UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROJECT 2 (Group)

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 06

SEMESTER II, SESSION 2019/2020

Lecturer: Dr. Izyan Izzati Kamsani

Team Name: Afflatus	
Name	Matric No.
Wong Hui Shi	A20EC0169
Soh Zen Ren	A20EC0152
Teoh Wei Jian	A20EC0229

Content

1.0 Introduction

The purpose of this study is to find out the mean of the worldwide happiness score and the factors that affect the happiness score. The research team is interested in this question because the result of this study can give a simple idea about which country is more happy and better for students to choose when they want to work overseas. At the end of this study, the result will show what are the factors that will affect the country's happiness score and what country's happiness score can be considered as a blessed country compared to other countries with lower happiness scores.

2.0 DataSet

All the data used for analysis in this project is collected from a secondary source, website Kaggle(*World Happiness Report*, 2019). Many data can be found on Kaggle and the data that has been chosen is World Happiness Report 2015. The World Happiness Report is a landmark survey about national happiness. The statistics are presented in the table. There are 158 samples in the data set. Among those samples, there are a few types of variables selected for analysis such as nominal, ordinal, ratio and interval. And those variables include country, region, happiness rank, happiness score, standard error, economy (GDP per capita), family, health (life expectancy), freedom, trust (government corruption), generosity and dystopia residual. 4 variables including region, happiness score, economy (GDP per capita and family are used in this study. The reason to choose these four variables is to investigate the relationship among the region, economy (GDP per capita, family and happiness score and how the happiness score has been affected by those variables. The data used is based on the statistics in 2015. The statistical test analysis related to the variables chosen are hypothesis test for 2 samples, correlation and regression which are compulsory tests, chi square and Anova which are optional tests. Many mediums such as R studio, SPSS and excel can be used for analysis of data. R studio is chosen and used in this project for the purpose to test and analyze the data. Besides, some of the findings are presented using the graph.

3.0 Data Analysis

3.1 Hypothesis Testing 1 Samples

From the report of sustainable development solutions network, we can know that the average 2015 world happiness score is 5.1 (Zhang, 2015).By using hypothesis testing 1 sample, we can check whether the mean of world happiness score in our data set is equal to or higher than 5.1. The population variances of the data sample are unknown. We assume that the confidence level is 95%.

Statement: the mean worldwide happiness score 2015 is greater than 5.1.
The null hypothesis is $H_0 : \mu = 5.1$
The alternative hypothesis is $H_1 : \mu > 5.1$
The mean of sample data happiness score can be calculated by R-software with formula, $\bar{x} = \frac{\Sigma x}{n}$ and the result is 5.3757. The standard deviation of the sample data happiness score can be calculated by using R studio with the formula,

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$$

and the result of s is 1.145. Since the number of data sample is 158 which is more than 30, the test statistic for mean can be calculated by using formula, $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$.

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$
$$= \frac{5.3757 - 5.1}{1.145 / \sqrt{158}}$$
$$= 3.027$$

where,
z = z score
$\bar{x}$ = mean
μ = population mean
s = standard deviation of sample
n = total sample

With the formula $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, the test statistic for mean can be obtained which is 3.027. The critical region, a, is equal to 0.05 and the critical value can be obtained from the Z-score table which is 1.645

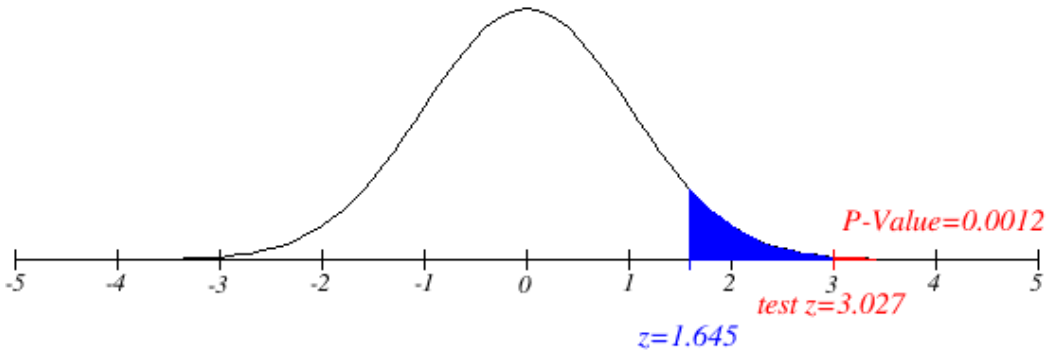

Figure 3.1

Figure 3.1 shows that the test statistic means happiness score falls within the critical region. Since the critical value, $Z_{0.05}$=1.645 < test statistic for mean, Z = 3.027, the null hypothesis is rejected. There is sufficient evidence to conclude that the mean worldwide happiness score 2015 is greater than 5.1

3.2 Correlation

In this correlation part, the relationship between economy (GDP per Capita) and happiness score are used. Hence, this correlation test is used to determine whether richer countries will lead to higher happiness scores. Obviously, the variable of economy (GDP per Capita) is considered as an independent variable while the variable of happiness score is a dependent variable. Thus, the economy (GDP per Capita) is initialized to x value and happiness score is initialized to y value. Due to the variables (economy (GDP per Capita), happiness score) are ratio type variables, therefore we use Pearson's product-moment correlation coefficient technique to analyze the correlation between variables.

R | Global Environment

Data	
WorldHappinessReport2015	158 obs. of 12 variables
Values	
r	0.780965526866021
sum_of_X	133.68968
sum_of_X2	138.6333123828
sum_of_XY	775.27491697
sum_of_Y	849.366
sum_of_Y2	4771.800404

- sum_of_X = 133.68968
- sum_of_Y = 849.366
- sum_of_X2 = 138.6333123828
- sum_of_Y2 = 4771.800404
- sum_of_XY = 775.27491697
- n denoted as total of country = 158

Figure 3.2.1 Data and variable in R studio for correlation

For this relationship between economy (GDP per Capita) and happiness score, x denoted as economy (GDP per Capita) and y denoted as happiness score where n = 158. By using Pearson technique, r denoted as the result of the correlation coefficient obtained using R studio, with the

formula: cor(x, y, method = "pearson"). And the result is r = 0.7809655, which is same as the formula such that:

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

$$r = \frac{775.27491697 - (133.68968 \times 849.366)/158}{\sqrt{[(138.6333123828 - \frac{(133.68968)^2}{158}][(4771.800404) - \frac{(849.366)^2}{158}]}}$$

$$r = 0.7810$$

where,
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

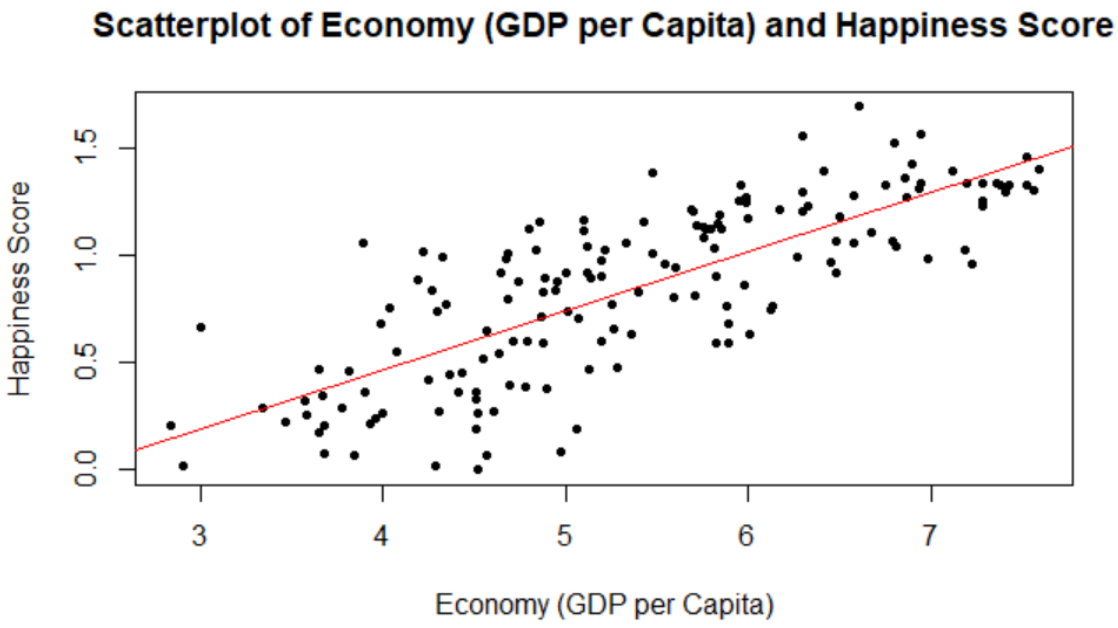

Figure 3.2.2 Scatterplot of happiness score versus economy (GDP per capita)

Significance Test for Correlation

Hypotheses: $H_0 : \rho = 0$ (no linear correlation)

$H_1 : \rho \neq 0$ (linear correlation exist)

Degree of freedom: df = 158 - 2 = 156

Assume $\alpha = 0.05$, $t_{0.05} = 1.6547$

Test statistics:

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

$$= \frac{0.7810}{\sqrt{\frac{1 - 0.7810^2}{158 - 2}}}$$

$$t = 15.6192$$

where,

t = T score

r = Sample correlation coefficient

n = Sample size

Conclusion:

Since r = 0.7810, it is closer to 1. Therefore, it is a strong positive linear relationship between the variables. By using r = 0.7810, the test statistics calculated using formula is 15.6192, which is greater than the critical value, 1.6547. Therefore, null hypothesis (H_0) is rejected. There is sufficient evidence of a linear relationship between economy (GDP per Capita) and happiness score. In this contect, the point of view is supported by an analytical study. Based on the analytical study, the results statistically show that there is significant evidence that the economy drives happiness at the global scale which means that the economy (GDP per Capita) and happiness score are closely related to each other (Nguyen & Le, 2021). Based on this result, the economy (GDP per Capita) is directly proportional to happiness score which means that when the economy (GDP per Capita) increases, the happiness score also increases. Therefore, in real life, people in richer countries tend to be happier than the people in poorer countries.

3.3 Regression

In this regression part, a measurement for the relationship between the family and happiness score is carried out. Hence, this regression test is used to determine whether a more harmonious family will lead to higher happiness scores. This is because most people, especially teenagers, are not happy and stressed when their families are not in harmony. Obviously, the

variable of family is considered as an independent variable while the variable of happiness score is a dependent variable. Thus, family is initialized to x value and happiness score is initialized to y value. The sample regression line provides an estimate of the population regression line.

$$Y_i = b_0 + b_1 x$$

where,
Y_i = Estimated (or predicted) Y value
b_0 = Estimate of the regression intercept
b_1 = Estimate of the regression slope
X = Independent variable

For the relationship between family and happiness score , n is 158 which limits the value x from 0 to 1.5 , x is family and y is happiness score to calculate the regression model using R studio. For this relationship, the line is coloured with red and the dots with black.

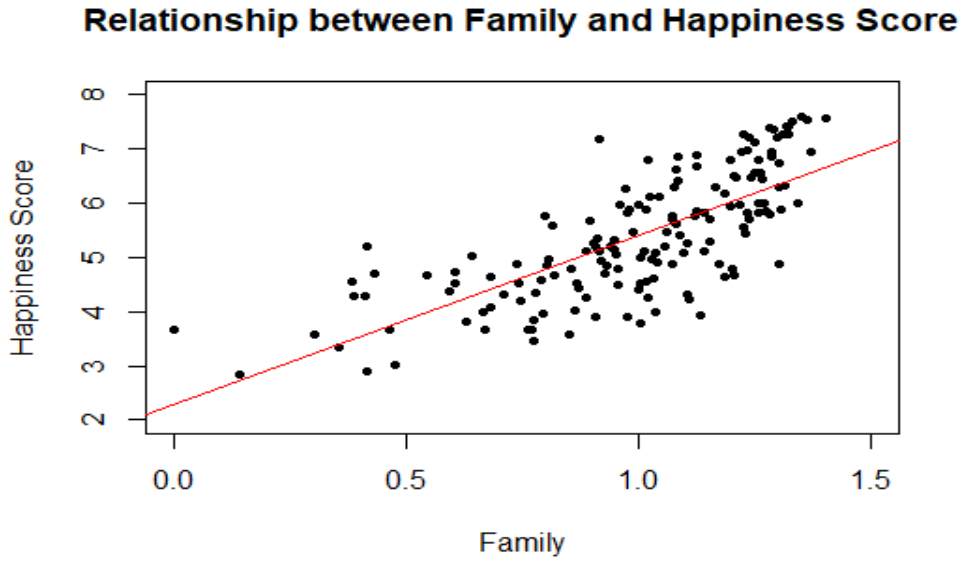

Figure 3.3.1 Scatter plot graph about the relationship between happiness score and family

```
> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88667 -0.47443 -0.02976  0.61584  2.04955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2902     0.2324   9.855   <2e-16 ***
x             3.1134     0.2262  13.766   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7718 on 156 degrees of freedom
Multiple R-squared:  0.5485,    Adjusted R-squared:  0.5456
F-statistic: 189.5 on 1 and 156 DF,  p-value: < 2.2e-16
```

Figure 3.3.2 The summary of scatter plot graph from Rstudio

From the summary of this relationship, the formula for the estimated regression model:

$$\hat{y} = 2.2902 + 3.1134x$$

When x equals 0, the estimated y value is 2.2902 that means there is no family that has 0 value and the happiness score within the range of size observed, 2.2902 is the portion of happiness score. The b1 is 3.1134 which means that the average value of happiness score is increased by 3.1134.

Coefficient of determination, R^2 is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

$$R^2 = SSR / SST \text{ where } 0<=R^2<=1$$

```
> # TODO: To get (SSE)
> sse<-sum((fitted(t_line)- y)^2 )
> sse
[1] 92.93512
> # TODO: To get (SSR)
> ssr<-sum((fitted(t_line)- mean(y))^2 )
> ssr
[1] 112.8994
> #TODO: To get  (SST)
> sst<-ssr+sse
> sst
[1] 205.8346
> # TODO: To summarize the relationship between x and y
> summary(lm(y ~ x))
```

Figure 3.3.3 The value of SSR,SSE & SST

Figure 3.3.3 is used to prove that the equation of R^2 is correct. From the equation that we obtain, we divide 112.8994 with 205.8346. From there, we can get 0.5485 for R^2 which is equal to the result from the summary. Since R^2 is not close to 0 or 1, it shows a weaker linear relationship between x and y.

Regression t Test

Hypotheses: $H_0 : \beta_1 = 0$ (no linear relationship)

$H_1 : \beta_1 \neq 0$ (linear relationship does exist)

$$t = \frac{b_1 - B_1}{S_{B_1}}$$

$$t = \frac{3.1134 - 0}{0.2262}$$

$$= 13.766$$

where,

B_1 = Hypothesized slope

b_1 = Sample regression slope coefficient

S_{B1} = Estimator of the standard error of the slope

Conclusion:

According to Figure 3.3.2 and the calculation above, the test statistic is 13.766. Since the critical value of t-test with the significance level of 0.05 is 1.6547, 13.766 is greater than 1.6547. Hence, we reject the null hypothesis, H_0. There is sufficient evidence that family affects the happiness score. In this context, this point of view is supported by the result of Wu that family is the main effect on the happiness of people, especially high-school students. According to the research, it is revealed that students that grow up in a harmonious family will easily get happiness compared to others (Wu ,2014).

3.4 Chi-Square Test of Independence

In this Chi-Square test, the test is regarding the relationship between two nominal variables, which is happiness score and continent .Hence, a modified dataset is prepared to measure this relationship. In this context, the happiness score is divided into 3 categories which are $2<x<4$, $4<x<6$ and $6<x<8$. Besides, the regions are divided into 5 available continents which are Africa, America, Asia , Australia and Europe to reduce the level of variables. So, in this test, we want to test whether there is a relationship between the happiness score and the continents.

Hypotheses: H_0 : Happiness score is independent with the continents
H_1 : Happiness score is dependent with the continents

```
> #TODO: Output the critical value
> print(x2.alpha)
[1] 16.91898
> #TODO: Output the chi-square value
> output$statistic
X-squared
 53.90738
> #TODO: Output the parameter of degree of freedom
> output$parameter
df
 8
> #TODO: Output the observed value table
> output$observed

       Africa America Asia Australia Europe
  2<x<4     19       0    2         0      0
  4<x<6     35       9   17         0     32
  6<x<8      6      15    3         2     18
> #TODO: Output the expected value table
> output$expected

          Africa   America      Asia Australia   Europe
  2<x<4  7.974684  3.189873  2.924051 0.2658228  6.64557
  4<x<6 35.316456 14.126582 12.949367 1.1772152 29.43038
  6<x<8 16.708861  6.683544  6.126582 0.5569620 13.92405
>
```

Figure 3.4.1 Calculation and contingency table from Rstudio

Two-ways contingency table is drawn using Rstudio. From the calculation above, the chi-square value is 53.90738 which is greater than the critical value calculated which is 16.91898.

Conclusion:
Since chi-square value is greater than critical value, the null hypothesis, H_0 is rejected. Therefore, there is strong evidence that there is a relationship between happiness score and the continents. This hypothesis can be supported by the research of Dulababu that every continent has a different level of happiness (T, 2017).

3.5 ANOVA

In this ANOVA test, the test is regarding the measurement of equality of 4 different regions with the happiness scores.(sub-Saharan Africa, Latin America and Caribbean, Southern Asia and Western Europe). From the previous chi-square independence, there is the relationship between continents and happiness score, thus a test on determining whether different types of regions can have the same mean of happiness score is carried out.

Hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : at least one mean is different

```
> summary(res.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
y            3  58.65  19.550   44.01 9.55e-15 ***
Residuals   56  24.88   0.444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.5.1: The result of ANOVA

where,

F= variance between samples / variance within samples

Based on the Figure 3.5.1, the numerator is k-1= 4-1=3. The denominator = k(n-1)=4(15-1)=56. The test statistic, F = 44.01 . P-value is the significance level of the f-test. P value is 9.55×10^{-15}.

Conclusion:

Since P-value of F is less than the significance level 0.05, the null hypothesis is rejected. There is insufficient evidence to claim that the different types of regions among the 4 regions have the same mean of happiness score.

4.0 Conclusion

As a conclusion, the research team has learned that there is a need to consider what analysis test will be conducted in this study but not the objective of study only when choosing the dataset. The research team needed to figure out first what type of data set they needed such as in nominal, ordinal, ratio or interval before they decided to use any dataset. When doing the pre-processing of data, the research team needs to always remember to stick to the objective of study or else the research team will face problems during the analysis process. In this study, the study shows that there is a strong relationship between the happiness score and economy (GDP per Capita).The result of this study also shows that family will also affect the happiness score. In the Chi-Square Test analysis process,it shows that the continents that have higher mean happiness scores are Europe and America while Africa have the lowest mean of happiness score. The study also gives a clue about why Africa has the lowest mean of happiness score which is their happiness score is strongly affected by their downturn country's economy and may be the not harmonious family situation of the people who live in Africa.

5.0 References

Nguyen, B. K. Q., & Le, V. (2021). The relationship between global wealth and happiness: An analytical study of returns and volatility spillovers. *Borsa Istanbul Review*. https://doi.org/10.1016/j.bir.2021.04.006

T, D. (2017). Global Happiness: Continental and Cross-Cultural Models Perspective. *Journal of Global Economics*, *05*(04). https://doi.org/10.4172/2375-4389.1000268

World Happiness Report. (2019, November 27). Kaggle. Retrieved: https://www.kaggle.com/unsdsn/world-happiness

Wu, Z. (2014). Family is the most influential factor on happiness in high school students. *Health*, *06*(05), 336–341. https://doi.org/10.4236/health.2014.65049

Zhang, K. (2015, April 23). World Happiness Report 2015 Ranks Happiest Countries. Retrieved: https://www.unsdsn.org/news/2015/04/23/world-happiness-report-2015-ranks-happiest-countries