Name:	**Kim Seang CHY**
Student No:	**998008**
Tutorial Day/Time:	**Thursday 10am**

Question 1:

a. State appropriate null and alternative hypotheses.

We can start by assuming the proportion of people who support bicycle lane does not differ between the cities and we want to look for evidence that might suggest the two proportions differ between the two cities.

$H_0: p_1 = p_2$ vs $H_1: p_1 \neq p_2$

b. Carry out a test that has $\alpha = 0.05$.

```
#Entering the sample data with y being the number of respondents, and n
being the sample size
y <- c(520, 600)
n <- c(800, 1000)

#Two-sided proportion test and produce a confidence level of 0.95
prop.test(y, n, alternative = "two.sided", correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  y out of n
## X-squared = 4.7269, df = 1, p-value = 0.02969
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.005118323 0.094881677
## sample estimates:
## prop 1 prop 2
##   0.65   0.60
```

Reading from proportion test summary, the p-value is 0.02969, which is lower than the significant level of 0.05. Implying there is **a significant evidence to reject null** indicating it is **highly likely** that the proportion of people who support cycle lane in City of Yarra was different from the proportion from the City of Moreland.

c. Give a 95% confidence interval for the difference in proportion.

Reading from the proportion test summary in part b, the 95% confidence interval for the differences in the proportion between the two cities is $(0.0051, 0.0949)$.

Question 2:

a. Calculate 95% confidence interval.

Let X = {12.1, 12.2, 17.4, 13.1, 17.8, 19.8, 13, 10.8, 18.4, 16}

A normal distribution with an unknow variance and small sample can be approximated using a student T distribution.

$$\frac{\bar{X}-\mu}{S_x/\sqrt{n}} \sim t_{n-1}$$

Where $\bar{X}, S_x$ is the sample mean and sample standard deviation.

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = \frac{150.6}{10} = 15.06$$

$$S_X^2 = \frac{1}{10-1}\sum_{i=1}^{10}(X_i - \bar{X})^2 = \frac{90.664}{9} = 10.0738$$
$$\Rightarrow S_X = 3.1739$$

The 95% confidence interval range

$$\phi^{-1}(0.025) < \frac{15.06-\mu}{\frac{3.1739}{\sqrt{10}}} < \phi^{-1}(0.975)$$

Where $\phi(t)$ is student T distribution with 9 degree of freedom.

$$15.06 - 1.0337 \cdot \phi^{-1}(0.975) < \mu < 15.06 - 1.0337 \cdot \phi^{-1}(0.025)$$
$$\Rightarrow 15.06 - 1.0337 \cdot 2.26 < \mu < 15.06 + 1.0337 \cdot 2.26$$
$$\Rightarrow 12.72 < \mu < 17.40$$

The 95% confidence interval for mean is (12.75, 17.40)

b. n for 95% confidence interval with known variance $\sigma_x^2 = 9$ to get width of 2.

For a normal distribution with known variance is $X \sim N\left(\bar{X}, \frac{\sigma_x^2}{n}\right) = N\left(15.06, \frac{9}{n}\right)$. Therefore, by symmetry the width of 95% confidence interval is given by:

$$\frac{3}{\sqrt{n}}\phi^{-1}(0.975) = \frac{3}{\sqrt{n}}\,1.96$$

Where $\phi(t)$ is the standard normal distribution.

Solve for n such that $1.96 \cdot \frac{3}{\sqrt{n}} = 1$ and we get an n of 34.57.

Therefore, the minimum number of observations required to get a width of least 2 is 35 observations.

c. Calculate 95% confidence interval

The distances from Queen Victoria market to Peter Hall is similar from state library to Peter Hall. However, similar distance may indicate same population mean, but it does not necessary indicate the same population variance. Therefore, we will assume the variance is differences for safe purpose.

Let Y = {20.1, 21.3, 20.4, 21.7, 20.3, 19.5, 19.4, 19.9}

$$\bar{Y} = \frac{1}{8}\sum_{i=1}^{8} Y_i = \frac{162.6}{8} = 20.325$$

$$S_Y^2 = \frac{1}{8-1}\sum_{i=1}^{8}(Y_i - \bar{Y})^2$$

$$= \frac{1}{7}\sum_{i=1}^{8}(Y_i - 20.325)^2 = \frac{4.615}{7} = 0.6593$$

$$S_y = 0.8120$$

For comparing two mean with difference variance can be done using Welch's approximation with the degree of freedom as:

$$df = \frac{\left(\frac{S_X^2}{10} + \frac{S_Y^2}{8}\right)^2}{\frac{S_X^4}{10^2(10-1)} + \frac{S_Y^4}{8^2(8-1)}}$$

$$= \frac{\left(\frac{10.07}{10} + \frac{0.6593}{8}\right)^2}{\frac{101.48}{900} + \frac{0.4347}{448}} = 10.4357$$

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{10} + \frac{S_Y^2}{8}}} \sim t_{10.8571}$$

$$\Rightarrow \frac{15.06 - 20.325 - (\mu_X - \mu_Y)}{\sqrt{\frac{10.07}{10} + \frac{0.6593}{8}}}$$

Therefore 95% confidence interval of $\mu_X - \mu_Y$ is given by:

$$-5.265 - \cdot\, \omega^{-1}(0.975) < \mu_X - \mu_Y < -5.265 - \cdot\, \omega^{-1}(0.025)$$

Where $\omega(t)$ is student-T distribution with 10.8571 degree of freedom

$$-5.265 - \cdot\ \omega^{-1}(0.975) < \mu_X - \mu_Y < -5.265 - \cdot\ \omega^{-1}(0.025)$$
$$\Rightarrow -5.265 - 1.0437 \cdot 2.2155 < \mu_X - \mu_Y < -5.265 + 1.0437 \cdot 2.2155$$
$$\Rightarrow -7.5775 < \mu_X - \mu_Y < -2.9525$$

d. 95% confidence interval for $\frac{\sigma_{X^2}}{\sigma_Y^2}$

For comparing sample variance, we can use F-distribution

$\frac{S_y^2/\sigma_y^2}{S_X^2/\sigma_X^2} \sim F_{m-1,n-1}$ where m is sample size for Y and n is the sample size for X.

$$\Rightarrow \frac{\sigma_X^2}{\sigma_y^2} \sim \frac{10.07}{0.6593}\ F_{7,9}$$

The 95% confidence interval for $\frac{\sigma_X^2}{\sigma_y^2}$ is given by:

$$\frac{10.07}{0.6593}\ F_{7,9}^{-1}(0.025) < \frac{\sigma_X^2}{\sigma_y^2} < \frac{10.07}{0.6593}\ F_{7,9}^{-1}(0.975)$$
$$\Rightarrow 15.27 \cdot 0.20733 < \frac{\sigma_X^2}{\sigma_y^2} < 15.27 \cdot 4.1970$$
$$\Rightarrow 3.17 < \frac{\sigma_X^2}{\sigma_y^2} < 64.13$$

e. 95% confidence interval for $\frac{\sigma_{X^2}}{\sigma_Y^2}$ using R

```r
#Conducting a hypothesis testing for comparing variance and it 95% CI
x <- c(12.1, 12.2, 17.4, 13.1, 17.8, 19.8, 13, 10.8, 18.4, 16)
y <- c(20.1, 21.3, 20.4, 21.7, 20.3 ,19.5, 19.4, 19.9)
var.test(x, y, alternative = "two.sided")

##
##  F test to compare two variances
##
## data:  x and y
## F = 15.28, num df = 9, denom df = 7, p-value = 0.00163
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   3.167976 64.130185
## sample estimates:
## ratio of variances
##           15.27984
```

Question 3:

a. Fit the sample and state the estimate

```
coffee <- read.csv("coffee.csv") #Opening up data
customer <- coffee$customer # Redefining varible
sale <- coffee$sales #Redefining varible

#Fit the model
model1 <- lm(sales ~ customer, data = coffee)

# Show result
summary(model1)

##
## Call:
## lm(formula = sales ~ customer, data = coffee)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116.51  -47.18    1.25   36.21  136.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.3442    62.2027   -0.52    0.609
## customer      6.4005     0.6345   10.09 7.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.6 on 18 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8413
## F-statistic: 101.7 on 1 and 18 DF,  p-value: 7.816e-09
```

The model has a beta hat value of 6.4 and alpha.hat value of -32.34. The standard error and p-value for beta.hat is 0.6345 and $7.82 \cdot 10^{-9}$, indicating that beta.hat value is statistically significant. The standard error, p-value for alpha.hat is 62.2027, 0.609, indicating that it is not statistically significant. The multiple r-squared is 0.8497, indicating that 84.97% of the sale is explained by the number of customers.

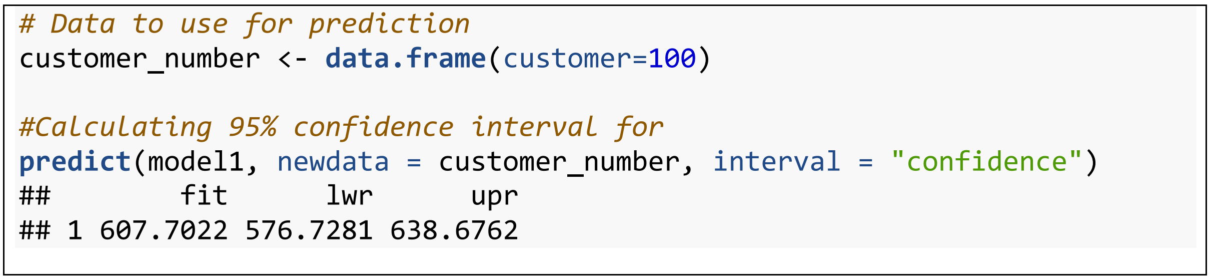
The sigma.hat is given by the residual standard error which is 64.6.

b. 95% confidence interval for regression coefficients.

```
#finding 95% confidence interval for regression coefficient
confint(model1)

##                   2.5 %    97.5 %
## (Intercept) -163.027250 98.338933
## customer       5.067368  7.733558
```

The 95% confidence interval for apha.hat is (-163.03, 98.34). The 95% confidence interval for beta.hat is (5.07, 7.73).

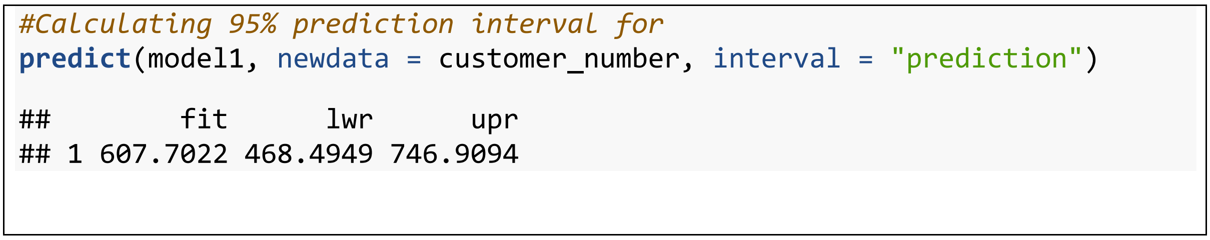
c. 95% confidence interval for mean of sale if customer equal 100.



```
# Data to use for prediction
customer_number <- data.frame(customer=100)

#Calculating 95% confidence interval for
predict(model1, newdata = customer_number, interval = "confidence")
##        fit      lwr      upr
## 1 607.7022 576.7281 638.6762
```

The 95% confidence interval for mean of sale if customer equal 100 is (576.73, 638.68)

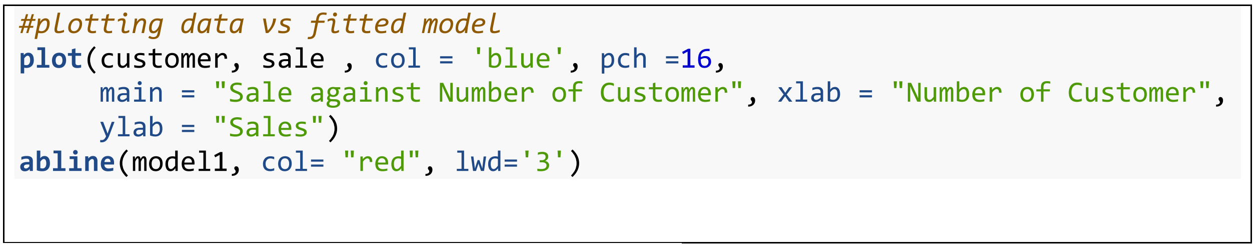
d. 95% prediction interval for mean of sale if customer equal 100.



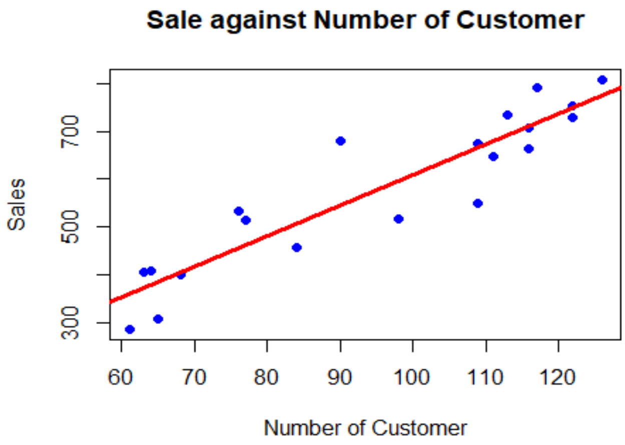
```
#Calculating 95% prediction interval for
predict(model1, newdata = customer_number, interval = "prediction")

##        fit      lwr      upr
## 1 607.7022 468.4949 746.9094
```

The 95% prediction interval for mean of sale if customer equal 100 is (468.49, 746.91).

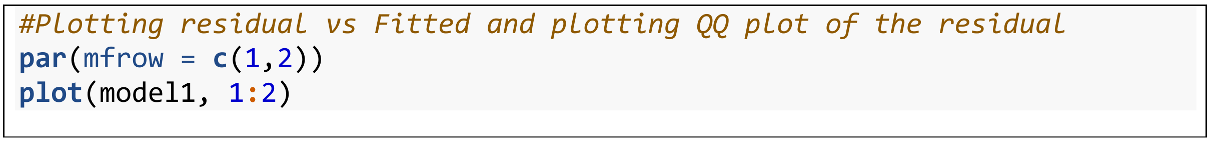
e. Regression assumption



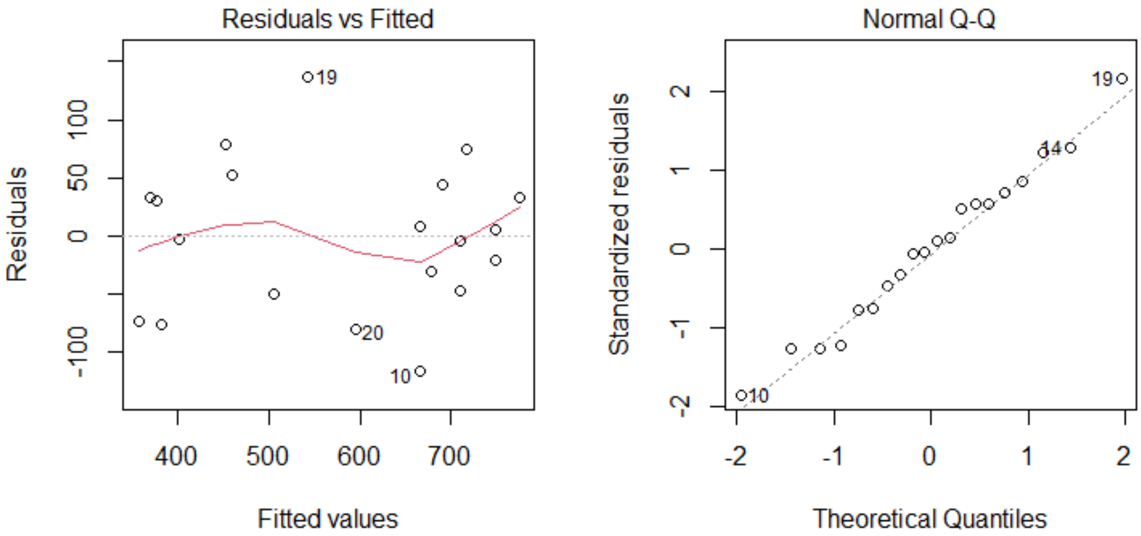
```
#plotting data vs fitted model
plot(customer, sale , col = 'blue', pch =16,
     main = "Sale against Number of Customer", xlab = "Number of Customer",
     ylab = "Sales")
abline(model1, col= "red", lwd='3')
```



Looking at data and fitted model the regression is a linear model, thus meeting assumption of linear model for the mean.



```
#Plotting residual vs Fitted and plotting QQ plot of the residual
par(mfrow = c(1,2))
plot(model1, 1:2)
```



Looking at the variability in residuals vs fitted, it is looked to be random not showing any pattern, thus the assumption of homoscedasticity will still hold. Similarly looking at the QQ plot, the normal distribution is a very good fit for the residual value, thus final assumption of the residual being normally distributed should hold.

Question 4:

For the parameter to be a pivot it must be independence from unknow parameters.

a. T_1

$$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$$

The parameter does depend on the unknow parameter σ, therefore it is not a pivot.

b. T_2

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Similar part a, the parameter depends on σ, thus it is not a pivot.

c. T_3

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

The parameter is independence of the two unknow parameters, thus it is a pivot.

d. T_4

$$\frac{\bar{X} - \mu}{S} \sim \frac{t_{n-1}}{\sqrt{n}}$$

Similar to part c, the parameter is independence of two unknown parameter, thus it is a pivot.

Question 5:

a. Type I Error

A type I error, refer to the error of rejecting the null when the null is true, which is given by

$$P(x > 4|\theta = 2) = \int_4^\infty \frac{e^{\frac{-x}{2}}}{2} dx$$

$$= e^{-2} = 0.1353$$

b. Type II Error

A type II error refer to the error of not rejecting the null when the null is false.

$$P(x \leq 4|\theta = 5) \int_0^4 \frac{e^{\frac{-x}{5}}}{5} dx$$

$$= 1 - e^{\frac{-4}{5}} = 0.5507$$

c. Power of the test

The power of the test is 1 – Type II error, which is given by

$$1 - \left(1 - e^{-\frac{4}{5}}\right) = e^{-\frac{4}{5}}$$

$$= 0.4493$$

d. Find a test

A single exponential distribution is just a Gamma distribution with n=1. Using this fact and the know pivot of gamma distribution with Chi-square distribution

$$2n\theta^{-1}\bar{X} \sim \chi_{2n}^2$$

$$\Rightarrow \frac{2}{2}\bar{X} \sim \chi_2^2$$

The hypothesis is:

$$H_0: \theta = 2 \; vs \; H_1: \theta > 2$$

For a critical value of 0.05, we reject the null if and only if $\bar{X} < c$.

$$\Pr(\bar{X} < c| \, H_0) = 0.05$$

$$\Rightarrow c = \omega^{-1}(0.05)$$

Where ω^{-1} is the inverse cdf of χ_2^2.

$$\omega^{-1}(0.05) = 0.1026$$

Therefore, we will reject the mean if the sample mean is less than 0.1026 for significant level at 0.05.