PSDA PROJECT 2

PROBABILITY AND STATISTICAL DATA ANALYSIS

PROJECT 2: COVID-19 IN THE USA

NAME: CHEONG CHIEN LI

MATRIC NO: A19EC0186

SECTION: 06

LECTURER: DR CHAN WENG HOWE

INTRODUCTION

On the 31st of December, 2019, the Wuhan Municipal Health Commission in China reported a cluster of cases of pneumonia in Wuhan, Hubei Province. This eventually led to the discovery of a new novel coronavirus called Covid-19. Before long, this virus had already spread outside of China into almost every part of the World. A global pandemic has been caused due to the outbreak of this virus. At first, this virus only seemed like a normal flu, however, it eventually started killing more and more of those who were infected with it. Since March 26th 2020, the United States of America had become the country that has the most confirmed cases in the world. Hence, the purpose of this study is to study the changes of the Covid-19 pandemic over time in USA, in terms of its severity and case fatality rate depending on the duration of the pandemic, and also to study the relationship between these variables. The target population of this study is the people of the United States of America who are infected by the Covid-19 virus from 31st December 2019 to 19th June 2020, and no further than that. Some of the variables that will be focused on more in this study are Duration of pandemic, Month, number of confirmed cases per day, number of confirmed deaths per day, and also case fatality rate. A few statistical tests will be carried out in this study in order to determine the validity of some claimed statements about Covid-19 in the USA. This study hopes to confirm or deny some claims about the virus in the USA and also to learn about the relationship between some of the variables that are related to this global pandemic.

CONTENT

Focus on topic

The few tests that will be carried out in this study are:

1) **Two samples mean hypothesis testing. (population variance unknown, assumed to have equal variance – t-distribution with standard error)**

   The purpose of this test is to determine whether or not the mean number of deaths per day has increased in the month of April compared to the month of March in the USA.

2) **Significance test for correlation using Pearson's correlation coefficient**

   Based on the conclusion of the previous test, the purpose of this test is to determine whether or not there exists a linear correlation between the duration of the pandemic (number of days) and the deaths per day, and if it exists, this test aims to determine the strength of that relationship.

3) **Regression analysis using simple regression model**

   Continuing from the previous test, the purpose of this test is to provide a model to predict the number of deaths per day after a specific duration of the pandemic that is not provided by the dataset, and also to explain the impact of the duration of the pandemic on the number of deaths per day.

4) **Chi-Square test for goodness of fit**

   Following the previous tests, the purpose of this test is to determine whether or not the case fatality rate (total confirmed deaths / total confirmed cases) has risen over time.

5) **<u>One-way ANOVA with equal sample sizes</u>**

The purpose of this test is to determine whether or not the mean number of deaths per day in the top 3 Covid-19 hotspot countries (USA, Russia, Brazil) are similar to each other.

Support for topic

NOTE: All calculations in this section are done using the RStudio software.

All tests done in this section assumes that the significance level, $\propto = 0.05$, which means that for all of the tests that will be done, we are 95% confident that the conclusions drawn after completing the tests, are true.

1) **<u>Two samples mean hypothesis testing. (population variance unknown, assumed to have equal variance – t-distribution with standard error)</u>**

In this test, we will be testing whether or not the mean deaths per day in the USA has increased in April since March. For this test, we do not know the actual value for the population variance, but it can be assumed that the population variances for both the March sample and the April sample are equal, because the data is basically collected from the same population, just from different timelines where one sample is collected during the month of March while the other sample is collected from the month of April.

Variables in this test are: Month (March and April), and Confirmed deaths per day.

Hence, a two-sample test on mean using the t-distribution with standard error will be used to test the following statements:

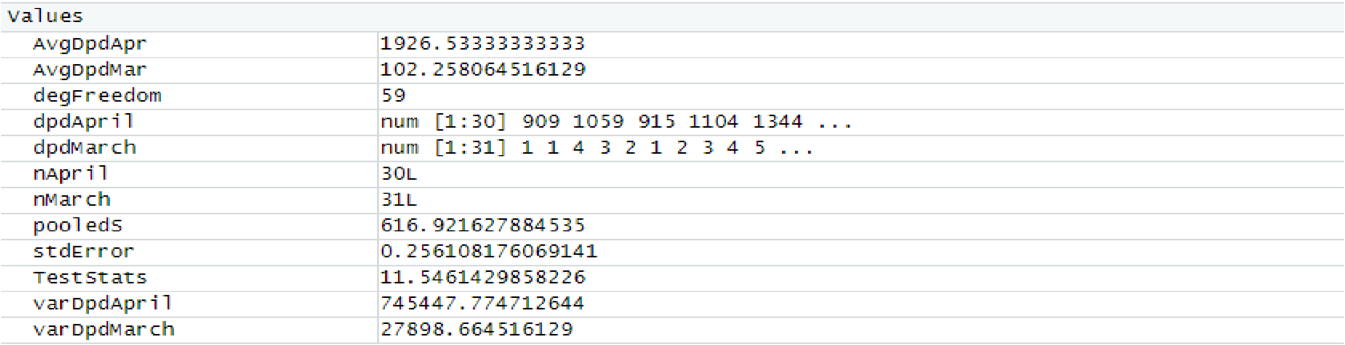
Null Hypothesis, H_0: The mean deaths per day in April is the same as in March in the USA.

$$H_0: \mu_a = \mu_m$$

Alternate Hypothesis, H_1: The mean deaths per day in April has increased since March in the USA.

$$H_1: \mu_a > \mu_m$$

After some calculations done using the RStudio software (RScript file provided in zip file), these are the values of the computed variables:



Values	
AvgDpdApr	1926.53333333333
AvgDpdMar	102.258064516129
degFreedom	59
dpdApril	num [1:30] 909 1059 915 1104 1344 ...
dpdMarch	num [1:31] 1 1 4 3 2 1 2 3 4 5 ...
nApril	30L
nMarch	31L
pooledS	616.921627884535
stdError	0.256108176069141
TestStats	11.5461429858226
varDpdApril	745447.774712644
varDpdMarch	27898.664516129

As we can see, the Test-Statistics value is 11.546, while the Critical value should be $t_{0.05,59}$

Due to the degrees of freedom being larger than 29, we had to refer to the Z-score to get the critical value, $Z_{0.05} \approx 0.13$ which is significantly smaller than 11.546. Since the test-statistics is larger than the critical value, it falls into the rejection region, thus the null hypothesis must be rejected.

Conclusion: H_0 is rejected. There is sufficient evidence to show that the mean deaths per day in the month of April has increased since March in the USA.

Hence, via this test, we were able to confirm the fact that the mean number of deaths per day has indeed increased in the month of April since the month of March.

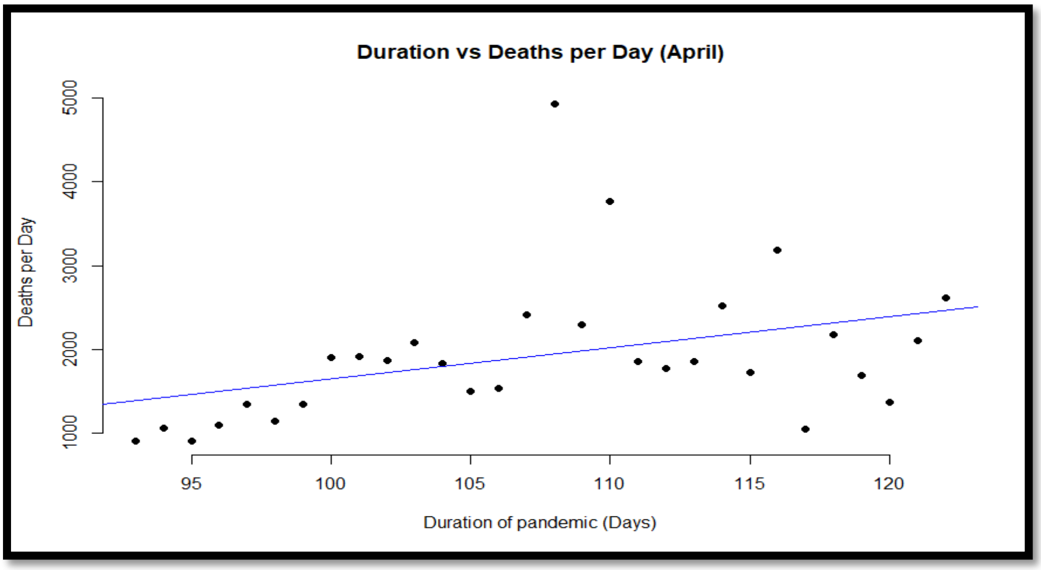
Since now that we know the number of deaths has increased in a month, we can perform a test to see whether or not there exists a relationship between the duration of the pandemic (measured in days starting from 31st of December 2019), and deaths per day. In order to determine the existence of a relationship between these two variables (duration, and deaths per day), we will carry out the next test:

2) Significance test for correlation using Pearson's correlation coefficient.

As mentioned above, the purpose of this test is to test whether or not there exists a linear relationship between the duration of the pandemic (Starting from 31st December 2019) and the deaths per day in the USA, and to determine the strength of this relationship (if any). Note that in this analysis, we are only focusing on determining the existence of a linear correlation, and determining the strength of the relationship (if any), thus we will not be explaining how the changes in an independent variable will affect the dependent variable in this analysis. The explanations will be given later during the regression analysis.

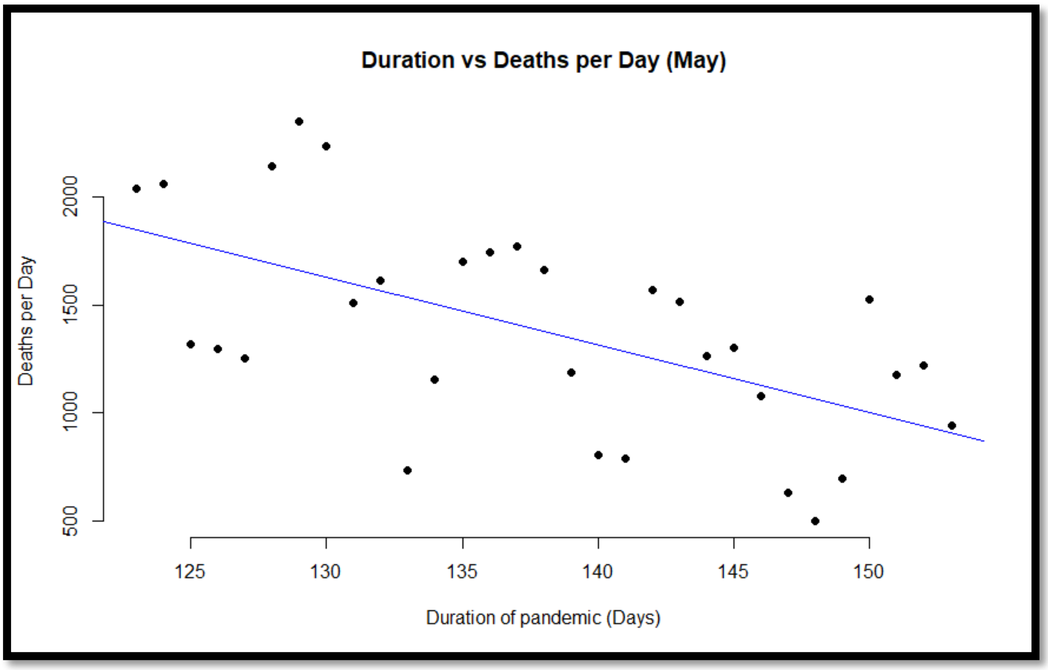
Before we are able to carry out the significance test to see whether or not there exists a relationship between the variables, we must first calculate the correlation coefficient, r.

Hence, to determine the value of r, this test will be carried out by firstly plotting a scatter plot to see whether or not there exists a linear correlation between the variables. We will first look at the scatter plot from the month of April:



From this scatter plot, we can see that there exists a positive linear correlation between the variables. Hence, we can utilise correlation analysis using Pearson's correlation coefficient in order to determine the strength of this relationship.

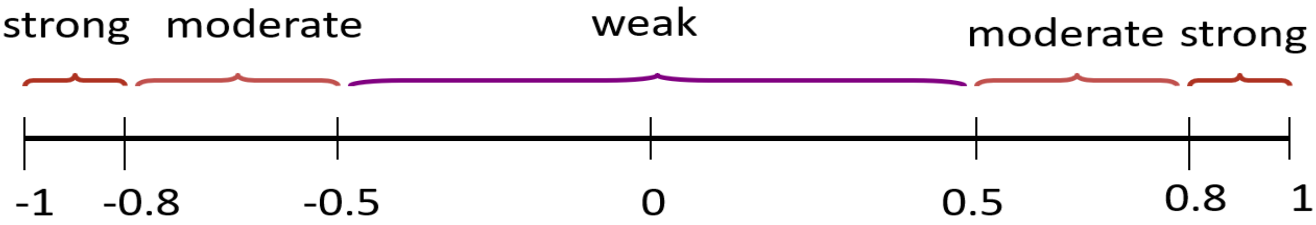
However, the scatter plot for the month of May is different from in April, in order for us to see the difference, we will also plot the scatter plot for the data collected in the month of May:



From this plot, we can clearly see a negative linear correlation between the variables. We will study the impact of the independent variable on the dependent variable later during the regression analysis. As for the current correlation analysis, we are only interested in determining whether there exists a linear correlation between the variables and also the strength of the relationship (if any). Nevertheless, we can still clearly see from the plot above that there exists a linear relationship between the variables. Hence, correlation analysis can be used to determine the strength of this linear relationship.

No matter positive or negative, we are still certain that a linear relationship exists between the duration of the pandemic and the deaths per day. Hence for this test, we will first perform correlation analysis using Pearson's correlation coefficient on the data in April in order to determine the strength of this relationship. Then, we will carry out the exact same test on the data in May in order to see whether or not the magnitude that indicates the strength of this relationship is similar in both of these months.

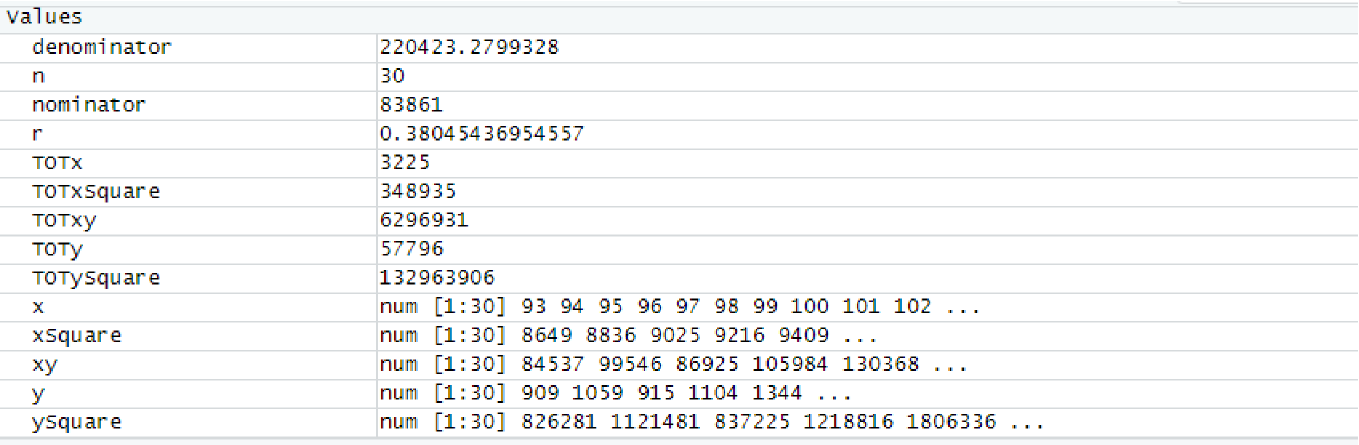
The magnitude of the correlation coefficients will be interpreted using the scale:



If the magnitude of correlation coefficient of the month of May falls into the same strength range as the magnitude in the month of April, then we can conclude that the strength of this relationship is relatively similar in both of these months.

Correlation analysis on the data in the month of April:

After some calculations using RStudio (RScript provided in zip file), these are the values:



```
Values
  denominator     220423.2799328
  n               30
  nominator       83861
  r               0.38045436954557
  TOTx            3225
  TOTxSquare      348935
  TOTxy           6296931
  TOTy            57796
  TOTySquare      132963906
  x               num [1:30] 93 94 95 96 97 98 99 100 101 102 ...
  xSquare         num [1:30] 8649 8836 9025 9216 9409 ...
  xy              num [1:30] 84537 99546 86925 105984 130368 ...
  y               num [1:30] 909 1059 915 1104 1344 ...
  ySquare         num [1:30] 826281 1121481 837225 1218816 1806336 ...
```

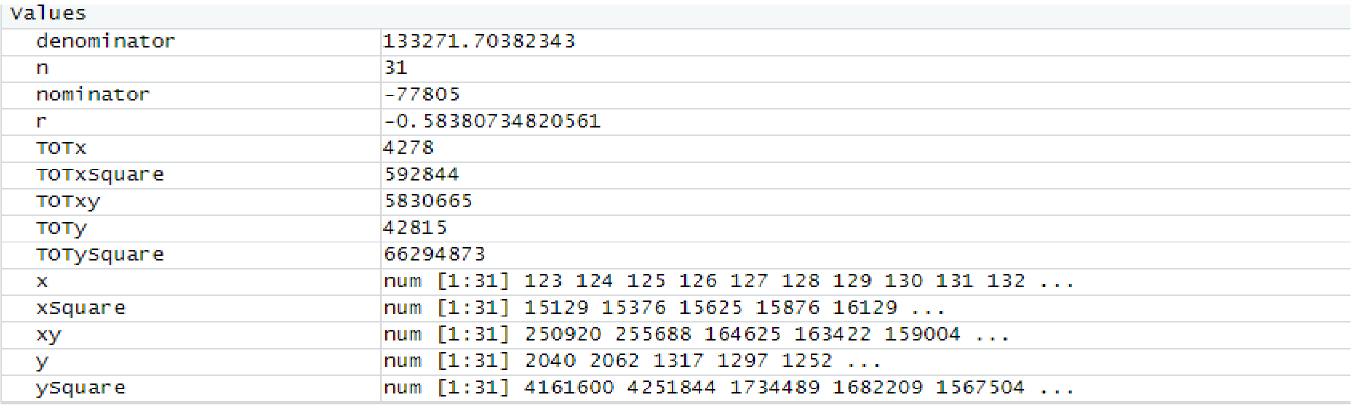
From this, we can see that the correlation coefficient value, r is equal to + 0.38

This r value falls within the range of 0 < r < + 0.5 which tells us that the strength of the positive linear correlation between duration of pandemic and deaths per day is relatively weak.

Next, we will perform the same calculations on the data in May to see whether or not the correlation coefficient falls into the same strength interval.

Correlation analysis on the data in the month of May:

After some calculations using RStudio, here are the values:



```
Values
  denominator     133271.70382343
  n               31
  nominator       -77805
  r               -0.58380734820561
  TOTx            4278
  TOTxSquare      592844
  TOTxy           5830665
  TOTy            42815
  TOTySquare      66294873
  x               num [1:31] 123 124 125 126 127 128 129 130 131 132 ...
  xSquare         num [1:31] 15129 15376 15625 15876 16129 ...
  xy              num [1:31] 250920 255688 164625 163422 159004 ...
  y               num [1:31] 2040 2062 1317 1297 1252 ...
  ySquare         num [1:31] 4161600 4251844 1734489 1682209 1567504 ...
```

From this, we can see that the correlation coefficient, r is equal to – 0.58

This r value falls within the range of – 0.8 < r < – 0.5 which tells us that the strength of the negative linear correlation between the variables is moderate.

From the r values calculated above, we can conclude that the positive linear correlation between the duration of the pandemic and the number of deaths per day is relatively weak before a peak is reached. However, once the peak has been reached, the strength of the negative linear correlation will be slightly stronger. Further explanations will be given in the regression analysis that will be done later in the study.

Now that we have calculated the r values of both of these months, we can now carry out the significance test to properly test and prove the existence of a linear correlation between the duration of the pandemic and the number of deaths per day. The significance test will be carried out twice; once for the month of April, and another for the month of May. This is to test whether the existence of the linear correlation is true for both of the months.

Significance test for correlation in the month of April:

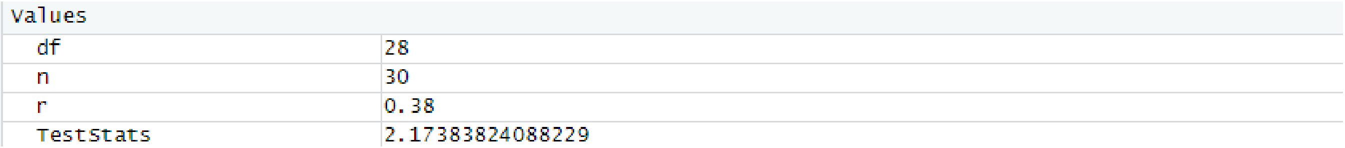
Null Hypothesis, $\boldsymbol{H_0}$: No linear correlation.

$$H_0 : \boldsymbol{\rho} = \ \mathbf{0}$$

Alternate Hypothesis, $\boldsymbol{H_1}$: Linear correlation exists.

$$H_1 : \boldsymbol{\rho} \neq \mathbf{0}$$

After some calculations done using RStudio (RScript file provided in zip file), here are the values:



```
Values
  df                            28
  n                             30
  r                             0.38
  TestStats                     2.17383824088229
```

As we can see, the test statistic is approximately equal to 2.174, while the critical value,

$$t_{\frac{\alpha}{2},df} = \ t_{0.025,28} = 2.084 < 2.174$$

Since the test statistics is larger than the critical value, it falls into the region of rejection. Hence, the null hypothesis must be rejected.

Conclusion for April: **$\boldsymbol{H_0}$ is rejected.** There is sufficient evidence to show that there exists a linear correlation between the duration of the pandemic and the number of deaths per day.

Next, we will perform the same significance test on the data of May to determine whether or not there also exists a linear correlation between the variables.

Significance test for correlation in the month of May:

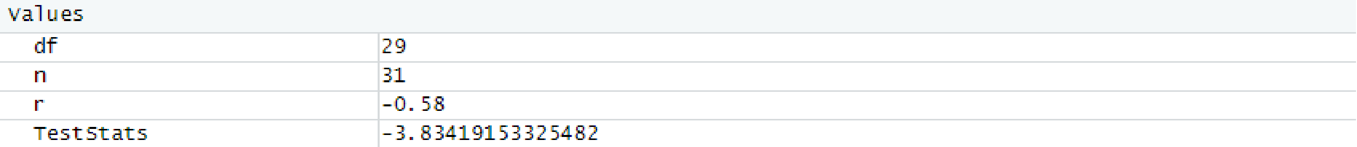
Null Hypothesis, $\boldsymbol{H_0}$: No linear correlation.

$$H_0: \boldsymbol{\rho} = \mathbf{0}$$

Alternate Hypothesis, $\boldsymbol{H_1}$: Linear correlation exists.

$$H_1: \boldsymbol{\rho} \neq \mathbf{0}$$

After some calculations done using RStudio (RScript file provided in zip file), here are the values:



Values	
df	29
n	31
r	-0.58
TestStats	-3.83419153325482

As we can see, the test statistic is approximately equal to -3.834, while the critical value,

$t_{\frac{\alpha}{2},df} = t_{0.025,29} = -2.045$ > -3.834

Since the test statistics is smaller than the critical value, it falls into the region of rejection. Hence, the null hypothesis must be rejected.

Conclusion for May: **$\boldsymbol{H_0}$ is rejected.** There is sufficient evidence to show that there exists a linear correlation between the duration of the pandemic and the number of deaths per day.

Hence, since we were able to prove the existence of a linear correlation between the variables for both of these months by using correlation analysis, it can be concluded that there exists a weak positive linear correlation between the duration of the pandemic and the number of deaths per day before a specific peak is reached, and a moderate (slightly stronger) negative linear correlation between the variables after the peak is reached. However, this does not tell us any information about which variable causes the change of which.

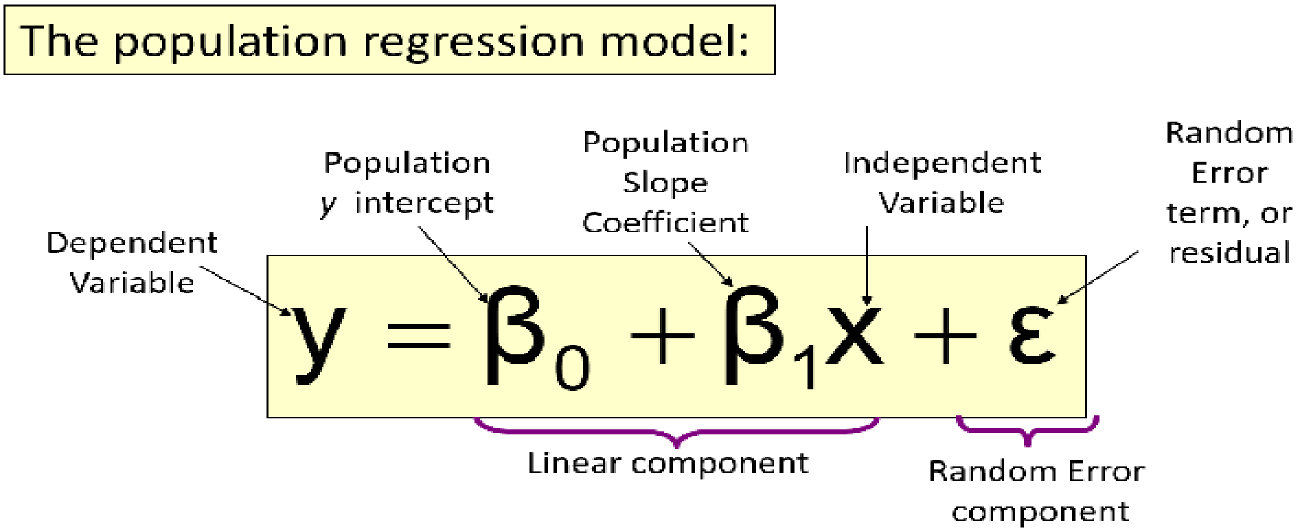
Next, we can carry out a regression analysis, which will mainly allow us explain the impact of a change in the independent variable on the dependent variable, and also to predict the value of the dependent variable according to the independent variable that might not be given by our dataset.

3) Regression analysis using simple regression model

In this analysis, it is to be noted that the independent variable is the duration of the pandemic (days), while the dependent variable is the number of deaths per day.

Now that we know that there exists a linear correlation between the duration of the pandemic and the number of deaths per day, we can perform regression analysis using simple regression model in order for us to properly explain the results in the correlation analysis, and predict the number of deaths per day after any specific duration of pandemic that might not be recorded in our dataset.

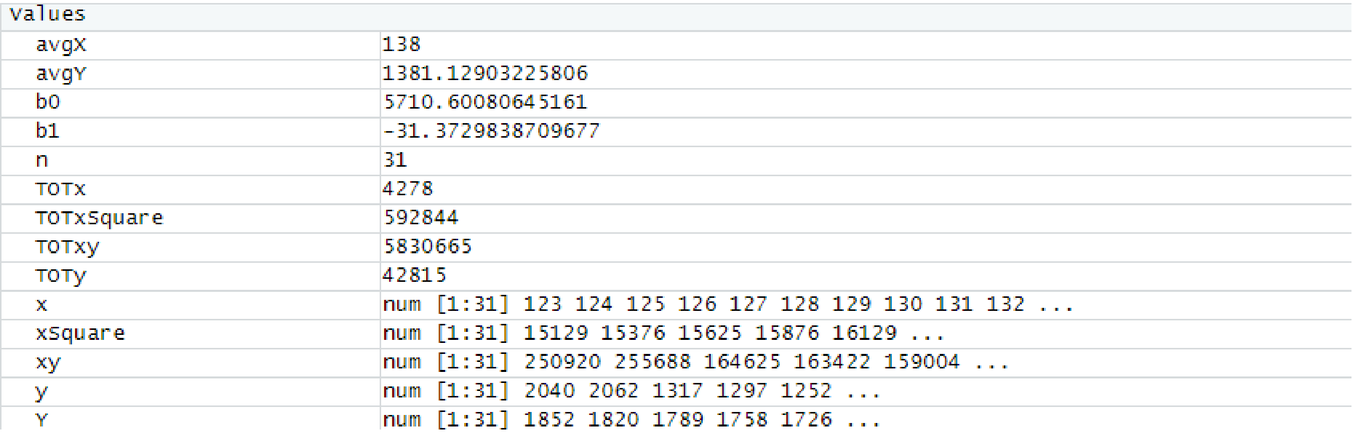
For regression analysis, we are basically trying to generate a sample regression line that is an accurate approximation of the population regression line. Knowing that the variables have linear correlation, the sample regression line will be generated according to the formula: $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ is the estimated (or predicted) y-value, b_0 is the estimate of the regression intercept, b_1 is the estimate of the regression slope, and x is the independent variable. This formula is used as an approximation to the population regression model, which is based on the formula:


The population regression model:
Population y intercept
Population Slope Coefficient
Independent Variable
Random Error term, or residual
Dependent Variable
$y = \beta_0 + \beta_1 x + \varepsilon$
Linear component
Random Error component

For this analysis, we will only be analysing the data in the month of May as it will allow us to predict the deaths per day on the duration after passing the peak mentioned in the correlation analysis section.

Regression analysis for the month of May:

After some calculations done according to the formulas using RStudio (RScript file provided in zip file), the values of the variables in the equation are as follows:



```
Values
  avgX           138
  avgY           1381.12903225806
  b0             5710.60080645161
  b1             -31.3729838709677
  n              31
  TOTx           4278
  TOTxSquare     592844
  TOTxy          5830665
  TOTy           42815
  x              num [1:31] 123 124 125 126 127 128 129 130 131 132 ...
  xSquare        num [1:31] 15129 15376 15625 15876 16129 ...
  xy             num [1:31] 250920 255688 164625 163422 159004 ...
  y              num [1:31] 2040 2062 1317 1297 1252 ...
  Y              num [1:31] 1852 1820 1789 1758 1726 ...
```

As we can see, the values of b_0 and b_1 are approximately 5710.6008 and -31.37298 respectively. Hence the sample regression line is defined by the equation: $\hat{y} = \ 5710.6008 - 31.37298x$

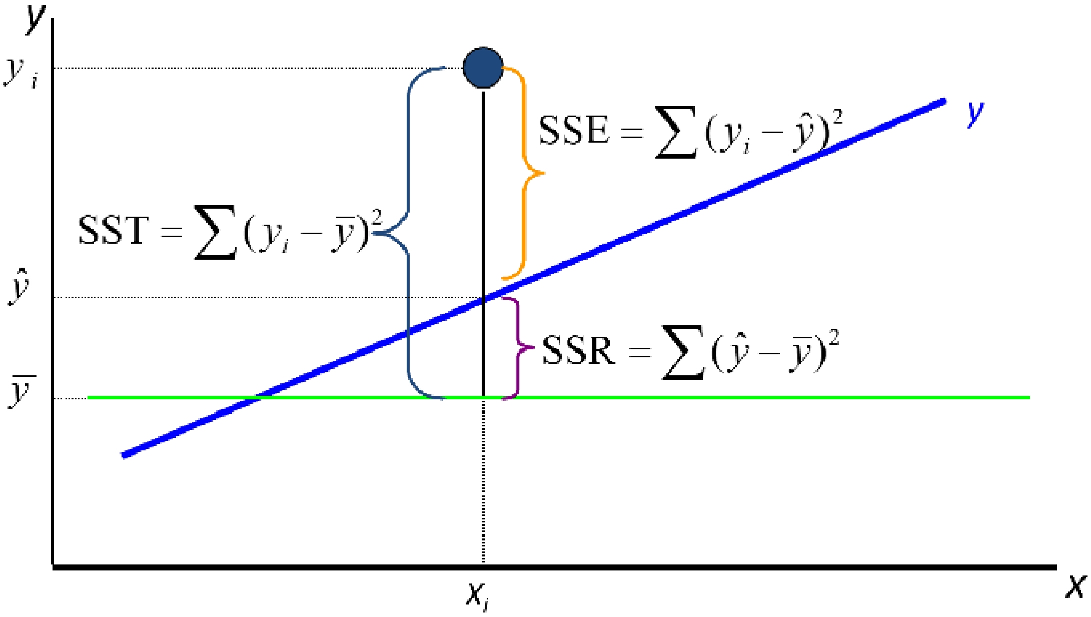
Since $b_1 = \ -31.37298$, this means that starting from the month of April, for every single day that passes, there are approximately 31 lesser deaths per day. And as for $b_0 = \ 5710.6008$, this just means that for the duration of the pandemic within the range of the month of May, 5710.6008 deaths per day is not explained by days of the pandemic in this particular month.

The next part of regression analysis is the section of explained and unexplained variations:

In this section, there are a few different measurements that has to be introduced, they are:

1) **Total sum of squares, SST**: Measures the variation of the observed y values around their mean.
2) **Error sum of squares, SSE**: Variation attributable to factors other than the relationship between x and y.
3) **Regression sum of squares, SSR**: Explained variation attributable to the relationship between x and y.

To picture their differences, refer to the image below:



y
y_i
$\hat{y}$
$\bar{y}$
$\text{SST} = \sum (y_i - \bar{y})^2$
$\text{SSE} = \sum (y_i - \hat{y})^2$
$\text{SSR} = \sum (\hat{y} - \bar{y})^2$
y
X_i
X

4) **Coefficient of determination, $\boldsymbol{R^2}$**: The portion of the total variation in the dependent variable that is explained by variation in the independent variable.

$$\boldsymbol{R^2 = \frac{SSR}{SST}}$$

` Since this is a single independent variable case, the $\boldsymbol{R^2}$ value calculated should be equal to the $\boldsymbol{r^2}$ value from the correlation analysis.

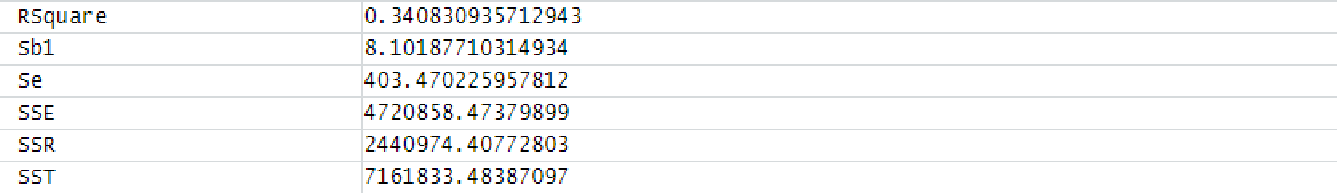
5) **Standard error of estimate, $\boldsymbol{S_\varepsilon}$:** The standard deviation of the variation of observations around the regression line.

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

6) **Standard deviation of the regression slope, $\boldsymbol{S_{b_1}}$:** The standard error of the regression slope coefficient (b_1).

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum(x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

After some calculations done using RStudio (RScript file provided in zip file), the values of the newly introduced measurements for the month of May are as follows:



RSquare	0.340830935712943
Sb1	8.10187710314934
Se	403.470225957812
SSE	4720858.47379899
SSR	2440974.40772803
SST	7161833.48387097

As we can see, the value of $\boldsymbol{R^2}$ is approximately equal to 0.341, which is between 0 and 1. This means that there exists a weak linear relationship between the duration of the pandemic and the number of deaths per day, and that roughly 34% of the variation in the number of deaths per day is explained by the duration of the pandemic.

Next, looking at the value of $\boldsymbol{S_\varepsilon}$, we can see that its value is quite large (403.47). This means that the variation between the observed number of deaths per day and the estimated number of deaths per day is roughly ±403 deaths per day. Meaning that if the estimated deaths for a specific day is 803 people, then the actual number of deaths on that day might be anywhere in the range of 400 to 1206 deaths. Hence, the estimated value is not very accurate.

Moving on, by looking at the value of $\boldsymbol{S_{b_1}}$**,** we can see that it is relatively large as well. This means that the estimation of the slope of the population regression line, b_1 might have a very different value depending on the samples used.

Lastly in this regression analysis, we can use the values calculated above to generate an inference about the slope (t-test for population slope). The purpose of this test is to confirm the existence of a linear relationship between the duration of the pandemic and the number of deaths per day.

t-test for population slope:

Null Hypothesis, $\boldsymbol{H_0}$**:** $\beta_1 = 0$ (No linear relationship.)

Alternate Hypothesis, $\boldsymbol{H_1}$**:** $\beta_1 \neq 0$ (Linear relationship exists.)

After some calculations done using RStudio, the value of the test statistic is approximately equal to -3.872

The critical value for this test is $t_{\frac{\alpha}{2},df} = t_{0.025,29} = -2.045$ > -3.872

Since the test statistic is smaller than the critical value, it falls into the area of rejection. Hence, the null hypothesis must be rejected.

Conclusion: $\boldsymbol{H_0}$ **is rejected.** There is sufficient evidence to show that there exists a linear relationship between the duration of the pandemic and the number of deaths per day.

The reason that the number of deaths per day had a positive linear correlation with the duration of the pandemic was because in the month of April, the population was still unprepared to handle the outbreak. Hence, the number of deaths per day kept increasing as the duration of the pandemic increases. However, after a certain time, the population starts to develop and carry out some safety measures such as undergoing quarantine in order to control the outbreak. This leads to less people getting infected, which directly leads to the decrease in the number of deaths per day as time passes, which was exactly why there was a weak positive linear correlation in the month of April, and a slightly stronger negative linear correlation in the month of May, when the population started to isolate themselves to control the outbreak.

After all the analysis done above, it should be clear by now that the number of deaths per day has a linear relationship with the duration of the pandemic. Before a certain peak, the number of deaths per day will increase as the duration of the pandemic increases. However, knowing that the number of confirmed cases also increases as the duration increases, it is only natural that the number of deaths per day will increase. Thus, a more accurate way of determining whether the virus has become deadlier over time, is to do some tests on its case-fatality rate, (CFR), which brings us to the next test.

4) Chi-square test for goodness-of-fit

Knowing that the number of deaths per day increases over time does not accurately convey how deadly the virus is. This is because as more people get infected, it is only natural that more people will die every day. Hence, a more accurate way of allowing us to measure the deadliness of the virus is by studying its Case-fatality rate (CFR). The case fatality rate of the virus is calculated by using the formula:

$$\boldsymbol{CFR = \frac{Total\ confirmed\ deaths}{Total\ confirmed\ Cases}}$$

By comparing the CFR of the virus across different durations of the pandemic, we will be able to more accurately determine whether or not the virus has become more dangerous or less dangerous.

Hence, in this section, we will be performing a Chi-square test for goodness-of-fit by using the case fatality rate by the end of March (31st March 2020) as the expected proportion, and using the observed total number of confirmed deaths, and the total number of confirmed cases at the end of April (30th April 2020) as the observed frequencies. This test aims to determine whether or not the Covid-19 virus has become deadlier by the end of April compared to during the end of March.

According to the dataset, the CFR on the 31st of March 2020 was approximately 0.019256

This case fatality rate will be used as the expected proportion of the total confirmed deaths over the total confirmed cases on the 30th of April.

In this case, the number of trials, n will be equal to the total number of confirmed cases on the 30th of April. This is because the case fatality rate tells us the proportion of total confirmed deaths to total confirmed cases, or simply meaning that the case fatality rate tells us the probability of a person dying if he/she has been infected with the virus. Hence, the it should be clear that the number of trials should be equal to the total number of confirmed cases.

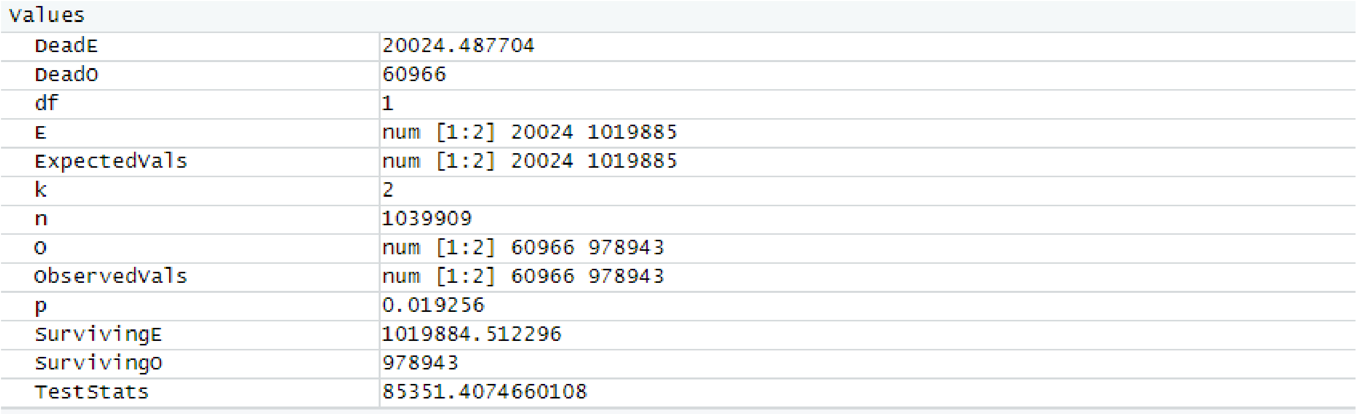
Null Hypothesis, $\boldsymbol{H_0}$: $p = 0.019256$ (The case fatality rate on 30th April is the same as on 31st March 2020.)

Alternate Hypothesis, $\boldsymbol{H_1}$: The case fatality rate on 30th April is different than expected (31st March).

According to the dataset, some calculations are done using RStudio (RScript file provided in zip file), the following table is plotted:

30th April 2020				
Total Confirmed Deaths		Total still Surviving		Total Confirmed Cases
Observed (O)	Expected (E)	Observed (O)	Expected (E)	
60966	20024.487704	978943	1019884.512296	1039909

These are the values of the calculations done using RStudio:



```
Values
  DeadE          20024.487704
  DeadO          60966
  df             1
  E              num [1:2] 20024 1019885
  ExpectedVals   num [1:2] 20024 1019885
  k              2
  n              1039909
  O              num [1:2] 60966 978943
  ObservedVals   num [1:2] 60966 978943
  p              0.019256
  SurvivingE     1019884.512296
  SurvivingO     978943
  TestStats      85351.4074660108
```

We can see that the test- statistic is approximately equal to 85351.41 while the critical value is $X^2_{0.05,1}$ = 3.841 < 85351.41 (Reject null hypothesis).

Since the test statistic is significantly larger than the critical value, it falls into the rejection region. Hence, the null hypothesis must be rejected.

Conclusion: H_0 is rejected. There is sufficient evidence to show that the case fatality rate of the virus on 30th April 2020 is different than on 31st March 2020.

According to the results that we found, we can confidently say that the case fatality rate has definitely increased by the end of April since the end of March, and the virus has become deadlier.

Now that we have done some in-depth study on the statistics of the Covid-19 virus in the USA, we will now compare the mean deaths per day in the USA with some other top countries in terms of confirmed cases, specifically in the month of May. This will be done by using the following test:

5) <u>One-way ANOVA with equal sample sizes</u>

The purpose of this test is to determine whether or not the mean number of deaths per day in the top 3 Covid-19 hotspot countries (USA, Russia, Brazil), is similar or different to one another, in the month of May.

It is to be noted that this test is done based on the assumptions that the populations have normal distribution and same variances, and the samples are random and independent of each other.

The mean deaths per day for the USA, Russia, and Brazil is summarised in the table below:

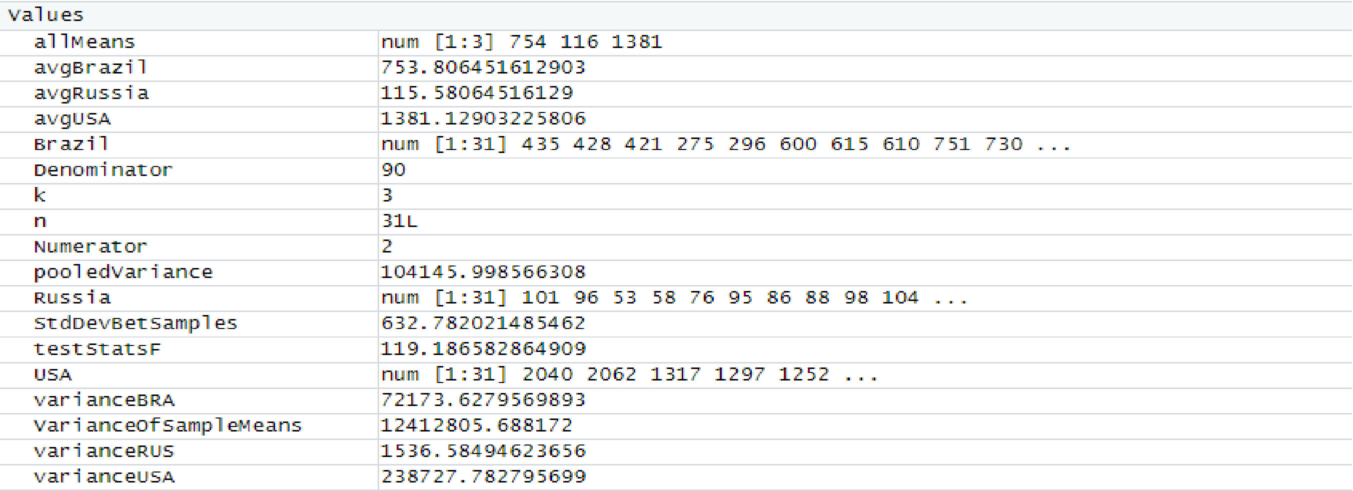
Mean deaths per day (May 2020)		
USA	Russia	Brazil
1381.129	115.5806	753.806

Null Hypothesis, H_0: The mean deaths per day for all 3 countries are equal.

$$H_0 : \boldsymbol{\mu_{USA}} = \boldsymbol{\mu_{BRA}} = \boldsymbol{\mu_{RUS}}$$

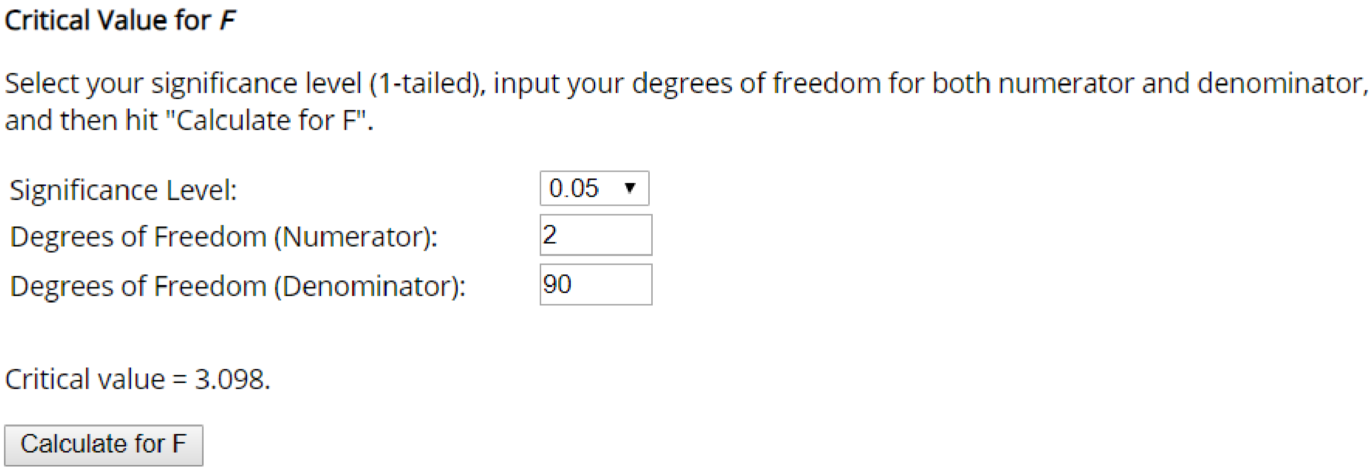
Alternate Hypothesis, H_1: At least one of the means are different.

After some calculations done using RStudio (RScript file provided in zip file), the values of the measurements are shown below:



```
Values
  allMeans                 num [1:3] 754 116 1381
  avgBrazil                753.806451612903
  avgRussia                115.58064516129
  avgUSA                   1381.12903225806
  Brazil                   num [1:31] 435 428 421 275 296 600 615 610 751 730 ...
  Denominator              90
  k                        3
  n                        31L
  Numerator                2
  pooledVariance           104145.998566308
  Russia                   num [1:31] 101 96 53 58 76 95 86 88 98 104 ...
  StdDevBetSamples         632.782021485462
  testStatsF               119.186582864909
  USA                      num [1:31] 2040 2062 1317 1297 1252 ...
  varianceBRA              72173.6279569893
  varianceOfSampleMeans    12412805.688172
  varianceRUS              1536.58494623656
  varianceUSA              238727.782795699
```

As we can see, the test statistic is approximately equal to 119.1866 while the critical value, $\boldsymbol{F_{0.05,2,90}} \approx \mathbf{3.098} < 119.1866$ (Reject null hypothesis)



Critical Value for *F*

Select your significance level (1-tailed), input your degrees of freedom for both numerator and denominator, and then hit "Calculate for F".

Significance Level: 0.05 ▾

Degrees of Freedom (Numerator): 2

Degrees of Freedom (Denominator): 90

Critical value = 3.098.

Since the test statistic is significantly larger than the critical value, it falls into the region of rejection. Thus, the null hypothesis must be rejected.

Conclusion: $\boldsymbol{H_0}$ is rejected. There is sufficient evidence to show that the mean deaths per day between the top 3 Covid-19 hotspot countries are different in the month of May.

Hence, we are now able to conclude with 95% confidence that the mean number of deaths per day is very different depending on the countries even though they are all hotspots for Covid-19 cases.

CONCLUSION

After carrying out all these tests, we were able to confirm some facts about the Covid-19 virus in the USA. Firstly, we can tell with 95% confidence that the average deaths per day in the USA has increased from March to April. This conclusion is made after completing the two samples mean hypothesis testing. Next, since we know that the average number of deaths per day increased in a month, we then carried out a correlation analysis to determine whether or not there exists a linear correlation between the duration of the pandemic and the number of deaths per day. From this analysis, we were able to confirm the fact that there exists a weak positive linear correlation between the duration of the pandemic and the number of deaths per day, before a certain point in reached. However, once the specific point had been reached, there will be a slightly stronger negative linear correlation between these two variables. This phenomenon is then analysed and explained by using the regression analysis that is done following the completion of the correlation analysis. Through this regression analysis, we were able to determine that the duration of the pandemic acts as an independent variable that will affect the number of deaths per day, which is the dependent variable. There is a change in direction of the linear relationship between these two variables because in the beginning of the pandemic, the population we unexpectedly faced with the outbreak of the virus. More and more people were getting infected which in turn causes the rise of the number of deaths per day with the increase in the duration of the pandemic. However, after a certain amount of time, the population has developed some safety measures and precautions in order to control and contain the outbreak. This leads to the decrease in the number of deaths per day after a certain peak is reached. During this period of decline in the number of deaths per day, our regression analysis tells us that roughly 31 less people die every single day, and this period of negative linear relationship between the variables started from the beginning of May. However, there were high standard errors to the estimations that are made by the regression line. Hence, we cannot assume the estimated number of deaths per day according to this analysis will be an accurate representation of the actual number of deaths per day in reality. Up until here, we were able to draw a clear conclusion that the number of deaths per day definitely has a relationship with the duration of the pandemic. However, the changes in the number of deaths per day is not an accurate measurement of how deadly the virus is. This is because when the number of infected people increases, it is only natural that the number of deaths per day will also increase. For example, if 10 people were infected on March, and 5 of them died, and 20 people were infected on April and 10 of them died, it is true that the number of deaths increased, but it does not mean that the virus was deadlier in April. Hence, in order to accurately determine whether or not the virus has become deadlier over time, we carried out a Chi-square test for goodness of fit. This test aims to determine whether or not the case fatality rate by the end of April was a good fit with the case fatality rate by the end of March. After this test, we were 95% confident that the case fatality rate of the Covid-19 virus has definitely risen by April since March in the USA. Lastly, an ANOVA test was carried out in order to compare the mean deaths per day on the month of May among the top 3 Covid-19 hotspot nations in the world, they are USA, Russia, and Brazil (as of June 2020). After this test, we were able to conclude that even though all 3 of these countries have the most cases of Covid-19, their average number of deaths per day are significantly different from each other, which could imply that some of these countries are not doing as well as the others at controlling the outbreak. These are all the facts that we have discovered after this study. It is hoped that every country will do their absolute best in order to contain this global pandemic as soon as possible.