| Name | Student ID Number | Tutor | Tutorial Day & Time | Tutorial Location |
|---|---|---|---|---|
| **Kim Seang CHY** | **998008** | **Charles Siriban** | **Fri 12:00pm** | **The Spot 3014** |
| | | | | |

| | |
|---|---|
| **Subject Code:** ECOM20001 | **Subject Name:** Econometric 1 |
| **Assignment Name:** Assignment 1 | |
| **Due Date and Time:** 8am, Monday, 08 April 2019 | |

1. Descriptive Statistics

Table 1: 2060 wines sample from wine regions in California and Washington on the west coast of the United States

| Data | Median | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| price | 28.56 | 34.67 | 21.31 | 6.9 | 20.7 | 40.8 | 150 |
| cases | 999 | 2405 | 3438 | 20 | 430 | 2500 | 20000 |
| score | 87 | 86.972 | 3.59 | 68 | 85 | 89 | 97 |
| napa | 0 | 0.3107 | 0.463 | 0 | 0 | 1 | 1 |
| sonoma | 0 | 0.251 | 0.434 | 0 | 0 | 1 | 1 |
| d1995 | 0 | 0.4806 | 0.5 | 0 | 0 | 1 | 1 |
| d1999 | 1 | 0.5194 | 0.5 | 0 | 0 | 1 | 1 |

Looking at the data sets, we can see a typical wine in California, and Washington's wine region has a price $34.67 per bottle, a WSM testing score of 86.97/100 per bottle and an average 2,405 cases per wine production. The data set also indicates that 31.07% of wine production occurred in the Napa Valley wine region while 25.1% of the productions occurred in the Sonoma wine region. Additionally, we can see that 48.06% of the wine production happened in 1995, and 51.94% of the wine production occurred in 1999.

1. Computing 99% confidence interval for:
   (i). Price:

| Statistic | Result |
|---|---|
| Mean ($\bar{X}$) | 34.67 |
| Sample Sizes ($n$) | 2060 |
| St. Dev. ($\sigma$) | 21.31 |
| Z-value for 99% | $\pm 2.58$ |

$$\left(\bar{X} - 2.56\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.56\frac{\sigma}{\sqrt{n}}\right) = \left(34.67 - 2.56\frac{21.31}{\sqrt{2060}}, 34.67 + 2.56\frac{21.31}{\sqrt{2060}}\right)$$

$$= (33.4587, 35.8817)$$

Thus, based on simple data the 99% confident interval for the price of the wine per bottle have an upper bound of $35.88 and a lower bound of $33.46.

(ii). Cases:

| Statistic | Result |
|---|---|
| Mean ($\bar{X}$) | 2405.07 |
| Sample Sizes ($n$) | 2060 |
| St. Dev. ($\sigma$) | 3438.29 |
| Z-value for 99% | $\pm 2.58$ |

$$\left(\bar{X} - 2.56\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.56\frac{\sigma}{\sqrt{n}}\right) = \left(2405.07 - 2.56\frac{3438.29}{\sqrt{2060}}, 2405.07 + 2.56\frac{3438.29}{\sqrt{2060}}\right)$$

$$= (2209.6239, 2600.5183)$$

Thus, based on simple data the 99% confident interval for the number of cases of wine produced per production have an upper bound of 2601 cases and lower bound of 2210 cases.

(iii). Scores:

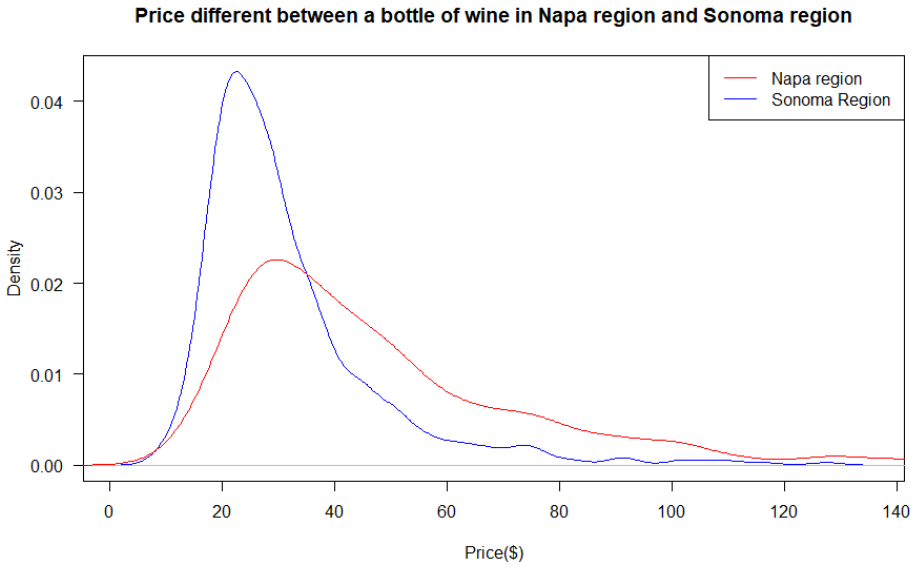
| Statistic | Result |
|---|---|
| Mean ($\bar{X}$) | 86.97 |
| Sample Sizes ($n$) | 2060 |
| St. Dev. ($\sigma$) | 3.59 |
| Z-value for 99% | $\pm 2.58$ |

$$\left(\bar{X} - 2.56\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.56\frac{\sigma}{\sqrt{n}}\right) = \left(86.97 - 2.56\frac{3.59}{\sqrt{2060}}, 86.97 + 2.56\frac{3.59}{\sqrt{2060}}\right)$$

$$= (86.7676, 87.1760)$$

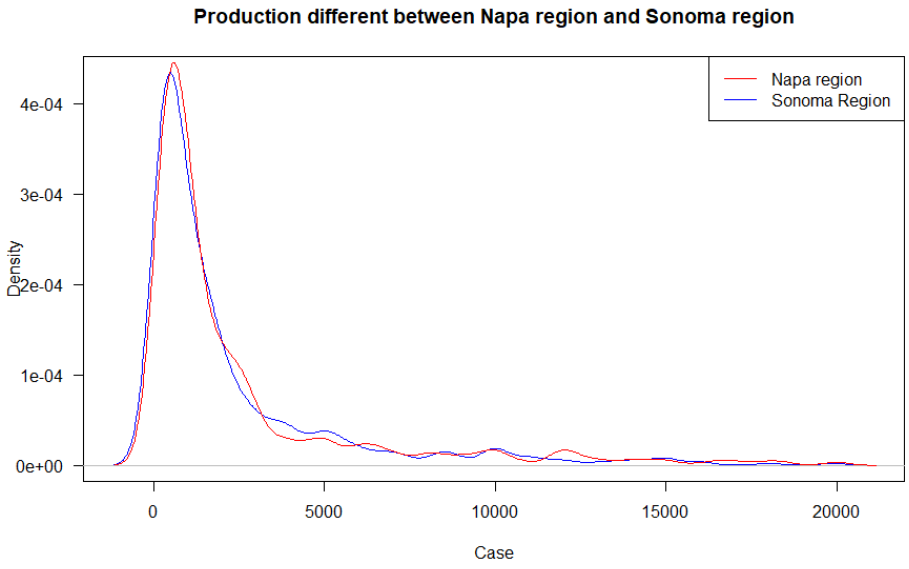
Thus, based on simple data the 99% confident interval for the WSM wine tasting score have an upper bound of 87.17 and lower bound of 86.77 cases.

2. Probability density graph:
   a. The price difference between a bottle of wine in Napa Valley and Sonoma Valley



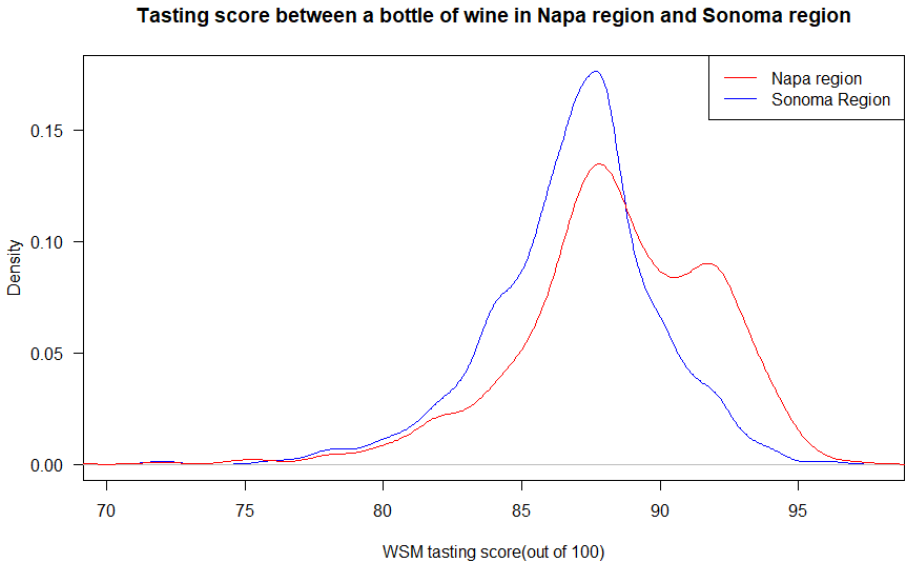

- The price of an individual wine bottles in Napa region have a larger variance when compared to the Sonoma region.
- From looking at the density plots, we can imply an average wine in Napa region are more expensive than wine in the Sonoma region
- The two density plots are right-skewed

b. Production different between Napa Valley and Sonoma Valley



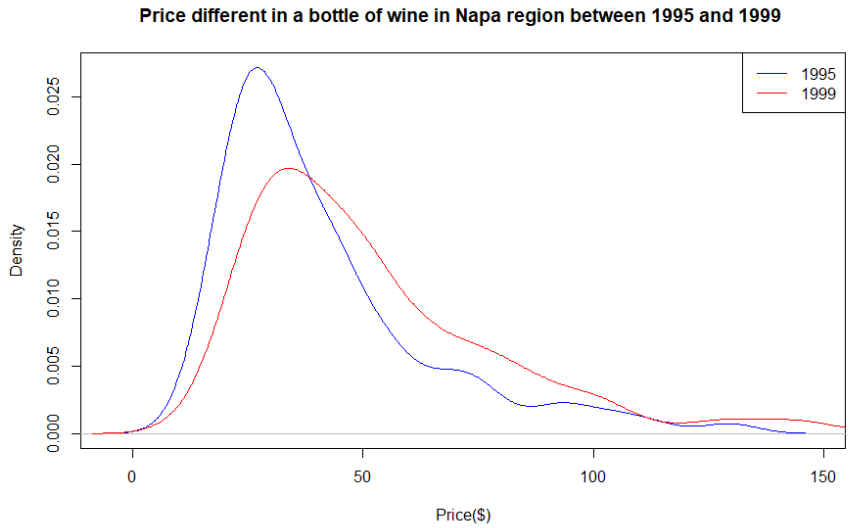

- From looking at density plots, the two regions have an identical variance for quantity of cases per production, which imply an average quality of cases production in identical to one another.
- Additionally, the two density plots for the quantity of production are right-skewed.

c. Tasting score different between Napa Valley and Sonoma Valley



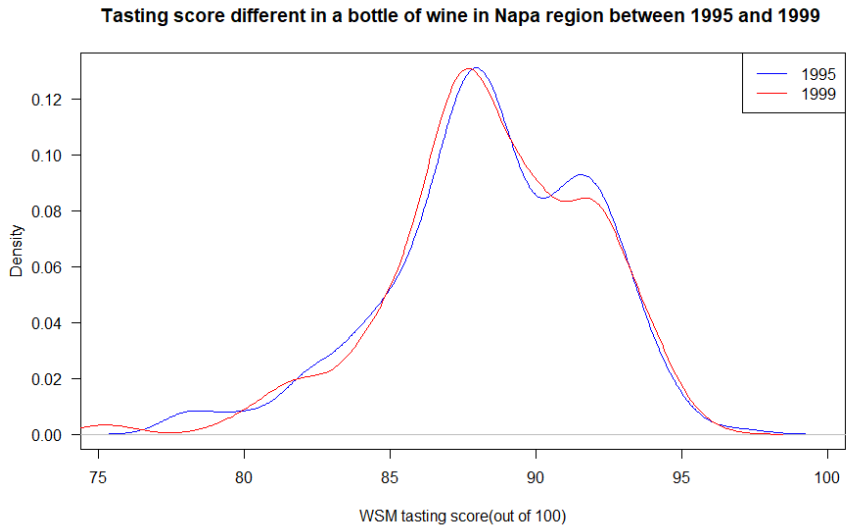

- A typical bottle wine in the Napa region will be of a higher quality than a typical wine in a Sonoma region, which is reflected in the larger density concentration of wines in Napa region for wines with a rating of 90 to 95 WSM score.
- The density plots for two region tasting score are left-skewed.

3. Plot density plot graph for price and score different in 1995 and 1999
   a. Price different in Napa wine region between 1995 and 1999



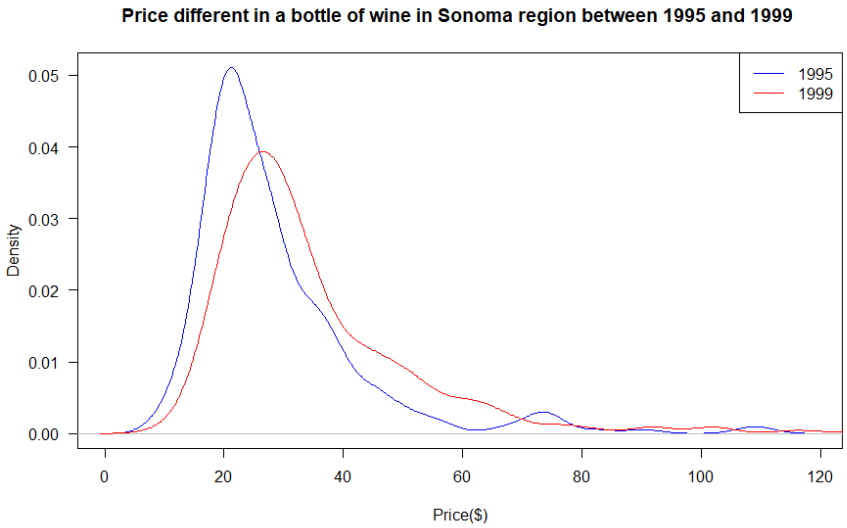

- The price variance and average prices of a typical bottle of wine produced in 1999 in Napa Valley has increased when compared in 1995 in Napa Valley.
- Looking at the density plot for the price of the 1999 and 1995 vintage from the Napa Valley wine region, we can see it is right-skewed.

   b. Score different in Napa wine region between 1995 and 1999



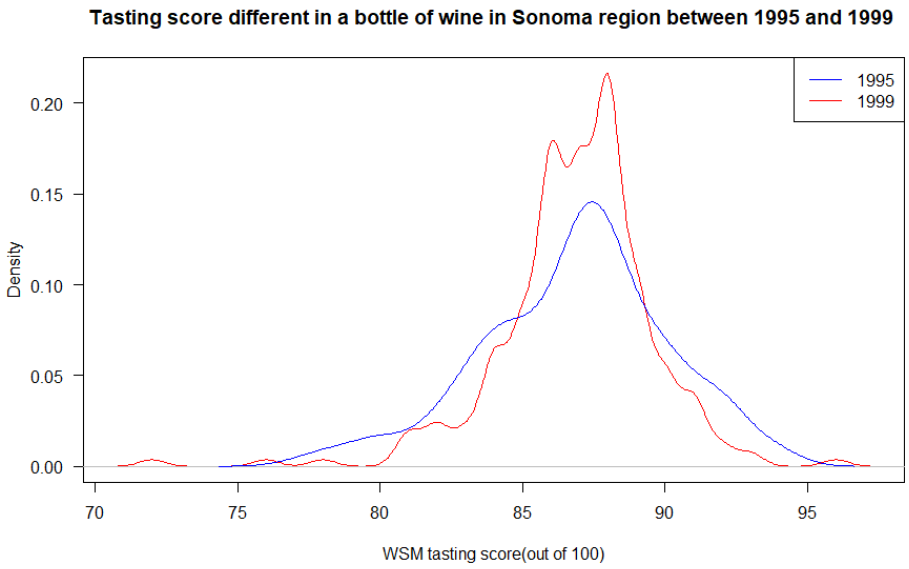

- The quality of wine in 1999 vintage in Napa wine region has an identical variance and mean to the quality of 1995 vintage wine in Napa region.
- Looking at the density plot for the quality of the 1999 and 1995 vintage from the Napa Valley wine region, we can see it is left-skewed.

c. Price different in Sonoma wine region between 1995 and 1999



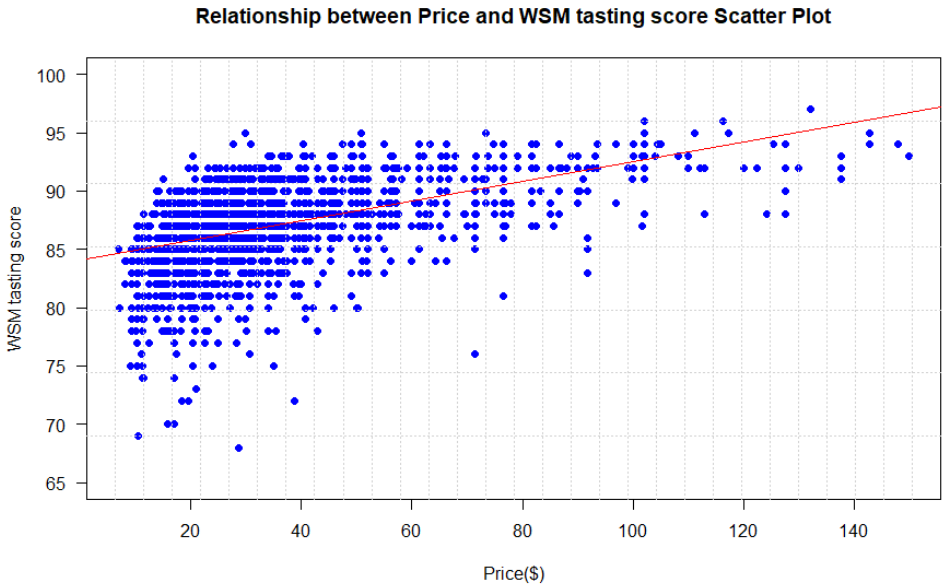

- The price variance and average prices of a typical bottle of wine produced in 1999 in Sonoma Valley has increased when compared in 1995 in Sonoma Valley.
- Looking at the density plot for the price of the 1999 and 1995 vintage from the Sonoma Valley wine region, we can see it is right-skewed.

d. Score different in Sonoma wine region between 1995 and 1999



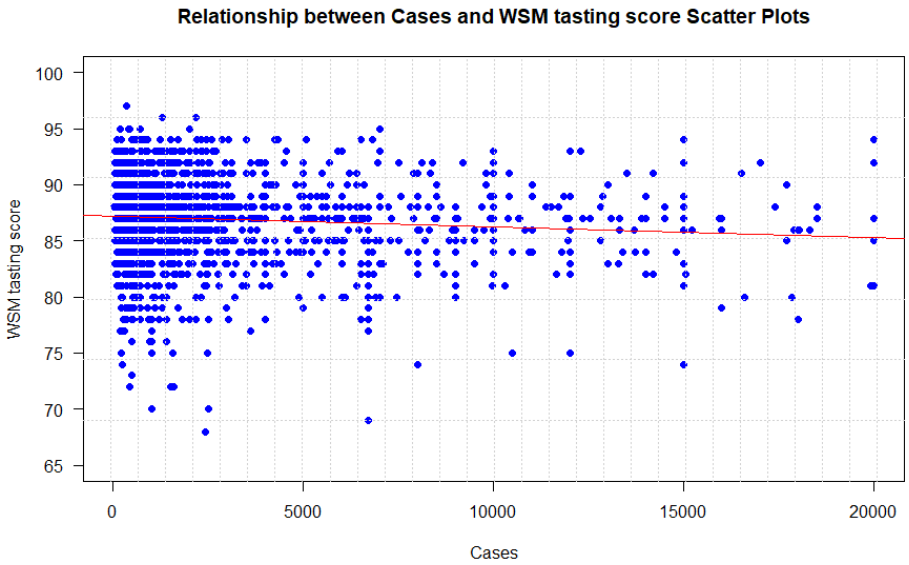

- The variance in the quality of the wine in Sonoma region has reduced significantly when comparing the 1999 vintage to the 1995 with more wine being concentrated around WMS tasting score of 85 to 90.
- The average quality of wine in Sonoma region have increased by a small amount.
- Looking at the density plot for the quality of the 1999 and 1995 vintage from the Napa Valley wine region, we can see it is left-skewed.

4. Plotting scatter plots
   a. The relationship between Price and WSM tasting score:




- The covariance and correlation between Price and the quality of the wine are 38.42 and 0.5018.
- Individual wine with a high WSM score will signal greater taste and higher quality control during production, thereby able to attract more wine enthusiasts enable the wine to be sold at a high price.

   b. The relationship between Quantity per production and WSM tasting score:




- The covariance and correlation between Quantity of wine per production and quality of wine are -1150.59 and -0.0321.
- There is an insignificant correlation between high-quality wine and a small amount of case per productions, indicating there are other factors beyond the quantity of wine per production that contribute to the high WSM tasting score.

5. Hypothesis testing using T-test for the following:
   a. Testing for prices different between Napa Valley region wine and non-Napa Valley region in California and Washington on the west coast of the United States. At 5% significant level.

| To test the hypotheses |
| --- |
| $H_0$: mean (price for Napa) = mean (price all other wine) |
| $H_1$: mean (price for Napa) ≠ mean (price all other wine) |

- The difference in the sample mean $\$47.40 - \$28.93 = \$18.47$
- The p-value for the test is less than $2.2\boldsymbol{e} - 16 \approx -10.01978$
- T-statistic for the test of difference in mean is 16.58 and |16.58| > 1.98; we can reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (16.28, 20.65), which indicate with 95% confident we cannot reject the price difference between Napa Valley region wine and non-Napa Valley region in California and Washington on the west coast of the United States is as low as $16.28 and high as $20.65.

   b. Testing for quality different between Napa Valley region wine and non-Napa Valley region in California and Washington on the west coast of the United States. At 5% significant level.

| To test the hypotheses |
| --- |
| $H_0$: mean (score for Napa) = mean (score all other wine) |
| $H_1$: mean (score for Napa) ≠ mean (score all other wine) |

- The difference in the sample mean $88.39 - 86.33 = 2.05$
- The p-value for the test is less than $2.2\boldsymbol{e} - 16 \approx -10.0978$
- T-statistic for the test of difference in mean is 12.17 and |11.17| > 1.98; we can reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (1.72, 2.38), which indicate with 95% confident we cannot reject the quality different between Napa Valley region wine and non-Napa Valley region in California and Washington on the west coast of the United States is as low as 1.72 and high as 2.38.

   c. Testing for prices different between Sonoma Valley region wine and non-Sonoma Valley region in California and Washington on the west coast of the United States. At 5% significant level.

| To test the hypotheses |
| --- |
| $H_0$: mean (price for Sonoma) = mean (price all other wine) |
| $H_1$: mean (price for Sonoma) ≠ mean (price all other wine) |

- The difference in the sample mean $\$31.68 - \$35.67 = \$3.99$
- The p-value for the test equal to $1.482\boldsymbol{e} - 5 \approx -0.9715$
- T-statistic for the test of difference in mean is -4.35 and |-4.35| > 1.98, we can reject the null hypotheses at 5% level significant that the difference mean is equals 0

- The 95% confidence interval is (-5.80, -2.19), which indicate with 95% confident we cannot reject the price difference between Sonoma Valley region wine and non-Sonoma Valley region in California and Washington on the west coast of the United States is as low as $-5.8 and high as $-2.19.

d. Testing for quality different between Sonoma Valley region wine and non-Sonoma Valley region in California and Washington on the west coast of the United States. At 5% significant level.

| To test the hypotheses |
| --- |
| $H_0$: mean (score for Sonoma) = mean (score all other wine) |
| $H_1$: mean (score for Sonoma) ≠ mean (score all other wine) |

- The difference in the sample mean $86.86 - 87 = 0.14$
- The p-value for the test equal to $0.3807$
- T-statistic for the test of difference in mean is -0.88 and |-0.88| < 1.98, we cannot reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (-0.46, 0.18), which indicate with 95% confidence we can reject the quality different between Sonoma Valley region wine and non-Sonoma Valley region in California and Washington on the west coast of the United States is as low as -0.46 and high as 0.18.

6. Hypothesis testing using T-test for the following:
   a. Testing for the price difference between the wine produced in 1995 and 1999 in the Napa Valley region. At 5% level significant.

| To test the hypotheses |
| --- |
| $H_0$: mean (price for Napa in 1999) = mean (price for Napa in 1995) |
| $H_1$: mean (price for Napa in 1999) ≠ mean (price for Napa in 1995) |

- The difference in the sample mean $\$51.78 - \$41.28 = \$10.50$
- The p-value for the test equal to $2.207\boldsymbol{e} - 07 \approx -1.0008$
- T-statistic for the test of difference in mean is 5.24 and |5.24| > 1.98, we can reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (6.56, 14.43), which indicate with 95% confident we cannot reject the price different in Napa Valley wine region between wine produced in 1999 and 1995 is as low as $6.56 and high as $14.43.

   b. Testing for the quality difference between the wine produced in 1995 and 1999 in the Napa Valley region. At 5% level significant.

| To test the hypotheses |
| --- |
| $H_0$: mean (score for Napa in 1999) = mean (score for Napa in 1995) |
| $H_1$: mean (score for Napa in 1999) ≠ mean (score for Napa in 1995) |

- The difference in the sample mean $88.36 - 88.42 = 0.06$
- The p-value for the test equal to $0.8309$

- T-statistic for the test of difference in mean is -0.21 and |-0.21| < 1.98, we cannot reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (-0.62, 0.50), which indicate with 95% confidence we can reject the quality difference in Napa Valley wine region between wine produced in 1999 and 1995 is as low as -0.62 and high as 0.50.

c. Testing for the price difference between the wine produced in 1995 and 1999 in the Sonoma Valley region. At 5% level significant.

| To test the hypotheses |
| --- |
| $H_0$: mean (price for Sonoma in 1999) = mean (price for Sonoma in 1995) |
| $H_1$: mean (price for Sonoma in 1999) ≠ mean (price for Sonoma in 1995) |

- The difference in the sample mean $\$34.81 - \$28.75 = \$6.06$
- The p-value for the test equal to $2.145\boldsymbol{e} - 05 \approx 0.8307$
- T-statistic for the test of difference in mean is 4.29 and |4.29| > 1.98, we can reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (3.29, 8.84), which indicate with 95% confident we cannot reject the price different in Sonoma Valley wine region between wine produced in 1999 and 1995 is as low as $3.29 and high as $8.84.

d. Testing for the quality difference between the wine produced in 1995 and 1999 in the Sonoma Valley region. At 5% level significant.

| To test the hypotheses |
| --- |
| $H_0$: mean (score for Sonoma in 1999) = mean (score for Sonoma in 1995) |
| $H_1$: mean (score for Sonoma in 1999) ≠ mean (score for Sonoma in 1995) |

- The difference in the sample mean $86.89 - 86.84 = 0.05$
- The p-value for the test is equal to $0.8632$
- T-statistic for the test of difference in mean is 0.17 and |0.17| < 1.98, we cannot reject the null hypotheses at 5% level significant that the difference mean is equals 0
- The 95% confidence interval is (-0.47, 0.56), which indicate with 95% confidence we can reject the quality difference in Napa Valley wine region between wine produced in 1999 and 1995 is as low as -0.47 and high as 0.56.

**Appendix**

```
#Pre Question
data=read.csv(file="as1_wine.csv") #open file


#defining data$name to a simple variable
price=(data$price)
cases=(data$cases)
score=(data$score)
napa=(data$napa)
sonoma=(data$sonoma)
d1995=(data$d1995)
d1999=(data$d1999)

#function for calculating the mean different between two data set
df_mean <- function(x,y){
  paste( mean(x), "-", mean(y),"=", abs(mean(x)-mean(y)))
}

# Q1.(3 marks) Report summary statistics for price, cases, score, napa, sonoma,d1995, d1999. Interpret each of the means in plain language, thereby characterising a typical wine in the dataset. Your answer should be no more than four sentences long.

#getting summary of data
summary(data) #Printing out summary data

#Getting Standard Deviation
sd(price)  #Standard Deviation for Price
sd(cases)  #Standard Deviation for Cases
sd(score)  #Standard Deviation for Score
sd(napa)   #Standard Deviation for Napa
sd(sonoma) #Standard Deviation for Sonoma
sd(d1995)  #Standard Deviation for d1995
sd(d1999)  #Standard Deviation for d1999


#Q2.(3 marks) Compute 99% confidence intervals for price, cases, score

#computing 99% confidence interval for price
x=mean(price)        # Sample mean of price
n=length(price)      # Number of sample(N)
std=sd(price)        # Sample standard deviation
```

```
m=std/sqrt(n)       # Marginal error of the sample mean
CI99_lower=x-2.58*m  # Lower bound of the 99% CI
CI99_upper=x+2.58*m  # Upper bound of the 99% CI


#Printing out 99% confidence interval result for price
paste("99% confidence interval for prices", "(" ,CI99_lower,  ",",CI99_upper, ")" )
#remove excess values
remove(x, n, std, m, CI99_upper, CI99_lower)


#computing 99% confidence interval for cases
x=mean(cases)       # Sample mean of price
n=length(cases)     # Number of sample(N)
std=sd(cases)       # Sample standard deviation
m=std/sqrt(n)       # Marginal error of the sample mean
CI99_lower=x-2.58*m  # Lower bound of the 99% CI
CI99_upper=x+2.58*m  # Upper bound of the 99% CI


#Printing out 99% confidence interval result for cases
paste("99% confidence interval for cases", "(" ,CI99_lower,  ",",CI99_upper, ")" )
#remove excess values
remove(x, n, std, m, CI99_upper, CI99_lower)


#computing 99% confidence interval for score
x=mean(score)       # Sample mean of price
n=length(score)     # Number of sample(N)
std=sd(score)       # Sample standard deviation
m=std/sqrt(n)       # Marginal error of the sample mean
CI99_lower=x-2.58*m  # Lower bound of the 99% CI
CI99_upper=x+2.58*m  # Upper bound of the 99% CI


#Printing out 99% confidence interval result for score
paste("99% confidence interval for score", "(" ,CI99_lower,  ",",CI99_upper, ")" )
#remove excess values
remove(x, n, std, m, CI99_upper, CI99_lower)


#3. Graphing different between the napa wine region and sonoma wine region

#Plotting graph for prices
plot(density(price[sonoma==1]), col="blue",lty=1,
```

```
    main="Price different between a bottle of wine in Napa region and Sonoma region",
    xlab="Price($)",las=1)
lines(density(price[napa==1]), col="red", lty=1)
legend("topright", legend=c("Napa region", "Sonoma Region"),
      col=c("red","blue"), lty=c(1,1))


#Plotting graph for cases
plot(density(cases[sonoma==1]), col="blue",lty=1,
    main="Production different between Napa region and Sonoma region",
    xlab="Case", las=1)
lines(density(cases[napa==1]), col="red", lty=1)
legend("topright", legend=c("Napa region", "Sonoma Region"),
      col=c("red","blue"), lty=c(1,1))

#Plotting graph for score
plot(density(score[sonoma==1]), col="blue",lty=1,
    main="Tasting score between a bottle of wine in Napa region and Sonoma region",
    xlab="WSM tasting score(out of 100)", las=1)
lines(density(score[napa==1]), col="red", lty=1)
legend("topright", legend=c("Napa region", "Sonoma Region"),
      col=c("red","blue"), lty=c(1,1))

#4 Graphing the different between wine price and quality produced 1995 and 1999

#Plotting graph for prices in napa valley between 1995 and 1999
plot(density(price[napa==1  & d1995==1]), col="blue",lty=1,main="Price different in a bottle of wine in Napa region between 1995 and 1999 ", xlab="Price($)")
lines(density(price[napa==1 & d1999==1]), col="red", lty=1)
legend("topright", legend=c("1995", "1999"),
      col=c("blue","red"), lty=c(1,1))

#Plotitng graph for prices in napa valley between 1995 and 1999
plot(density(score[napa==1  & d1995==1]), col="blue",lty=1,
    main="Tasting score different in a bottle of wine in Napa region between 1995 and 1999 ",
    xlab="WSM tasting score(out of 100)", las=1)
lines(density(score[napa==1 & d1999==1]), col="red", lty=1)
legend("topright", legend=c("1995", "1999"),
      col=c("blue","red"), lty=c(1,1))

#Plotting graph for prices in sonoma valley between 1995 and 1999
plot(density(price[sonoma==1  & d1995==1]), col="blue",lty=1,
```

```
    main="Price different in a bottle of wine in Sonoma region between 1995 and
1999 ",
    xlab="Price($)", las=1)
lines(density(price[sonoma==1 & d1999==1]), col="red", lty=1)
legend("topright", legend=c("1995", "1999"),
      col=c("blue","red"), lty=c(1,1))

#Plotting graph for score in sonoma valley between 1995 and 1999
plot(density(score[sonoma==1  & d1999==1]), col="red",lty=1,
    main="Tasting score different in a bottle of wine in Sonoma region between 1995
and 1999 ",
    xlab="WSM tasting score(out of 100)", las=1)
lines(density(score[sonoma==1 & d1995==1]), col="blue", lty=1)
legend("topright", legend=c("1995", "1999"),
      col=c("blue","red"), lty=c(1,1))


#5 Plotting scatter plot

# score on the vertical axis and price on the horizontal axis
x = price
y = score

plot(x, y,
    main="Relationship between Price and WSM tasting score Scatter Plot",
    xlab= "Price($)", ylab="WSM tasting score",
    pch=19, las=1, ylim=c(65,100), col="blue")
grid(nx=30, ny=7, col="lightgray", lty = "dotted")
abline(lm(y ~ x, data = data), col = "red", lwd=1) #linear regression line
paste("Covariance=", cov(x,y)) #printing out covariance between price and score
paste("Correlation=", cor(x,y)) #printing out correlation between price and score
remove(x, y)

# score on the vertical axis and cases on the horizontal axis
x = cases
y = score

plot(x, y,
    main="Relationship between Cases and WSM tasting score Scatter Plots",
    xlab= "Cases", ylab="WSM tasting score",
    pch=19, las=1, ylim=c(65,100), col="blue")
grid(nx=30, ny=7, col="lightgray", lty = "dotted")
abline(lm(y ~ x, data = data), col = "red", lwd=1) #linear regression line
paste("Covariance=", cov(x,y)) #printing out covariance between cases and score
paste("Correlation=", cor(x,y)) #printing out correlation between cases and score
```

```
remove(x, y)


#6 Testing the following

#Test 6a T-test mean price different between napa and overall price
t1=price[napa==1]
t2=price[napa==0]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#Test 6b T-test mean score different between napa overall score
t1=score[napa==1]
t2=score[napa==0]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#Test 6c T-test mean price different between sonoma and overall price
t1=price[sonoma==1]
t2=price[sonoma==0]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#Test 6d T-test average score different between sonoma and overall score
t1=score[sonoma==1]
t2=score[sonoma==0]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#7 Testing the following


#7a T test for Nappa mean price different between 1995 and 1999
t1=price[napa==1 & d1999==1]
t2=price[napa==1 & d1995==1]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#7b T-test for Nappa mean score different between 1995 and 1999
t1=score[napa==1 & d1999==1]
t2=score[napa==1 & d1995==1]
```

```
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#7c T-test for Sonoma mean price different between 1995 and 1999
t1=price[sonoma==1 & d1999==1]
t2=price[sonoma==1 & d1995==1]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable

#7d T-test for sonoma mean score different between 1995 and 1999
t1=score[sonoma==1 & d1999==1]
t2=score[sonoma==1 & d1995==1]
df_mean(t1, t2)   #Calculating the mean different using user-made defined previously
t.test(t1, t2)     #two tailed t-test for different in sample mean of 0
remove(t1, t2)     #remove excess variable
```