COVID-19-Geographic-Distribution-Worldwide

Amit Hasan Sadhin(A19EC4006)

Section 07

Lecturer: Dr. Aryati Bakri

School of computing, Faculty of Engineering

Abstract- **At the end of 2019 and the beginning of 2020 world has been faced with a global pandemic called COVID-19. All of all the country in the entire world is struggling with this invisible enemy of human beings. The virus becomes so deadly that it's causing thousands of deaths. According to the world health organization, COVID-19 has a mortality rate of 3.4% which is much more dangerous than we thought in this age of science and technology. Here we calculate all the country data from the beginning of the pandemic to 5th of June 2020.**

I. INTRODUCTION

As we know that a pandemic is defined as "an epidemic occurring worldwide, or over a very wide area, crossing international boundaries and usually affecting a large number of people". The classical definition includes nothing about population immunity, virology or disease severity. Coronavirus appears to tick all of those boxes to become a global pandemic. With no vaccine or treatment that can prevent it yet, containing its spread is vital.

The title of this project-2 is "COVID-19-geographic-distribution-worldwide". I decided to conduct an ongoing study that can help us to understand the geographical road map for COVID-19. As we all that the coronavirus started in Wuhan in Hubei province in China, but after that, it spread all over the world. My study aims are to know which continent and which country are facing form cases and death. I also want to know is there any relation between geographical locations and the cases that increase day by day. As WHO said the mortality rate is 3.4%, I am curious to see the actual death and cases ration by analyzing all

the data that are in the Excel sheet, and in the end, I want to give a clear idea to the country who can be the next target.

II. METHODOLOGY

Total 16322 data for 5 months from the starting of the pandemic were collected from 195 countries in the world. The record was collected by European Union. This data aim is to clarify the next hotspot for the virus as Italy is suffering in Europe.

The parameters and the variables of the data was given below.

Data collected	Data type	Data collected name
date	Ratio & interval	3/11/2020, 5//11/2020 etc.
cases	Ordinal	3,10, 20, etc.
deaths	Ordinal	4,6,8, etc.
Countries And Territories	Nominal	Malaysia, Brunei, etc.
popData2018 (Population)	Ratio	12345632. 32435422, etc.
continentExp	Nominal	Asia, Europe, etc.

III. Result and Discussion

a. Hypothesis testing

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process as our data for the **"COVID-19 Geographical distribution"**. Here, in hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the

opposite of a null hypothesis; e.g., the population mean return is not equal to zero. Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

However, here we take 95% confidence level so for this the null hypothesis would be represented as Ho: mu= 195 .. P = 0.95. The alternative hypothesis would be denoted as "Ha" and be identical to the null hypothesis, except with the equal sign struck-through, meaning that it does not equal 95%.

The R-script for Hypothesis testing,

For one sided:

```
#Ho: mu< 195
# One-sided 95% confidence interval for mu
t.test(cases,
mu=195,alternative="less",conf.level = 0.95)
```

output:

```
        One Sample t-test

data:  cases
t = 4.1712, df = 16321, p-value = 1
alternative hypothesis: true mean i
s less than 195
95 percent confidence interval:
     -Inf 270.2396
sample estimates:
mean of x
   248.96
```

As our confidence level is 95% and total data is 16322, so degree of freedom is 16321. According to the R-studio output file we got population mean 248.96. On the other hand, `t = 4.1712` and p-value = 1 proved that we can't reject the null hypothesis.

For two sided:

```
t.test(cases,
mu=195,alt="two.sided",conf.level = 0.95)
```

output:

```
Two Sample t-test

data:  cases
t = 4.1712, df = 16321, p-value = 3
.046e-05
alternative hypothesis: true mean i
s not equal to 195
95 percent confidence interval:
 223.6034 274.3166
sample estimates:
mean of x
   248.96
```

Two-sided Hypothesis test also proved that; we can't reject null hypothesis.

b. Correlation:

Correlation shows the strength of a relationship between two variables and is expressed numerically by the correlation coefficient. The correlation coefficient's values range between -1.0 and 1.0. A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one security moves, either up or down, the other security moves in lockstep, in the same direction. A perfect negative correlation means that two assets move in opposite directions, while a

zero correlation implies no linear relationship at all. In short Correlation, is a statistic that measures the degree to which two securities move in relation to each other. For example, in our project (**COVID-19 Geographical distribution**) we use case and deaths as two variables.

R-Script:

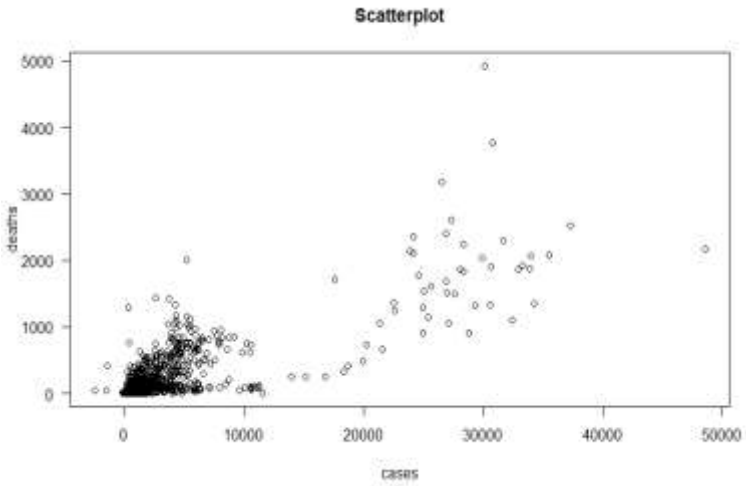
```
class(cases)
class(deaths)
plot(cases, deaths, main= "Scatterplot", las=1)
cor(cases, deaths, method = "pearson")
cor.test(cases, deaths, method = "pearson")
```

Output:

```
Pearson's product-moment correlation

data:  cases and deaths
t = 201.91, df = 16320, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8406187 0.8493915
sample estimates:
      cor
0.845062
```

As we can see that the correlation coefficient is 0.845062 which is a positive correlation and provide that means when the cases increase the deaths also increase significantly.



As per we see in the R-output code that the correlation coefficient is positive and now in the scatterplot we can see that when the cases are increasing the deaths numbers are significantly increases. At the beginning of the scatter plot we see that almost all the countries cases and deaths ratio are decently but as the time passed the case are more likely to increases as well as the deaths numbers.

c. Regression:

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A.

The regression equation representing how much y-axis changes with any given change of x-axis can be used to construct a

regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right.

R-script:

```
class(cases)
plot(cases, deaths)
cor(cases, deaths)
mod <- lm(cases ~ deaths)
summary(mod)
abline(mod)
abline(mod, col=2, lwd=3)
```

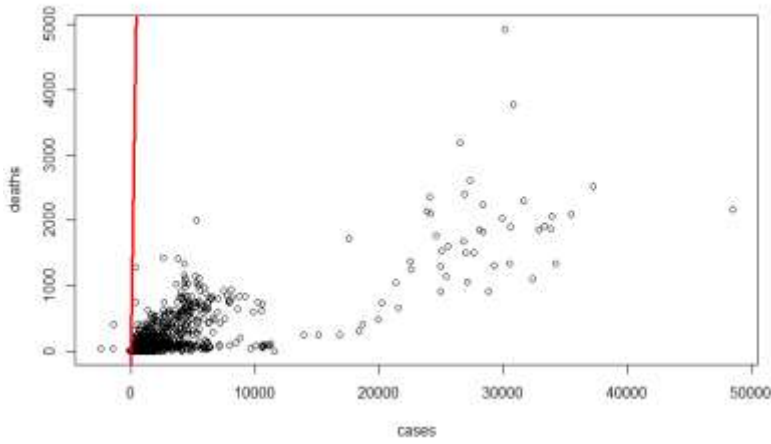
Output:

```
Call:
lm(formula = cases ~ deaths)

Residuals:
     Min       1Q   Median       3Q      Max
-24808.7    -56.3    -56.3    -41.3  24275.5

Coefficients:
            Estimate Std. Error t value Pr(>
|t|)
(Intercept) 56.31558    6.98236   8.065  7.8
e-16 ***
deaths      11.14051    0.05517 201.913  < 2
e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 883.7 on 16320 degr
ees of freedom
Multiple R-squared:  0.7141, Adjusted R-squar
ed:  0.7141
F-statistic: 4.077e+04 on 1 and 16320 DF,  p
-value: < 2.2e-16
```



As we see that our correlation is positive and its almost the highest value of correlation now its clear that our regression will be positive and the straight line will be straight to the up. As we can see in the graph which is a straight line.

d. Chi-square test:

There are basically two types of random variables and they yield two types of data: numerical and categorical. A chi square (X2) statistic is used to investigate whether distributions of categorical variables differ from one another. Basically, categorical variable yield data in the categories and

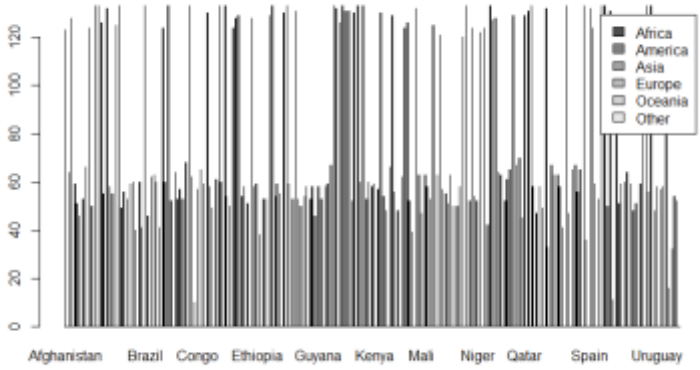
numerical variables yield data in numerical form.

R-script:

```
class(continentExp)
class(countriesAndTerritories)
levels(continentExp)
levels(countriesAndTerritories)
table(continentExp,
countriesAndTerritories)
Tab<- table(continentExp,
countriesAndTerritories)
barplot(Tab, beside = T, legend=T)
chisq.test(Tab, correct = T)
CHI<- chisq.test(Tab, correct = T)
CHI$expected
```

Output:

```
countriesAndTerritories
continentExp  Venezuela    Vietnam Western_Sahara      Yemen     Zambia   Zimbabwe
     Africa  10.8345791 24.0975983      2.98884941  5.9776988 10.0873667  9.7137606
     America 11.0229139 24.5164808      3.04080382  6.0816076 10.2627129  9.8826124
     Asia    14.8677858 33.0680064      4.10145815  8.2029163 13.8424213 13.3297390
     Europe  19.0431320 42.3545521      5.25327778 10.5065556 17.7298125 17.0731528
```



As we are taking 195 countries data from all geographic location of the earth. That's the reason we divided the data into Continent wise as Africa, America, Asia, Europe, Oceania and Others. Here we have two types of data Countries and territories and the data Continent which we have overall 6. The output of the data describes about the continent based Covid-19 spreading demography. As we know that the virus started at Asia and then spread all over the world but the graph shows totally different things which proved that the Europe is the worst victims from the starting of May 2020.

e. ANOVA:

Analysis of Variance (ANOVA) is a parametric statistical technique used to compare datasets. This technique was invented by

R.A. Fisher, and is thus often referred to as Fisher's ANOVA, as well. It is similar in application to techniques such as t-test and z-test, in that it is used to compare means and the relative variance between them. However, analysis of variance (ANOVA) is best applied where more than 2 populations or samples are meant to be compared as in our data **"COVID-19 Geographical distribution".**

For ANOVA after the hypothesis test we just need to write ANOVA command on r-script which is **anova(mod)**

Output:

```
Analysis of Variance Table

Response: cases
             Df     Sum Sq    Mean Sq F
value    Pr(>F)
deaths        1 3.1836e+10 3.1836e+10
40769 < 2.2e-16 ***
Residuals 16320 1.2744e+10 7.8089e+05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*
' 0.05 '.' 0.1 ' ' 1
```

Here, Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

As we can see in the `Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`

Which is gradually increases up to 1 which means we can't reject null Hypothesis.

IV. Conclusion

Based on the tests that we have done in this project **"COVID-19 Geographical distribution"** we can clearly say that the virus COVID-19 is spreading faster day by day. As we saw in the data while at the beginning of the year 2020 the number of cases is more likely in a very small number. As we saw from the correlation scatter plot that as much as the cases increase the number of deaths is also increasing which can cause a huge impact in the near future of this pandemic. The global death percentages are gradually increased as time passed.

As we first mentioned that the world health Organization said that the mortality rate of COVID-19 is 3.4% but after our analysis, we can see that 3.4% is the general number in the entire world whereas we have some countries where there were far difference that this. So, 3.4% can't be a mortality rate for this virus. Because in some countries like Italy has a very higher mortality than another countries.

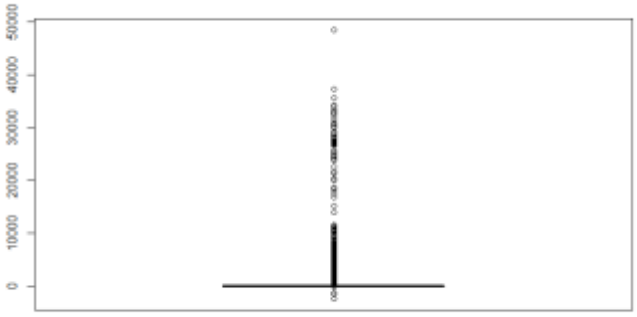
As shown in the scatterplot of regression we saw, although Asia is the origin of the virus, Europe is the worst sufferer of the pandemic.

But this is not finished yet, as shown in the correlation the next hotspot can in Asia because the graph gradually increasing over time because of the high population's country like India, China, Indonesia, Bangladesh, etc.

As shown in the Chi-square test the country which has more population the percentages of spreading the virus is high.

In the end, we can say that the problem is a global disaster and we are living in a world which connected to every nation. We are living in a global village where we need to co-operate with each other and we have to find out the solution to our problems. As the virus is spreading faster, we should take the precautionary arrangement to fight with this invisible enemy.

V. Appendix



This the boxplot of our data as the data is for 195 countries so this is a generalized box-plot for the entire survey data.

R-script:

```
class(cases)
table(cases)
mean(cases)
boxplot(cases)
```

Reference

Data collected from https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-disbtribution-worldwide-2020-05-12.xlsx

- Barron's AP Statistics, 8th Edition;Written by-Martin Sternstein, PhD.
- Statistics; Written by-Robert S. Witte and John S. Witte
- OpenIntro Statistics;Written by-David M Diez, Mine Çetinkaya-Rundel, and Christopher D Bar
- Head First Statistics – A Brain-Friendly Guide;Written By – Dawn Griffiths
- All of Statistics A Concise Course in Statistical Inference ; Written by: Larry Wasserman
- Encyclopedia of Statistical Sciences, A to Circular Probable Error by Samuel I. Kotz (Editor-In-Chief); Norman L. Johnson; Campbell B. Read (Associate Editor)
- Discovering Statistics Using R by Andy Field; Jeremy Miles; Zoe Field
- Statistics in Plain English, Fourth Edition by Timothy C. Urdan