes(sample = ei_suncor)) +  stat_qq(col='blue') + stat_qqline(col='red') + ggt
```



Looking the normality probability plot, the residuals do seem to be approximately normally distributed and therefore so is the dependent variable (Suncor stock returns). The normality of residuals condition holds.

Homoscedasticity:

```
ggplot(q7_diagnostic.df, aes(x = predicted.return.suncor, y = ei_suncor)) +  geom_point(size=2, col='blu
```



Looking the plot of fits to residuals, the residuals do seem to be evenly distributed over the return of Suncor's stock. We can they say that the condition of homoscedasticity holds.

Since both conditions hold, our linear regression model is valid.

f)

To test whether the monthly return of Suncor stock can be expressed as a positive linear function of the monthly return of the TSE Index, we must test whether the linear regression coefficient of the TSE Index is different from 0. The null and alternative hypothesis are below:

Null hypothesis: $H_0 : \beta_{TSE} = 0$

Alternative hypothesis: $H_A : \beta_{TSE} \neq 0$

We will set the alpha value to 0.05.

We can use a t test to check our claim.

```
coef(summary(predicted_return))
```

```
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.01664794 0.01177393 1.413966 0.162810655
## TSE.Index   0.53869099 0.19177963 2.808906 0.006797419
```

From our t-test, we get a test statistic of 2.8089 and a p-value of 0.006797. This is smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient for TSE Index is 0. Therefore, we can conclude that the monthly return of Suncor's stock can be expressed as a linear function of the monthly return of the TSE Index, and since $\beta_{TSE} > 0$, we can say it is also positive.

g)

```
confint(predicted_return, conf.level=0.95)
```

```
##                     2.5 %     97.5 %
## (Intercept) -0.006928949 0.04022482
## TSE.Index    0.154658904 0.92272309
```

For β_1, the coefficient for TSE.Intercept, the 95% confidence interval is **(0.1547, 0.99227).** This means that we can say with 95% confidence that the linear coefficient for TSE.Intercept, β_1, is between 0.1547 and 0.9227.

In other words, when the TSE Index has a return of 1%, we can say with 95% confidence that the return of the Suncor stock will be between 0.1547% and 0.9227%.

h)

```
predict(predicted_return, newdata=data.frame(TSE.Index = 0.03), interval="conf") #compute the 95% CI fo
```

```
##          fit         lwr        upr
## 1 0.03280867 0.007660256 0.05795708
```

The 95% confidence interval for the mean monthly return for Suncor stock when the TSE Index monthly return is 3% is **(0.00766, 0.05796)**. This means we can say with 95% confidence that the mean monthly return of Suncor stock will be between 0.766% and 5.796% when the monthly return of the TSE Index is 3%.

i)

```
predict(predicted_return, newdata=data.frame(TSE.Index = 0.0116), interval="predict") #compute the 95%
```

```
##          fit        lwr       upr
## 1 0.02289675 -0.1587618 0.2045553
```

We get a 95% confidence interval of **(-0.1588, 0.2046)**. This means that in the month of September when the return of the TSE Index was 1.16%, we can say with 95% confidence that the return of the Suncor stock is between -15.88% and 20.46%.

j)

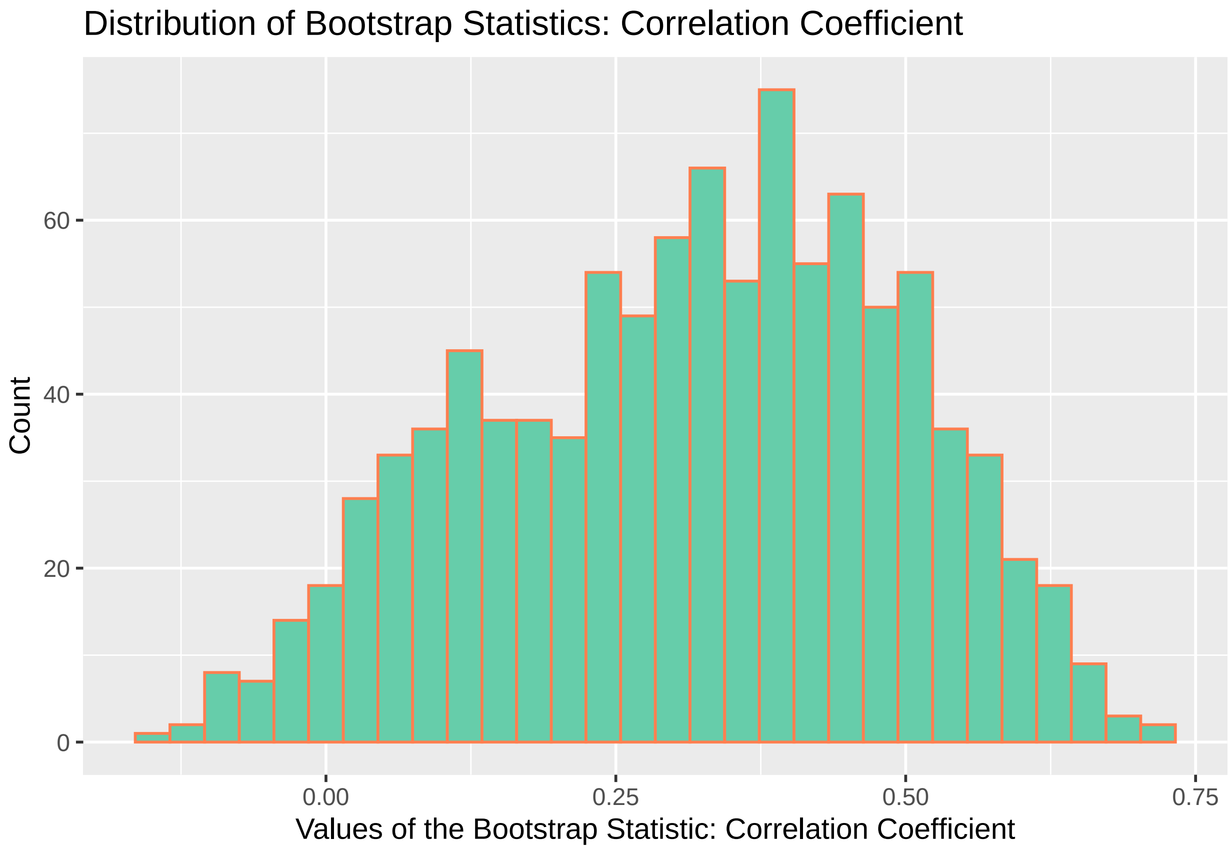
```
Nbootstraps = 1000
cor.boot = numeric(Nbootstraps)
nsize = dim(capmdata)[1]
for(i in 1:Nbootstraps){

    index = sample(nsize, replace=TRUE)
    suncor.boot = capmdata[index, ]
    cor.boot[i] = cor(~Suncor, ~TSE.Index, data=suncor.boot)
}

q7_bootstrapresultsdf = data.frame(cor.boot)
qdata(~cor.boot, c(0.025, 0.975), data=q7_bootstrapresultsdf)
```

```
##        2.5%       97.5%
## -0.03443019  0.62289245
```

From our bootstrap, we get a confidence interval of **(-0.0275, 0.6342)**. This means we can say with 95% confidence that the linear correlation between the monthly return of Suncor's stock and the TSE index is between -0.0275 and 0.6342.

```
ggplot(q7_bootstrapresultsdf, aes(x = cor.boot)) + geom_histogram(col="coral", fill="aquamarine3") + xla
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Question 8

a)

We want to test whether there is s relationship between GunLaw and SpendSci. To do this we will doa chi-squared test of independance since this data is categorical. Below are the null and alternative hypothesis:

Null hypothesis: GunLaw and SpendSci are independent (Not related)

Alternative hypothesis: GunLaw and SpendSci are dependent (Are related)

We will set alpha to 0.05.

```
gss = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/GSS2002.csv")
gun_and_spend = gss[, c("GunLaw", "SpendSci")]
```

b)

First we must convert the data into a contingency table, and then we can conduct the chi-squared test.

```
q8_counts = tally(~GunLaw + SpendSci, data=na.omit(gun_and_spend))
q8_counts
```

```
##          SpendSci
## GunLaw   About right Too little Too much
##   Favor          166        117       42
##   Oppose          35         37       12
```

```
chisq.test(q8_counts, correct=FALSE)
```

```
##
##  Pearson’s Chi-squared test
##
## data:  q8_counts
## X-squared = 2.4447, df = 2, p-value = 0.2945
```

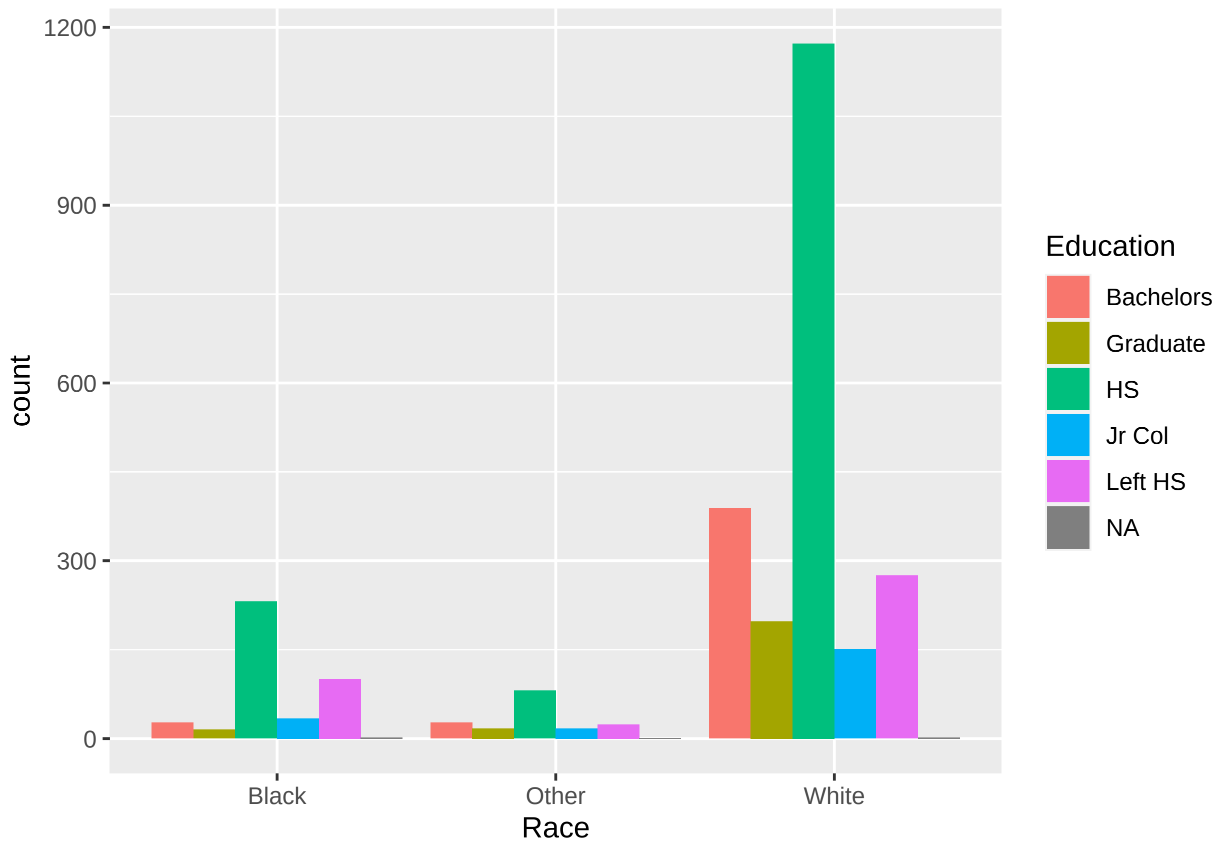
From our chi-squared test we get a test statistic of 2.4447 and a p-value of 0.2945.

c)

Because out p-value is larger than our set alpha value of 0.05, we fail to reject our null hypothesis and conclude that there is no relationship between one’s support for gun laws and their opinion on government spending on science with a significance level of 0.05.

d)

```
# Bar graph from Assignment 1 Q9
ggplot(data=gss, aes(x = Race, fill=Education)) + geom_bar(position = "dodge", na.rm=TRUE)
```



To test whether one's level of education is independent of their race, we will again conduct a chi-squared test of independence.

Null hypothesis: Education and Race are independent (Not related)

Alternative hypothesis: Education and Race are dependent (Are related)

We will set alpha to 0.05.

```
# Creating contingency table and conducting ch-squared test
education_and_race =  gss[, c("Education", "Race")]
q8d_counts = tally(~Education + Race, data=na.omit(education_and_race))
q8d_counts
```

```
##              Race
## Education    Black Other White
##   Bachelors     27    27   389
##   Graduate      15    17   198
##   HS           231    81  1173
##   Jr Col        34    17   151
##   Left HS      101    24   275
```

```
chisq.test(q8d_counts, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  q8d_counts
## X-squared = 79.05, df = 8, p-value = 7.59e-14
```

From our chi-squared test of independence, we get a chi-squared value of 79.05 and a p-value of 7.59e-14. Since out p-value is smaller than our set alpha value of 0.05, we can reject the null hypothesis that education and race are independence and have no relationship. Therefore, we can conclude with a 0.05 significance level that one's level of education and race is dependent of one another.

Question 9

```
# Creating contigency table
q9_treatment = rbind(c(15,7,3,15), c(22, 7,3,11))
rownames(q9_treatment) = c("Fluvoxamine", "Placebo")
colnames(q9_treatment) = c("No Response", "Moderate Response", "Marked Response", "Remission")
q9_treatment
```

```
##             No Response Moderate Response Marked Response Remission
## Fluvoxamine          15                 7               3        15
## Placebo              22                 7               3        11
```

To test whether there is association between the type of treatment a patient recieves and the patient's response, we conduct a chi-squared test of independence on the above contingency table containing the type of treatment and patient's response to the treatment. We will begin by setting up out null and alternative hypothesis:

Null hypothesis: Treatment and Response are independent (Not related)

Alternative hypothesis: Treatment and Response are dependent (Are related)

We will set alpha to 0.05.

Conducting chi-squared test: (since there are some cells with only 5 observations, we set "simulate.p.value" to TRUE)

```
chisq.test(q9_treatment, correct=FALSE, simulate.p.value=TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  q9_treatment
## X-squared = 1.8337, df = NA, p-value = 0.6237
```

From our chi-squared test of independence, we get a p-value of 0.6327. This is much larger than our set alpha value of 0.05, so we fail to reject the null hypothesis of a patient's treatment and response being independent. We then conclude with a significance level of 0.05 that the patient's treatment is not associated with their response.

Question 10

```
Ass5ques5data = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/bondsdata.csv")
# Remove the data in 2001.
Ass5ques5data_no_2001 <-  subset(Ass5ques5data, season != "2001")
head(Ass5ques5data_no_2001)
```

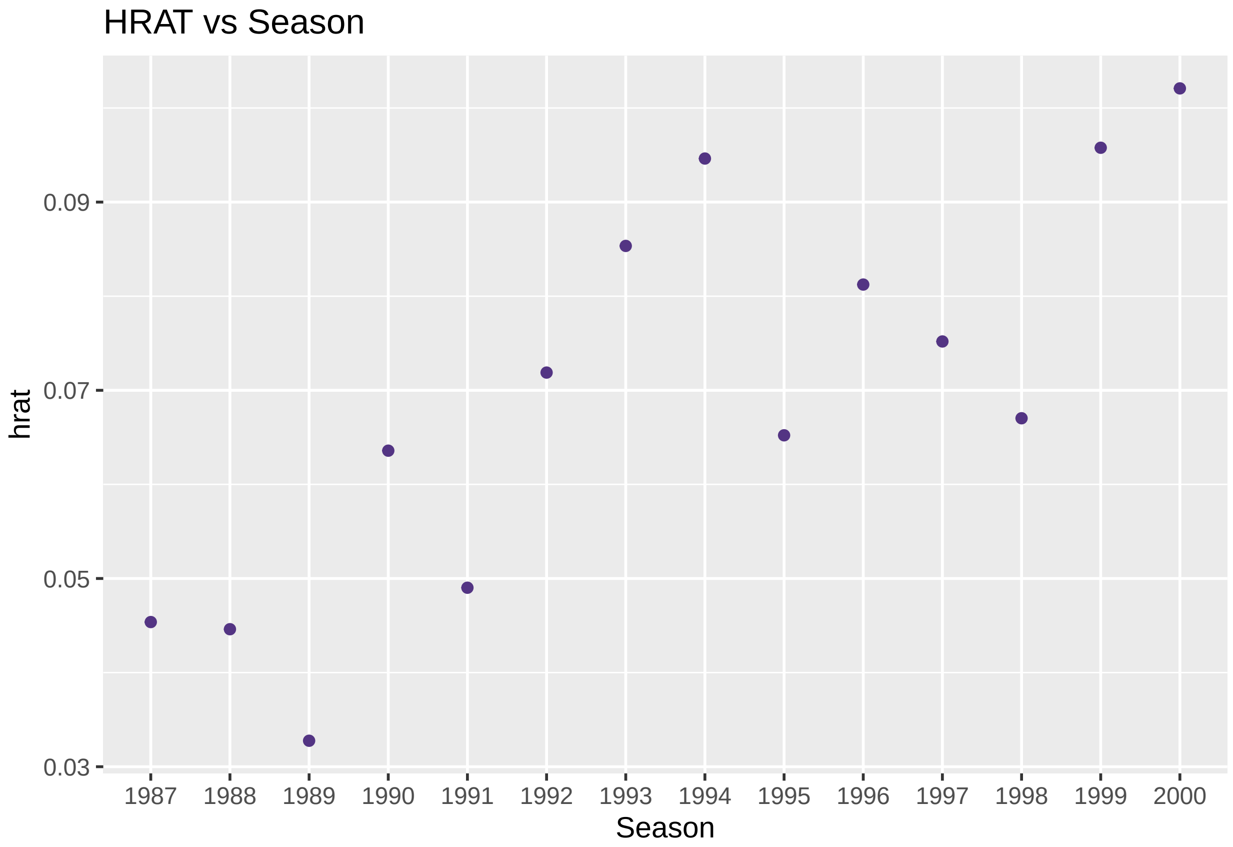
```
##   season     hrat
## 1   1987 0.045372
## 2   1988 0.044610
## 3   1989 0.032759
## 4   1990 0.063584
## 5   1991 0.049020
## 6   1992 0.071882
```

```
tail(Ass5ques5data_no_2001)
```

```
##    season     hrat
## 9    1995 0.065217
## 10   1996 0.081238
## 11   1997 0.075188
## 12   1998 0.067029
## 13   1999 0.095775
## 14   2000 0.102083
```

We want to build a linear regression model in the form: $HRAT_i = A + B * Year_i + e_i$. Lets start by looking at a plot of the data to get a better understanding.

```
ggplot(Ass5ques5data_no_2001)  + geom_point(aes(x = factor(season), y = hrat), color = "#533483") + xlab
```



There seems to be a general positive correlation with hrat and season.

Creating our linear regression model:

```
predicted_hrat = lm(hrat ~ season, data=Ass5ques5data_no_2001)
predicted_hrat$coef
```

```
##  (Intercept)       season
## -7.992499290  0.004044169
```

From our regression, we get the equation: $\hat{HRAT_i} = -7.9925 + 0.004044 * Year_i$. The interecept does not make much sense in the context of our data, so we will ignore any interpretation of it. As for the regression coefficient of Year, this says that as each season goes by, Barry Bonds' hrat increases by 0.004044.

Before we can use this model to make predictions, we must check two conditions to ensure it is valid.

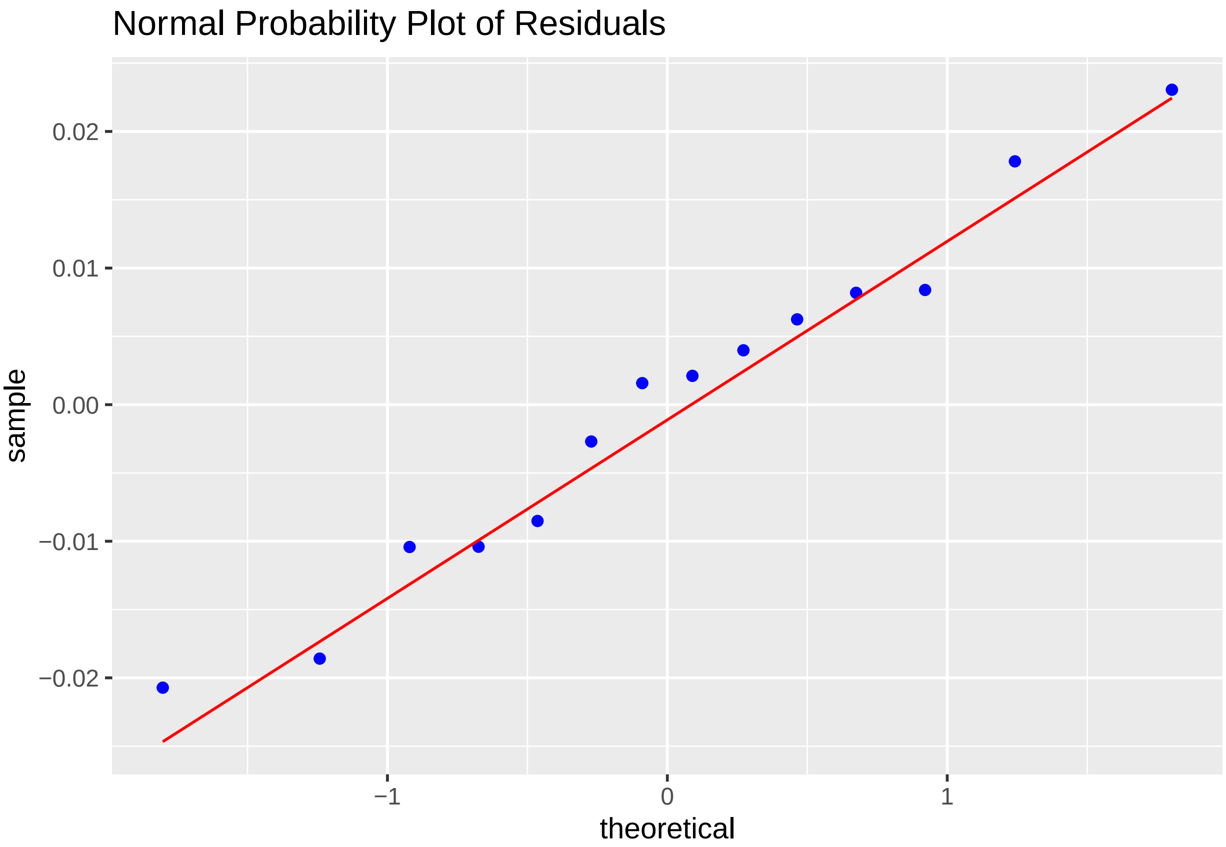
1. Normality of residuals: The dependent variable (Barry Bonds' HRAT) must be normally distributed with a mean of μ and standard deviation of σ. To check this we will plot a stat_qq plot of the residuals since $e_i = y_i - \hat{y}_i$, if y is normally distributed, so will the residuals.

2. Homoscedasticity: For each distinct value of the independent variable (Season), the dependent variable (Barry Bonds' HRAT) has the same standard deviation σ. To check this, we will plot a scatter plot of the fitted values and the residuals.

```
# Get the and residuals fitted values
predicted.hrat.fitted = predicted_hrat$fitted.values
ei_hrat = predicted_hrat$residuals
q10_diagnostic.df = data.frame(predicted.hrat.fitted, ei_hrat)
```

Normality of Residuals Plot:

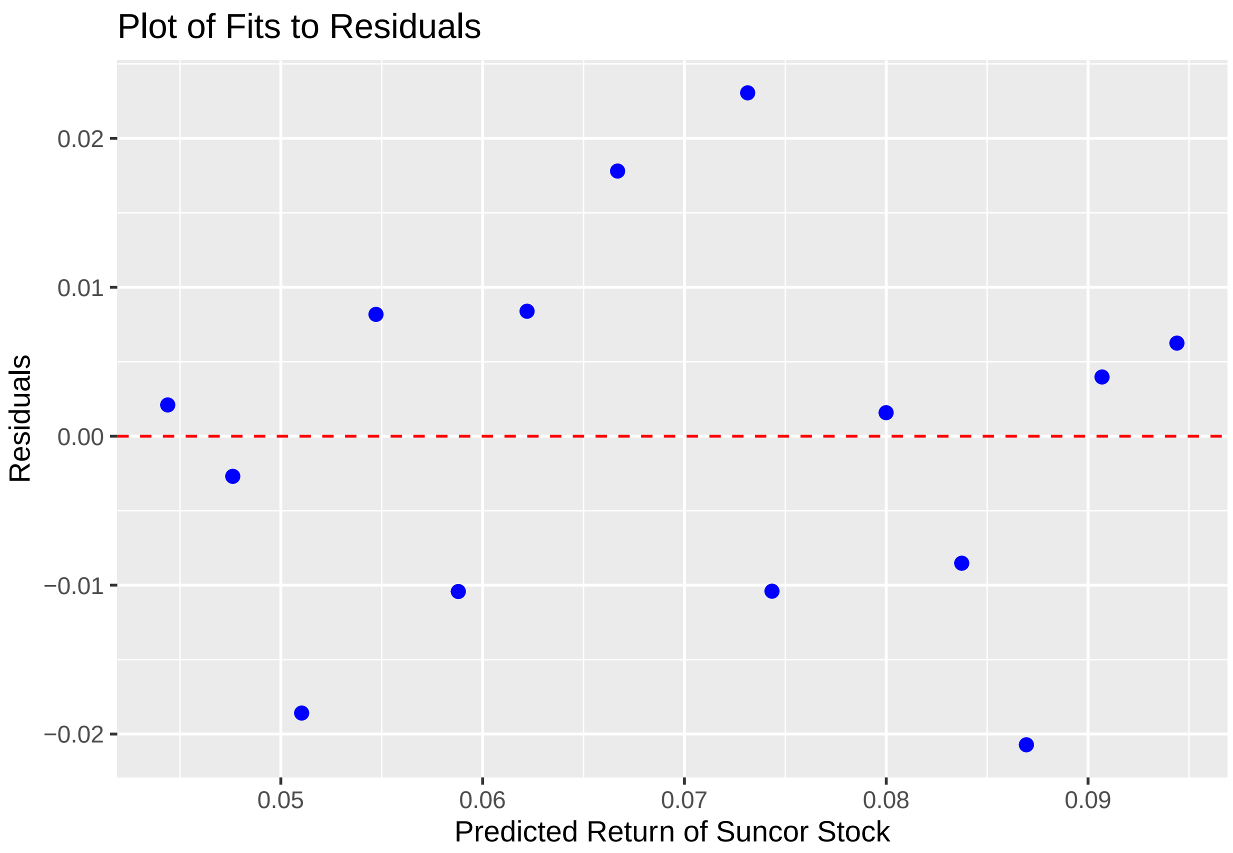
```
ggplot(q10_diagnostic.df, aes(sample = ei_hrat)) +  stat_qq(col='blue') + stat_qqline(col='red') + ggti
```



Based on the above normal probability plot, the residuals appear to be approximately normally distributed. Therefore, the normality of residuals condition holds.

Homoscedasticity:

```
ggplot(q10_diagnostic.df, aes(x = predicted.hrat.fitted, y = ei_hrat)) +  geom_point(size=2, col='blue'
```



Looking the plot of fits to residuals, the residuals seem to be evenly distributed over the HRAT. Therefore, we can they say that the condition of homoscedasticity holds.

Since both conditions hold, our linear regression model is valid and we can use it to predict values of HRAT based on season.

We will now do a hypothesis test to test whether the regression coefficient of Season is different from 0. Since we are only testing the coefficient of one variable, we will use a t-test.

Null hypothesis: $H_0 : \beta_{Season} = 0$

Alternative hypothesis: $H_A : \beta_{Season} \neq 0$

We will set the alpha value to 0.05.

```
summary(predicted_hrat)
```

```
##
## Call:
## lm(formula = hrat ~ season, data = Ass5ques5data_no_2001)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020722 -0.009931  0.001841  0.007701  0.023055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.9924993  1.7566775  -4.550 0.000666 ***
## season       0.0040442  0.0008812   4.589 0.000622 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.01329 on 12 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.6068
## F-statistic: 21.06 on 1 and 12 DF,  p-value: 0.0006222
```

From our t-test, we get a test statistic of 4.589384 and a p-value of 0.0006222474. This is smaller than the set alpha value of 0.05, so we reject our null hypothesis that the linear regression coefficient for Season is 0. And therefore, we can conclude that Barry Bonds’ HRAT can be expressed as a linear function of the season, and since $\beta_{Season} > 0$, we can say it is also positive.

We also get an R-squared value of 0.6371, meaning out independant variable explains approx. 63.71% of the variance in the dependent variable, which is quite good.

```
predict(predicted_hrat, newdata=data.frame(season = 2001), interval="predict")
```

```
##          fit        lwr       upr
## 1 0.09988334 0.06662845 0.1331382
```

Using our linear regression model, we get a predicted HRAT of 0.099883 for Barry Bonds’ 2001 season. How accurate is this to the real HRAT? Barry Bonds’ actual HRAT for the 2001 season was 0.153400, so our predicted error is $0.153400 - 0.099883 = 0.053517$.

We also calculated a 95% confidence interval of HRAT for the 2001 season. **(0.06663, 0.13314)**. This means we can say with 95% confidence that the value of Barry Bonds’ HRAT for the 2001 season is between 0.06663 and 0.13314, Interestingly, Barry Bonds’ actual HRAT is not in this range.

From this, we cannot conclude whether or not Barry Bonds was on steroids in the 2001 season, but it does seem like there was a factor that attributed to a much larger HRAT in the 2001 season than what was expected.

Session Info:

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-conda-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 LTS
##
## Matrix products: default
## BLAS/LAPACK: /opt/conda/lib/libopenblasp-r0.3.21.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] resampledata_0.3.1 mosaic_1.8.3       ggridges_0.5.3     mosaicData_0.20.2
##  [5] ggformula_0.10.1   ggstance_0.3.5     Matrix_1.4-1       lattice_0.20-45
##  [9] dplyr_1.0.9        ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] ggrepel_0.9.1     Rcpp_1.0.9        tidyr_1.2.0       assertthat_0.2.1
##  [5] digest_0.6.29     utf8_1.2.2        ggforce_0.3.4     R6_2.5.1
##  [9] plyr_1.8.7        backports_1.4.1   labelled_2.9.1    evaluate_0.16
## [13] highr_0.9         pillar_1.8.1      rlang_1.0.4       rstudioapi_0.14
## [17] rmarkdown_2.15    labeling_0.4.2    splines_4.1.3     readr_2.1.2
## [21] stringr_1.4.1     htmlwidgets_1.5.4 polyclip_1.10-0   munsell_0.5.0
## [25] broom_1.0.0       compiler_4.1.3    xfun_0.32         pkgconfig_2.0.3
## [29] mgcv_1.8-40       htmltools_0.5.3   tidyselect_1.1.2  tibble_3.1.8
## [33] gridExtra_2.3     mosaicCore_0.9.0  fansi_1.0.3       crayon_1.5.1
## [37] tzdb_0.3.0        withr_2.5.0       MASS_7.3-58.1     grid_4.1.3
## [41] nlme_3.1-159      gtable_0.3.0      lifecycle_1.0.1   DBI_1.1.3
## [45] magrittr_2.0.3    scales_1.2.1      cli_3.3.0         stringi_1.7.8
## [49] farver_2.1.1      leaflet_2.1.1     ellipsis_0.3.2    ggdendro_0.1.23
## [53] generics_0.1.3    vctrs_0.4.1       tools_4.1.3       forcats_0.5.2
## [57] glue_1.6.2        tweenr_2.0.1      purrr_0.3.4       hms_1.1.2
## [61] crosstalk_1.2.0   fastmap_1.1.0     yaml_2.3.5        colorspace_2.0-3
## [65] knitr_1.39        haven_2.5.0
```