COVID-19 Daily Case Analysis in Sri Lanka (2020–2021)

Name: Kalu Thota Hewage Tharushi Nethma Dilhari

Email: tharushinethma392@gmail.com

Phone: 0719434274

Abstract

This study analyzes daily reported COVID-19 cases in Sri Lanka from the outbreak of the pandemic until May 2021. The dataset, obtained from Kaggle, was cleaned and processed using R. Statistical methods were applied to compare daily cases between April and May 2020. Normality and homogeneity tests indicated that assumptions for parametric tests were not satisfied. therefore, a non-parametric Wilcoxon rank-sum test was conducted. The results revealed a significant difference between the two months. Additionally, visualizations such as line plots, histograms, and boxplots were generated to observe the trend and distribution of daily cases. The study highlights the importance of applying correct statistical approaches to real world health data.

Introduction

The COVID-19 pandemic, caused by the coronavirus (SARS-CoV-2), created one of the most serious health challenges in recent history. Since the first case was reported in Sri Lanka in early 2020, the country has faced many difficulties in controlling the spread of the virus. Daily reporting of new cases became very important for the government, health workers, and the public, because these numbers helped guide decisions about lockdowns, hospital preparation, and vaccination plans.

Studying the daily number of cases is useful because it shows how the disease spreads over time. By looking at trends, we can identify periods when cases increase quickly and compare them with times when the situation is more stable. This type of analysis helps health officials prepare for future outbreaks and take the right actions to protect people.

This thesis focuses on the daily COVID-19 cases reported in Sri Lanka, with special attention to the months of April and May 2020. These two months are important because Sri Lanka experienced a sharp increase in the number of cases during this time. Comparing the case patterns in April and May can help us understand how the spread changed and whether the rise was statistically significant.

To study this, we use statistical methods and data analysis in R programming. The analysis includes normality tests, variance tests, and a non-parametric Wilcoxon test to compare the two months. We also use graphs such as line plots, boxplots, and histograms to clearly show the patterns in the data.

The main aim of this thesis is to show how statistical tools can be used to understand real health problems. By analyzing COVID-19 case numbers, this study hopes to provide insights into how data can support public health decision-making in Sri Lanka.

Objectives

The main objectives are,

- To study the daily reported COVID-19 cases in Sri Lanka from the start of the outbreak to May 2021.
- To compare the daily new cases between April 2021 and May 2021.
- To check if there is a significant difference in the distribution of cases between the two months.
- To use statistical methods and graphs to better understand the spread of the virus.
- To show how data analysis can help in making decisions about public health.

Methodology

Dataset

This study uses secondary data collected from the Kaggle website, which contains the official daily COVID-19 situation reports for Sri Lanka. The dataset includes the number of new cases reported each day, total confirmed cases, and other related information up to May 2021

Variables: Date, daily confirmed cases, total confirmed cases, and other related attributes

Sri Lanka COVID-19 dataset

Data Preprocessing

The analysis was done using **R programming language** because it provides powerful tools for statistics and data visualization. The steps followed are

- Imported using read.csv in R.
- Converted date column into proper Date format.
- Extracted month information for April and May subsets.
- Renamed variables for easier usage

Statistical Analysis

- **Shapiro-Wilk Test:** Tested for normality.
- **Bartlett’s Test:** Checked for homogeneity of variances.
- **Mann–Whitney U Test.:** Applied as a non-parametric alternative to t-test.

Visualization

- **Line plots:** Show total confirmed cases trend.
- **Boxplots:** Compare distribution of daily cases.
- **Histograms:** Show frequency distribution.

Results

Normality Test

Before comparing April and May COVID-19 cases, it was important to check whether the data followed a **normal distribution**, because many statistical tests assume normality. I used these three methods to examine normality,

<u>Shapiro-Wilk normality test</u>

This is a formal statistical test to check if a dataset is normally distributed.

Hypothesis:

Null Hypothesis: The daily COVID-19 cases in April and May 2020 are normally distributed

Alternative Hypothesis: The daily COVID-19 cases in April and May 2020 are not normally distributed

Interpretation:

- If the p-value > 0.05, we fail to reject H_0, meaning the data is approximately normal.
- If the p-value ≤ 0.05, we reject H_0**,** meaning the data is not normal.

The results of the Shapiro-Wilk test for each month are presented below.

April (2020)

p-value = 6.418e-05 < 0.05

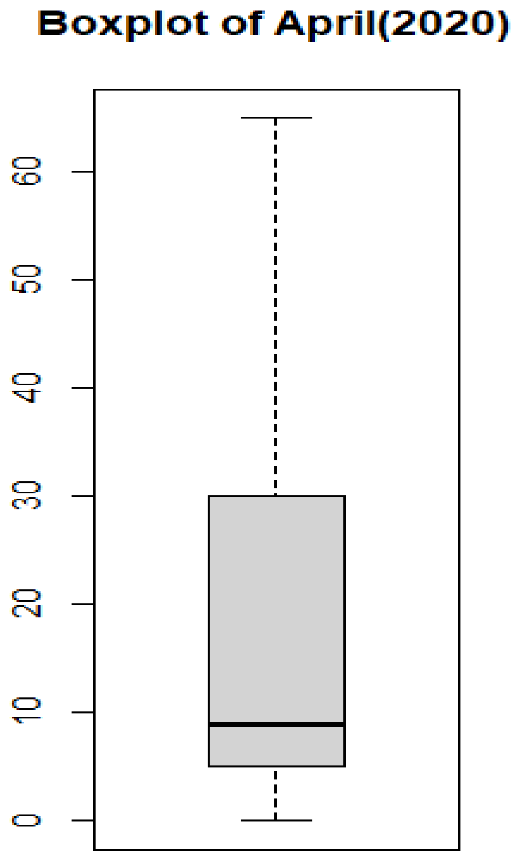
this means we can reject null hypothesis.

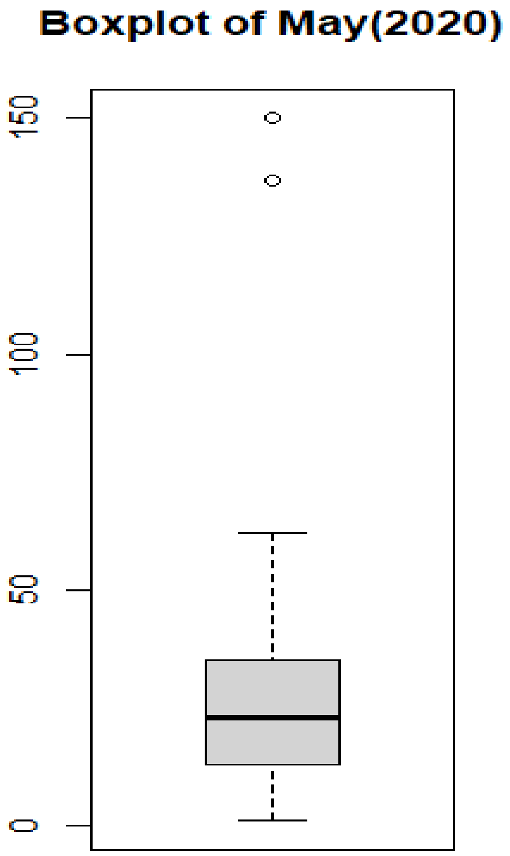
May (2020)

p-value = 5.539e-07 < 0.05
this mean we can reject null hypothesis.

These results indicate whether the daily COVID-19 cases for April and May are not normally distributed.

Boxplot





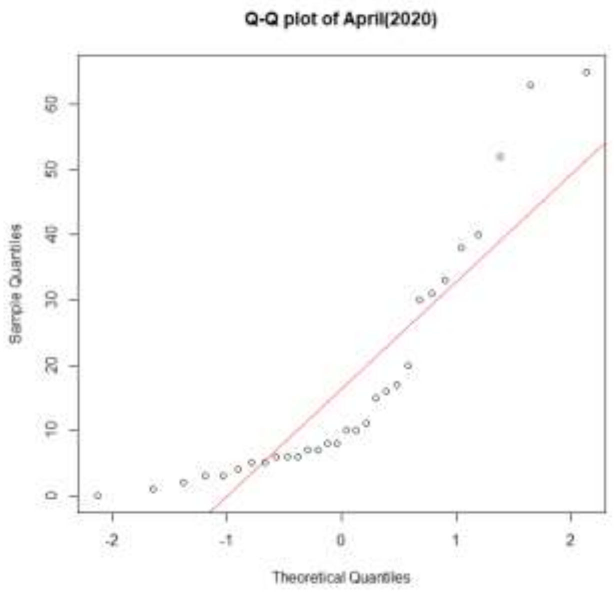
Boxplot Interpretation (April 2020):

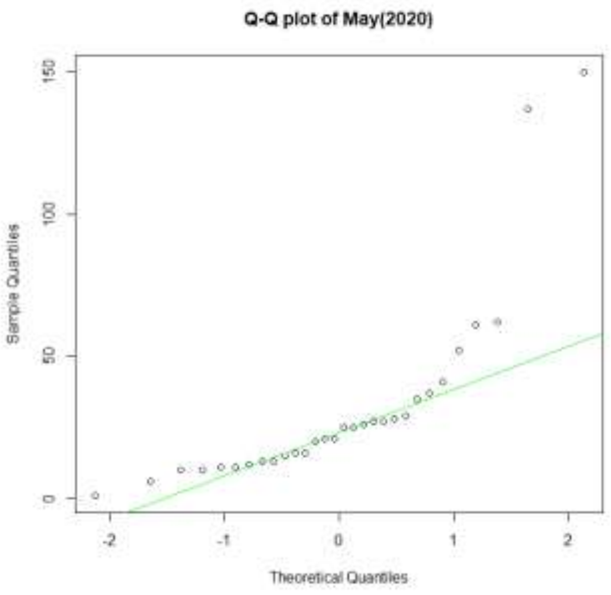
The median line is not centrally located within the box, and the plot shows a longer right whisker compared to the left. This indicates right-skewness in the data distribution. Since the boxplot does not appear symmetric and the whiskers are of unequal length, the data for April 2020 does not follow a normal distribution.

Boxplot Interpretation (May 2020):

The median line is not centered within the box, and the distribution appears right-skewed. The whiskers are of unequal length, further indicating asymmetry. Additionally, the presence of outliers shows that some days had unusually high numbers of cases compared to the general trend. These features suggest that the data for May 2020 does not follow a normal distribution.

Q-Q plot





Q-Q Plot Interpretation:

In both Q-Q plots, the data points deviate noticeably from the reference straight line. Several points show larger deviations, especially at the tails, indicating departures from normality. Therefore, these two datasets do not follow a normal distribution.

Checked for homogeneity

April and May cases did not have equal variances.

Variance (April)= 335.2828
Variance (May)= 1145.168

Bartlett's test

Null Hypothesis: The data sets are Homogenous

Alternative Hypothesis: The data sets are not Homogenous

p-value = 0.001452 < 0.05

This shows that the spread of cases in the two months is different.

Wilcoxon Rank Sum Test

Since the data was not normal and variances were unequal, the Wilcoxon test was used.

Null Hypothesis: There is **no significant difference** in the distribution of daily COVID-19 cases between April 2020 and May 2020.

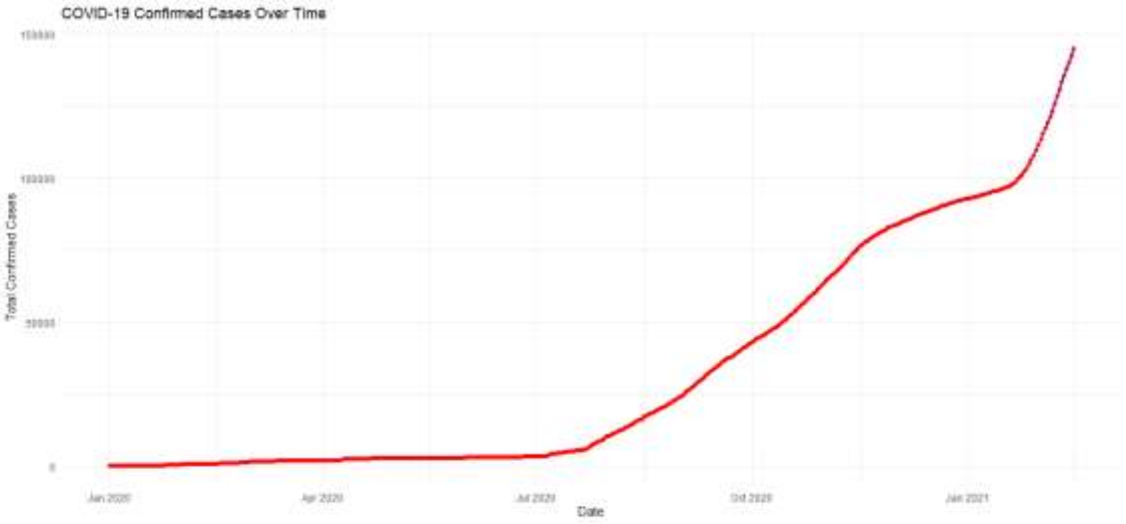
Alternative Hypothesis: There is a **significant difference** in the distribution of daily COVID-19 cases between April 2020 and May 2020.

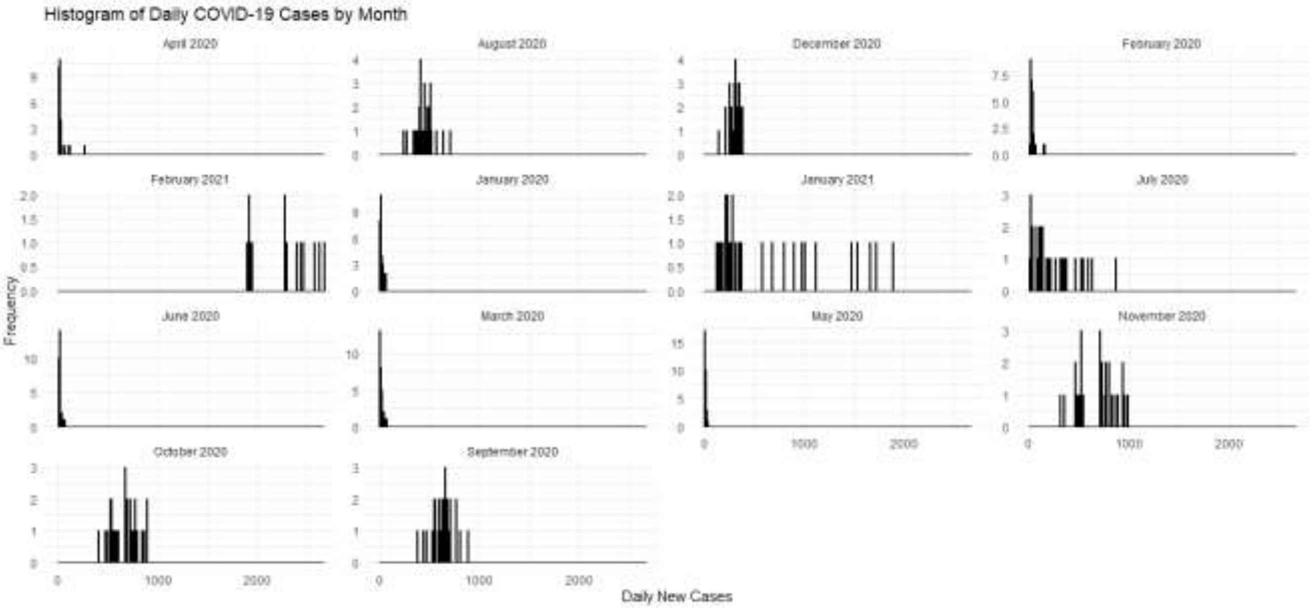
The test showed a significant difference between April and May cases, p-value = 0.0081 < 0.05

so, we **reject H₀** and conclude that May had significantly higher daily COVID-19 cases compared to April. This means that the daily cases in May were statistically higher than in April.

Data Visualization:

Line Plot




Histogram of Daily COVID-19 Cases by Month
April 2020
August 2020
December 2020
February 2020
February 2021
January 2020
January 2021
July 2020
June 2020
March 2020
May 2020
November 2020
October 2020
September 2020
Frequency
Daily New Cases
0
3
6
9
1
2
4
0.0
2.5
5.0
7.5
0.5
1.0
1.5
2.0
5
10
15
1000
2000

Discussion

The results confirm that case numbers in Sri Lanka rose significantly between April and May 2021. This aligns with reports of increased community transmission during that period. The rejection of normality assumptions emphasizes the importance of selecting robust statistical tests in epidemiological research.

This study demonstrates how statistical programming tools such as R can be used to draw meaningful insights from health data. The methodology used here can be applied to other countries or different time periods for comparative analysis.

Conclusion and Future Work

This thesis analyzed Sri Lanka's daily COVID-19 cases from January 2020 to May 2021 using R. The study found that April and May cases differed significantly, reflecting the rapid escalation of the pandemic in 2021.

Future directions include:

- Applying time-series forecasting models .
- Extending the dataset to include vaccination data.
- Comparing Sri Lanka's trend with other South Asian countries.

References

- Kaggle COVID-19 Dataset (Sri Lanka)
- World Health Organization (WHO) Reports on COVID-19
- R Documentation: Shapiro-Wilk Test, Bartlett's Test, Wilcoxon Rank-Sum Test

THANK YOU