MATH7343 Project - Coronavirus Epidemic and Vaccination Study

Team E: Ruobing Bai, Yahan Yang, Wenrui Zhao

ABSTRACT

As a worldwide pandemic since 2019, COVID-19 has influenced life for all human beings. The vaccination and epidemiology research and study for COVID-19 is crucial to the world. Therefore, this project aims to use statistical methods to study the vaccination and epidemiology for COVID. In this report, we use the T-test to compare and estimate the world increase rate for infection of COVID between different periods. The COVID lethal rates between genders will be compared using χ^2test. On the vaccination side, different progress on vaccination between countries will be evaluated with the Wilcoxon rank-sum test and ANOVA. Besides, by using χ^2test and spearman correlation analysis, the association and correlation between vaccination and infection will be analyzed. In conclusion, COVID is under control in April 2021 compared with October 2020, and the lethal rate between genders is different. On the other side, different countries have different vaccination progress, but the United States and the United Kingdom have more advanced vaccination progress. There is statistical significant showing for the association between vaccination and infection, and it is proved that these two variables are negatively correlated.

INTRODUCTION

Coronavirus disease 2019(COVID-19) is a contagious disease caused by the SARS-CoV-2 virus that could cause severe respiratory infections. According to the CDC, the first case of COVID-19 was discovered in Wuhan China in December 2019, then quickly spread to the world. After more than one year of research on the disease, the vaccines for COVID-19 were approved by different countries to control this global pandemic. Based on IBM COVID vaccination news, pharmaceutical companies such as Pfizer, Moderna, and AstraZeneca have been approved to provide vaccines for COVID since 2020. Due to the scarceness of vaccine resources, the vaccination rates are variable among different countries. As a global epidemic disease, vaccination could be critical to the pandemic. Thus, under this situation, this project will use different statistical methods to explore topics regarding COVID-19 disease infection and vaccination.

This project will research the following perspectives:

❏ *Is the increase rate on Oct. 10th, 2020 higher than the increase rate on Apr. 10th, 2021?*

The increased rate is calculated using deaths over confirmed cases. The paired t-test will be applied to see if those 187 countries have higher increase rates on Oct. 10th, 2020 compared to Apr. 10th, 2021.

❏ *Which countries are more advanced in vaccination? And which are not?*

Based on the 6 countries we chose from our dataset, the project compares the progress of vaccination for any two countries. And by having an overview of the confirmation and death proportions, the project determines if the two proportions have a critical influence on vaccination progress.

- ❑ *Does COVID-19 have the same lethal rates between genders?*

  Since some diseases have different lethal rates between genders, the project evaluates if the death rates between the male population and female population are the same.

- ❑ *Association and correlation between vaccination and infection.*

  Back in the beginning, the vaccination for COVID-19 became a critical method to control the spread of the disease. Thus, the association coefficient between the U.S vaccination and infection data is evaluated in this project. As proof of the results for association analysis, correlation analysis will be applied to the world dataset between the same variables.

According to the research problems above, some key statistical tests including the two-sample T-test, ANOVA analysis, Wilcoxon rank-sum test, χ^2test will be applied to the specific datasets. Since the datasets are collected from observational studies, we are trying to find associations between factors other than causation. This paper will report the project in 5 parts: a review of data, research methods regarding different problems listed above, analysis and results, discussion for further improvement, and conclusion.

REVIEW OF DATA

The project uses four major datasets: *COVID-19 world vaccination progress* dataset which is collected from Our World in Data GitHub repository, merged and uploaded on Kaggle's; *real-time COVID-19* dataset which is collected and maintained by Johns Hopkins University Center for Systems Science and Engineering; *the COVID-19 sex-disaggregated data tracker project* which is developed by Global Health 50/50, the African Population and Health Research Center and the International Center for Research on Women; and *the confirmed cases versus the not infected people under different condition of vaccination(fully two doses, one dose and not vaccinated)*, the dataset is collected and researched by CDC.

The major dataset we used in the project is the *world vaccination progress* dataset and the *real-time COVID-19* dataset. Because the majority vaccines take more than 14 days to generate antibodies, we combine the data for March 24th, 2021 from the infection dataset and the data for March 1st, 2021 from the vaccination dataset by country, and plot the data with pairs.panel function by R. As Figure 1 shows below, the diagonal graphs are histograms for each variable, the lower left side graphs are the scatter plots for each pair of variables, and the upper right side are the Pearson's correlations between variables. The columns from left to right (and the rows from up to bottom) correspond to confirmed cases, recovered cases, deaths, increase rates which are calculated using deaths over the confirmed cases (which come from the infection dataset), and total vaccinations, people vaccinated, people fully vaccinated, daily vaccination raw, daily vaccinations, total vaccinations per hundred, people vaccinated per hundred, people fully vaccinated per hundred, and daily vaccination per million (which come from the vaccination dataset).

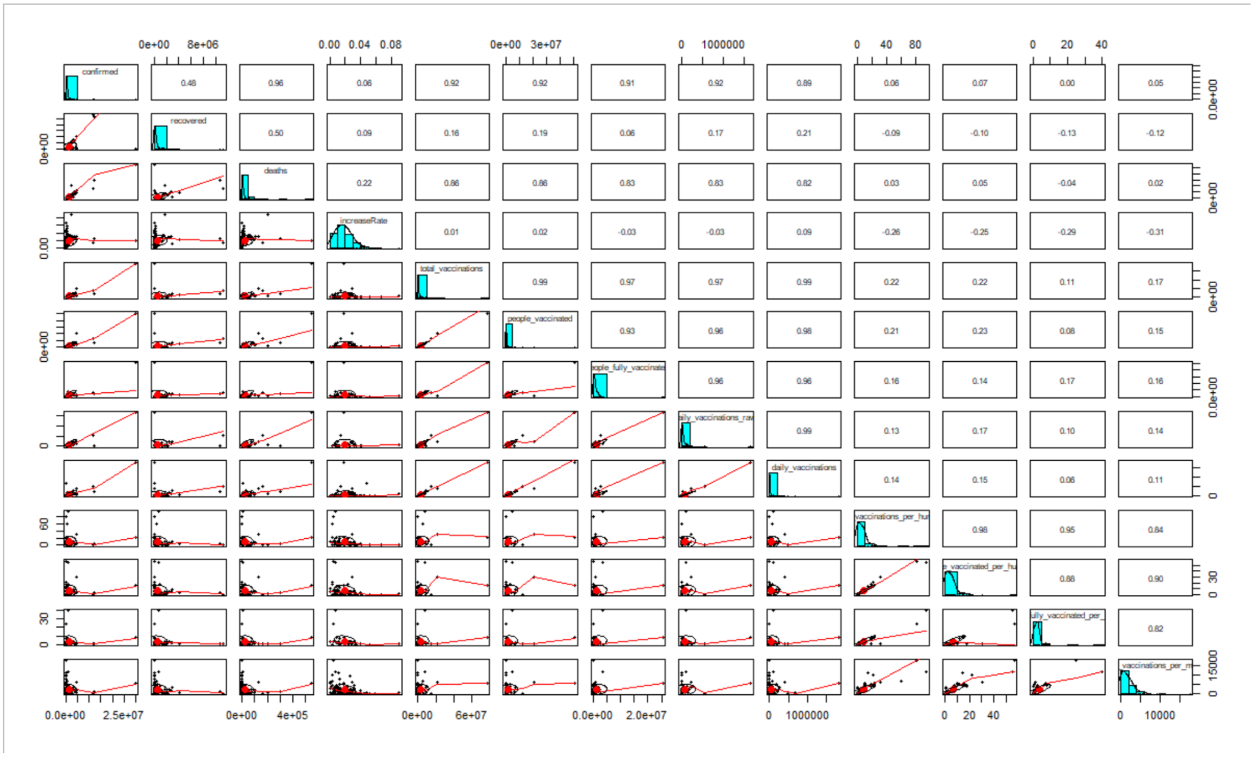


Figure 1: Scatter Plot, Histogram and Pearson Correlation between Variables

Based on the histograms, we conclude that the increase rate on March 24th and daily vaccination per million on March 1st are approximately normally distributed. Thus, the research for other variables should use non-parametric methods.

As for the gender death frequency dataset, we plot the bar plot as shown in Figure 2, there are slightly more Male death cases compared with females'. Thus, it is important to use statistical methods to prove if there is a significant difference in the death frequency between genders.

Apart from the review as above, other details for the datasets will be discussed in the following part.

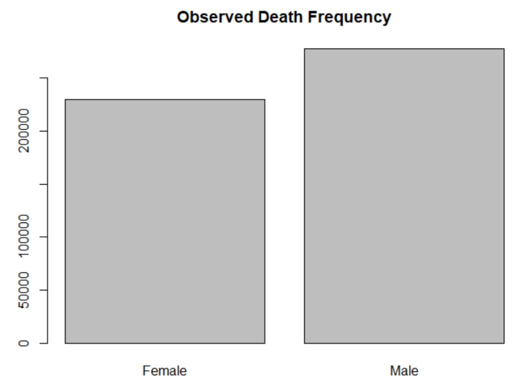


Figure 2: Barplot for the U.S observed Death Frequency(Until 3/6/2021)

RESEARCH PROBLEMS: METHODS, ANALYSIS, AND RESULTS

Q1: Is the increase rate on Oct.10th, 2020 higher than the increase rate on Apr. 10th, 2021?

<u>Method</u>

We aim to determine if all countries have higher increase rates on Oct. 10th, 2020 compared to Apr. 10th, 2021, and we use the increase rate from the 187 countries on those two dates to test the null hypothesis that the mean increase rate on Oct. 10th, 2020 is less or equal to the mean increase rate on Apr. 10th, 2021.

We choose to apply the paired t-test because we have the data for the increase rate of each country at a different time, which will be dependent, and the data for each date is approximately normally distributed. As Figure 3 shows below, two histograms are approximately normally distributed with minor outliers and skewness. Thus, we applied the paired t-test to the data.

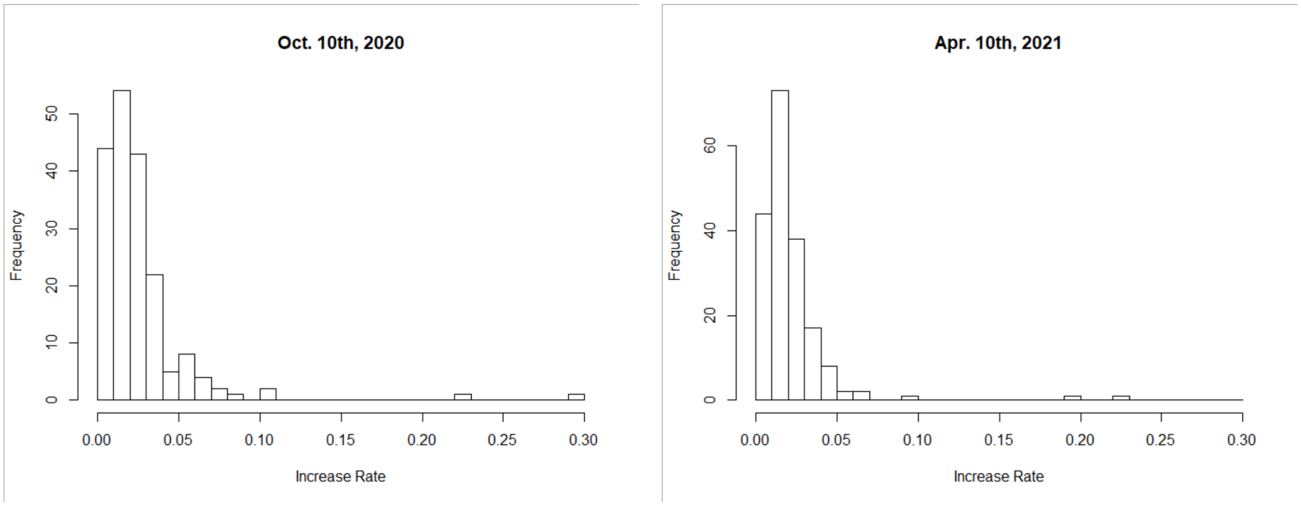


Figure 3: Histogram for Infection Increase Rate on Oct 10th, 2020 and April 10th, 2021

Analysis

For the one-sided paired t-test, we have data from 187 countries, so the degree of freedom is 186. Moreover, from R we get t = 4.1401 and p-value $= 2.63 * 10^{-5} < 0.05$, so we reject the null hypothesis.

Result

Based on the results, we conclude that the mean increase rate on Oct. 10th, 2020 is higher than the mean increase rate on Apr. 10th, 2021. In other words, among the confirmed cases, they have a lower death rate on Apr. 10th, 2021 compared to Oct. 10th, 2020. Thus, we believe that the pandemic is getting better since last year.

Q2: Vaccination Progress Study

Methods

We chose 6 countries (China, US, UK, Australia, India, and Japan) out of 126 countries to do tests on the proportions of confirmed cases, death, daily vaccinations; ratios of confirmed cases to daily vaccinations, and deaths to daily vaccinations.

1. ANOVA

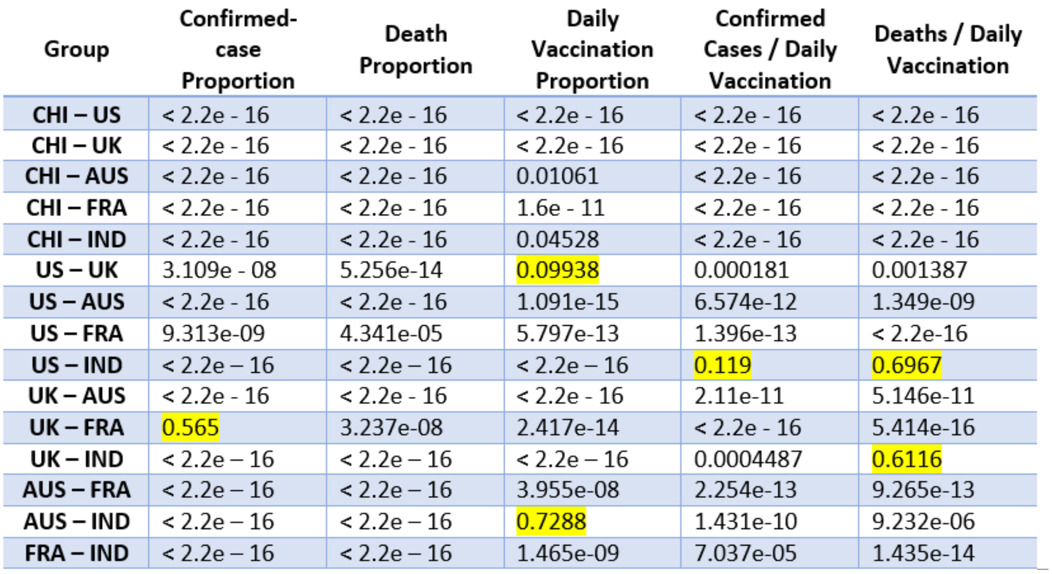
   We checked and found that the proportions and ratios mentioned above are not normally distributed. So, we used the log function, sqrt function, and sin-1sqrt function to normalize the data. However, all of the transformed data rejected Shapiro-Wilk null hypothesis and are not normally distributed. The reason behind this may be the lack of data, which is unsolvable at this time because we are just in the middle of vaccination progress so we can only obtain more data as the vaccination progress advances.

2. Wilcoxon Rank Sum Test

   When our samples are small and non-normal, the advantages of the Wilcoxon Rank Sum Test are highlighted. The Wilcoxon Rank Sum Test is often described as the non-parametric version of the two-sample t-test. It does not assume our data have a known distribution. And since the Wilcoxon Rank Sum Test does not assume known distributions, it does not deal with parameters, and therefore we call it a non-parametric test.

Analysis

Subject to the normal distribution problem, we cannot use ANOVA, but we can choose tests that don’t need a normal distribution assumption. So, we used the Wilcoxon Rank-sum test. The results are listed in Table 1 below.



Group	Confirmed-case Proportion	Death Proportion	Daily Vaccination Proportion	Confirmed Cases / Daily Vaccination	Deaths / Daily Vaccination
CHI – US	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16
CHI – UK	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16
CHI – AUS	< 2.2e - 16	< 2.2e - 16	0.01061	< 2.2e - 16	< 2.2e - 16
CHI – FRA	< 2.2e - 16	< 2.2e - 16	1.6e - 11	< 2.2e - 16	< 2.2e - 16
CHI – IND	< 2.2e - 16	< 2.2e - 16	0.04528	< 2.2e - 16	< 2.2e - 16
US – UK	3.109e - 08	5.256e-14	0.09938	0.000181	0.001387
US – AUS	< 2.2e - 16	< 2.2e - 16	1.091e-15	6.574e-12	1.349e-09
US – FRA	9.313e-09	4.341e-05	5.797e-13	1.396e-13	< 2.2e-16
US – IND	< 2.2e – 16	< 2.2e – 16	< 2.2e – 16	0.119	0.6967
UK – AUS	< 2.2e - 16	< 2.2e - 16	< 2.2e - 16	2.11e-11	5.146e-11
UK – FRA	0.565	3.237e-08	2.417e-14	< 2.2e - 16	5.414e-16
UK – IND	< 2.2e – 16	< 2.2e – 16	< 2.2e – 16	0.0004487	0.6116
AUS – FRA	< 2.2e – 16	< 2.2e – 16	3.955e-08	2.254e-13	9.265e-13
AUS – IND	< 2.2e – 16	< 2.2e – 16	0.7288	1.431e-10	9.232e-06
FRA – IND	< 2.2e – 16	< 2.2e – 16	1.465e-09	7.037e-05	1.435e-14

Table 1: Wilcoxon Rank Sum Test P-values Between Countries

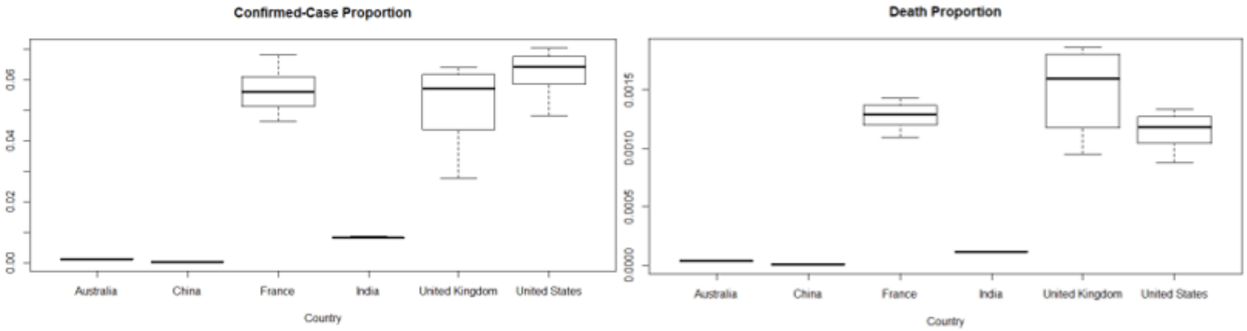


Figure 4: Boxplot for World Confirmed Case Proportion and Death Proportion

Combined with boxplots shown in Figure 4, we can find that the UK and France have the same mean confirmed case proportion. It may be due to the similar population density, so the possibility of people getting infected is close. About the other four countries, the US has a larger mean proportion than the other three countries. But when it comes to the mean death proportion, all countries have significantly different mean death proportions. We noticed that the death proportion in the UK is bigger than in France, which states that France has a more advanced medical system to prevent people from dying after being infected. And obviously, the US has a more efficient way to save confirmed patients' lives, so that it has a larger confirmation proportion than the UK and France but has a smaller death proportion.

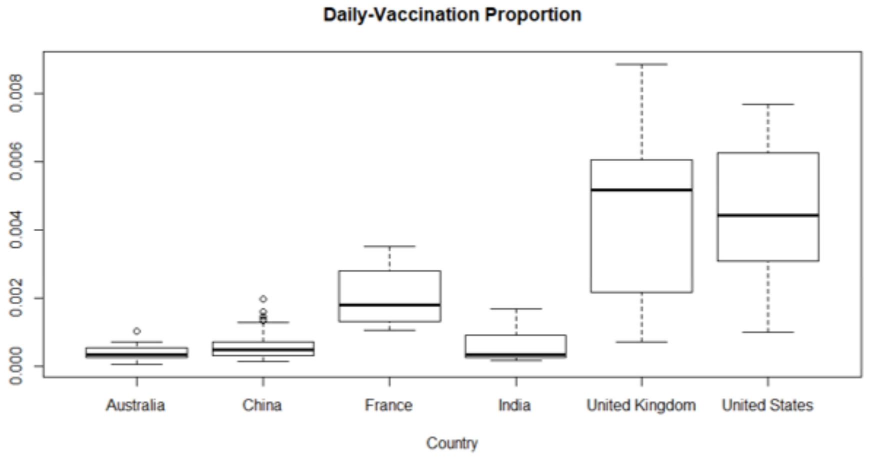


Figure 5: Boxplot for World Daily Vaccination Proportion

And for the daily vaccination proportions, based on Figure 5, the test results show that India-Australia, US-UK both have the same mean daily vaccination proportion. Although the mean confirmed case proportion in France is the same as it is in the UK, the daily vaccination proportion is smaller than in the UK, which means that the UK is trying harder than France to lower the damage of pandemic to people by increasing the number of vaccinations.

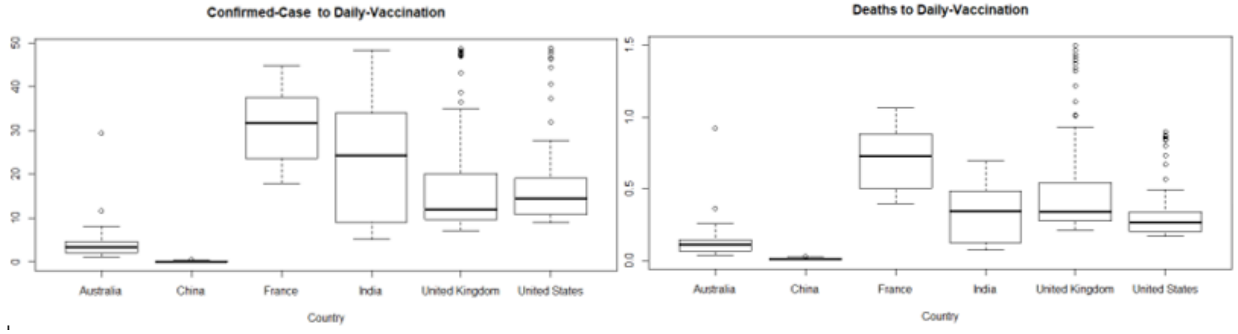


Figure 6: Boxplot for Ratios of Confirmed-Case and Deaths to Daily vaccinations

About the ratios of confirmed cases and death cases to daily vaccination, as shown in Figure 6, the US and India have the same mean ratios, the UK and India have the same second mean ratio. Different from the performances in the first three proportions, India has lower vaccinations compared to other countries, but India does not have high confirmation cases and death proportions. It indicates that apart from vaccination, other methods can also be efficient to control the epidemic. Besides, there are plenty of outliers for the US and the UK, but the high ratio data is only located in the early stage. Combined with confirmation and death proportion in the two countries, we believe that in the US and the UK, vaccination made confirmed cases and death proportions drop significantly.

Result

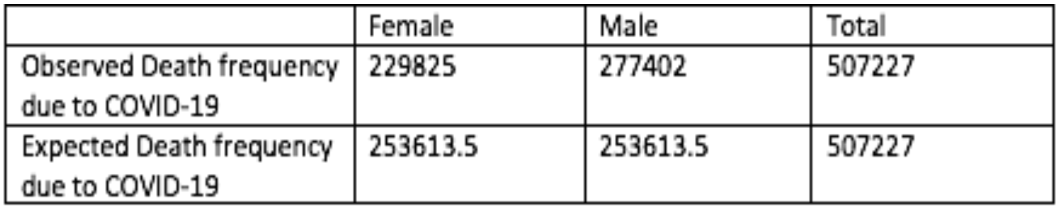
In conclusion, France, the US, and the UK have high confirmed cases and death proportions, but they also made significant success in controlling epidemics through efficient vaccination progress. Australia and China have good mechanisms to not only control confirmed cases but also carry out the vaccination. India has a low confirmed case proportion and death proportion, but they need to take more effort on vaccination.

Therefore, the US and UK have the most advanced vaccination progress, and India has the least among the 6 chosen countries. Moreover, the high confirmed case proportion and death proportion did promote vaccination progress in that country, but we cannot conclude that the country with less advanced vaccination progress is under a hostile epidemic environment.

Q3: COVID-19 death frequency between different genders is different.

Method

A solution to this problem is to test if the distribution for female and male death frequency due to COVID-19 fits the estimation that they are the same. The dataset used for this problem is *the sex, gender and data tracker for COVID-19*. The data are obtained from a random sample and the data between different genders could be considered as i.i.d data. Build Table 2 shown below, the expected frequencies for different genders are larger than 5. Thus, χ^2Test for Goodness of fit could be applied to this problem.



	Female	Male	Total
Observed Death frequency due to COVID-19	229825	277402	507227
Expected Death frequency due to COVID-19	253613.5	253613.5	507227

Table 2: Observed and Expected Values for COVID Death Frequency between Genders

Analysis

Expected values for male and female death frequency = 507227 $\div\ 2 = 253613.5$, conduct the null hypothesis and alternative hypothesis as below:

H_o: The death frequencies of COVID for females and males equals the expected values.

H_A: The death frequencies of COVID for females and males are different from the expected values.

With degree of freedom = 2-1 = 1, let α= 0.05, the critical value is 3.84.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} = \frac{(229825 - 253613.5)^2}{253613.5} + \frac{(277402 - 253613.5)^2}{277402} = 4462.64$$

Result

Based on the above equation, 4462.64 > 3.84, so reject the null hypothesis, there is enough evidence to show that the death frequency of COVID between genders is not as expected, which is different.

Q4: Association between vaccination and infection cases

Method

To evaluate the association between vaccination and infection cases, a χ^2Test for the association could be conducted. The sample data used for this problem is the frequency of confirmed cases and not infected frequency for different conditions of vaccination in the United States. The contingency table is shown below.

	Confirmed cases	Not infected	Total
Two doses (full)	3	78899	78902
One dose	8	41848	41856
Not vaccinated	161	116496	116657
Total	172	237243	237415

Table 3: Contingency Table for Vaccination vs Confirmed cases

Analysis

H_o: There is an association between vaccination and infection cases.

H_A: There is NOT an association between vaccination and infection cases.

Using R to compute the χ^2coefficient based on the contingency table with degree of freedom = (3-1)(2-1) = 2, get p-value < 2.2e-16.

Result

P-value < 2.2e-16, which is less than $\alpha = 0.05$, so reject the null hypothesis, there is enough evidence showing that there is an association between vaccination and infection of COVID.

Q4: Infection cases are negatively correlated to vaccination frequency.

Method

Based on the above problem and results, we could know there is an association between vaccination and frequency, but we do not know if it is a negative relationship or a positive relationship. Thus, the problem to evaluate the correlation between infection cases and vaccination frequency is necessary. Spearman correlation analysis is used to solve this problem because the vaccination frequency and infection cases are not normally distributed.

To prove the results from the previous problem based on the data only in the United States is also true for the world at the same time, we used a different dataset from the previous problem. In this problem, we used the *world vaccination progress* dataset and *world real-time COVID-19* dataset. Since most of the vaccines take more than 14 days to create antibodies, it is important to use the vaccination data at least 14 days before the infected increase rate data. Thus, we extracted the COVID increase rate data for 115 countries around the world on the date of March 24, combined with the daily vaccination rate for the same 115 countries on the date of March 1, and conducted the analysis as below.

Analysis

Conduct the Spearman correlation analysis hypothesis:

H_0: Infection of COVID is not correlated to vaccination rate

H_A: Infection of COVID is correlated to vaccination rate

Based on the output from R, p-value = 0.03 smaller than the significant level 0.05, so reject the null hypothesis, we could believe there is a correlation between the COVID confirmed case and vaccination values. To find if the correlation is positive or negative, based on R output, the estimated $\rho =$ -0.25. Thus, by this correlation analysis, we conclude that there is a negative correlation relationship between the increased rate of infection and the vaccination rate.

DISCUSSION

While doing hypothesis tests on time series data, the first thing is to test if the data is stationary series so that we can predict future conditions based on the history. However, for our data, vaccination data is not a stable series since the countries all over the world haven't finished this progress no matter where they are (coming to an end or just at the beginning stage). And due to the vaccination, the other data also changes by dates such as confirmed case number and death number. The non-normally distributed data makes us unable to do tests based on the normal distribution assumption such as ANOVA. Fortunately, other non-parametric methods, such as the Wilcoxon test can help to evaluate relationships between countries' vaccination progress. For further improvement for the project, as the vaccination process is moved forward, we can obtain more data about vaccination. And when we get a stationary series, we can forecast the future trend of vaccination.

For helping to better control and understand the pandemic, we also tried to fit a linear regression model with daily death cases as the response variable and same-day vaccination frequency as the explanatory variable, because same-day vaccination frequency could be considered as independent with the death cases. The p-value is relatively smaller than the significant level 0.05 with a negative estimated coefficient of vaccination frequency. But $r^2$for the model is fairly low and the scatter plot does not show a strong linear relationship between those two variables. Thus we dropped the regression part from this project. We believe it is due to the problem of the dataset and the variable contains missing values. With the data collection developed as the control of the pandemic goes further, a delicate regression model could be built to help on epidemiology researches.

CONCLUSION

The project conducts several tests to study present pandemic conditions and the influence of vaccination on the pandemic situation. First of all, based on the paired t-test results, we conclude that the global death rate for the confirmed cases on Oct. 10th, 2020 has significantly decreased compared to the one on Apr. 10th, 2021. In other words, the medical environment treated to confirmed patients has been improved significantly since October 2020 in general. Furthermore, by applying the χ^2test, we conclude that the death rates between genders are different.

Thirdly, for world vaccination progress, the US and the UK have the most advanced vaccination progress, and India has the least advanced among the 6 chosen countries. Moreover, the high confirmed case proportion and death proportion did promote vaccination progress in that country, but we cannot conclude that the country with less advanced vaccination progress is under a hostile epidemic environment. Last but not least, we prove there is an association between infection and vaccination. To be more specific, infection cases are negatively correlated to vaccination rates.

Due to the specialty of the period and limitation of personal medical data resources, the analysis of this project has constraints, and many statistical methods could not be applied due to non-normal distributed data. With the control for the pandemic moving forward, we believe the analysis could involve more research topics with more data.

REFERENCE

1. COVID-19 overview and infection prevention and Control priorities in non-us HEALTHCARE SETTINGS. (n.d.). Retrieved April 17, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/hcp/non-us-settings/overview/index.html
2. Dutta, G. (2021, April 13). Real-time Covid 19 data. Retrieved April 17, 2021, from https://www.kaggle.com/gauravduttakiit/covid-19
3. Nabil, E. (2020, June 19). Countries population by year 2020. Retrieved April 17, 2021, from https://www.kaggle.com/eng0mohamed0nabil/population-by-country-2020

4. Mark G. Thompson PhD, Jefferey L.Burgess MD, et al; Interim Estimates of Vaccine Effectiveness of BNT162b2 and mRNA-1273 COVID-19 Vaccines in Preventing SARS-CoV-2 Infection Among Health Care Personnel, First Responders, and Other Essential and Frontline Workers — Eight U.S. Locations, December 2020–March 2021. MMWR 2021, April 2, Vol.70 No.13. https://www.cdc.gov/mmwr/volumes/70/wr/pdfs/mm7013e3-H.pdf
5. Omna Sharma (2020, October 14). A Review of the Progress and Challenges of Developing a Vaccine for COVID-19.
6. Preda, G. (2021, April 14). Covid-19 world vaccination progress. Retrieved April 17, 2021, from https://www.kaggle.com/gpreda/covid-world-vaccination-progress
7. The covid-19 sex-disaggregated data tracker. (n.d.). Retrieved April 17, 2021, from https://globalhealth5050.org/the-sex-gender-and-covid-19-project/the-data-tracker/?explore=country&country=USA#search