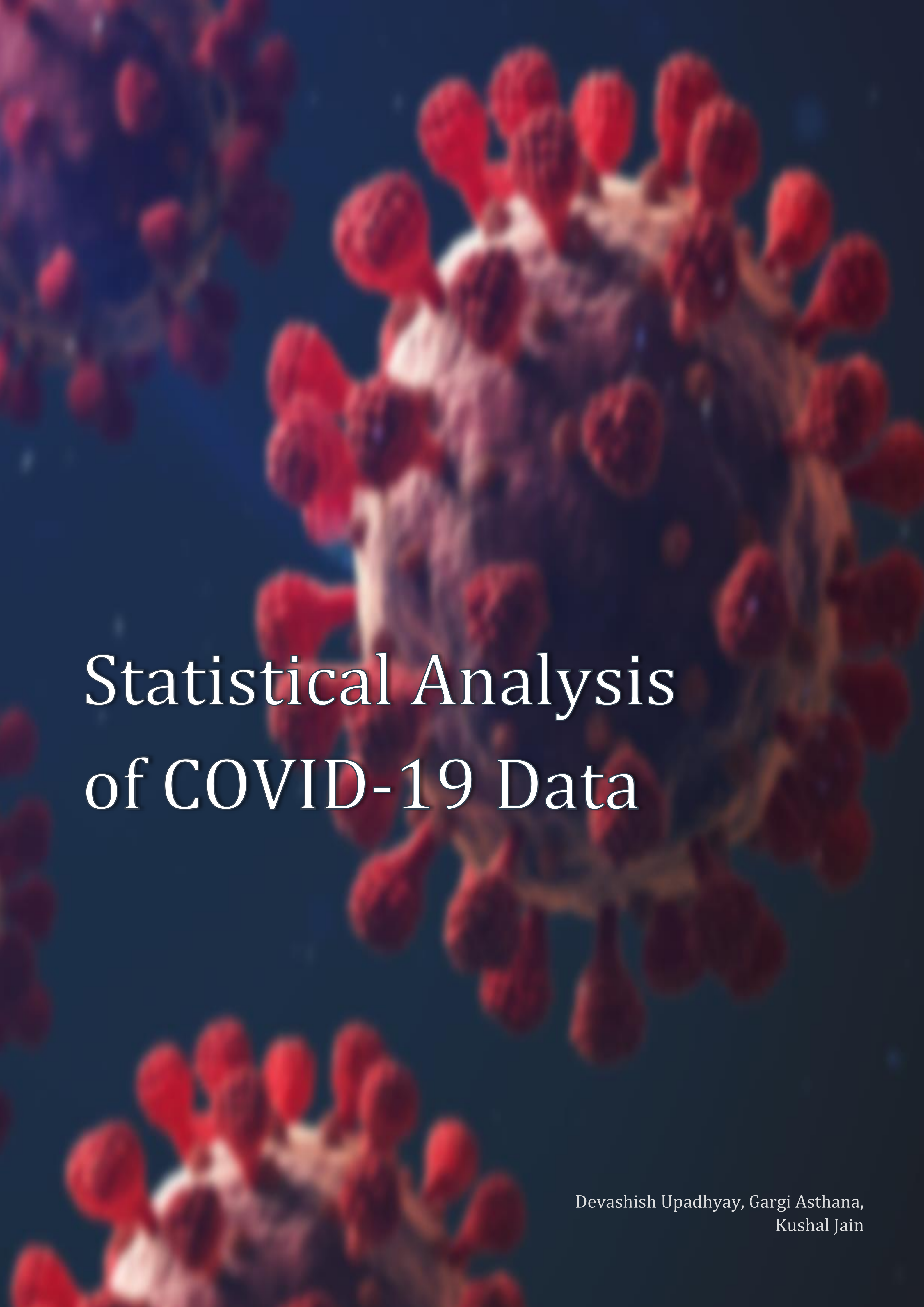


Statistical Analysis of COVID-19 Data

Devashish Upadhyay, Gargi Asthana,
Kushal Jain

Contents

Introduction

8.52 crore people have been infected by the coronavirus until 4th January,2021, causing about 19 lakh deaths worldwide. Between April and June of 2020, the ILO (International Labour Organisation) estimated the loss of about 400 million full time jobs, and a 10% decline in the income of all the workers worldwide in just the first 9 months of 2020, a loss equivalent to about 3.5 trillion dollars or about 255 lakh crores. We have never faced such an economic crisis since the great depression. The first case was reported in Wuhan in early December and after that it has only grown bigger. India, a country with about 135 crore people, reported its first case on 30th January,2020. After that point not only, the cases have gone up but also a lot data has been collected, but a true data scientist would know collected data is useless until it has been analysed correctly and the analysis has been used fruitfully. Since covid-19 has been the talk of the town for over a year now our group decided to work on the governments data and see if we could get useful information out of it. We have used several statistical concepts like box plots, histograms, quantile-quantile plot, hypothesis, and several more to analyse and answer the questions we have. Before we start analysing, we questioned the credibility of our dataset, and using the Benford's law we have got our answer.

Benford's Law

In 1938, a scientist named Frank Benford who worked on this law, which is the mathematical theory of leading digits. This law is also known as Newcomb-Benford's law and first digit law. It is an observation of the leading digits of the real-world numerical datasets. Intuitively, we may think all first digits would occur 1/9th or 11.1% of the time while, according to Benford's law number 1 occurs 31.1% of the time which is almost thrice as much than expected, number 2 occurs 17.6% of the time and it goes on till number 9 which occurs 4.6% of the time which is below 11.1%. This theory also covers the first digit, second digit, last digit, first 2 digits and various combinations of digits in the dataset. The law does not work for all sets of numbers, the numerical dataset must naturally occur. From the beginning Benford's law has always been a key tool in detecting frauds.

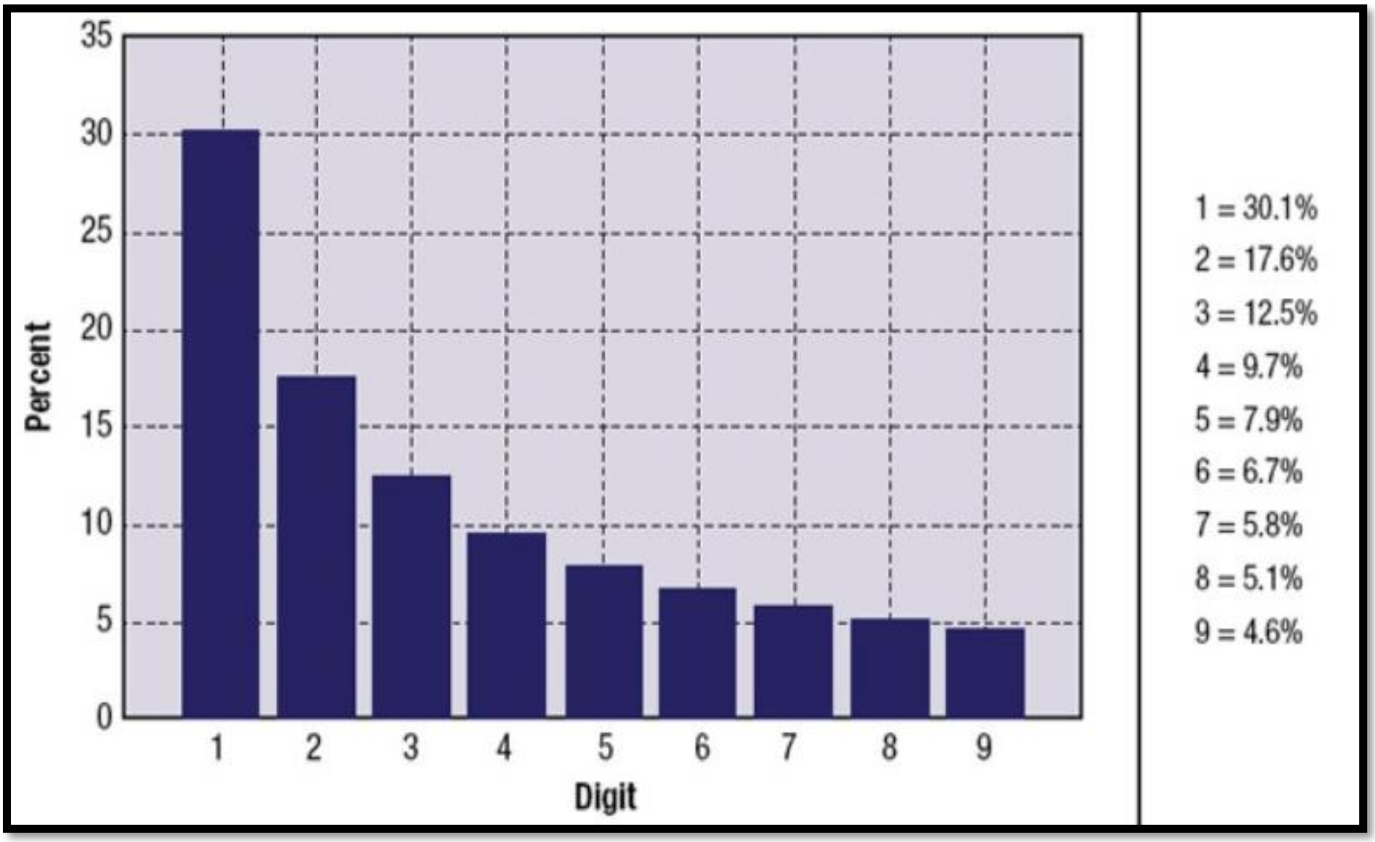


Figure A

Benford's Law Distribution Leading Digit

Working of Benford’s Law

Suppose we take 5 fair dices and multiply the outputs of all dices. The final output would range between 1 and 6^5 which is 7776. If we take about 500 random outputs and plot a graph of the first number of the final output, the graph will follow the same pattern as the Benford’s law, with 1 occurring the about 30% of the time whereas 9 only occurring about 5% of the time.

Usage

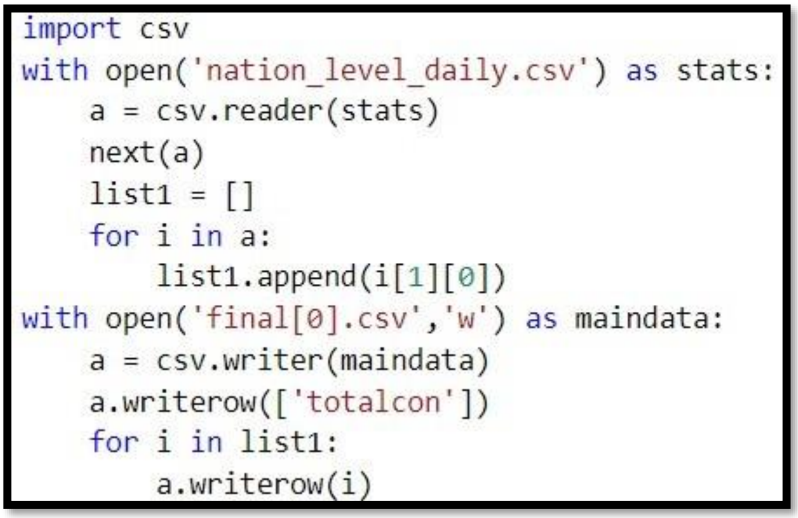
This law can be applied where datasets describe similar data, the data must be large enough, about 500 samples for accurate result, the dataset must include a wide variety of numbers in figures, i.e., tens, thousands, tens of thousands and there must be no maximum and minimum limit.
Examples of confirming datatypes - Credit card transaction, Loan data, stock prices, disturbances, journal entries, populations.

Using the Benford’s Law on our numeric dataset

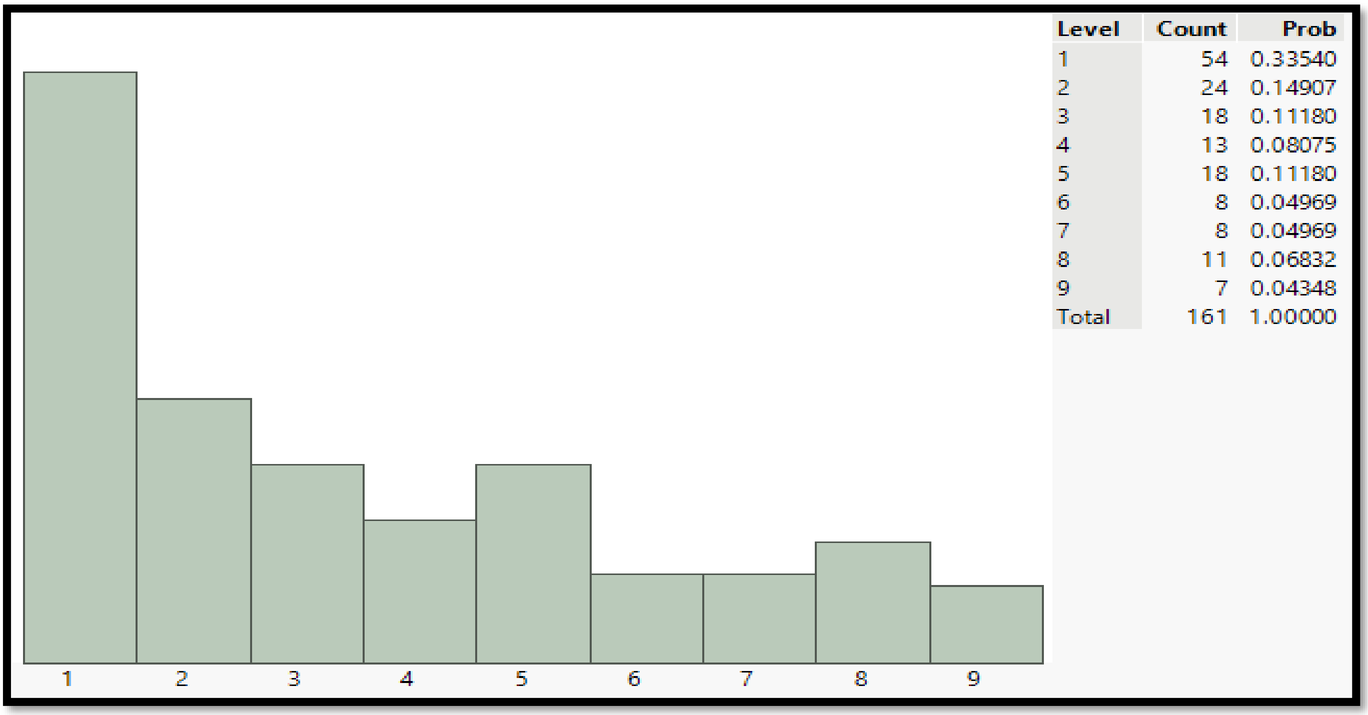
The dataset contains multiple columns of information namely:
i) Date, ii) Daily Confirmed, iii) Total Confirmed, iv) Daily Recovered, v) Total Recovered, vi) Daily Deceased and vii) Total Deceased.

We cleaned the dataset devoid of blank elements, irrelevant non-numerical inputs, etc. Then saved it in a CSV (Comma-separated values) format. For applying Benford’s Law, we selected “**Daily confirmed**” column in the data set and using Python programming language and CSV library we extracted the first digit of each row in the column and saved it in another CSV (Comma-separated values) format.

Python Program



```python
import csv
with open('nation_level_daily.csv') as stats:
    a = csv.reader(stats)
    next(a)
    list1 = []
    for i in a:
        list1.append(i[1][0])
with open('final[0].csv','w') as maindata:
    a = csv.writer(maindata)
    a.writerow(['totalcon'])
    for i in list1:
        a.writerow(i)
```



Level	Count	Prob
1	54	0.33540
2	24	0.14907
3	18	0.11180
4	13	0.08075
5	18	0.11180
6	8	0.04969
7	8	0.04969
8	11	0.06832
9	7	0.04348
Total	161	1.00000

Figure B

The First digits of the discrete data have been plotted in form of a histogram as shown in Figure B.

The Benford's law has always been a key tool in detecting frauds. In **Figure B** it is observed that the Count percent of each digit approximately according to the Benford law as shown in Figure A. By this we can conclude that the data set does not have intentionally manipulated numbers and the data can be used for applying further statistical methods for analysis.

About the dataset

The dataset contains multiple columns of information namely:
i) Patient number, ii) Date Announced, iii) Age Bracket, iv) Gender, v) Detected City, vi) Detected State, vii) Current Status, viii) Type of Transmission.
The dataset initially had 200,000+ entries. However, after cleaning the missing data points and mess/unordered data, the dataset was left with 107260 entries to analyse.

Age Bracket

For analysis, the age bracket data is of continuous form.

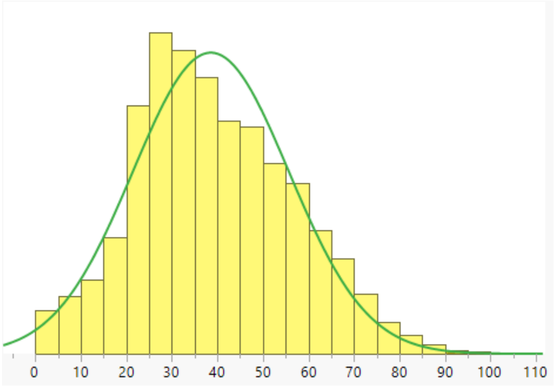


Figure 1.1

The age bracket of the observed patients has been plotted in the form of a histogram as shown in Figure 1.1. As it is seen, the distribution resembles a Normal Distribution.

Cumulative Distribution Function for Age Bracket

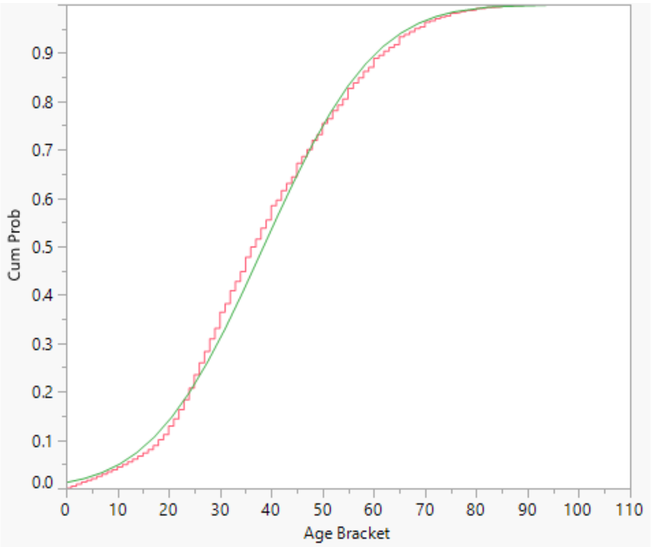


Figure 1.2
*Cum Prob – Cumulative Probability

Quantiles for Distribution of Age Bracket

100.0%	maximum	105
99.5%		84
97.5%		74
90.0%		62
75.0%	quartile	50
50.0%	median	37
25.0%	quartile	26
10.0%		18
2.5%		7
0.5%		2
0.0%	minimum	0

Statistical Inferences

Mean	38.538563
Std Dev	17.110028
Std Err Mean	0.0522434
Upper 95% Mean	38.64096
Lower 95% Mean	38.436167
N	107260
Variance	292.75305
Skewness	0.2896962
Kurtosis	-0.291862
Median	37
Mode	30

Quantile-Quantile Plot

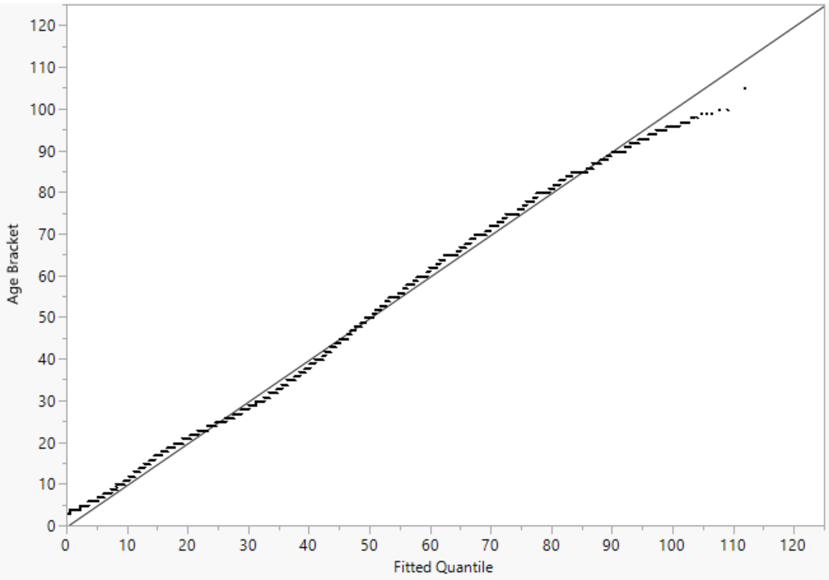
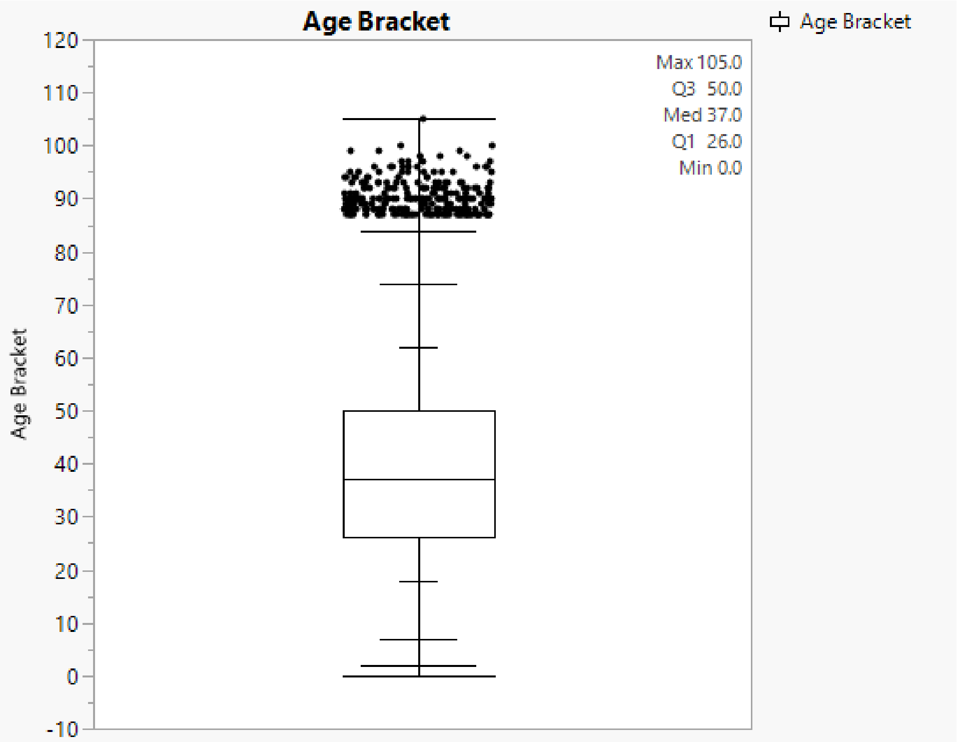


Figure 1.3

Here, the X-Axis plots the Normal Quantiles and the Y-Axis Plots the Data Quantiles.

From the graph it is evident that many points are on the line. Thus, it can be concluded that the data is normally distributed.

Box Plot



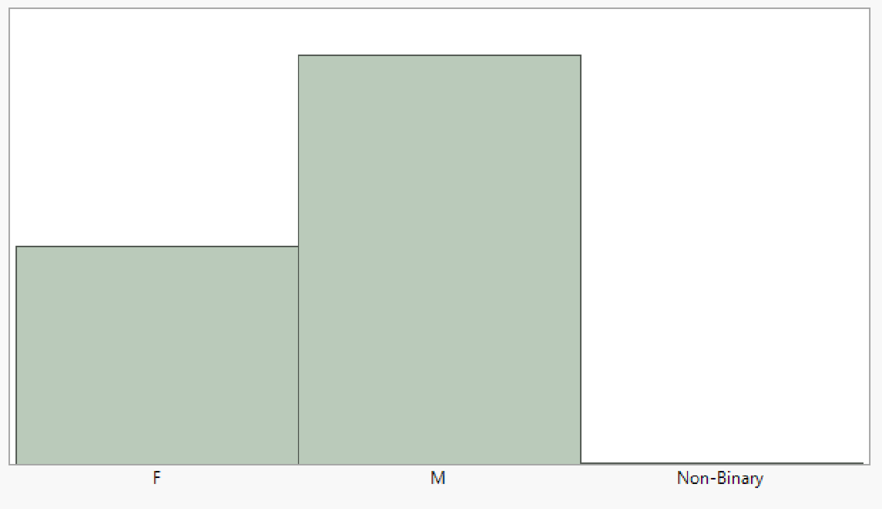
100%: 105
99.5%: 84
97.5%: 74
90%: 62
75%: 50
50%: 37
25%: 26
10%: 18
2.5%: 7
0.5%: 2
0%: 0
N: 107260

Figure 1.4

The Age Bracket for observed patients has been plotted in the form of a Box plot as shown in Figure 1.4

Gender

The gender data has three discrete point, namely: i) Male, ii) Female and iii) Non-Binary. The histogram for the data is plotted below.



Level	Count	Prob
F	37187	0.34670
M	70055	0.65313
Non-Binary	18	0.00017
Total	107260	1.00000

Figure 2.1

$$Percentage = \frac{Count\ of\ Desired\ Outcome}{Total\ Count} \times 100$$

Level	Percentage
F	34.6%
M	65.3%
Non-Binary	0.017%

As observed, males are more likely to be infected by the COVID-19 virus as compared to females.

For further analysis, the data set was split by Gender to yield Age Bracket columns separately

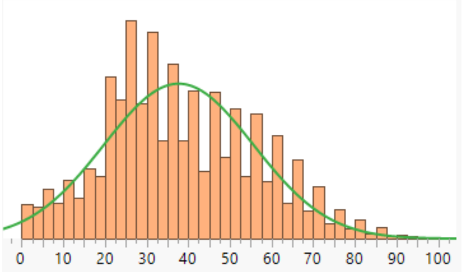
<table>
<tr><th>Graph</th><th>Female Distribution</th><th>Male Distribution</th></tr>
<tr><td>Normal Curve</td><td>


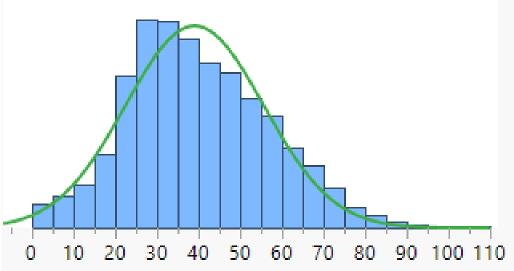
Figure 2.2</td><td>


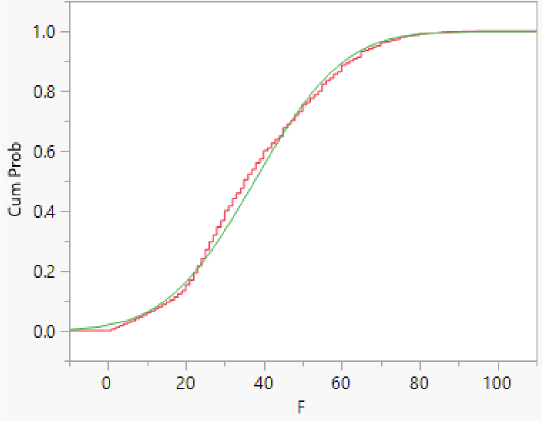
Figure 2.3</td></tr>
<tr><td>Statistical Inference</td><td>
<table>
<tr><td>Mean</td><td>37.677477</td></tr>
<tr><td>Std Dev</td><td>17.924671</td></tr>
<tr><td>Std Err Mean</td><td>0.0929513</td></tr>
<tr><td>Upper 95% Mean</td><td>37.859664</td></tr>
<tr><td>Lower 95% Mean</td><td>37.495289</td></tr>
<tr><td>N</td><td>37187</td></tr>
<tr><td>Median</td><td>35</td></tr>
<tr><td>Mode</td><td>30</td></tr>
</table>
</td><td>
<table>
<tr><td>Mean</td><td>38.994504</td></tr>
<tr><td>Std Dev</td><td>16.642709</td></tr>
<tr><td>Std Err Mean</td><td>0.0628788</td></tr>
<tr><td>Upper 95% Mean</td><td>39.117746</td></tr>
<tr><td>Lower 95% Mean</td><td>38.871261</td></tr>
<tr><td>N</td><td>70055</td></tr>
<tr><td>Median</td><td>37</td></tr>
<tr><td>Mode</td><td>30</td></tr>
</table>
</td></tr>
<tr><td>CDF Plot</td><td>


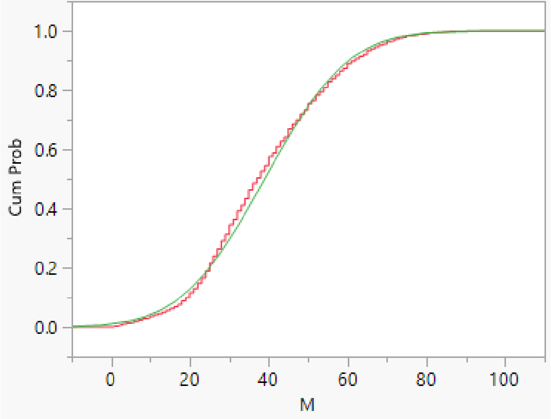
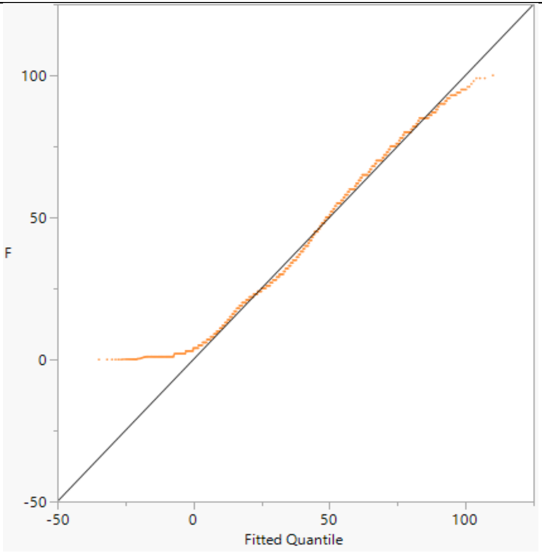
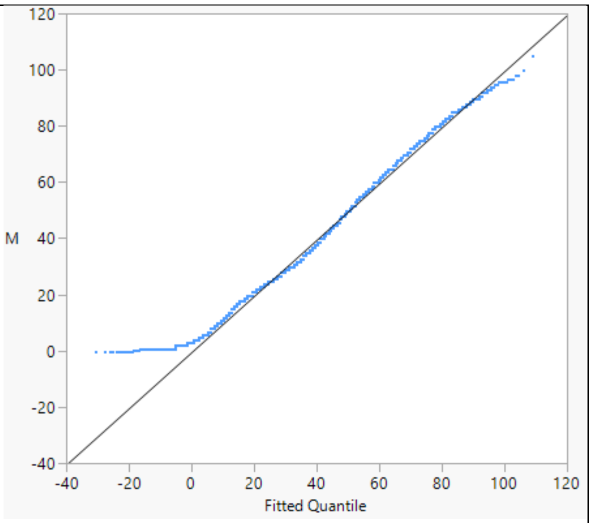
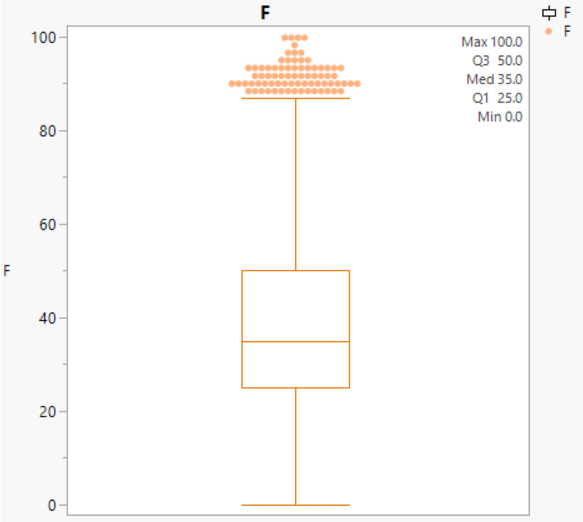
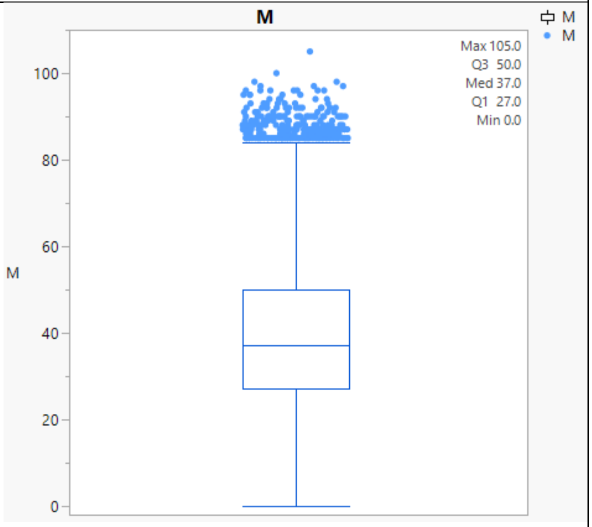
Figure 2.4</td><td>


Figure 2.5</td></tr>
</table>

<table>
<tr>
<td rowspan="2">QQ Plot</td>
<td>
Figure 2.6</td>
<td>
Figure 2.7</td>
</tr>
<tr>
<td colspan="2">Most points lie on the line; hence the dataset is normally distributed for both male and female population.</td>
</tr>
<tr>
<td>Box Plot</td>
<td>
Figure 2.8
Maximum, excluding outliers: 87
75%: 50
50%: 35
25%: 25
Minimum, excluding outliers: 0
N: 37187</td>
<td>
Figure 2.9
Maximum, excluding outliers: 84
75%: 50
50%: 37
25%: 27
Minimum, excluding outliers: 0
N: 70055</td>
</tr>
</table>

F-Test 2-Sided (To check Equality of Variances)

$$\alpha = 0.05$$

Null Hypothesis (H_o)	Alternate Hypothesis (H_a)
$\sigma_1^2 - \sigma_2^2 = 0 \;, \quad \therefore \; \sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$

F-Ratio	DF Number	p-Value
1.1600	37186	5.5543293897e-61

Since the P-Value < 0.5, our hypothesis does not hold true. Hence, it can be concluded that the null hypothesis has been rejected as there is not enough evidence that suggests equality of the variances from the two samples.

$$\therefore \sigma_1^2 \neq \sigma_2^2$$

t-Test for Two Means – Sample Assuming Unequal Variances

$$\alpha = 0.01$$

Null Hypothesis (H_o)	Alternate Hypothesis (H_a)
$\mu_1 - \mu_2 = 0 \,, \quad \therefore \mu_1 = \mu_2$	$\mu_1 \neq \mu_2$

	Female	*Male*
Mean	37.67748	38.9945
Variance	321.2938	276.9798
Observations	37187	70055
Hypothesized Mean Difference	0	
df	71103	
P(T<=t) two-tail	8.92E-32	
t Critical two-tail	2.575898	

Since the P-Value < 0.5, our hypothesis does not hold true. Hence, it can be concluded that the null hypothesis has been rejected as there is not enough evidence that suggests equality of the means from the two samples.

$$\therefore \; \mu_1 \neq \mu_2$$

Similarly, a hypothesis test can be conducted for the data given below.

Status of Patients

The dataset contained namely four statuses of the patients: i) Recovered, ii) Deceased, iii) Migrated, iv) Hospitalized.

Out of 4234 recorded status points, 3750 were confirmed to be deceased. By further splitting the dataset, it was analysed.

Status of Patients (Based on Age Bracket)

Status - Deceased

Distribution of Deceased – Age Bracket

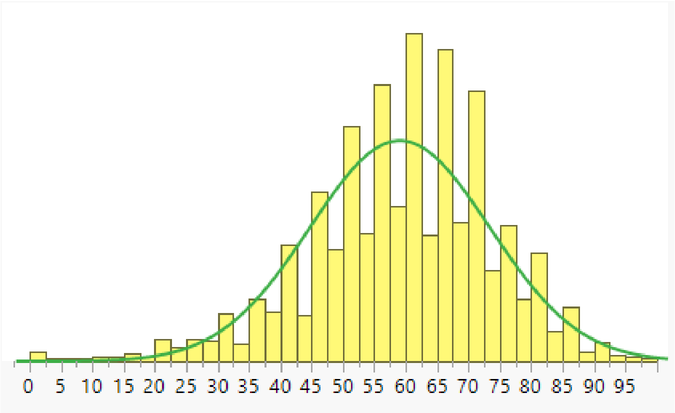


Figure 3.1

The Age Bracket of the observed deceased patients in the country has been plotted in the form of a histogram as shown in Figure 6.1.

Q-Q Plot

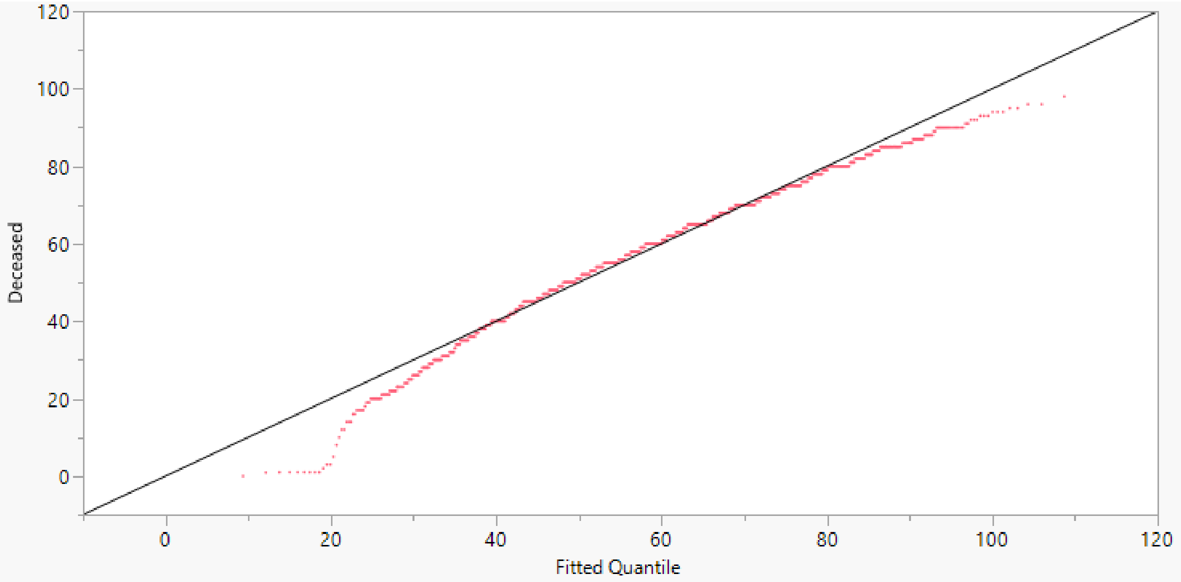


Figure 3.2

As observed many points deviate from the line, thus the curve is not ideally normal. However, for the purpose of the analysis, it has been assumed to be normal.

Quantiles for Distribution of Age Bracket

100.0%	maximum	98
99.5%		90
97.5%		85
90.0%		76
75.0%	quartile	69
50.0%	median	60
25.0%	quartile	50
10.0%		40
2.5%		27
0.5%		14
0.0%	minimum	0

Statistical Inferences

Mean	59.046373
Std Dev	14.342675
Std Err Mean	0.2342149
Upper 95% Mean	59.505574
Lower 95% Mean	58.587172
N	3750

CDF Plot

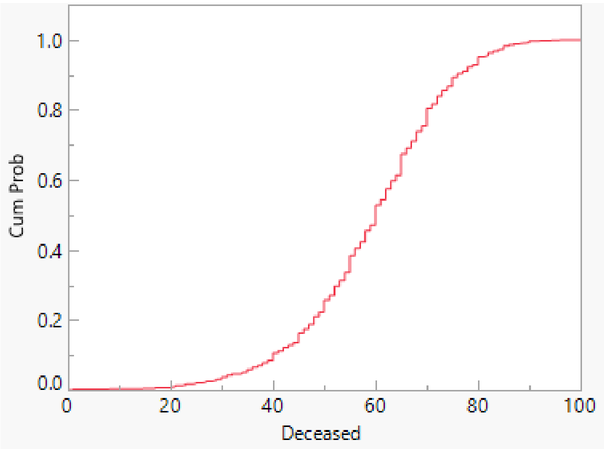


Figure 3.3

The CDF plot in Figure 3.3 shows that the cumulative probability starts increasing at the higher rate for age above 55-60 (Median being 60). This means there are more chances of fatalities above the age of 60 years as suggested by the slope of the graph.

Status - Recovered

Distribution of Recovered – Age Bracket

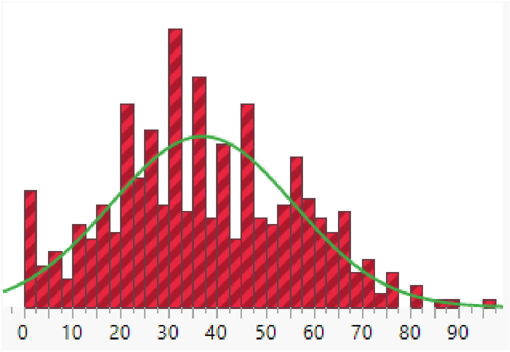


Figure 3.4

The Age Bracket of the observed Recovered patients in the country has been plotted in the form of a histogram as shown.

Q-Q Plot

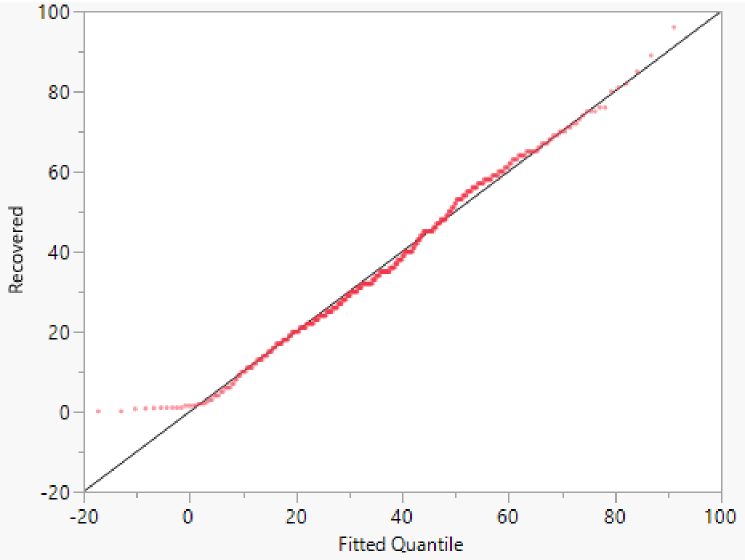


Figure 3.5

As suggested by the Q-Q Plot, most points are on the line. Hence, the given distribution is normal in nature.

Quantiles for Distribution of Age Bracket

100.0%	maximum	96
99.5%		87.38
97.5%		73.975
90.0%		63
75.0%	quartile	51
50.0%	median	35
25.0%	quartile	23.25
10.0%		13
2.5%		1.5
0.5%		0.343
0.0%	minimum	0.1

Statistical Inferences

Mean	36.956667
Std Dev	18.897472
Std Err Mean	0.8625476
Upper 95% Mean	38.651511
Lower 95% Mean	35.261822
N	480

CDF PLOT

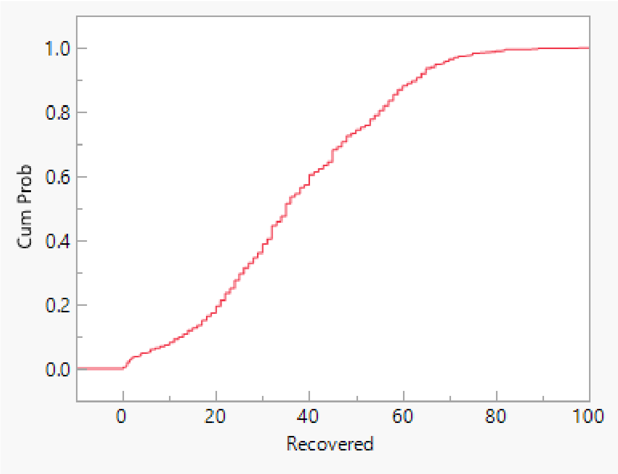


Figure 3.6

The CDF plot in Figure 10.2 shows that the cumulative probability of Recovered is increasing at a decreasing rate for the people more than the age of 40 (Median being 35). This shows that the people above the age of 40 have severe symptoms of ncov-2 and less chances of being recovered from disease as compared to younger age groups.

Status of Patients (Based on Gender)

Status - Deceased

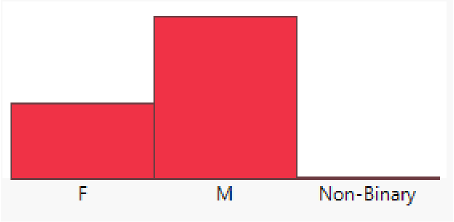


Figure 4.1

The gender of the observed deceased patients in the country has been plotted in the form of a histogram as shown in Figure 5.1. It has been noticed that there are more Male fatalities as compared to female.

Probability Distribution

Level	Count	Prob
F	1182	0.31520
M	2566	0.68427
Non-Binary	2	0.00053
Total	3750	1.00000

The Probability distribution in Figure 5.2 shows that the probability of male casualties are more than two times the probability of female casualties. This shows that this strain of ncov-2 is more contagious and catastrophic for males when compared to females.

Status Recovered

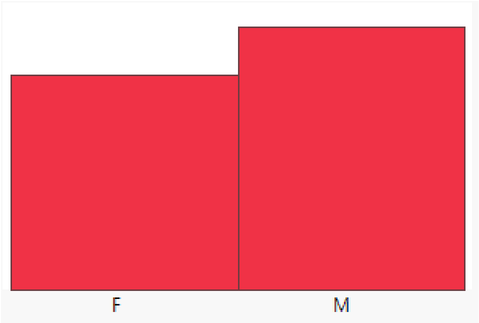


Figure 4.2

The gender of observed recovered patients in the country has been plotted in the form of a histogram as shown in Figure 7.1. It has been noticed that there are more Male recoveries as compared to Female recoveries.

Probability Distribution

Level	Count	Prob
F	216	0.45000
M	264	0.55000
Total	480	1.00000

The Probability distribution in Figure 7.2 shows that the probability of Male recoveries are more than the Probability of Female recoveries.

References

- https://www.isaca.org/resources/isaca-journal/past-issues/2011/understanding-and-applying-benfords-law

- https://brilliant.org/wiki/benfords-law/

- https://www.kaggle.com/imdevskp/corona-virus-report

Software Used

- Microsoft Excel
- JMP Pro 15
- Python Programming Language