UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING

Faculty of Engineering

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

Project 2

SECTION : 04 – 1SECR

COURSE NAME : BACHELOR OF COMPUTER SCIENCE – COMPUTER NETWORKS & SECURITY

NO.	NAME	STUDENT ID
1	MUHAMMAD RAFIQ REDHA BIN RUSHIDI	A19EC0106

LECTURER'S NAME : DR. SUHAILA BINTI MOHAMAD YUSUF

DATE OF SUBMISSION: 27th June 2020

TABLE OF CONTENT

INTRODUCTION

The Olympics is one of the largest sports tournaments in the world where the tournament is played only once every 4 years. Selecting a host for this tournament is very random and will only be determined by the International Olympic Committee. The athletes competing are not ordinary where they will be training to win medals for their country. Therefore, I will do a little research on them where the data set is based on the weight, height, age and medals obtained by the Olympics athletes. This data set contains a list of athletes' information and I only select 100 athletes of which 50 are men and 50 are women. This study is to look at weight, height and age for athletes only. The purpose of this study was to see if athletes' weight gain was influenced by age factors.

STATISTICAL TEST ANALYSIS

Hypothesis testing 1-samples

$$H_1 : \mu = 80$$

$$H_2 : \mu \neq 80$$

```
x = Athlete$weight
mean(x)
y = Athlete$weight
sd(y)

z = (mean(x)-80)/(sd(y)/sqrt(100))
alpha=0.05
z.alpha = qnorm(1-alpha/2)

c(-z.alpha, z.alpha)
```

By using the Rstudio and the coding in above, the samples of weight are obtained by the excel file. Since only 1 samples, the allowed value of alternative hypothesis is one of the two sides which not equal while assuming the variances and the pair is not equal with 0.95 confidence level.

```
> x = Athlete$weight
> mean(x)
[1] 70.07
> y = Athlete$weight
> sd(y)
[1] 13.06337
> z = (mean(x)-80)/(sd(y)/sqrt(100))
> alpha=0.05
> z.alpha = qnorm(1-alpha/2)
> c(-z.alpha, z.alpha)
[1] -1.959964  1.959964
```

z	-7.60140803645023

$\bar{x}$ = 70.07

Degree of freedom, df = 99

$criticalvalue_{(0.025,99)}$ = -1.959964

- $criticalvalue_{(0.025,99)}$ = 1.959964

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Test statistic, z = -7.601408

P-value = 0.0001

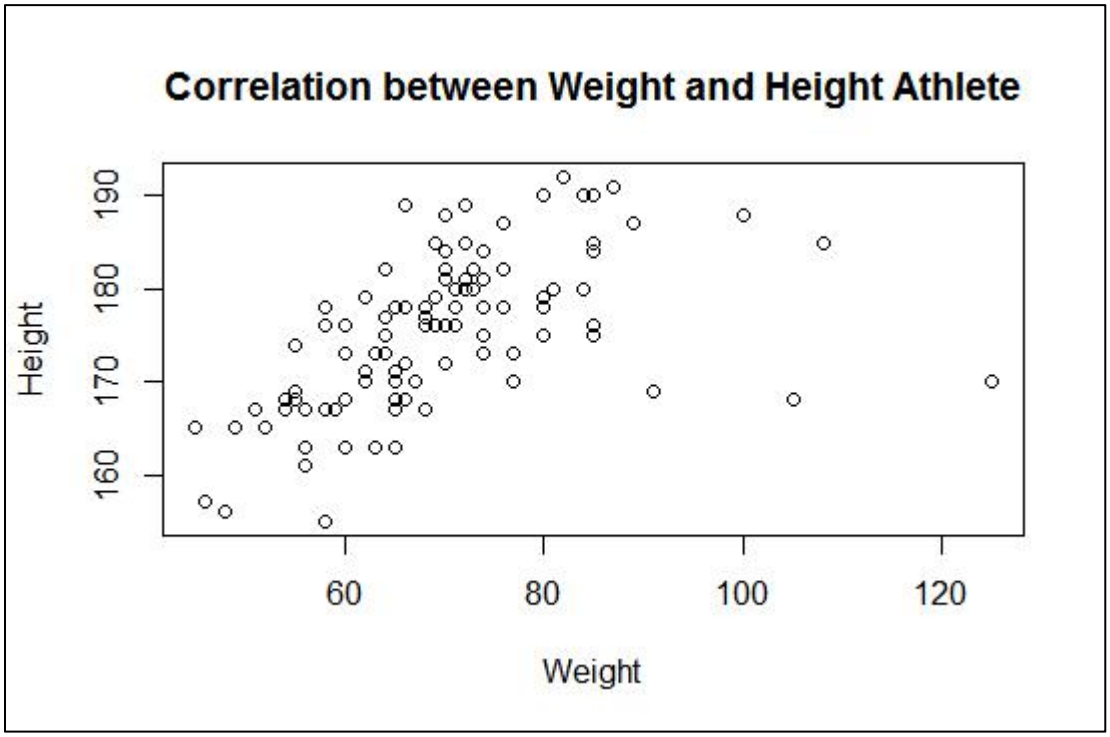
Decision : Since p-value is 0.0001 is greater than test statistic,z (-7.601408). We Fail to rejected null hypothesis.

Conclusion : At 95% confidence level. There is sufficient evidence to conclude that $\mu = 80$. There results suggest that the mean weight of athlete is 80.

Correlation

```
#correlation between Weight and Height Athlete
x = Athlete$Weight
y = Athlete$Height
plot(x,y,main='Correlation between Weight and Height Athlete',xlab = 'Weight', ylab = 'Height')
cor.test(x,y)
```

Based on this study, this correlation is about relationship between Weight and Height athlete. From the Rstudio, The dataset of Weight and Height athlete is obtain from file “Athlete.xlsx”, where variable x is for athlete's weight and variable y is athlete's height. Next, the distribution diagram will be compiled and cor.test works to calculate the correlation between Weight and Height athletes.




```
        Pearson's product-moment correlation

data:  x and y
t = 6.1098, df = 98, p-value = 2.023e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3666073 0.6541430
sample estimates:
      cor
0.5252062
```

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$H_1 : p = 0$ (No linear correlation)

$H_2 : p \neq 0$ (Linear correlation exists)

r = 0.5252062 t = 6.1098

Degree of freedom = 98 $\alpha = 0.05$

p-value = 2.023×10^{-8}

Decision : Since p-value is 2.023×10^{-8} is less than significance level (0.05). The null hypothesis (H0) will be rejected.

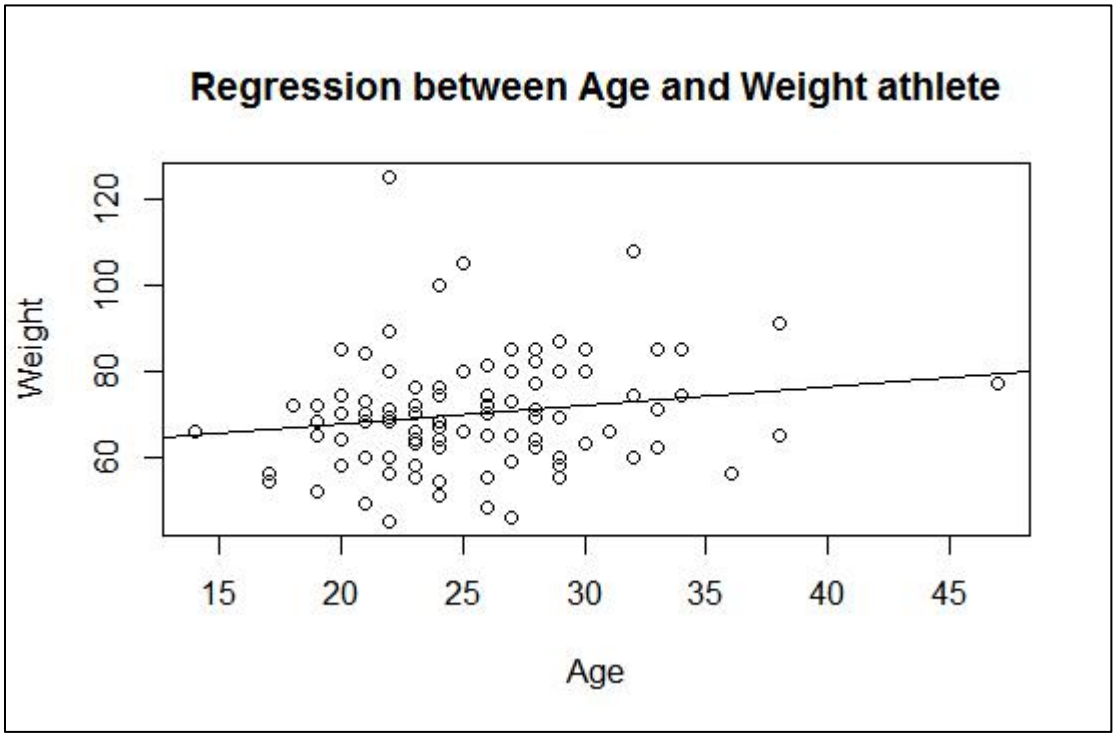
Conclusion : There is sufficient evidence at 95% confidence level that there is correlation between weight and height of athletes.

Based on graph scatter plot of correlation between weight and height athlete, the data shows a the pattern goes from bottom to top. The value of r is always and must between +1 and -1, the value of r is 0.5252062. This show the correlation is moderate positive linear relationship. This shows that there is a positive relationship between weight and height of an athlete where if the weight of an athlete increases, their height will increase. And it can be proven that this relationship has a correlation with p-value is 2.023×10^{-8} is less that significance level (0.05).

Regression

```
#regression between age and weight athlete
x = Athlete$Age
y = Athlete$Weight
dat =lm(y~x)
dat
summary(dat)
plot(x,y,main='Regression between Age and Weight athlete',xlab = 'Age',ylab = 'Weight')
abline(dat)
```

Based on this study, this regression is about relationship between Age and Weight athlete. From the Rstudio, The dataset of Age and Weight athlete is obtain from file "Athlete.xlsx", where variable x is for athlete's age and variable y is athlete's weight. The linear model function is used to create a simple regression model graph and is then stored in the dat variable. The summary function is used to produce the result of the regression model graph. Then, the graph will plotted with abline function that can add the regression line on the graph.



```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    59.2443       0.4274

> summary(dat)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-24.784  -8.752  -0.219   6.281  56.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.2443     6.5664   9.022  1.6e-14 ***
x             0.4274     0.2541   1.682   0.0958 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.94 on 98 degrees of freedom
Multiple R-squared:  0.02805,   Adjusted R-squared:  0.01813
F-statistic: 2.828 on 1 and 98 DF,  p-value: 0.09582
```

$H0 : \beta 1 = 0$ p-value (intercept) = 1.6×10^{-14}

$H1 : \beta 1 \neq 0$ p-value (slope) = 0.0958

$$\hat{y}_i = b_0 + b_1 x$$

$$\hat{y} = 59.2443 - 0.4274x$$

Decision : Since p-valuesof intercept is less than significance level, α = 0.05, the null hypothesis(H0) is to be rejected.

Conclusion : There is sufficient evidence at 95% confidence level that the weight of athlete will be affected because their age.

Based on the graph regression between Age and weight of athletes as their age increases, their weight will increase as well. The large p-values of intercept, 1.6×10^{-14} and slope,0.0958 indicates that the null hypothesis is rejected which means there is have relationship between the age and weight of athletes.

Anova

```
Var1=aov(Weight~Team)
Var1
summary(Var1)
```

```
> plot(TukeyHSD(Var1), las=1)
> Var1=aov(Weight~Team)
> Var1
Call:
   aov(formula = Weight ~ Team)

Terms:
                    Team Residuals
Sum of Squares  8022.54   8871.97
Deg. of Freedom      46        53

Residual standard error: 12.93815
Estimated effects may be unbalanced
```

Based on this study, this Anova is about relationship between weight and team of athlete. From our study, we were only taught to use one-way analysis of variance (ANOVA). The one- way analysis of variance (ANOVA) is an extension of independent two sample t-test for comparing means in a situation where there are more than two groups or types. Using Rstudio, I test and code the weight of their athletes and their teams, their country.

```
> summary(Var1)
            Df Sum Sq Mean Sq F value Pr(>F)
Team        46   8023   174.4   1.042   0.44
Residuals   53   8872   167.4
```

$H0$: The mean of weight of athletes is same for every team

$H1$: At least one mean is different

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{ns_{\bar{x}}^2}{s_p^2}$$

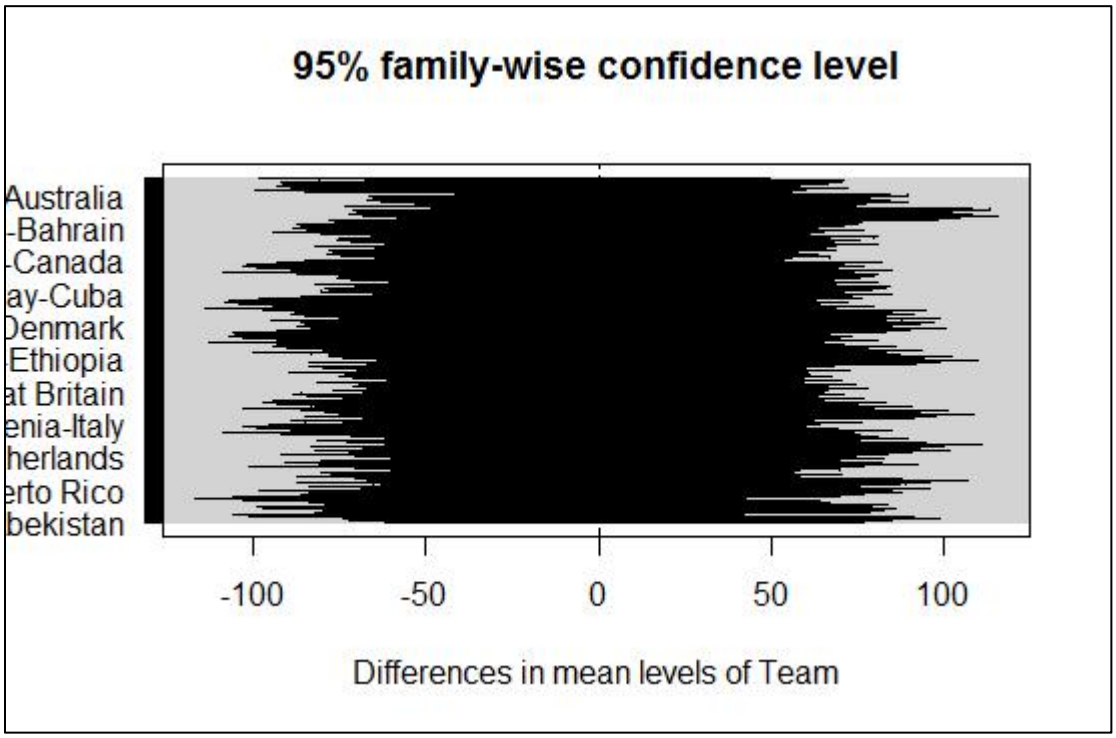
F value= 1.042 p-value= 0.44 $\alpha = 0.05$

Discussion : Since p-value is 0.44 is greater than signifincance level is 0.05. We fail to rejected the null hypothesis.

Conclusion : There is sufficient evidence to claim that the mean of weight athletes is same for every team.

```
plot(TukeyHSD(var1), las=1)
```

From the table of Tukey multiple comparisons of means 95% family-wise confidence level, We can conclude that there is a significant that the mean of weight athletes is same for every team because the p-values is greater than the significance level, $\alpha = 0.05$. By plotting TUKEYHSD(VAR) using Rstudio, I can visualize and analysis the significant difference. From the graph below, we can see that nothing is significant different because all the lines for each pair cross zero values.



DISCUSION

First of all, the mean weight for athletes is 70.07. This mean is the mean of the sample mean where it is found that the mean of the athletes for the Olympics is 70.07. Using 1-sample hypothesis test we conclude that mean of atletes is equal to 80 kg.

Next, the correlation between weight and age athlete are tested using Pearson's product-moment technique. Based on calculation using Rstudio, the result of correlation coefficient r obtained is 0.5252062, which is a moderate positive correlation linear relationship between the two variables it is weight and height. Therefore, if the athlete's weight increases, then the athlete's height is also high.

Moreover, the main purpose of this study this regression is about relation betweeen age and weight of athlete where this is dependent variable. Besides, the age of athlete is affected to athlete gain their weight. As an outcome, there is have regression relationship between both variables, and both variable is dependent.

Lastly, I conducted analysis of variance (ANOVA) to compare the mean weight of athletes separated by team. The result can conlude that the mean weight of the athletes is the same for every team. This shows that the mean of each athlete for each team is the same because what the Olympics contend for is a sport that should have a fit body shape to avoid being easily injured.

CONCLUSION

After all the insights I got from the statistical analysis tests I did, I can conclude that weight gain for athletes is due to their age factors. This is also due to the fact that they are undergoing training due to age factors and are not too late to engage in strenuous activity. In addition, the weight of athletes increases due to uncontrolled dietary control which causes them to lose weight as they age. At the same time, the weight of the athletes in each country is the same as this shows that the average athlete competing in the Omlipics is the same weight.

If we are a disciplined athlete then increasing age is not a barrier for us to stay fit and energetic and be able to control our diet and regularly set daily routines to maintain our existing weight. In addition, to maintain the weight of each athlete, sports management should play a role where they must ensure that the weight of the athletes is within a healthy environment and remain fit to prepare for the international sporting event.