SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143-07 KEBARANGKALIAN STATISTIK & ANALISIS DATA (PROBABILITY & STATISTICAL DATA ANALYSIS)

SEMESTER 2, 2019/2020

Individual Project 2

NAME: Abdellah Oussama

MATRIC NO.: A19EC4041

SUBMITTED TO:

DR. Aryati Binti Bakri

ABSTRACT
This research mainly focuses on the statistical analysis of the world happiness report, providing some insights towards the distribution of the countries based on the happiness of the people, and towards the impact of multiple social and material factors on the happiness of distinct nations in different parts of the world. Numerous statistical methods are used to assure that the conclusion are as accurate as possible.

Introduction

The data used in this case study was released by the UN in 2015. Lately, it has gained recognition from the vast majority of governments and organizations around the world, as it helped them to make some critical decisions and gave them a wide perspective through which they can have a clear understanding of the key factors that affect the prosperity and well-being of their people.

The happiness score presented in the data was obtained by adding up the seven factors following it (Economy – family – Health – Freedom – Trust in government…) These seven factors represent the extent to which they contribute to the calculation of the happiness score. To illustrate this, let's take the example of the economy score in Switzerland : When we say that the GDP score is 1.39651, it doesn't mean that the GDP score is 1.39, it means that this factor (GDP) contributes : (1.396551 / 7.587) * 100 % = **25.52** % to the happiness score.

Methodology

Several tests were used in this case study to find out whether the happiness score varies from region to region or not, and also to determine whether there is a relationship between different factors, in terms of their impact on the calculation of the happiness score. The software tool used in the study is the R programming language, with RStudio as an IDE.

Hypothesis testing - Discussion

1 – Two Samples Test :

In this test, we will try to compare two socially and economically different parts of the world : **Arab world** & **Western Europe,** and the variable we will focus on is the happiness score. The comparison of the mean score of the two regions is actually inefficient, since the result might be trivial. Hence we will try to compare the **standard deviation** of the happiness score in the two zones.
To perform this test, we will grab two samples from the data of equal sizes :
n = 14.
The chosen significance level **α = 0.05.**

Western Europe sample

Country	Region	Happiness.Rank	Happiness.Score
Switzerland	Western Europe	1	7.587
Iceland	Western Europe	2	7.561
Denmark	Western Europe	3	7.527
Norway	Western Europe	4	7.522
Finland	Western Europe	6	7.406
Netherlands	Western Europe	7	7.378
Sweden	Western Europe	8	7.364
Austria	Western Europe	13	7.200
Luxembourg	Western Europe	17	6.946
Ireland	Western Europe	18	6.940
Belgium	Western Europe	19	6.937
United Kingdom	Western Europe	21	6.867
Germany	Western Europe	26	6.750
France	Western Europe	29	6.575

Arab world sample :

Country	Region	Happiness.Rank	Happiness.Score
United Arab Emirates	Middle East and Northern Africa	20	6.901
Oman	Middle East and Northern Africa	22	6.853
Qatar	Middle East and Northern Africa	28	6.611
Saudi Arabia	Middle East and Northern Africa	35	6.411
Kuwait	Middle East and Northern Africa	39	6.295
Bahrain	Middle East and Northern Africa	49	5.960
Libya	Middle East and Northern Africa	63	5.754
Algeria	Middle East and Northern Africa	68	5.605
Turkey	Middle East and Northern Africa	76	5.332
Jordan	Middle East and Northern Africa	82	5.192
Morocco	Middle East and Northern Africa	92	5.013
Lebanon	Middle East and Northern Africa	103	4.839
Tunisia	Middle East and Northern Africa	107	4.739
Palestinian Territories	Middle East and Northern Africa	108	4.715

Hypothesis :

H0 : σ1 = σ2
H1 : σ1 > σ2

σ1 : standard deviation of the happiness score in all of the Arab world
σ2 : standard deviation of the happiness score in all of Western Europe

F Test :

After conduction the F test using the R language, here are the results obtained :

F = 5.3779
P-value = 0.004712 (α = 0.05)
df (numer / denom) = 13

```
> Arab <- c(6.901, 6.853, 6.611, 6.411, 6.295, 5.960, 5.754, 5.605, 5.332, 5.192, 5.013, 4.839, 4.739, 4.715)
> West <- c(7.587, 7.561, 7.527, 7.522, 7.406, 7.378, 7.364, 7.200, 6.946, 6.940, 6.937, 6.867, 6.750, 6.575)
> var.test(Arab , West)

        F test to compare two variances

data:  Arab and West
F = 5.3779, num df = 13, denom df = 13, p-value = 0.004712
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.726445 16.752465
sample estimates:
ratio of variances
          5.377937
```

Discussion :
Since the p-value < 0.05, **we reject H0.** Hence, there is enough evidence to conclude that, the standard deviation of the happiness score in the Arab world, is higher than that of the Western world at 0.05 significance level.

Interpretation :
Considering that the standard deviation is a measure of the average distance between the values of the data in the set and the mean, we can say that the happiness scores of countries in the western world, tend to be centered around the mean value, counter to the Arab world, where the happiness scores of different Arab countries tend to be significantly different from each other. In other words, we can say that the people of the Western world, generally have roughly the same level of happiness, which is by the way very high. However, in the Arab world, you can find people in some countries (mostly rich) lead a very happy life, and other people in some neighbor countries lead a very sad and pathetic life unfortunately.

2 – Correlation analysis:

In this test, we will try to see if there is a relationship between contribution of **trust** and **GDP** to the happiness score, in 30 countries, which have the highest happiness scores. The target population in this test is all the countries with high level of happiness.
The chosen significance level **$\alpha = 0.05$.**

Hypothesis :

H0 : $\rho = 0$ (no linear correlation)
H1 : $\rho \neq 0$ (linear correlation exist)

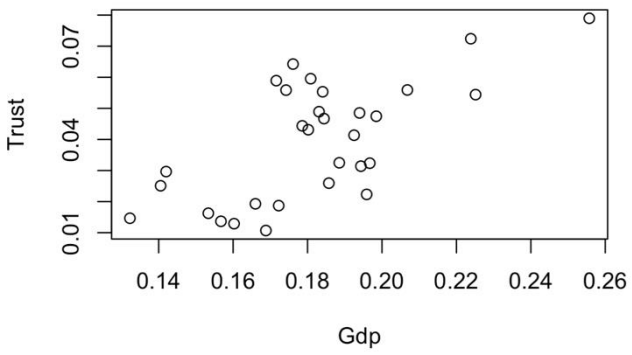
Test :
After conduction the test using R language, here are the results :

- **r = 0.6757853 (Pearson's product-moment correlation coefficient)**
- **t = 4.8513**
- **df = 28**
- **P-value = 0.00004164 ($\alpha = 0.05$)**

```
> # correlation test :
> Gdp <- head(data$Economy..GDP.per.Capita./ data$Happiness.Score, n= 30L)
> Trust <- head(data$Trust..Government.Corruption./ data$Happiness.Score, n= 30L)
> plot(Gdp,Trust)
> cor.test(Gdp,Trust)

        Pearson's product-moment correlation

data:  Gdp and Trust
t = 4.8513, df = 28, p-value = 4.164e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4170539 0.8331998
sample estimates:
      cor
0.6757853
```



Scatter plot

Discussion :

Since the p-value < 0.05, **we reject H0.** Thus, there is enough evidence to conclude that there is a linear relationship between contribution of **trust** and **GDP** to the happiness score, at 0.05 significance level, in the countries of high level of happiness.

And considering that Pearson's product-moment correlation coefficient is positive, we can say that this relation is positive, indicating that the contribution of GDP increases with that of trust and vice versa.

Interpretation :

The positive linear relationship between the contribution of GDP and trust to the calculation of the happiness score in the first ranked countries in terms of happiness score, shows that happy people in happy countries usually associate their good economy with a high level of trust in their governments, and they are very concerned of the level of corruption, because they know that a corrupted government can lead to a disaster, and can alter their level of happiness and satisfaction dramatically.

3 – Regression analysis :

In this test we will try to examine the relationship between the happiness score, and the contribution of the economy to the calculation of this score. Our target population is all the countries of the world, thus we will take a big sample of 158 countries. The purpose of this test is to see whether a high contribution of GDP to the calculation of the happiness score, implies a high value of this score. The chosen significance level **$\alpha = 0.05$**

Hypothesis :

H0 : $\beta = 0$ (no linear relationship)
H1 : $\beta \neq 0$ (linear relationship exist)

Test :

After conduction the test using R language, here are the results :

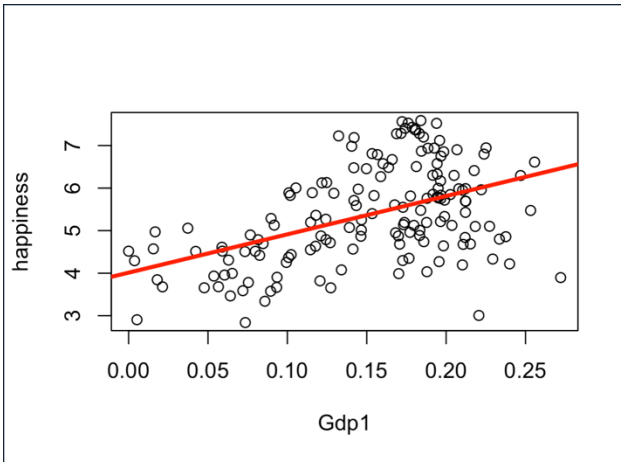
- **t = 18.222**
- **df = 156**
- **P-value < $2 * 10^{-16}$ ($\alpha = 0.05$)**
- **b0 = 4.01**
- **b1 = 9.0055**
- **sb1 = 0.22**
- **$R^2 = 0.22$**

```
> Gdp1 <- c(data$Economy..GDP.per.Capita./ data$Happiness.Score)
> happiness <- c(data$Happiness.Score)
> regression <- lm(happiness ~ Gdp1)
> summary(regression)

Call:
lm(formula = happiness ~ Gdp1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.99273 -0.65636 -0.00859  0.79897  2.02296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0119     0.2202  18.222  < 2e-16 ***
Gdp1          9.0055     1.3528   6.657 4.49e-10 ***
```



R code

Discussion :

We've chosen the happiness score as a dependent variable (y), and the contribution of GDP as the independent variable (x).

The equation we obtained is :

y = 4.01 + 9.0055 x

This equation can actually help us to estimate the value of the happiness score, when the contribution of the economy to the calculation of the score is known. For instance, a country with GDP contributing 20 % to the calculation of the happiness score, have on average a score of :

4.01 + 9.0055 * 0.2 = 5.81

But since the coefficient of determination is $R^2 = 0.22$, the equation is valid for only 22 % of the data, considering that only 22 % of the variation in happiness score is explained by the variation of the contribution of the GDP.

For the t test, we found that **P-value << 0.05** and thus **we reject H0**.

So there is enough evidence to conclude that the change in the contribution of the GDP affects the happiness score

(if we calculate Pearson's product-moment correlation coefficient we find r = 0.47, which indicates a positive relationship).

Interpretation :

Since we found that the change in the extent to which the GDP contributes to the calculation of the happiness score can affect the score itself, we can say that, we are 95 % confident that a country with a high contribution of GDP to the happiness of the people is likely to have a high happiness score; this can mean that if the economy of some country constitute a big part of the happiness of its people, it's very likely that this country has a high level of happiness.

4 – ANOVA :

In this test, we will try to compare the means of the happiness score in three different parts of the world :
Sub-Saharan Africa ; Western Europe ; Eastern Asia
The chosen samples are of a size of n = 6

Finland	Western Europe	6	7.406
Netherlands	Western Europe	7	7.378
Sweden	Western Europe	8	7.364
Austria	Western Europe	13	7.200
Luxembourg	Western Europe	17	6.946
Ireland	Western Europe	18	6.940

Cameroon	Sub-Saharan Africa	133	4.252
Angola	Sub-Saharan Africa	137	4.033
Mali	Sub-Saharan Africa	138	3.995
Congo (Brazzaville)	Sub-Saharan Africa	139	3.989
Comoros	Sub-Saharan Africa	140	3.956
Uganda	Sub-Saharan Africa	141	3.931

Taiwan	Eastern Asia	38	6.298
Japan	Eastern Asia	46	5.987
South Korea	Eastern Asia	47	5.984
Hong Kong	Eastern Asia	72	5.474
China	Eastern Asia	84	5.140
Mongolia	Eastern Asia	100	4.874

Hypothesis :

H0 : μ1 =μ2 =μ3
H1 : At least one mean is different

μ1 : the mean of happiness score in Western Europe
μ2 : the mean of happiness score in Sub-Saharan Africa
μ3 : the mean of happiness score in Eastern Asia

Test :

After conduction the test using R language, here are the results :

- **df (num) = 2**
- **df (denom) = 15**
- **F = 124**
- **P-value = $4.7 * 10^{-10}$ (α = 0.05)**

```
> Europe <- c(7.406, 7.378, 7.364, 7.200,6.946,6.940)
> Africa <- c(4.252, 4.033, 3.995, 3.989, 3.956,3.931)
> Asia <- c(6.298, 5.987, 5.984, 5.474, 5.140, 4.874)
> combined <- data.frame(cbind(Europe, Africa, Asia))
> stacked <- stack(combined)
> ANOVA <- aov(values ~ ind , data = stacked)
> summary(ANOVA)
            Df Sum Sq Mean Sq F value  Pr(>F)
ind          2 30.331  15.166     124 4.7e-10 ***
Residuals   15  1.835   0.122
```

Discussion :

Since P-value << 0.05 (α) , we reject H0. Therefore, there is enough evidence at 0.05 significance level to conclude that different regions have different means of happiness scores.

Interpretation :

We've found that the mean happiness score changes from region to region. Hence it's reasonable to say that the geographic location of a country may significantly affect the happiness of its people. This phenomenon can be the result of different political and economic factors.

Overall conclusion :

The results of the four conducted tests, may seem at first glance unrelated to each other, however if we deeply examine them, we can clearly see a strong relationship, because they all give us some very important criteria for every nation to be happy and prosperous :

- The first one is that solidarity and unity between neighbor countries and the countries that have roughly the same cultures or religion is vital to achieve happiness. This conclusion is drawn from the first test which showed us how the western countries have nearly the same level of happiness because they are helping each other and supporting each other. On the other hand the Arab world is completely disunited, and there you can find some neighbor countries with a huge economic gap ; I think this can present a big controversy, because the Arab world is Muslim, and as we all know our religion strongly emphasizes the importance of solidarity and unanimity, so the only explanation of that, is that we are deviating away from our religion unfortunately.

- The second criterion can be drawn from the second, third and fourth tests, and it's that happy people are often those who fight for their rights, and those who very much care about their government to be uncorrupted and trustworthy.