
# Alexandria University
# Faculty of Engineering
# Department of Electrical Engineering
# Communications & Electronics Program

**Second Year 2020**

**Course: Mathematics IX (Probability)**

# Covid-19 Data Analysis I

**Submitted To: Prof.Dr / Yasmine Abouelseoud**

**Submitted From: Ismail Mohamed Farahat**
**Section: 2**
**ID: 51**
**Date: 25/3/2020**

# Content:

# 1.Data Wrangling

## 1.1. Info about Covid-19

The 2019–20 coronavirus pandemic is a pandemic of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was first identified in Wuhan, Hubei, China, in December 2019, and was recognized as a pandemic by the World Health Organization (WHO) on 11 March 2020. As of 25 March, more than 446,000 cases of COVID-19 have been reported in more than 190 countries and territories, resulting in more than 19,800 deaths and more than 113,000 recoveries.

## 1.2. Gathering the data

**In this step, I gathered the data from multiple sources of trusted organizations and websites:**

1- **World Health Organization (WHO)**
2- **Chinese Center for Disease Control and Prevention**
3- **European Centre for Disease Prevention and Control**
4- **Data world website**
5- **Kaggle website**
6- **Worldometers**

I used only the data that will be benefit in this report and I used the other sources to make sure that I gathered trusted data…but mainly I used Kaggle website in this report.

## 1.3. Cleaning the data

After gathering the data, we should clean the data first before doing any analysis and that's because the data have null values, repeated values, useless columns and discrete data that should be in one column.

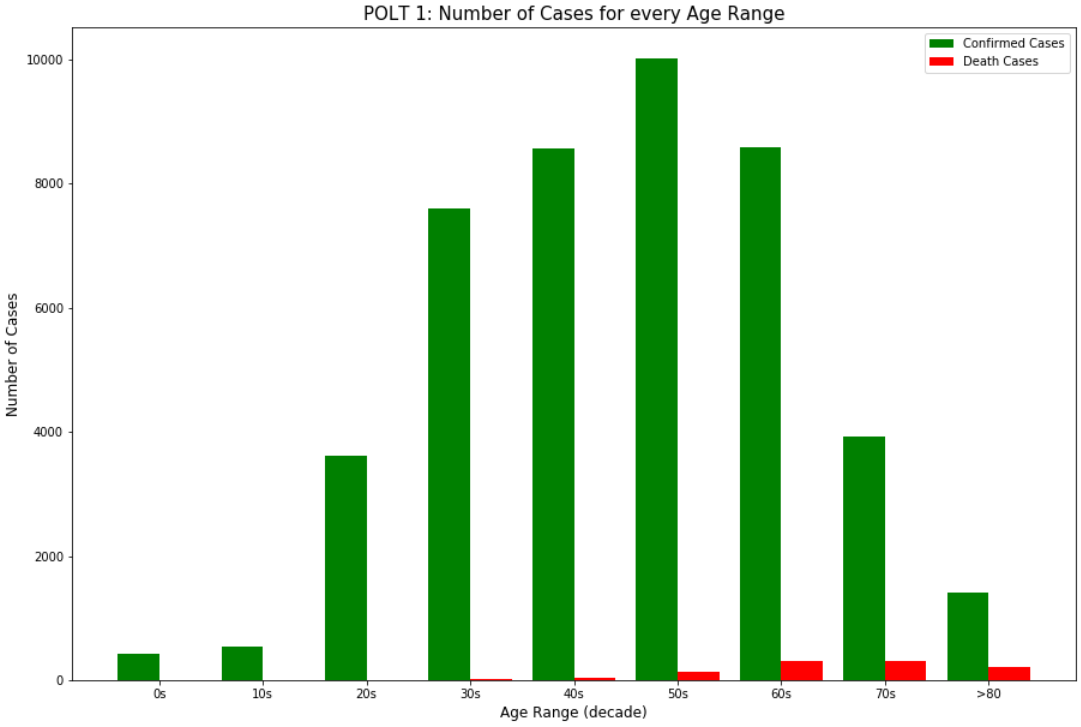
I did clean the data using python programming language and excel to make the data useful in my analysis…but this not the point of this report so I am not going to discuss the techniques of this step.

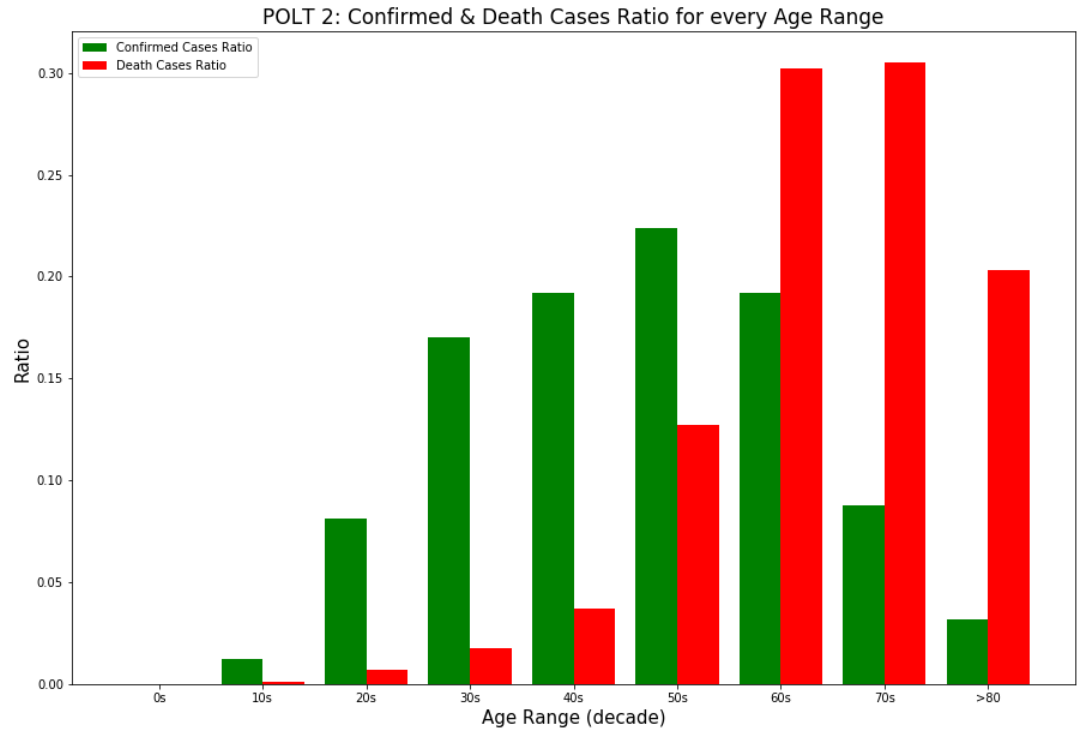
In the end, we have three datasets that will be used in our analysis:

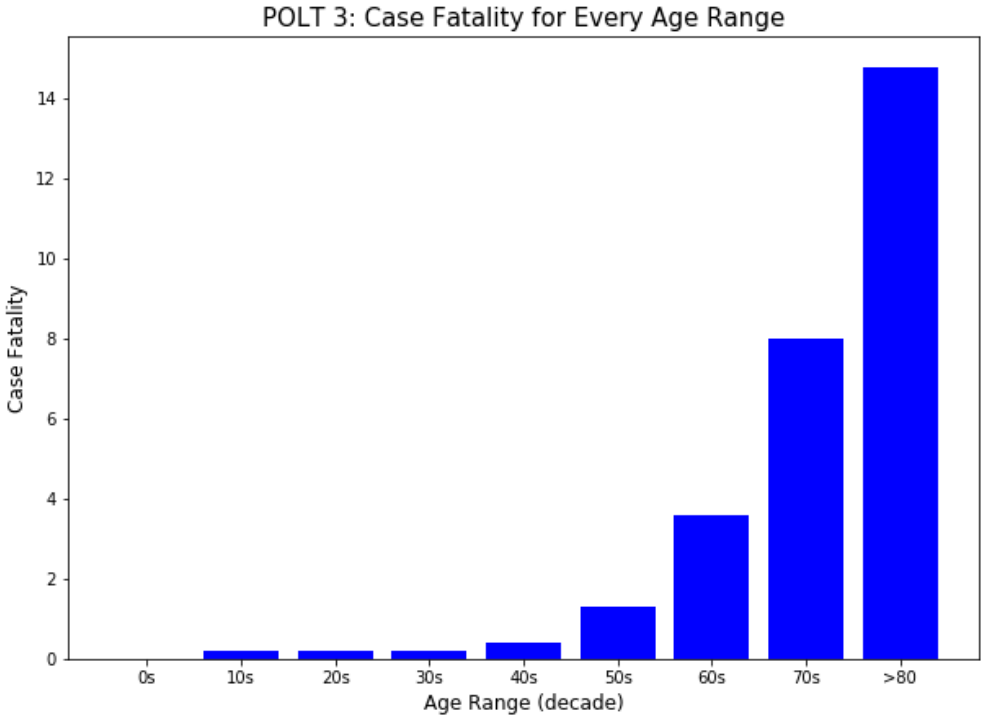
1- age_data_covid19
2- incubation_data_covid19
3- total_cases_covid19

# 2. Age Probability Distribution

## 2.1. Histogram






## 2.2. Comment on the Plots

**PLOT 1:** From this plot, we could see that the number of death cases is more fewer than the confirmed cases but this plot is not good enough to estimate the danger of the virus or even to make a good analysis so we are going to plot the ratio instead of cases number in PLOT 2.

**PLOT 2:** From this plot, we could see that the ratio of death cases is smaller than the ratio of confirmed cases until the age 60 and after age 60 and older the ratio of death cases is higher than the confirmed one…which means that the possibility that you are going to die is high if your age is higher than 60.

In this plot, we could see that the confirmed cases have almost a **normal distribution function** over different ages of the patients…and the death cases have a **negatively skewed distribution function** (which could be fitted the best by the Dagum distribution or the Gompertz distribution).

**PLOT 3:** From this plot, we could see that case fatality, which is the proportion of deaths from a certain disease compared to the total number of people diagnosed with the disease for a certain period of time, is highly increased by increasing the age….this plot represents the dangerous of the Covid-19 virus on older people very well.

## 2.3. Theoretical Formula of Confirmed Cases Distribution

### The equation of normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

### Calculations of distribution parameters

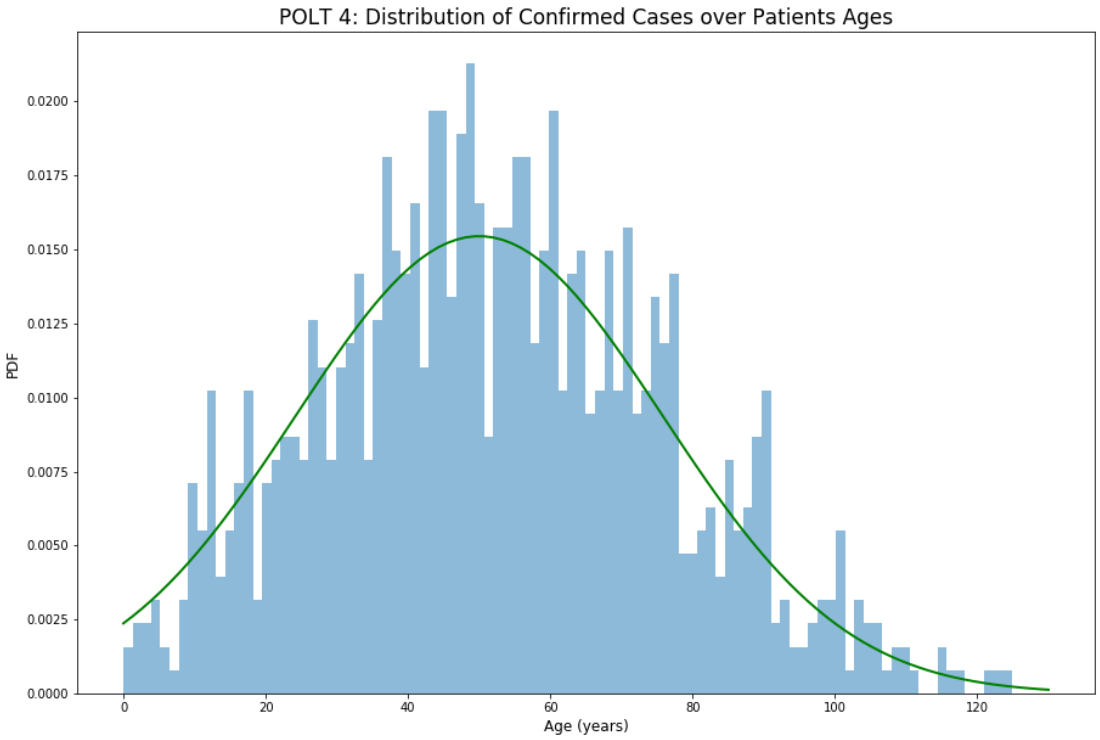
Mean(μ) = 50

Standard deviation (σ) = 25.82

### The equation of confirmed cases normal distribution

$$f(x) = 0.0155 * e^{\frac{(x-50)^2}{1333.35}}$$

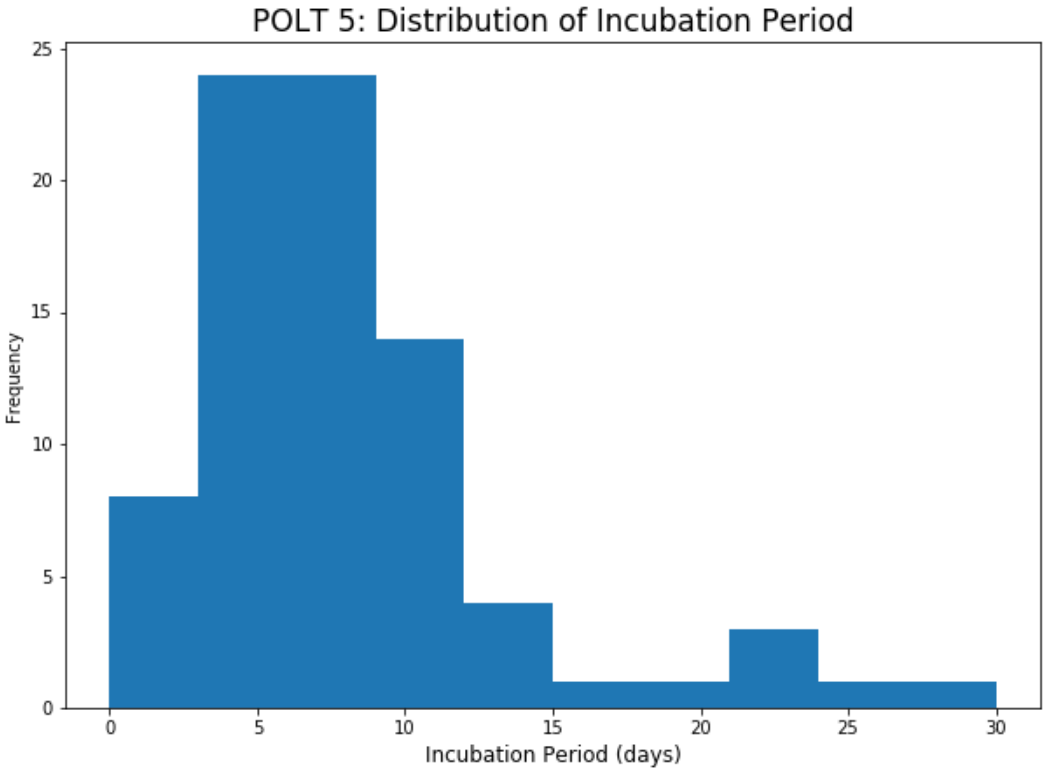
# 2.4. Probability Distribution Curve with Python.




**Note: the code that used to plot this figure is in the jupyter workspace at the end of this report.**

# 3. Incubation Probability Distribution

## 3.1. Histogram




## 3.2. Comment on the Plots

**PLOT 5:** From this plot, we could see that most of the data is concentrated in range from 0 to 15 days and we could estimate that the best probability distribution is **gamma distribution.**

## 3.3. Theoretical Formula of the Distribution

**The equation of gamma distribution**

$$f(x) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\; x^{\alpha-1}\; e^{\frac{-x}{\beta}}$$

**Calculations of distribution parameters**

Mean (μ) = 7.642

Standard deviation (σ) = 5.428

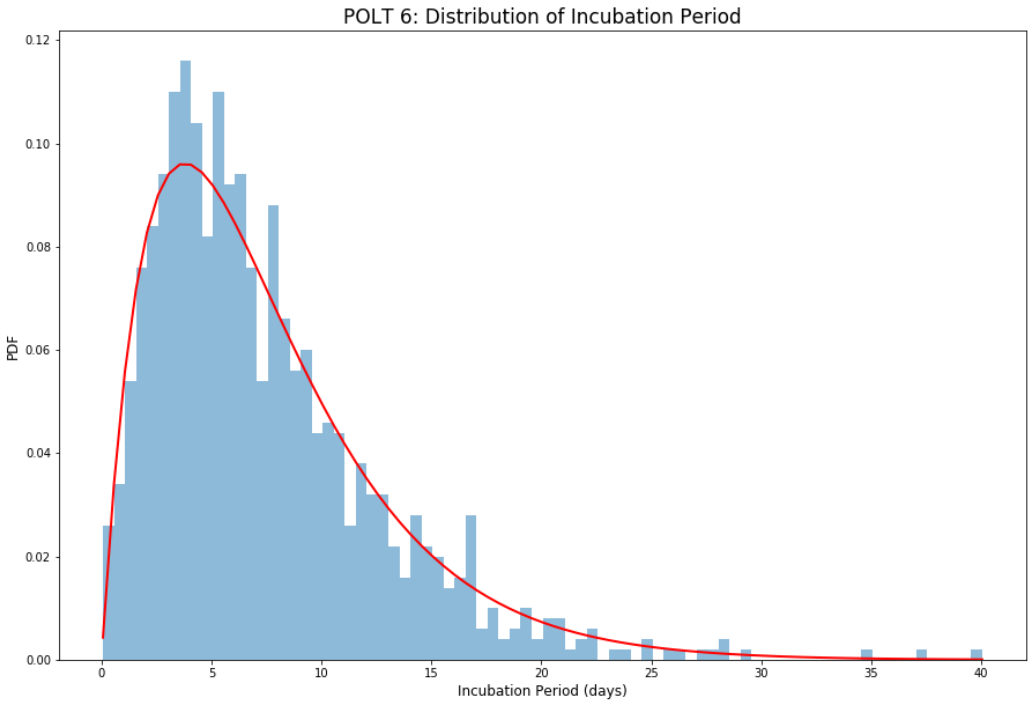
Shape (α) = $(\frac{\mu}{\sigma})^2$ = 1.982

Scale (β) = $\frac{\sigma^2}{\mu}$ = 3.856

**The equation of incubation period normal distribution**

$$f(x) = 0.0694\; x^{0.982}\; e^{\frac{-x}{3.856}}$$

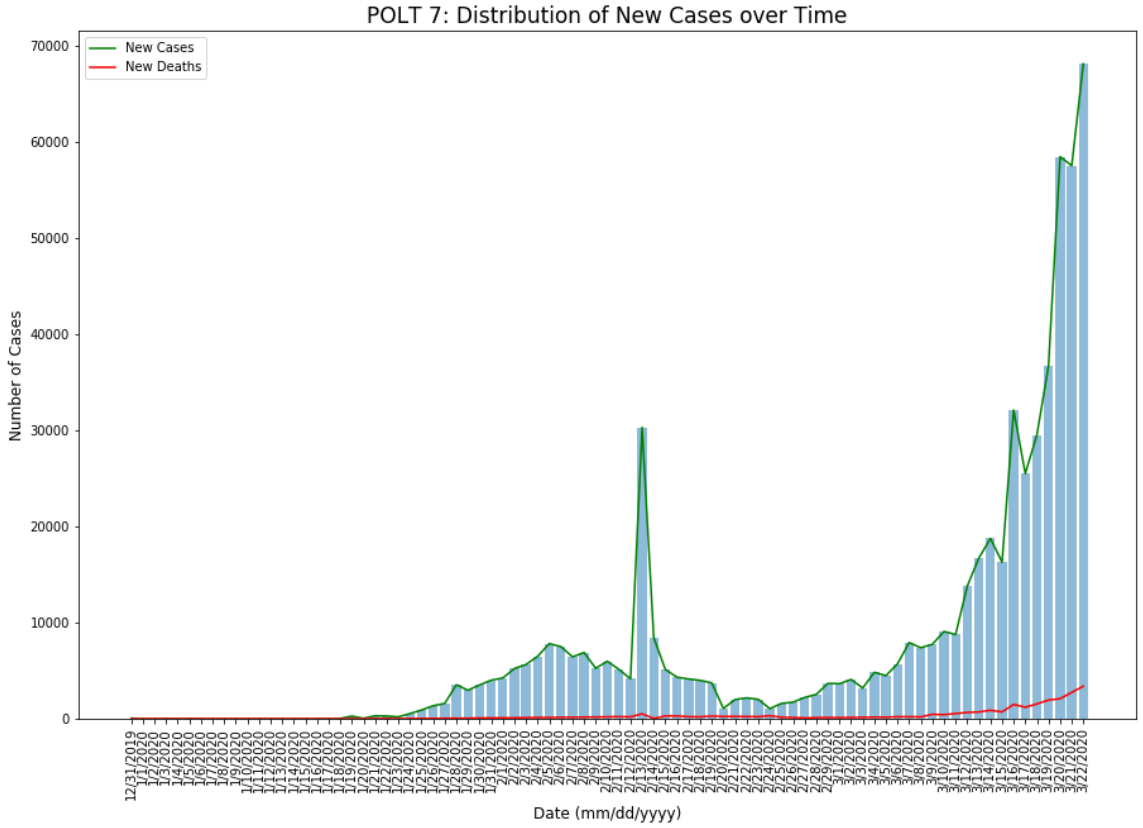
## 3.4. Probability Distribution Curve with Python

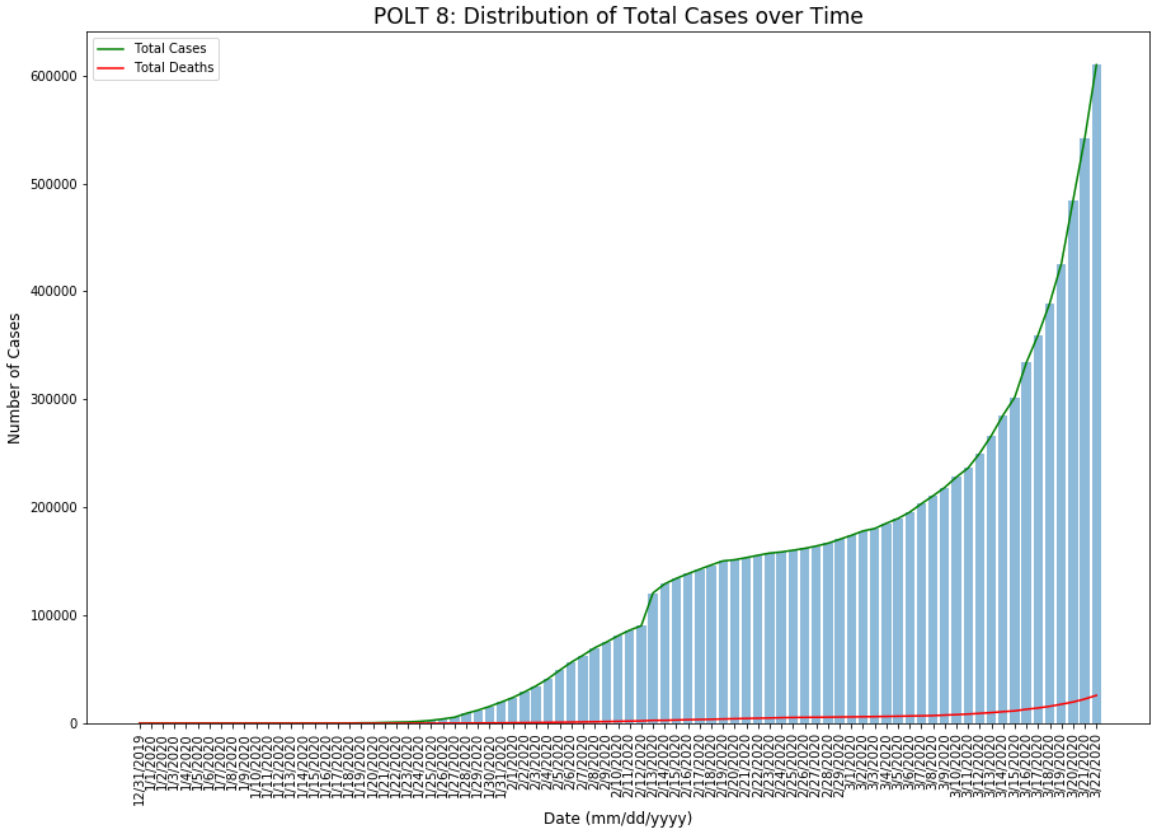



**Note: the code used to plot this figure is in the jupyter workspace at the end of this report.**

# 4. Spread Probability Distribution

## 4.1. Histogram


POLT 7: Distribution of New Cases over Time
New Cases
New Deaths
Number of Cases
Date (mm/dd/yyyy)


POLT 8: Distribution of Total Cases over Time
Total Cases
Total Deaths
Number of Cases
Date (mm/dd/yyyy)

## 4.2. Comment on the Plots

**PLOT 7:** From this plot, we could see that the new cases and new deaths grow exponentially…which means that the spread of the virus is very serios matter that should be taken into consideration even if the new deaths is very few compared to the confirmed new cases.

**PLOT 8:** From this plot, we notice the same note from the previous plot that the total cases and total deaths grow exponentially.

## 4.3. Fitting Probability Distribution

From previous plots, we know how the virus spread exponentially…but the question now how the virus is going to spread in the future.

The answer of this question is very easy, as any pandemic disease…in the beginning, the number of infected cases is small then the number of cases grows exponentially until reach to the maximum point of patients, then after infecting a lot of people by the disease…there is no chance for anyone to be infected any more so the number of new cases is going to decrease…and as we know this distribution fits the best normal distribution so the distribution of the pandemic disease over time is **normal distribution** as described in FIGURE 9 (every dieses has its own mean and standard deviation).

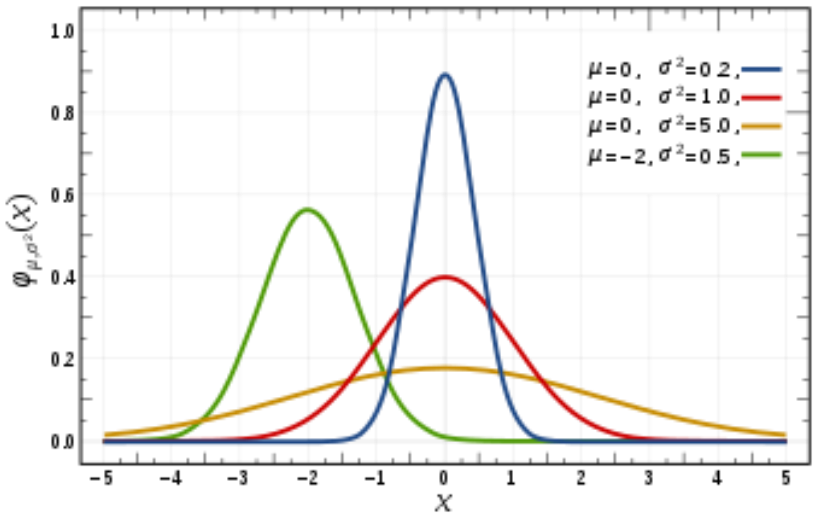



Figure 9: the different curves of normal distributions

In the present, the world stands in the blue region in FIGURE 10 and it is going to be at different areas of this distribution over time until the end of this pandemic disease and the end of this distribution curve.

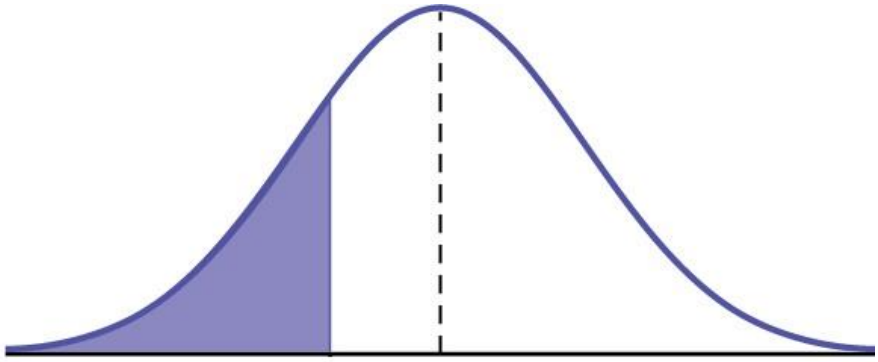


Figure 10: the blue region represents the region that we are into today

Today, china has achieved this distribution and now there is no new cases in the country (Wuhan)…but unfortunately, Italy & Spain & France & USA and many other countries, they are in the first region of the curve.

## 4.4. Flattening the Curve

Today, all the countries are trying to flatten the curve of this normal distribution by taking a lot of serious decisions like closing the universities and schools, making people stay at their home and other decisions….the plot in FIGURE 11 describes this situation.

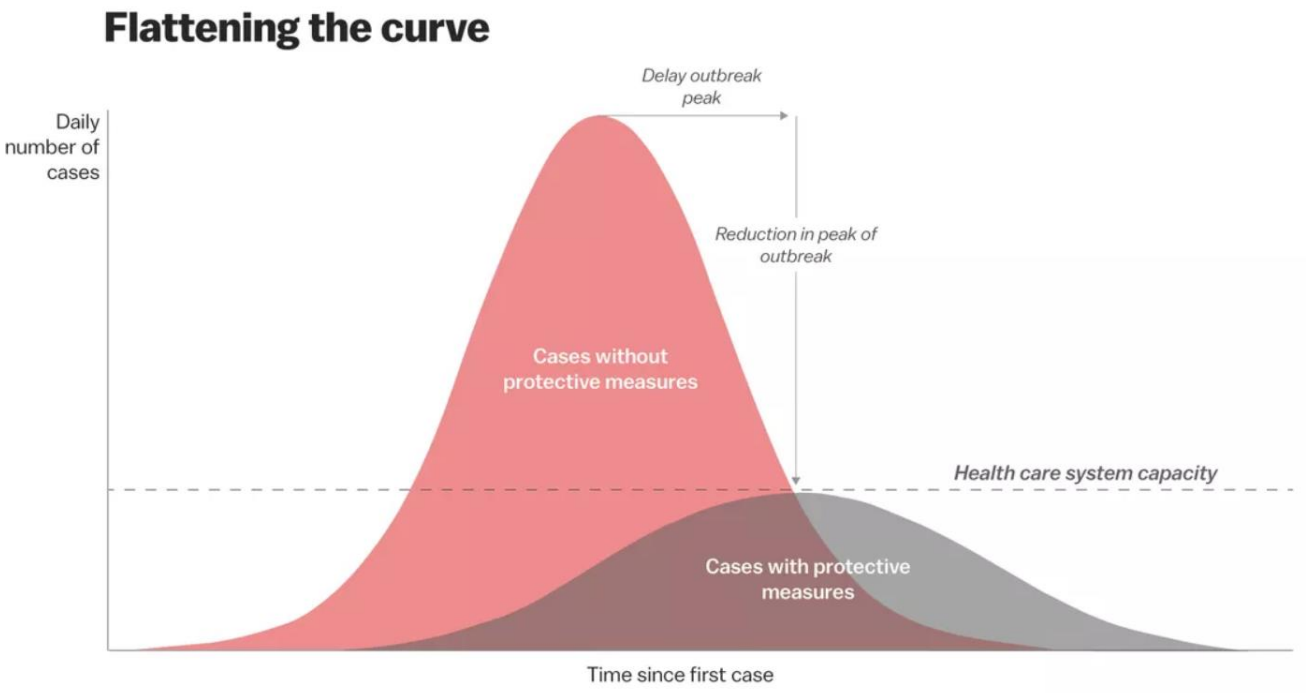



Figure 11: Flattening the Curve

The natural distribution of the disease path is described by the red curve (without Quarantine) …this curve has small mean (means that the time of the virus spread will be very short time) and small standard deviation…but the blue curve (with Quarantine) has high value for the mean (means that the time of the virus spread will be longer) and high value for the standard deviation.

The purpose of this step is to reduce the number of infected people in a given time to be suitable with the health care system capacity and distribute the number of cases over longer time than the natural time taken by the disease spread without Quarantine.