UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

PROJECT 2

2019 WORLD HAPPINESS REPORT

SECTION 02

DR CHAN WENG HOWE

GROUP MEMBERS:

NAME	MATRIC NO
AMIRAH RAIHANAH BINTI ABDUL RAHIM	A20EC0182
CHLOE RACQUELMAE KENNEDY	A20EC0026
LEE CAI XUAN	A20EC0062
MYZA NAZIFA BINTI NAZRY	A20EC0219

INTRODUCTION

Happiness is what everyone wants to get. It will make us happy and help us forget our worries and pains. Everyone has different definitions of happiness. Some people say that eating delicious food is a kind of happiness; others say that family reunion is a kind of happiness; others say that the birth of a new life is a kind of happiness; some people even say that being praised by others is a kind of happiness.

However, many people think that happiness is momentary. This is because we always build happiness on external objects. In life, as long as you can constantly surpass and break through yourself, this will make your life full of hope and happiness, so that you can spend every day with a positive mindset. As long as everyone jumps out of the box and takes a closer look at yourself, you will understand that all the suffering in life, tiredness, happiness all depends on a person's mindset.

At this moment, we are curious about what are the factors that could influence our happiness. The objective of the project is to explore the relationship between happiness and the factors. Besides, we expect to see the influence of different factors on happiness. In this report, we will start with an introduction about happiness and the reason we chose this dataset. Then, we will continue this report with the data analysis of 4 different tests which are hypothesis sample test, correlation test, regression test and ANOVA test. Lastly, this report will end with a conclusion of this project.

DATASET

For our project, we decided to use the 2019 World Happiness Report which is about the score of happiness in 156 countries. In the datasets, there are six factors that affect happiness such as social support, healthy life expectancy and generosity. We chose this data because we find it interesting as we never know that there are happiness scores for each country. Apart from that, we are eager to find the relationship between all of the factors with the happiness score given. Below are the datasets we use;

Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393
2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41
3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
4	Iceland	7.494	1.38	1.624	1.026	0.591	0.354	0.118
5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
6	Switzerland	7.48	1.452	1.526	1.052	0.572	0.263	0.343
7	Sweden	7.343	1.387	1.487	1.009	0.574	0.267	0.373
8	New Zealand	7.307	1.303	1.557	1.026	0.585	0.33	0.38
9	Canada	7.278	1.365	1.505	1.039	0.584	0.285	0.308
10	Austria	7.246	1.376	1.475	1.016	0.532	0.244	0.226
11	Australia	7.228	1.372	1.548	1.036	0.557	0.332	0.29
12	Costa Rica	7.167	1.034	1.441	0.963	0.558	0.144	0.093
13	Israel	7.139	1.276	1.455	1.029	0.371	0.261	0.082
14	Luxembourg	7.09	1.609	1.479	1.012	0.526	0.194	0.316
15	United Kingdom	7.054	1.333	1.538	0.996	0.45	0.348	0.278
16	Ireland	7.021	1.499	1.553	0.999	0.516	0.298	0.31
17	Germany	6.985	1.373	1.454	0.987	0.495	0.261	0.265
18	Belgium	6.923	1.356	1.504	0.986	0.473	0.16	0.21
19	United States	6.892	1.433	1.457	0.874	0.454	0.28	0.128
20	Czech Republic	6.852	1.269	1.487	0.92	0.457	0.046	0.036
21	United Arab Emirates	6.825	1.503	1.31	0.825	0.598	0.262	0.182
22	Malta	6.726	1.3	1.52	0.999	0.564	0.375	0.151

Figure 1 data sets on 2019 World happiness report

In addition we have to do a few statistical tests which are hypothesis one sample test, correlation, regression and anova. Due to that, we have chosen the parameters to come out with a hypothesis statement. First of all for hypothesis one sample test, we chose the happiness score to compare with the population mean of the scores because we want to prove that the mean population is wrong. We predict that the sample score is larger than the population mean score. Then, to investigate the relationship between happiness score and factors that affect happiness, we used correlation tests and chose GDP per capita as we wanted to know whether income can affect happiness or not. We predict that there will be a strong relationship between the two variables. Next, we wish to investigate whether health life expectancy can affect happiness score or not. We use regression to test this as it can indicate the impact of a change between dependent and independent variables. We predict that health life expectancy will affect happiness scores. Lastly, we wish to check the similarity between factors that affect happiness. We chose the means for social support, freedom to make life choices and generosity. We predict that there will be differences in the mean value. The following will reveal our calculations and outcomes, as well as whether or not they will match our expectations.

DATA ANALYSIS

1.1 Hypothesis Sample test

To test whether the mean happiness score for each country is greater than the claimed population mean which is 4.5.

$$H_0: \mu = \mathbf{4.5}$$
$$H_1: \mu > \mathbf{4.5}$$

$$\alpha = 0.05$$
$$\overline{x} = 5.41$$
$$s = 1.11$$
$$n = 156$$

$$z = \frac{\overline{x} - \mu}{s / \sqrt{n}} = 10.178$$

$$c.v.z_{0.05} = 1.645$$

$$p(z > 10.178) = 0$$

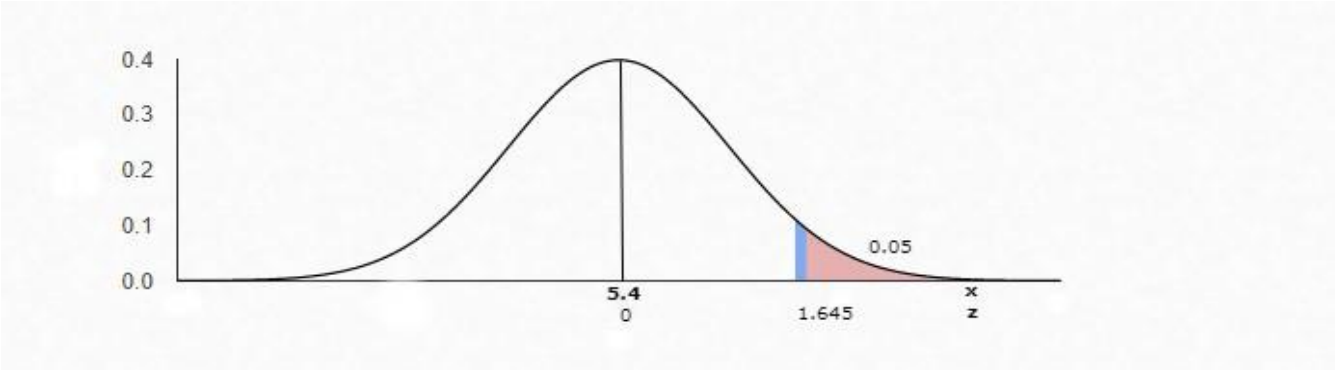


Figure 1

Conclusion :

From the dataset we chose ,the size of the sample is > 30 with variance unknown, so we decided to use Z-test since it is normally distributed. The result of Z-test shows that, test statistics = 10.178 > 1.645, reject H_0. There is sufficient evidence to conclude that the mean happiness score for each country is greater than the claimed population mean which is 4.5. Based on figure 1, the

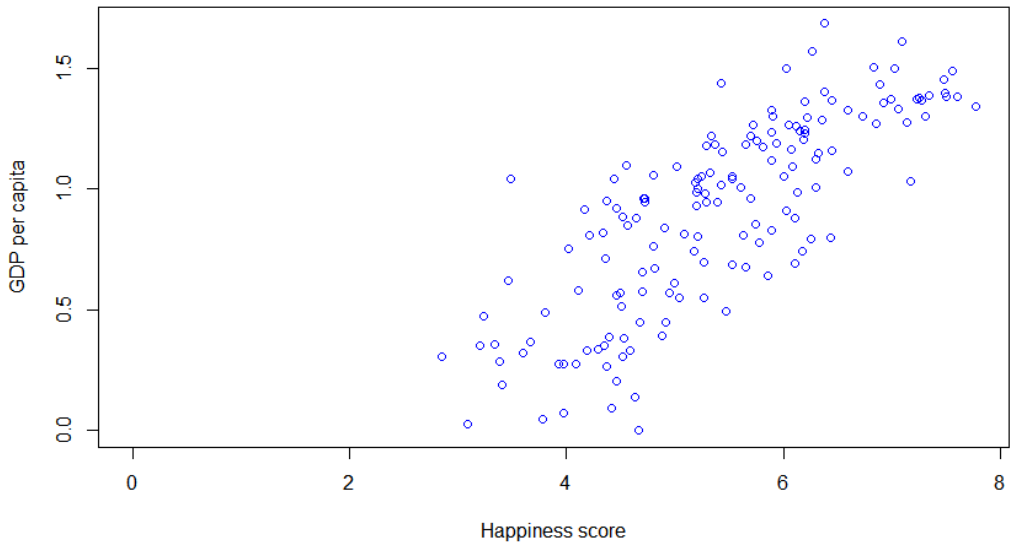
shaded region is a critical region where we reject H_0. Since 10.178 belong in the shaded region, this clearly shows that we should reject H_0.

1.2 Correlation

To measure the strength of the linear relationship between happiness score and GDP per capita.

$$H_0 : \rho = 0$$

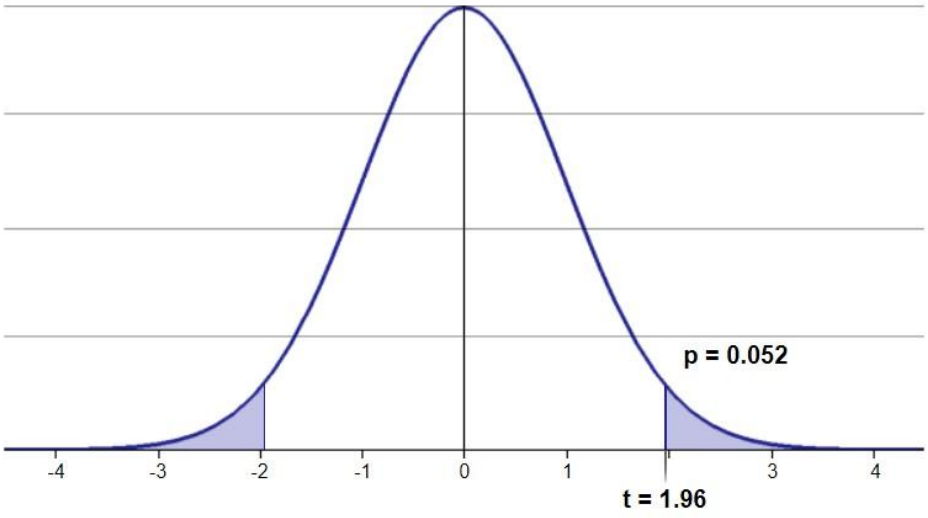
$$H_1 : \rho \neq 0$$



r = 0.7938829

$\alpha = 0.05$

df = n-2 = 154



$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = 16.202$$

p-value = 0

$c.v.\ t_{df,-a/2} =$ -1.960

$c.v.\ t_{df,-a/2} =$ 1.960

Conclusion:

It can be seen that the happiness score increases as the GDP per capita increases. The scatter plot and correlation analysis of the data, r = 0.7549057 is close to 1 and thus indicates that there is a moderate positive linear relationship between happiness score and GDP per capita. Since statistic value = 16.202 > 1.960 and p-value < 0.052. We reject H_0 at a significant level of 0.05. There is sufficient evidence of a linear relationship between happiness score and GDP per capita.

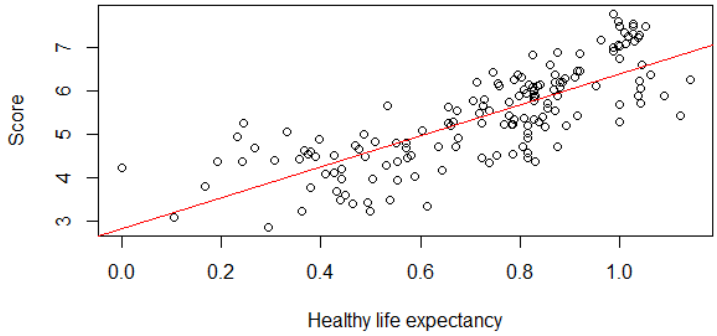
1.3 Regression

To examine the relationship between healthy life expectancy and the score in all the countries.

- Dependent variable: Score
- Independent variable: Healthy life expectancy

$$H_0: B_1 = 0$$

$$H_1: B_1 \neq 0$$



$\hat{y} = 2.807 + 3.585x$

$b_0 = 2.807$

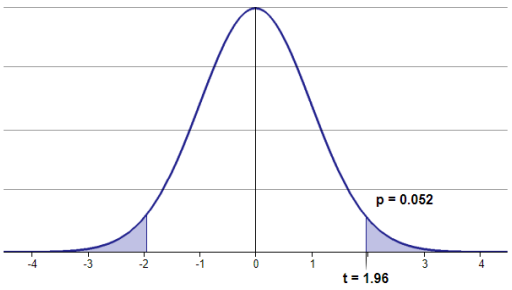
$b_1 = 3.585$

$R^2 = 0.6082$

$S_{b1} = 0.2319$

$\alpha = 0.05$

$df = n - 2 = 154$



$$t = \frac{b_1 - B_1}{S_{b_1}} = 15.46$$

$t_{-a/2} = -1.960$

$t_{a/2} = 1.960$

P-value = 0

Conclusion:

Since the statistic value = 15.46 > 1.96 and the P-value = 0 < 0.052, the H_0 is rejected at the significance level of 0.05. There is sufficient evidence that healthy life expectancy affects the score in all the countries.

1.4 ANOVA Test

To test the significant differences between the means for social support, freedom to make life choices and generosity for different countries.

Source of Variation	Sums of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	F
Between Treatments	$SSB = \Sigma n_j (\bar{X}_j - \bar{X})^2$	k-1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
Error (or Residual)	$SSE = \Sigma\Sigma (X - \bar{X}_j)^2$	N-k	$MSE = \frac{MSE}{N-k}$	
Total	$SST = \Sigma\Sigma (X - \bar{X})^2$	N-1		

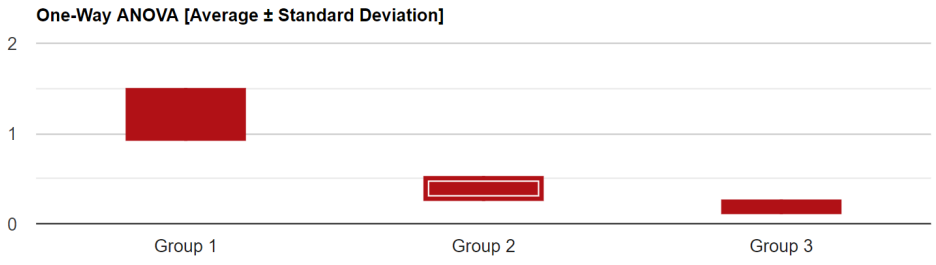
where

- X = individual observation,
- $\bar{X}_j$ = sample mean of the j^{th} treatment (or group),
- $\bar{X}$ = overall sample mean,
- k = the number of treatments or independent comparison groups, and
- N = total number of observations or total sample size.

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

H_1 : at least one mean is different

$$\alpha = 0.05$$

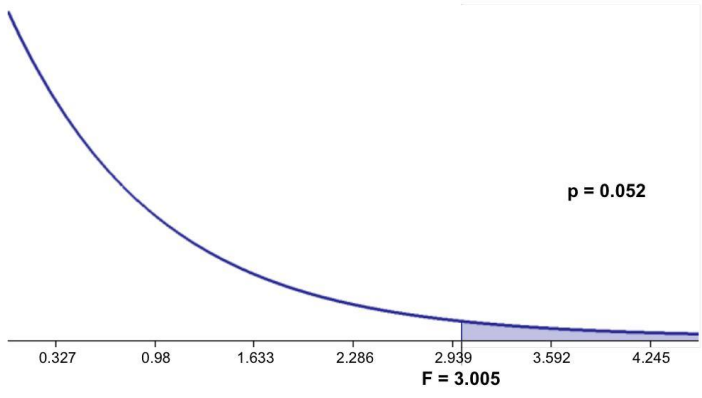


DATA SUMMARY				
GROUPS	**N**	**MEAN, μ**	**STD. DEV.**	**STD. ERROR**

Group 1	156	1.2088	0.2992	0.024
Group 2	156	0.3926	0.1433	0.0115
Group 3	156	0.1848	0.0953	0.0076

ANOVA SUMMARY					
SOURCE	**DEGREE OF FREEDOM, DF**	**SUM OF SQUARES, SS**	**MEAN SQUARE, MS**	**F-STAT**	**P-VALUE**
Between Groups	k-1= 3-1= 2	91.4128	45.7064	1150.9317	0
Within Groups	N-k = 468 - 3= 465	18.4663	0.0397		
Total:	467	109.8792			

k= 3
N= 468



F critical value = 3.0054

Conclusion:

Since the $F_{\text{Test statistic}} > F_{\text{Critical value}}$ (1150.93 > 3.0054) and the P-value = 0 < 0.00001, the null hypothesis, H_0 is rejected at the significance level of 0.05. There is enough evidence that at least one means for social support, freedom to make life choices and generosity for different countries are different. .

CONCLUSION

Based on the data analysis that our group has conducted, we have discovered a lot of interesting findings from our results. From our analysis where we have done the hypothesis sample test, we have discovered that the mean happiness score for each country is greater than the claimed population mean which is 4.5 hence this has proven that our initial hypothesis where we intended to prove that the mean population is wrong. Next, we also conducted a correlation test where we intended to study the relationship between happiness score and the factors that can affect happiness. By using the correlation test with GDP per capita and the happiness score as our variables to prove whether income affects happiness, we discovered that there is a moderate positive linear relationship between happiness score and GDP per capita. Hence, this shows that our initial hypothesis that factors that affect happiness affects the happiness score. After that, we also conducted a regression test to investigate whether health life expectancy can affect the happiness score as it could indicate the impact of change between dependent and independent variables with the prediction that health life expectancy will affect the happiness score. By doing the regression test, we come across that health life expectancy does affect the happiness score which proves our initial prediction. Lastly, we did an ANOVA test to investigate the significant differences between the means for social support, freedom to make life choices and generosity for different countries with a prediction that there will be a difference in the mean values between the three variables. After conducting the test, the ANOVA test results have proven that there is a difference in the mean values for different countries.

In this project, we learn a lot of things and understand the process of different analysis. We had successfully done all four analyses which are hypothesis sample test, correlation, regression and ANOVA analysis. Before we start to do the analysis, we revise the steps to calculate all the variables that we learnt from the class. Then, we start the analysis process which includes all the calculations and write codes using Rstudios. We get to improve our skill to use Rstudio to calculate all the variables and plot all the graphs. From this project, we can see that there are many factors affecting the happiness score in different countries. There might be unforeseen variables that affect the result of the analysis and we need to be concerned about it.