
UTM

UNIVERSITI TEKNOLOGI MALAYSIA

School of Computing

**SECI2143 - Probability and Statistical Data Analysis**

Section 7

Semester 2, Session 2021/22

**Project 2 :**

**Inferential Statistics**

Lecturer : Dr. Nor Azizah Ali

Team : Spongebob & Jojo


| Name | Matric No. |
|---|---|
| 1. Wong Li Jie | A21EC0238 |
| 2. Ng Qian Hui | A21EC0212 |
| 3. Toh Kang Lun | A21EC0234 |
| 4. Ho Wei Chun | A21EC0184 |

# Table of Content

## 1.0 Introduction

The background of the study is to investigate the relationship between a country's GDP per capita, brain drain rate and inflation rate and unemployment rate. Also, we aim to study the monthly average income of different countries in order to obtain better insights on the country's performance. We are interested to learn how various economic environments can affect the performance of a country and its netizens.

The data collected from different online sources are organized neatly according to our studies. In order to carry out data analysis, our team has utilized R-Programming for graph visualizations and calculations. Several statistical methods are also used in the estimation of monthly average income and unemployment rate of each country.

## 2.0 Data Description

The dataset that we obtained from online sources is secondary data. After that, we processed the data and randomly picked 10 countries' which is United States of America, Australia, South Korea, Malaysia, China, Indonesia, Laos, Cambodia, Pakistan and Nepal as our research countries.

| Variable | Method to Test | Explanation |
|---|---|---|
| Average Income | Hypothesis 1 Sample | To investigate whether the average monthly income is greater than USD 800. |
| GDP per capita<br>Brain Drain Rate | Correlation | To investigate whether gdp affects brain drain rate, we expect that a higher GDP leads to a lower brain drain rate. |
| GDP per capita<br>Inflation Rate | Regression | To figure out whether Inflation rate affects GDP per capita, we expect that a higher inflation rate gets a lower GDP per capita. |
| GDP per capita<br>Unemployment Rate | Chi-Square Test of Independence | To survey whether there exists a relationship between two nominal variables, which are status of countries and unemployment rate. |

## 3.0 Data Analysis

### 3.1 Hypothesis Testing

Based on the data, we want to study whether the average monthly income of 10 different countries is greater than USD 800. So, hypothesis testing using one sample is carried out. We assumed that the variance is unknown. Since the number of observations for this test is 10, the test statistic formula: $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ is used, where:

t = test statistic
$\bar{x}$ = sample mean
μ = population mean
s = standard deviation
n = number of observations

Hypothesis Testing using single sample:

Statement: The average monthly income is greater than USD 800.
Significance level, $\alpha = 0.05$
$H_0$ : μ = USD 800
$H_1$ : μ > USD 800

Degree of freedom = 10 - 1 = 9

$t_{0.05, 9} = 2.262$

$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{1292.02 - 1000}{1486.06 / \sqrt{10}} = 1.046$

Based on the calculation, we obtained the t value of 1.046 as the test statistic value. The value is smaller than $t_{0.05, 9} = 2.262$, 1.046 < 2.262. Therefore, there is not sufficient evidence to believe that the average monthly income is greater than USD 800.

∴ Fail to reject $H_0$ since $t < -t_{0.05, 9}$. There is not sufficient evidence to prove that the average monthly income of 10 different countries is greater than USD 800 at significance level of 0.05.

## 3.2 Correlation

Sample correlation coefficient :

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

r = sample correlation coefficient,
n = 10,
x = GDP per capita,
y = Brain drain rate.

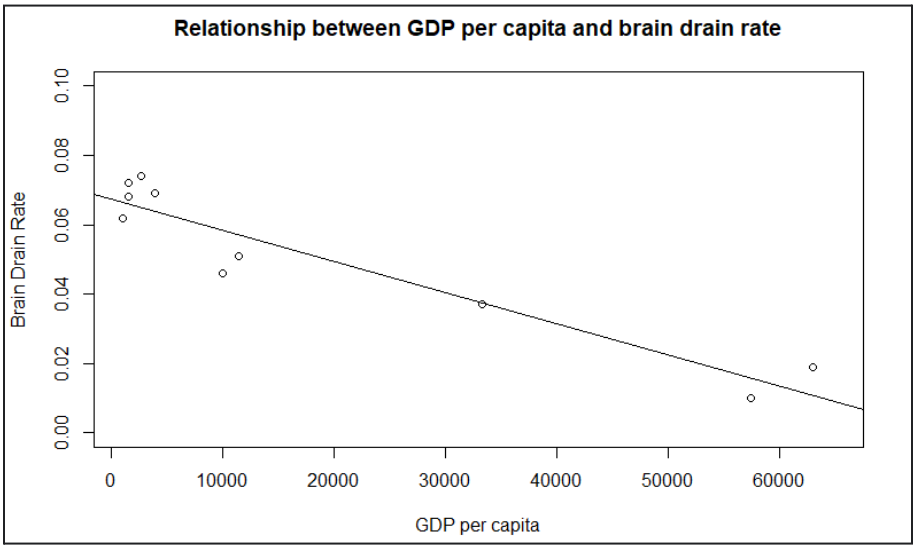


Figure 3.2.1 : Relationship between GDP per capita and brain drain rate

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = -8.518, df = 8, p-value = 2.772e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9881871 -0.7936812
sample estimates:
       cor
-0.9490472
```

Figure 3.2.2 : Result of correlation coefficient

Based on the calculation, we obtained a value of -0.9490472 as the correlation coefficient. The scatter plot and correlation analysis of the data indicates that there's a negative relationship between the GDP per capita and brain drain rate. As the GDP per capita increases, the rate of brain drain decreases. This is also a strong negative relationship because r falls within -1 and -0.8, such that $-1 < r < -0.8$.

Significance Test for Correlation :

Significance level, $\alpha = 0.05$

H0 : ρ = 0 (no linear correlation)

HA : ρ ≠ 0 (linear correlation exists)

From Figure 3.2.1,p-value = $2.772 \times 10^{-5}$,

Degree of freedom = 10 - 2 = 8

$t_{0.025, 8} = \pm 2.306$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.9490472}{\sqrt{\frac{1-(-0.9490472)^2}{10-2}}} = -8.51799$$

∴ Reject H0 since $t <$ - $t_{0.025, 8}$ , there's sufficient evidence that a linear relationship exists between GDP per capita and brain drain rate at significance level of 0.05.

## 3.3 Regression

Estimated Regression Model:

$$\overline{Y} = b_0 + b_1 X$$

Where :

Y = Estimated (or predicted) Y value

$b_0$ = Estimate of the regression intercept

$b_1$ = Estimate of the regression slope

X= Independent variable

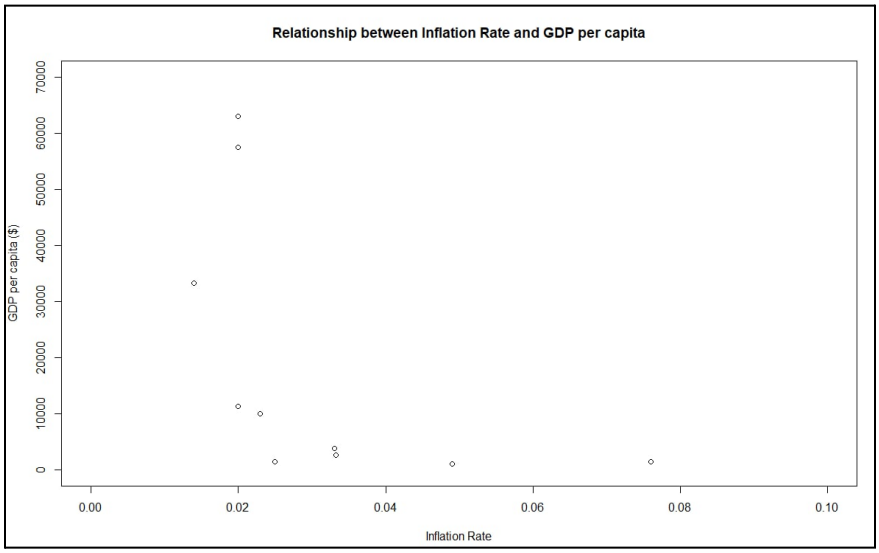


Figure 3.3.1: Result of GDP per capita against Inflation Rate

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-21251 -14542  -9666  10270  36912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    39373      14072   2.798   0.0233 *
x            -664411     391524  -1.697   0.1281
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21830 on 8 degrees of freedom
Multiple R-squared:  0.2647,    Adjusted R-squared:  0.1728
F-statistic:  2.88 on 1 and 8 DF,  p-value: 0.1281
```

Figure 3.3.2: The summary of the graph by using RStudio

From the summary, we can get the formula for estimated regression model is:

$$\overline{Y} = 39373 - 664411\,x$$

The coefficient of determination:

$$R^2 = \frac{SSR}{SST} = \frac{sum\ of\ square\ explained\ by\ regression}{total\ sum\ of\ squares}$$

We get that the coefficient of determination, $R^2 = 0.2647$. Since, $0 < R^2 < 1$, shows weaker linear relationship between x and y. Some but not all the variation in y is explained by variation in x.

Test Statistical of Regression :

$H_0 : \beta_1 = 0$ (non linear relationship)

$H_1 : \beta_2 \neq 0$ (linear relationship)

Test statistic, $t = \frac{b_1 - \beta_1}{S_{b_1}}$

$= -1.69699$

Degree of freedom = 10-2 =8

Where:

$b_1$ = Sample regression slope coefficient

$\beta_1$ = Hypothesized slope

$S_{b_1}$ =Estimator of the standard error of the slope

$\therefore$ Fail to reject $H_0$. Based on the figure, test statistic, t = -1.69699. P-value is the significance level of the t-test. Since P-value is 0.1281 is more than significance level 0.05, there's sufficient evidence that a non-linear relationship exists between inflation rate and GDP per capita.

### 3.4 Chi-Square

Chi-Square Test of Independence:

$$x^2 = \Sigma \frac{(O-E)^2}{E}$$

Where,
O= observed values
E= expected values

Test hypothesis:
$H_0$: Unemployment rate is independent with the countries status
$H_1$: Unemployment rate is dependent with the countries status

```
> # output critical value
> print(x.alpha)
[1] 16.91898
> # output the chi-square value
> output$statistic
X-squared
 4.222222
> # output the parameter of degree of freedom
> output$parameter
df
 4
> # output the observed value table
> output$observed

      Developed Developing Least developed
  0<x<2         0          0               2
  2<x<4         2          1               3
  4<x<6         1          1               0
> # output the expected value table
> output$expected

      Developed Developing Least developed
  0<x<2       0.6        0.4               1
  2<x<4       1.8        1.2               3
  4<x<6       0.6        0.4               1
```

Figure 3.4.1 : Calculation and contingency table from R Studio

In this Chi-Square test of independence, the test is conducted to examine whether there exists a relationship between two nominal variables, which are status of countries and unemployment rate. In this context, a modified dataset is prepared as the unemployment rate is divided into 3 categories which are 0<x<2, 2<x<4 and 4<x<6. Also, the countries are also classified into 3 status which are developed, developing and least developed according to the GDP per capita of a country.

From the calculation observed, the chi-square value is 4.22222. By using a significance level of 0.05, degree of freedom of 4 , the critical value calculated is 16.91898.

∴ Since the test statistic value or the calculated chi-square value is less than the critical value, we fail to reject the null hypothesis. Therefore, there is evidence that the status of countries and the unemployment rate are dependent.

## 4.0 Conclusion

In conclusion, our team has applied different data analysis theories and techniques in this study. In the phase of choosing a suitable dataset, we have learned to select a dataset that can fulfill the objective of performing each particular data analysis test so that the requirements of this project can be achieved. From the data pre-processing stage, we also gained some data pre-processing knowledge such as data cleaning and data transforming. In the data analysis process, we have successfully drawn some important conclusions in this study. After conducting the hypothesis testing, the hypothesis test shows that there is sufficient evidence to prove that the average monthly income is greater than USD 800 based on the 10 countries selected. From the correlation test, we can conclude that there is sufficient evidence that a linear relationship exists between GDP per capita and brain drain rate which means when GDP per capita increases, the rate of brain drain decreases. In the regression test, it shows that there exists a non-linear relationship between inflation rate and GDP per capita. Based on the chi-square test of independence, we can conclude that the status of countries and unemployment rate are dependent, meaning that the unemployment rate is dependent on the status of the countries which is categorized according to the GDP per capita of the countries.

## 5.0 Appendix

| Team Member | Task Distribution |
|---|---|
| Wong Li Jie | • Introduction<br>• Correlation |
| Ng Qian Hui | • Conclusion<br>• Chi Square Test of Independence |
| Toh Kang Lun | • Dataset Description<br>• Regression<br>• Video Editing |
| Ho Wei Chun | • Dataset Description<br>• Hypothesis Testing |

(20/6/2022, Monday)

Source of data :

Unemployment rate :
https://ourworldindata.org/grapher/unemployment-rate?tab=chart&country=~MYS

Average income :
https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=105

Brain drain rate :
https://tcdata360.worldbank.org/indicators/ha03234ca?country=BRA&indicator=43980&viz=line_chart&years=2006,2021

GDP per capita :
https://statisticstimes.com/economy/world-gdp-capita-ranking.php

Inflation Rate :
https://www.inflation.eu/en/inflation-rates/cpi-inflation-2019.aspx