UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION-02

SECI2143

20202021/2

PROJECT TITLE :

COUNTRIES DATA

LECTURER'S NAME :

DR CHAN WENG HOWE

GROUP MEMBERS :

NAME	MATIRC NO.
HONG PEI GEOK	A20EC0044
LOW JUNYI	A20EC0071
TAN YONG SHENG	A20EC0157
VINCENT BOO EE KHAI	A20EC0231

Table of Contents

1.0 INTRODUCTION

Countries data can be known as the data that has been collected about the country such as population and growth of a country, life expectations and mortality of the citizens. It is some useful data that is very important to a country because there are relationships between the data collected by the country and the future planning of the country. We will take vaccination for example, in order to get all citizens to be vaccinated, we will need to know the relationship between the vaccine needed and the population of the country. In this study, countries's data was selected in order to conduct our testing.

There are several purposes of this study which are applying the concept of hypothesis testing that we learnt no matter one sample, two sample or chi-square tests. Next, we would like to investigate the relationship between some variables of countries' data such as relationship between population and mortality and also predicting the useful value based on the data we have.

We are interested as we think that this is one of the most important and practical questions that we might be able to apply in the future society because based on population, we can predict and do further study. Furthermore, through countries' data, we also get to know the growth, life expectations and mortality which can help us to form a great image about the world, at the same time we could know the world better. In this study, we expect to know the accuracy of this data and also understand the relationship between variables based on the data we have.

2.0 DATASET

The dataset that we chose for this project is countries data. The reason our group chose this data is because we want to investigate the relationship between the countries and the variety aspects of countries. The dataset includes the countries all around the world and each of the countries provided with the data population, growth of GDP, percentage of population under 15, life expectancy and mortality.

The variable population describes the estimated total number of citizens in each country. The next variable on the dataset is growth. The variable growth provided contains decimal and sign which indicates that the growth rate of each country can be increasing or decreasing according to the poverty rate and food production. Next, the variable is under 15. Under 15 means the percentage of the population which is under 15 years old and this shows the maturity of the countries and the overall birth rate of the countries. Therefore the population can be greatly affected by the percentage of under 15. The next variable given is life expectancy with one decimal point and this variable indicates that the average number of years an individual can live in their specific countries. The life expectancy is based on the particular population of the average age when they die. Lastly, the variable provided is mortality and it contains one decimal point as well. Mortality means that a person who is alive has been subjected to death. There are a variety of factors that affect the mortality of the countries such as disease, medical facilities, education level and food supply. Therefore, there are a total of 5 variables in this dataset and we are going to determine the relationship of each variable by using R studio.

Next, we start to do data pre-processing by importing all the crucial libraries that are needed in the project. Afterwards, we start to import the dataset given by Dr in e-learning into R studio. Since the dataset given does not contain any missing value, we are free from handling missing value. The last step is feature scaling which marks the end of the data pre-processing. This is used to standardize the range of certain independent variables. This is done to easily compare the variables on common ground.

For this project, we use some statistical test to investigate the difference between the mean population and mean sample. The statistical analysis that we are going to use are hypothesis 1 or 2 sample test, correlation, regression, Goodness of fit, and Chi-square. The

possible outcome of the test based on each variable is expected to be different. However, we expect there will be relationships between variables of country data.

3.0 DATA ANALYSIS

3.1 Hypothesis Testing

3.1.1 Hypothesis Testing 1-Sample

In this part, we wish to determine if there is any difference between the sample mean of Asia population and the claim of population mean under 0.05 significance level and unknown variance.

u = Population Mean of Asia population

$\overline{x}$ =Mean of the sample of Asia population

s = Standard deviation of the sample of Asia population

n =Size of the sample of Asia population

$H_0 : u = 4641054775$

$H_1 : u \neq 4641054775$

α= 0.05

$t_0 = \frac{\overline{x}-u}{\frac{s}{\sqrt{n}}}$

= -27.7029

P-value = 5.0515e-10

$t_{\frac{\alpha}{2},df} = 2.262$

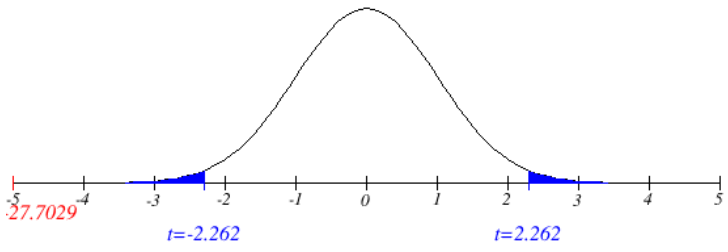

Figure 3.1.1 Critical Region of Hypothesis Testing 1 Sample

Reject H_0 if $t_0 > 2.262$ or $t_0 < -2.262$

Since $t_0 = -27.7029 < -2.262$, thus reject H_0 and there is sufficient evidence to conclude that the mean of the sample population is different from the claim population mean.

Discussion of Hypothesis Testing 1 Sample

We have selected 10 samples of Asia countries to determine whether the sample mean of population is equal to the population mean which is 4641054775. This claim value is obtained from the statista (Statista, 2021). We have got the result of rejecting the null hypothesis at 0.05 significance level. This means that the mean population of Asia countries is not equal to 4641054775. The reasons are insufficient sample size selected and there is a dramatic increase in Asia population in recent years due to economic, geography and culture.

3.1.2 Hypothesis Test 2-Sample

We wish to determine whether there is any difference between the mean of Asia population and the mean of Europe population under 0.1 significance level. We assume that both variances of both variables are unequal.

u_1 = Mean of Asia population

u_2 = Mean of Europe population

$\overline{x_1}$ =Mean of the sample of Asia population

$\overline{x_2}$ =Mean of the sample of Europe population

s_1= Standard deviation of the sample of Asia population

s_2= Standard deviation of the sample of Europe population

n_1 =Size of the sample of Asia population

n_2 =Size of the sample of Europe population

$H_0 : u_1 = u_2$

$H_1 : u_1 \neq u_2$

α= 0.10

$$t_0 = \frac{\overline{x_1} - \overline{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

= 2.0301

P-value = 0.0729

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$= 9$

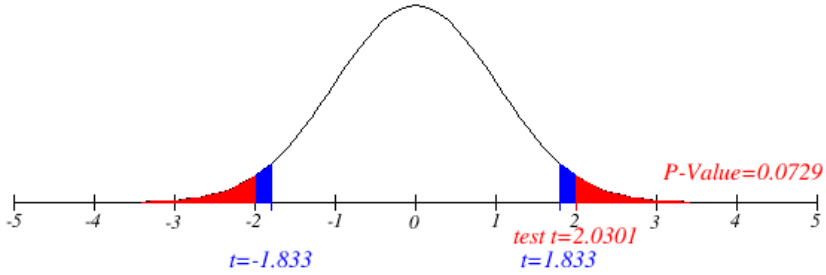

Figure 3.1.2 Critical Region of Hypothesis Testing 2-Sample

Reject H_0 if $t_0 > t_{\frac{\alpha}{2}, df = 1.833}$ or $t_0 < t_{\frac{\alpha}{2}, df = -1.833}$

Since $t_0 = 2.0301 > 1.833$, thus to reject H_0 and there is sufficient evidence to conclude that the mean of Asia population is not equal with the mean of Europe population.

Discussion of Hypothesis Testing 2 Sample

We have selected 10 samples from Asia countries and also 10 samples from Europe countries. This is because we want to determine if the mean of Asia population is the same with the mean of Europe population at 0.1 significance level. Next, we got the result that the mean of Asia population is different from the mean of Europe population which led us to reject the null hypothesis. This result is not surprising and it matches the real world condition. The reason might be that the size of Asia continent is larger than the European continent.

3.2 Correlation Test

Correlation is the measure of strength of linear relationship between two variables. Therefore in this part we will measure the strength of the linear relationship between the variables, **population** and **mortality**. If the correlation is positive then the mortality will increase if the population increases.

Below is the guideline which identifies the strength of correlation of coefficients.

Size of correlation coefficient	Strength of correlation
1 - 0.8 or -1 - -0.8	Strong
0.8 - 0.5 or -0.8 - 0.5	Moderate
0.5 - 0 or -0.5 - 0	Weak

In this case we need to find the correlation of coefficients before we can conclude the relationship between two variables. The correlation coefficient is known as r in this case.

Hypothesis statement :

H_0 : $p = 0$ (No linear correlation)

H_1 : $p \neq 0$ (There exist linear correlation)

Test statistic :

Find r :

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

where:

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

According to the result from R studio, the **correlation of coefficient is 0.00767431**. Therefore, the relationship between population and mortality has a weak positive correlation.

Afterwards, we are going to use the t value formula to do test statistics.

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

Then we use R studio to code for getting the t value.

Therefore the value of t we got from R studio is 0.1074435.

P-value = 0.9145

The critical value of $t_{a/2}$ = 0.025 and df = 196 is 1.9721

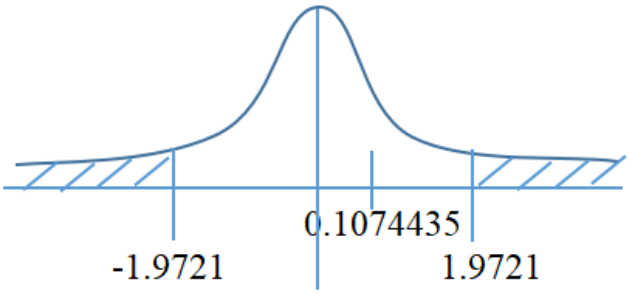

Figure 3.2.1 : Critical Region of Correlation Test

Reject H_0 if t > 1.9721 or t < -1.9721

Therefore, since -1.9721 < t = 0.1074435 < 1.9721, we fail to reject the null hypothesis. There is sufficient evidence of no linear relationship between population and mortality at the 5% significant level.

Below are the scatter plot hypotheses in order to provide a more clearer relation between the variables. This is done by using R studio.

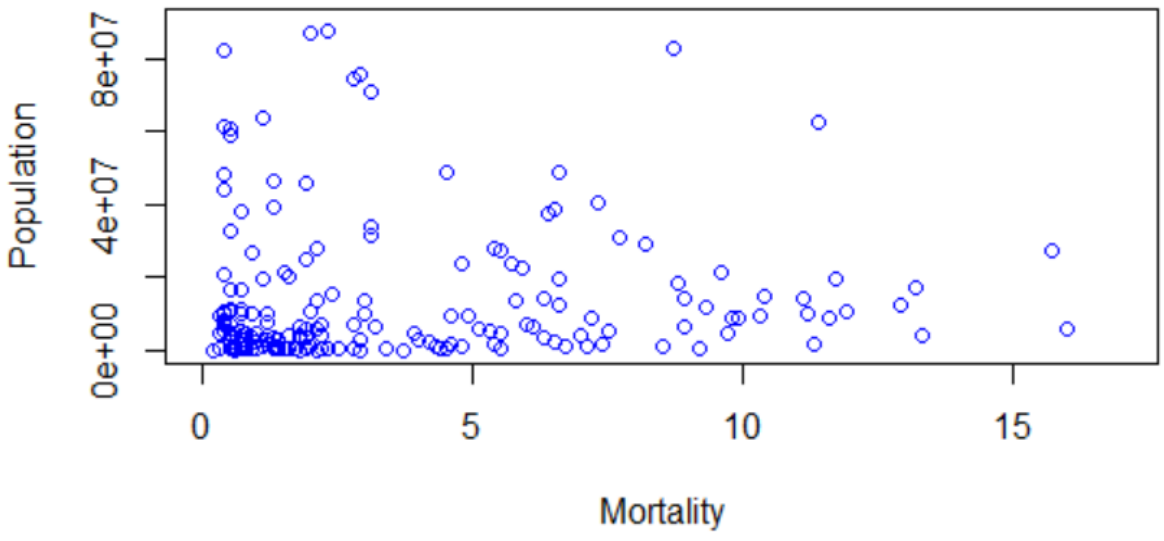

Figure 3.2.2 : Scatter plot of mortality against population

Discussion of Correlation Test (Population vs Mortality)

We have used 198 samples of countries to test the relationship between population and mortality at 0.05 significance level. We obtained a result which fails to reject the null hypothesis and this means there is no relationship between population and mortality. From figure 3.2, we can see that there is no specific pattern between these two variables. Thus, we know that the population will not be affected by the mortality rate.

The next relationship that I want to test is the relationship between population and percentage of population under 15. If the correlation is positive then the **population** will increase when **under 15** increase.

First, we need to find the correlation of coefficients before we can make the conclusion between the relationship of two variables. The formula that will be used is the same as the previous one.

Hypothesis statement :

H_0 : p = 0 (no linear relationship)

H_1 : p ≠ 0 (there exist linear relationship)

Find r :

According to the result from R studio. The **correlation of coefficient is -0.0473475**. Therefore, the relationship between the population and under 15 is a weak negative relationship.

Afterwards, the value of t according to the calculation of R studio is -0.6636093. The t value using back the previous one.

P-value = 0.5077

The critical value of $t_{a/2}$ = 0.025 and df = 196 is 1.9721

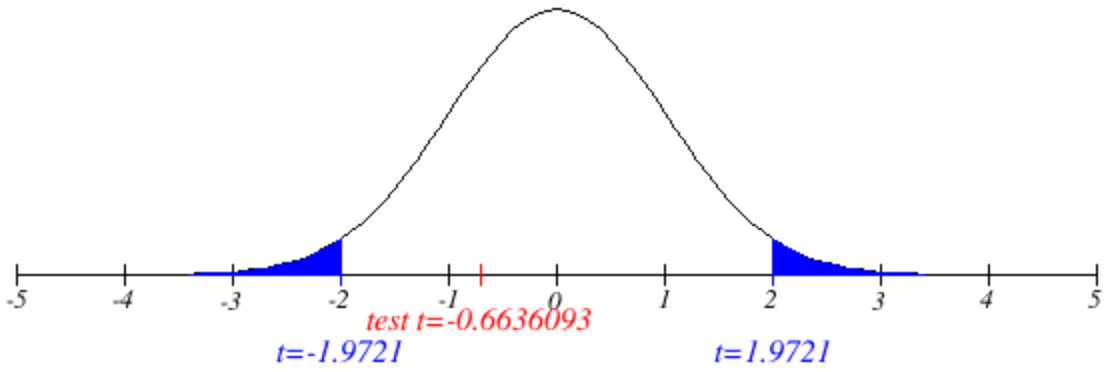

Figure 3.2.3 : Critical Region of Correlation Test

Reject H_0 if t > 1.9721 or t < -1.9721

Therefore, since the -1.9721 < t = -0.6636093 < 1.9721, we fail to reject the null hypothesis. There is sufficient evidence of no linear relationship between population and under 15 at the level of 5% significant level.

Below is the scatter plot hypothesis that is done by using R studio.

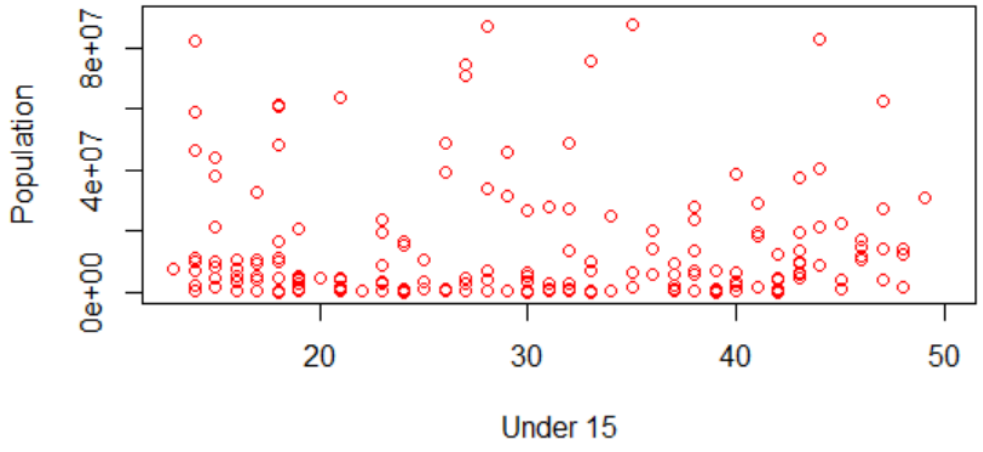

Figure 3.2.4 : Scatter plot of population against under 15

Discussion of Correlation Test (Population vs Under 15)

We have used 198 samples of countries to test the relationship between population and under 15 at 0.05 significance level. We obtained a result which fails to reject the null hypothesis and this means there is no relationship between population and mortality. From figure 3.2, we can see that there is no specific pattern between these two variables. Hence, we know that the population will not be affected by the percentage of population under 15.

3.3 Regression Test

Linear regression is a measure for the relationship between an independent variable and dependent variable. For this case, we will predict the value of growth based on mortality for part 1 and growth based on life expectancy for part 2.

Part 1:

Dependent variable: Growth

Independent variable: Mortality

Part 2:

Dependent variable: Growth

Independent variable: Life Expectancy

Hypothesis statement :

H_0 : $\boldsymbol{\beta_1} = 0$ (No linear relationship)

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

Growth Vs Mortality : $\hat{y} = 0.72265 + 0.17888x$

$b_0 = 0.72265$

$b_1 = 0.17888$

$R^2 = 0.256689$

$\alpha = 0.05$

$df = n - 2 = 196$

$S_{b_1} = 0.02174$

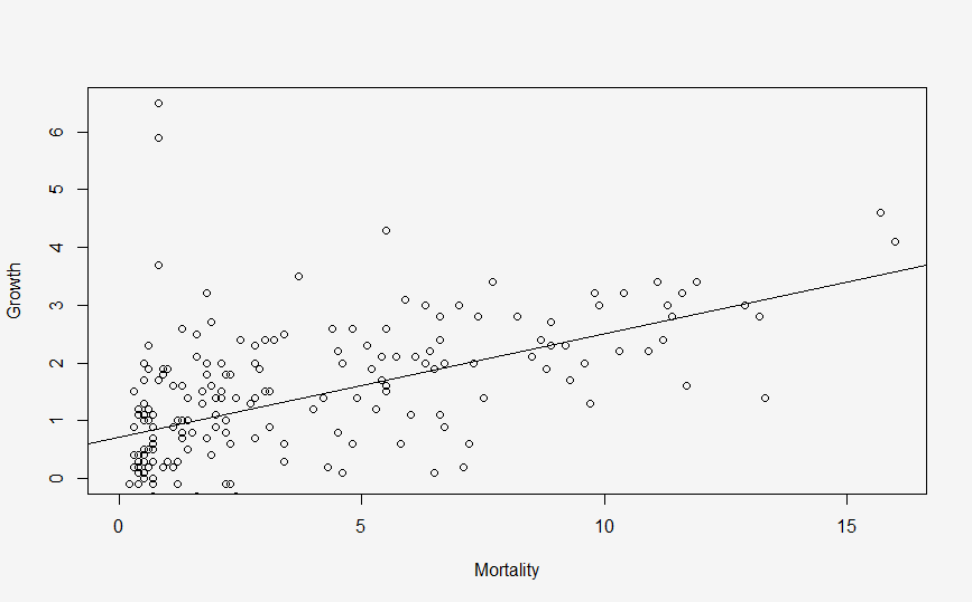

Figure 4.1.1: Mortality VS Growth

t Test,t = 8.227

P-value = 2.625e-14

c.*v.t df*,−*α*/2 = -1.972

c.*v.t df*, *α*/2 = 1.972

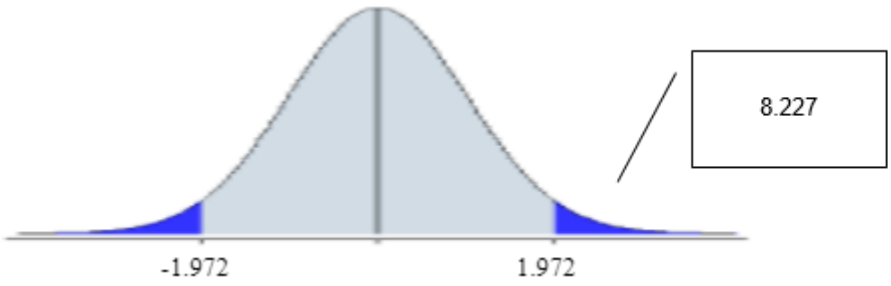

Figure 4.1.2: Critical Region

Since statistic value = 8.227 > 1.972 and P-value < 0.05. We reject H_0 at a significance level of 0.05. There is sufficient evidence that the variable that mortality affects growth.

Discussion of Regression Test (Growth vs Mortality)

We have used 198 samples of countries to test the relationship between mortality and growth at 0.05 significance level. We obtained a result which rejects the null hypothesis and this means there is a relationship between mortality and growth. From figure 4.1.1, we can see that there is a positive linear relationship between these two variables. Thus, we know that the growth will be affected by mortality. This is true because of the statistical value related to the new world. As currently there are many cases in which people would work hard until they can't manage their healthy lifestyle which led to an increase of mortality rate but on the other side increase the growth rate.

Growth Vs Life Expectancy : $\hat{y} = 4.442731 - 0.044979x$

$b_0 = 4.442731$

$b_1 = -0.044979$

$R^2 = 0.149006$

$a = 0.05$

$df = n - 2 = 196$

$S_{b_1} = 0.007678$

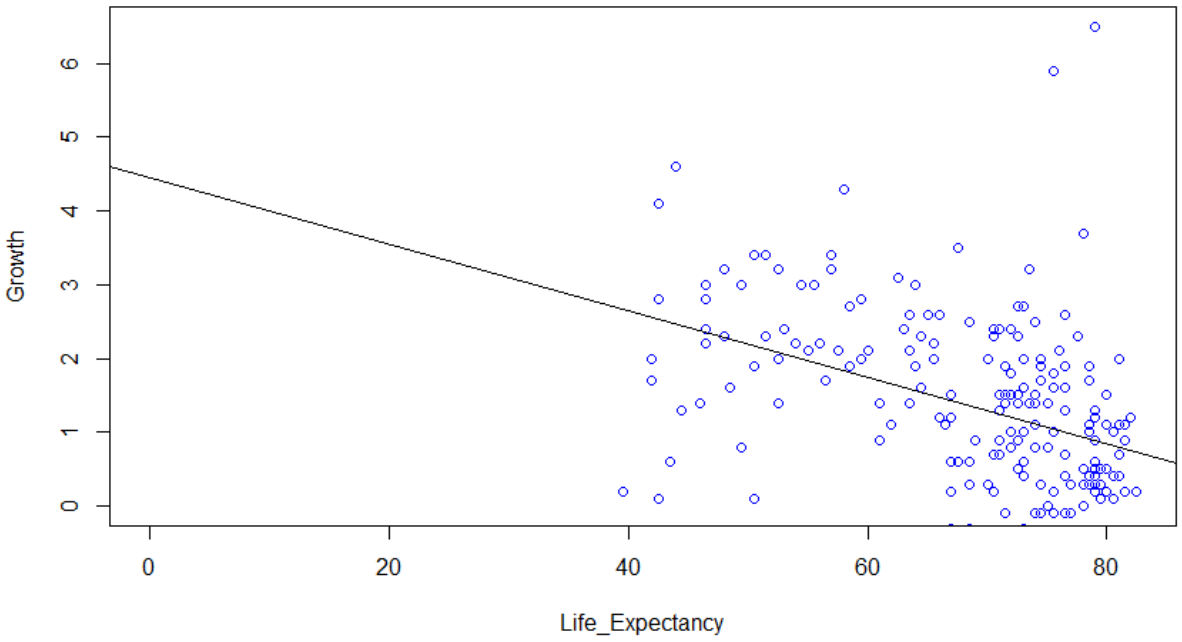

Figure 4.2.1: Life Expectancy VS Growth

t Test,t = -5.858

P-value = 1.948e-08

c.*v.t df*,$-\alpha/2$ = -1.972

c.*v.t df*, $\alpha/2$ = 1.972

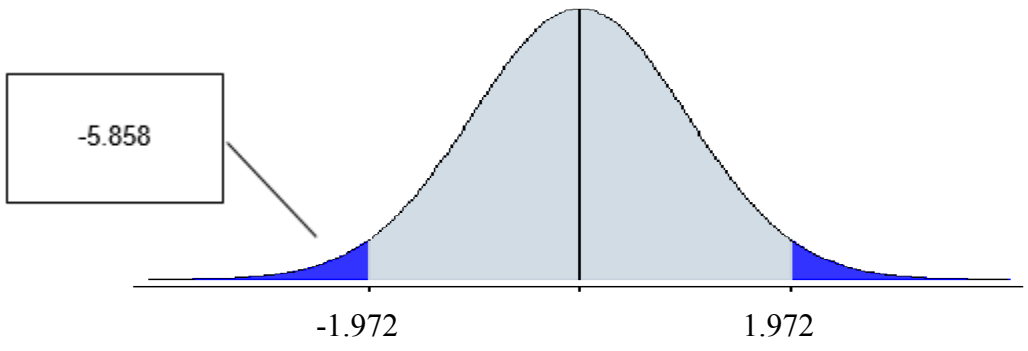

Figure 4.2.2: Critical Region

Since statistic value = -5.858 < -1.972 and P-value < 0.05. We reject H_0 at a significance level of 0.05. There is sufficient evidence that the variable that the life expectancy affects growth.

Discussion of Regression Test (Growth vs Life Expectancy)

We have used 198 samples of countries to test the relationship between life expectancy and growth at 0.05 significance level. We obtained a result which rejects the null hypothesis and this means there is a relationship between life expectancy and growth. From figure 4.2.1, we can see that there is a negative linear relationship between these two variables. Thus, we know that the growth will be affected by life expectancy. This is due to when the life expectancy is higher, there will be more older people and they have less ability to contribute to the country's economy. This will definitely decrease the growth rate of the country.

3.4 Goodness of Fit Test

A sample data of 198 countries was selected and we wish to test if the life expectancy is equal for all the 198 countries at 0.05 significance level.

p = Proportion of life expectancy of each country.

n = Total number of trials

k = Number of different countries

O = Observed frequency of life expectancy

E = Expected frequency of life expectancy

$H_0 : p_1 = p_2 = \ldots = p_{198}$

H_1 : At least one of the proportion is different from other

α= 0.05

$E = \frac{n}{k} = 68.3$

$x^2 = \sum \frac{(O-E)^2}{E} = 336.04$

P-value = 2.473e-09

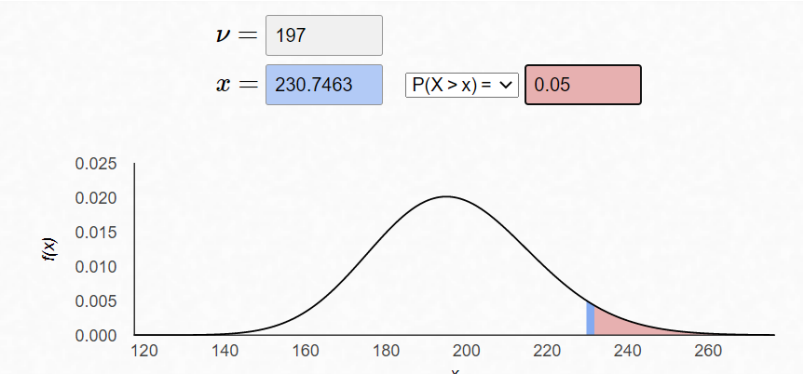

Figure 3.4.1 Critical Region

$X^2_{df,\alpha} = 230.75$

Reject H_0 if $x^2 > 230.75$

Since $x^2 = 336.04 > 230.75$, thus reject H_0 and there is sufficient evidence to conclude that the life expectancy of countries are different.

Discussion of Goodness of Fit Test

We have chosen a sample data with 198 countries in order to determine if the life expectancy of these countries are the same or not. We got the result that rejects the null hypothesis at 0.05 significance level. This shows that the life expectancy of countries are not the same. It is true because the life expectancy will vary according to the health facilities, medicine level and also the supporting plans provided by the government. Therefore, it is impossible that all the selected countries have the same life expectancy.

3.5 Chi Square Test of Independence

In the Chi-Square test of independence, 198 countries were used and we want to test the independence relationship between these countries and its population.

H_0 : Countries are independent on age group population .

H_1 : Countries are dependent on age group population.

Test statistic,

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

Result,

x^2 = 2.6064e +10

Degree of Freedom, df = 197

P-Value = 0

α= 0.05, $X^2_{df,\alpha}$ = 230.75

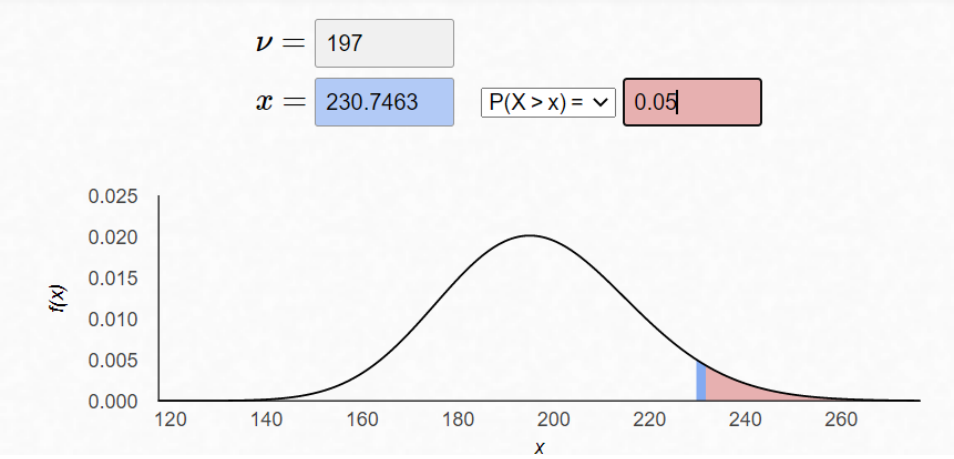

Figure 3.5.1 Critical Region

Reject H_0 if x^2 > 230.75

Since x^2 = 2.6064e +10 > 230.75, thus reject H_0 and there is sufficient evidence to conclude that the countries are dependent on population.

Discussion of Chi-Square Test of Independence

We have used 198 countries in this test just to know are these selected countries independent or dependent on their population. We got the result to reject the null hypothesis at 0.05 significance level with 197 degree of freedom. This shows that these countries are dependent on age group population. Based on the calculated result, we can know that it is true because as we can see, every country has a different population as only larger countries are allowed to have a high population due to land area.

4.0 CONCLUSION

As a summary of all the testing that we have done, after having applied the concept of hypothesis testing, the result of the hypothesis testing is that the mean of the Asia population is different from the mean of the European population and this result matches the real world condition. In the correlation test, the relationship between population and mortality has a weak positive correlation due to the result obtained from R studio. After using R studio for coding and calculations, t=0.1074435>1.9721 and so we fail to reject the null hypothesis. There is sufficient evidence of no a linear relationship between population and mortality at the 5% significant level. From another correlation test, we know that there is also no linear relationship between population and percentage of population under 15. In a regression test, 198 samples of countries were used to indicate the relationship between an independent variable and dependent variable at 0.05 significance level. We found that growth of GDP will be affected by mortality and also life expectancy. We use the goodness of fit test to calculate life expectations with 198 countries as sample data. After the calculations, the result obtained is to reject the null hypothesis at 0.05 significance level, which means the life expectations of each country are different. Next, for the Chi Square Test of Independence, 198 countries were used in this test. After the calculation, the null hypothesis at 0.05 significance level and 197 degree of freedom was rejected. In this test, the results obtained show that countries are dependent on their population.

We have obtained some of the results that are surprising or not the same as our expectation. For example, before conducting the test, we think there will be a relationship between population and mortality because mortality is about the rate of death while population is the number of people in a country. When people die, there will be a decrease in population. However, the result shows that there is no linear relationship between population and mortality.

I think we have learned a lot of things about how to conduct hypothesis testing throughout this project and also the skill of using R studio. It is not an easy task if without any effort to learn the process of testing and also coding.

5.0 REFERENCES

Goyal, K. (2020, January 22). *Data Preprocessing in Machine Learning: 7 Easy Steps To Follow*. UpGrad Blog. https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/.

.

Statista. (2021, June 16). *Total population of the ASEAN countries from 2011 to 2021*. https://www.statista.com/statistics/796222/total-population-of-the-asean-countries

6.0 APPENDIX

Video & Dataset & R Script Link:

https://drive.google.com/drive/folders/1GNdFbplKURbLLOnOjyRY7C0zJ_R0ftLV?usp=sharing