Statistical Analysis
Of
Covid 19 Deaths and Vaccinations

Prepared For

Dr. Ranjib Banerjee
Professor
School of Engineering and Management

By

Naman Jain
<210C2030101>
Dhaval Pathak
<210C2030089>
Bhaskar Vivek Agarwal
<210C2030051>

BML MUNJAL UNIVERSITY™

SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY, GURGAON

Acknowledgement

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. We give special gratitude to our Mathematics for Engineers II faculty, Dr. Ranjib Banerjee, whose contribution in stimulating suggestions and encouragement helped us to coordinate our project especially in writing this report. We feel thrilled that under his mentorship we were able to complete this report. Last but not least, many thank goes to our team members, without them this report would not have been completed. We also have to appreciate the guidance given by other supervisors as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comments and advice.

Sincerely,

Naman Jain

Dhaval Pathak

Bhaskar Vivek Agarwal

Table of Content

Abstract

This report deals on the various hypothesis analysis of covid 19 virus across the globe. The analysis is done on the worldwide data of covid-19 available to us. The content of report focuses on the contrast in immunities regarding the two ethnicities (North Americans and South Americans).

The report illustrates this hypothesis through R programming. Study suggests that North Americans are genetically well built as compared to other races, we try to accept/reject this study through statistical analysis.

Researchers' claim has been examined in our report step by step. The belief of genetic superiority has been analyzed deeply through pure statistics and mathematics. Even scientists use this mathematics to use for analysis.

Introduction

The widespread of covid 19 virus made researchers and analysts get to work. The researchers gathered the facts of this outbreak and analysts read and comprehended this data to provide trends and future possible hypothesis. This report also does the work of analysts and focuses on the prospects of immunities affecting the deaths of covid positive patients. This hypothesis is thoroughly analyzed and is based on mathematics. The null and alternate hypothesis is used for analysis. The claim of researchers has been accepted or rejected in the data analysis. The study is done through hypothesis testing, and claims are accordingly accepted or rejected.

Symptoms of COVID19 include fever, cough, chest tightness or pain, fatigue, and sore throat. However, asymptomatic infections have also been reported. In several studies, it was found that during infection, the patient's white blood cell count was normal (71.4%) or decreased (28.6%) in nearly 70% of patients, and leukopenia was observed. Demand is found in 50% (7/14) of them (Su et al., 2020). Dr. Sakoulas estimates that 80% of COVID19 participants will not require medical care, 15% may require non-ICU medical care, and 5% may require an ICU stay. Pneumonia puts the patient's life in danger. COVID19 appears to be milder in children than in adults. Approximately 90% of pediatric patients are diagnosed with asymptomatic, mild or moderate disease. According to statistics, a quarter of deaths from COVID19 occur in people between 70 and 79 years old. Up to two-thirds occur in people over 80 years of age, regardless of incidence or the extent to which deaths are recorded in countries.

Another analysis of research is done over the vaccination status versus the number of cases rising. We hypothesized that increase in vaccination drive should decrease the number of covid cases. However, this hypothesis may be true or not.

Both the hypothesis is globally relevant as it tells the final outcome of the research papers published. The usage of R programming is utilized here to get desired outcomes.

The Study

Source of Data

The data used in this report is obtained from the public database "Our World in Data". The data is public and available at www.ourworldindata.org/covidvaccination. The data includes the number of confirmed cases of COVID-19, the number of deaths, and the number of vaccinations administered to people in 48 Asian countries between 24 February 2020 and 26 September 2021. However, the database is updated daily on the Web site and includes all information related to the COVID pandemic. This dataset contains information on deaths and vaccinations that occurred around the world between February 2020 and May 2022. Data cleaning and pre-processing is an important step before starting any analysis as it makes our data reliable for further calculations. Data from online databases provide structured data on a daily basis. All variables were converted to monthly data, i.e. COVID cases, deaths and vaccinated people. The data were normalized to fit a range and statistical calculations were performed using the standard scale and the minimum scale. After normalizing and normalizing the data, logarithmic transformations are also performed so that the data can become reliable for analysis. Following the data cleaning, statistical calculations was performed Z score, t test, p value method and statistical graph to get the maximum insight on the number of deaths and vaccination rate during the pandemic era. All analysis was performed on the R studio platform.

Study

Statistical analysis is essential in verifying assumptions and demonstrating them to create a concrete conclusion about a study. This study focuses on the immunity power (Death rate) between the North America and South America.

Hypothesis testing works by collecting data and measuring the probability of a particular set of data (assuming the null hypothesis is true), when studied on a randomly selected representative sample. The null hypothesis assumes that there is no relationship between the variables in the population from which the sample was selected.

The structure of hypothesis testing will be formulated with the use of the term null hypothesis, which refers to any hypothesis we wish to test and is denoted by H0. Upon rejection of H0, an alternative hypothesis is accepted, denoted by H1. In order to understand the rudiments of hypothesis testing, it is crucial to understand the different roles played by the null hypothesis and alternative hypothesis. It is important to define the alternative hypothesis H1 because it refers to the question to be answered or the theory to be tested. The null hypothesis H0 opposes H1 and is often a logical complement to H1.

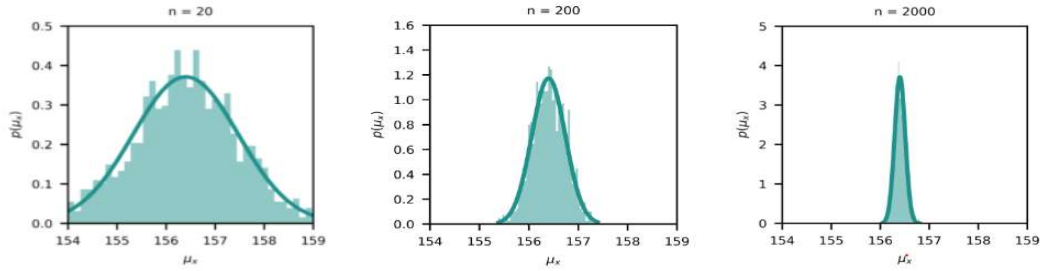


Figure 1: Distribution of sample means as a function of sample size. The standard deviation of this distribution becomes narrower with more samples (source: author).

When choosing or rejecting a hypothesis, one can led to one of two erroneous conclusions if there is a decision error. In fact, the new vaccine was no more beneficial than the currently used (H0 true), but in a group of 10 randomly selected, more than 8 people survived Corona virus episodes, despite the fact that the new vaccine may not be as effective as the old one. H0 is true and we would make a mistake to reject it in favour of H1. This is known as a type I error. A type I error is when the null hypothesis is rejected when it is true.
A second kind of error is committed if 8 or fewer of the group survive the Corona virus waves and we are unable to conclude that the new vaccine is better than the old vaccine (H1 true). A type II error occurs when we fail to reject H0 when H0 in fact is false.
Nonrejection of the null hypothesis when it is false is called a type II error.

p-Value
As mentioned above, statistical hypothesis testing involves comparing groups and the aim is to assess whether the differences between groups are significant based on estimated sample statistics. For this purpose, the complete statistics, their respective confidence intervals and values are calculated. The value is the probability associated with the t-statistic using the t-distribution, similar to the probability associated with the z-score score and the Gaussian distribution.

False positives and false negatives
If the value is below a certain threshold α, the difference is said to be statistically significant. The rejection of the null hypothesis when it is indeed true is called a type I error (false positive), and the probability of a type I error is called the significance level ("some threshold") α. Accepting the null hypothesis when it is false is called a type II error (false negative) and its probability is denoted by β. The probability, that the null hypothesis is rejected when it is false is called the power of the test and is equals 1β. By being stricter with the significance level α, the risk for false positives can be minimized. However, tuning for false negatives is more difficult because the alternative hypothesis includes all other possibilities.

Data Analysis

- According to several reports it was stated that immunity power of North America is highest followed by South America. Upon this data it was inferenced that number of covid related deaths of North Americans will be lesser than or equal to South Americans. But some researchers raised their concerns that covid related deaths will be more among South Americans than North America. So, we need to analyze the real time death rate and state whether the hypothesis stated is rejected or not.

According to reports the stated null hypothesis is: -

$$\mathbf{H_0: \mu_1 <= \mu_2,}$$

μ_1 = Population Mean of Deaths among North Americans
μ_2 = Population Mean of Deaths among South Americans

According to Concerned Researchers the stated alternative hypothesis is: -

$$\mathbf{H_a: \mu_1 > \mu_2,}$$

μ_1 = Population Mean of Deaths among North Americans
μ_2 = Population Mean of Deaths among South Americans

Now, we need to check whether the stated alternative hypothesis is true or not and based upon the results we need to conclude that whether we have rejected null hypothesis or we failed to reject null hypothesis.
Here, we used covid 19 deaths and vaccination data to test our hypothesis to reach towards our desired output.

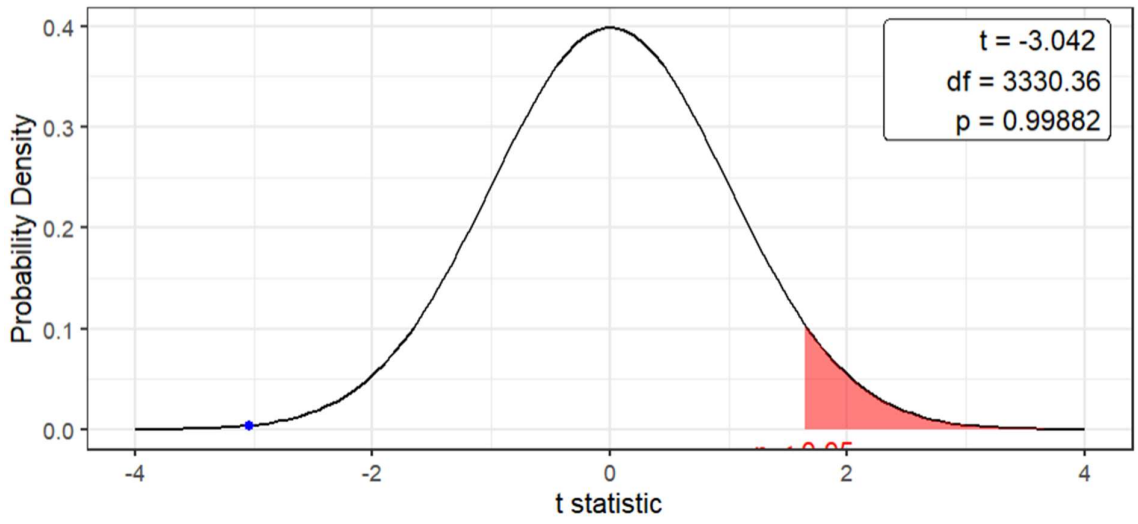
Here, we have performed t testing as our population std. deviations are unknown and sample size is less than 30.
We have taken significance level 0.05 for our data analysis.

```
> # Null Hypothesis: mu1 <= mu2
> # Alternative Hypothesis: mu1 > mu2
>
> library(webr)
> data <- read.csv("C:/Users/naman/Downloads/covid_data_cleaned1.csv")
> data$new_deaths <- as.integer(data$new_deaths != 0)
> data$new_deaths[is.na(data$new_deaths)] = 0
> m1 = sum(data$new_deaths) / nrow(data)
>
> s1 <- subset(data$new_deaths,data$continent == "North America")
> mu1 <- mean(s1, na.rm = TRUE)
> s2 <- subset(data$new_deaths,data$continent == "South America")
> mu2 <- mean(s2, na.rm = TRUE)
>
> x = t.test(s1, s2, alternative="greater", conf.level = 0.95)
> print(x) # t test reults

        Welch Two Sample t-test

data:  s1 and s2
t = -3.0422, df = 3330.4, p-value = 0.9988
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.05077567         Inf
sample estimates:
mean of x mean of y
0.8575739 0.8905273

> plot(x) # Hypothesis graph
```



- ❖ According to several reports it was stated that Vaccination drive will decrease the no. of new cases but some people say that vaccination will not affect the no. of new cases found and it will keep om increasing. So, our task is to analyze the give the final verdict whether the vaccination drive will affect the no of new cases found or not.

According to reports the stated null hypothesis is: -

$$\mathbf{H_0: \mu_1 <= \mu_2,}$$

μ_1 = Population Mean of new cases found
μ_2 = Population Mean of no of people vaccinated

According to Concerned Researchers the stated alternative hypothesis is: -

$$\mathbf{H_a: \mu_1 > \mu_2,}$$

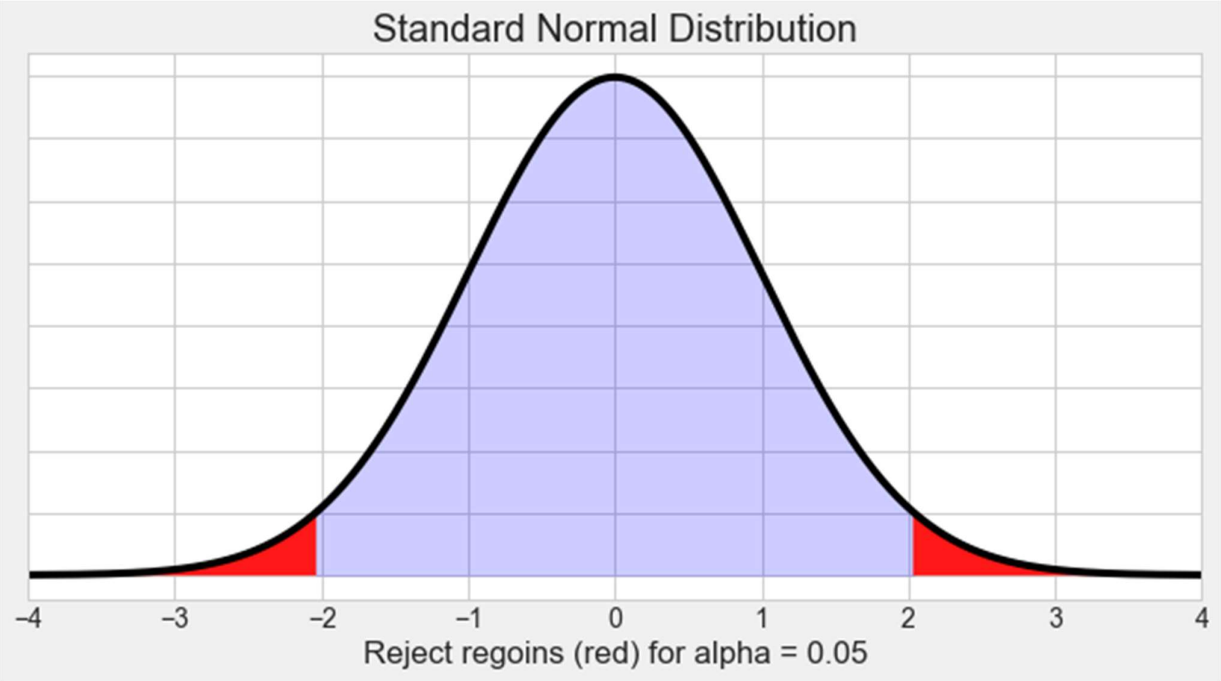
μ_1 = Population Mean of new cases found
μ_2 = Population Mean of no of people vaccinated

Now, we need to check whether the stated alternative hypothesis is true or not and based upon the results we need to conclude that whether we have rejected null hypothesis or we failed to reject null hypothesis.
Here, we used covid 19 deaths and vaccination data to test our hypothesis to reach towards our desired output.
Here, we have performed Z testing as our population std. deviations are known and sample size is greater than 30.
We have taken significance level 0.05 for our data analysis.

```
> # Null Hypothesis: mu1 <= mu2
> # Alternative Hypothesis: mu1 > mu2
>
> library(webr)
> library(BSDA)
> data <- read.csv("C:/Users/naman/Downloads/covid_data_cleaned1.csv")
> data$new_cases[is.na(data$new_cases)] = 0
>
> s1 <- data$new_cases
> mu1 <- mean(s1, na.rm = TRUE)
> std1 <- sd(s1)
> s2 <- data$new_vaccinations
> mu2 <- mean(s2, na.rm = TRUE)
> std2 <- sd(s2)
>
> x = z.test(s1,s2,sigma.x = std1,sigma.y =  std2, conf.level = 0.95, alternative = "greater")
> print(x) # Z test results

        Two-sample z-Test

data:  s1 and s2
z = -40.466, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -178350.6        NA
sample estimates:
mean of x mean of y
 17365.37 188749.54
```



Conclusion

- ✓ As Our P value = 0.9988 which is greater than 0.05 and we failed to gather sufficient statistical evidences that's' why we failed to reject null hypothesis. By using various tools of hypothesis testing, we came to understand that the research claim is invalid and thus, Final Verdict is that the researchers were wrong and deaths among North Americans is greater than the deaths among South Americans.
- ✓ As Our P value = 1 which is greater than 0.05 and we failed to gather sufficient statistical evidences that's' why we failed to reject null hypothesis. By using various tools of hypothesis testing, we came to conclusion that those people were wrong who said that vaccination will not affect rate of new cases and the reports we correct.

In both the hypothesis we failed to rejected the null hypothesis and nullified the alternative hypothesis. This proves that our assumed null hypothesis was correct. Emphasis on tests of significance and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective.

References

- https://www.kaggle.com/datasets/digvijaysinhgohil/covid19-data-deaths-and-vaccinations
- https://www.cdc.gov/coronavirus/2019-ncov/index.html/
- Ancker, J. S. 2020. “The COVID-19 Pandemic and the Power of Numbers.”
- Numeracy 13(2): Article 2. https://doi.org/10.5038/1936-4660.13.2.1358
- Ancker, J. S., and M. D. Begg. 2017. “Using Visual Analogies to Teach
- Introductory Statistical Concepts.” Numeracy 10(2): Article
- https://doi.org/10.5038/1936-4660.10.2.7
- Best, J. 2020. “COVID-19 and Numeracy: How about Them Numbers?”
- Numeracy 13(2): Article 4. https://doi.org/10.5038/1936-4660.13.2.1361
- De Veaux, R. D., P. F. Velleman, and D. E. Bock. 2020. Stats: Data & Models, 5th edition. Hoboken, NJ: Pearson