PROPOSAL OF PROJECT 2

PROBABILITY AND STATISTICAL DATA ANALYSIS

SECI / SCSI 2143

Name: **Cheong Chien Li** Matric No: **A19EC0186**

Source of dataset (URL)	:	https://ourworldindata.org/grapher/covid-19-total-confirmed-cases-vs-total-confirmed-deaths Click the "DATA" tab to download spreadsheet.
Dataset description (collected by who and for what purpose?)	:	The dataset is collected by the European Centre for Disease Prevention and Control (ECDC). The data was used by Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell with the purpose of assessing the mortality risks of COVID-19.
Variables (name of variables and type – categorical / ordinal / interval / ratio)	:	1) Country (categorical) 2) Date (ordinal) 3) Month – Derived from Date (ordinal) 4) Duration – Derived from Date (ratio) 5) Confirmed Infected Cases (ratio) 6) Confirmed Death Cases (ratio) 7) Case Fatality Rate (ratio) 8) Severity – Derived from Confirmed death and infected cases. (ordinal)
Description of purpose of study	:	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.
Specification of target population	:	The target population is the people of the United States of America who are infected by the COVID-19 virus from 31st December 2019 to 22nd May 2020, and no further than that.
Selection of variables (potential variables that will be selected for analysis)	:	The variables that will be most focused on are Month, Duration, Case Fatality Rate, and also Severity.
Proposed analysis (potential statistical test analysis related to the variables chosen)	:	1) Two samples proportions test - using samples from two different months to test whether the case fatality rate of the virus has increased in April since March 2020. 2) Significance test for correlation using Pearson's correlation coefficient – to test the existence of a correlation between the Case Fatality Rate (Dependent) and the Duration of the pandemic (Independent) 3) Regression analysis using simple regression model – using only one independent variable (Duration), and the dependent variable (Case Fatality Rate). A positive linear relationship would be the expected regression model for this test where the case fatality rate of the virus increases as the duration of the outbreak increase. The least squares equation will be found and analyzed to

	potentially estimate the case fatality rates for the duration that are not covered by the sample. 4) Chi-square test for goodness of fit – to test whether the severity (ratio of dead to recovered) of the virus has improved or worsened in April since March 2020, by using the severity in March as expected proportions, and the data of recovered and dead in April as the observed values. Expected values of dead and recovered in April will be calculated by multiplying the total infected cases with the proportions in March, in order to determine whether or not the sample in April is a good fit with the sample in March. 5) Chi-square test for independence – to test the existence of a relationship between the duration of the pandemic (independent variable) and the severity of the virus (ratio of dead to recovered), which is the dependent variable.
Expected outcome for analysis :	After the series of tests, we would expect to find a correlation between the duration of the outbreak and the case fatality rate of the virus, where the case fatality rate increases as the duration of the outbreak increases (assuming nothing is done to stop it). We would also expect to see a relationship between the severity of the virus and the duration of the outbreak, where the severity of the virus to worsen from March to April, meaning that the dead to recovered ratio will increase as long as the duration of the outbreak increases (assuming that nothing is done to stop it).

Note: Submit this proposal within 2 weeks after briefing done by the lecturer.