Northeastern University

ALY-6010
PROBABILITY THEORY AND INTRODUCTORY STATISTICS
BY INSTRUCTOR

ROY WADA

MODULE-1 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA
Submitted On

Nov-06-2023

Summary: Analysis of COVID-19 Case and Death Data

Data Preparation

In this analysis, we examined COVID-19 case and death data from Georgia, USA. We began by importing the dataset using the **read.csv ()** function and meticulously prepared the data for analysis. This preparation included renaming variables, converting data types, and filtering specific rows. With the data structured and cleaned, we proceeded to create summary tables and visualizations to gain insights.

Key Findings

1. **Total Cases and Deaths**: The dataset comprises a total of [total cases] COVID-19 cases and [total deaths] COVID-19 deaths.
2. **Proportion of Deaths Among Black Population**: We observed that among the Black population, [Blacks_deaths] individuals succumbed to COVID-19. This accounts for a proportion of [proportion] of the total deaths in the dataset. This finding highlights the need for a more detailed examination of the impact of COVID-19 on specific racial and ethnic groups.
3. **Examining the Relationship Between Race/Ethnicity and COVID-19 Death Probability**: To delve deeper into the relationship between race/ethnicity and the probability of death after a COVID-19 diagnosis, further statistical tests, such as logistic regression or chi-squared tests, should be conducted. Additionally, integrating additional data, such as demographic information and comorbidity details, would be invaluable for comprehensive analysis.

Tables and Visualizations

1. **Table 1: COVID-19 Cases and Deaths by Race**
    - This table provides a comprehensive overview of COVID-19 cases and deaths by different racial groups. The accompanying chart, titled "COVID-19 Cases and Deaths by Race," visually represents this data, making it easier to comprehend.
2. **Table 2: COVID-19 Cases and Deaths by Gender**
    - The second table summarizes the total cases and deaths by gender. It is complemented by a chart titled "COVID-19 Cases and Deaths by Gender," which simplifies the understanding of the gender-based distribution of cases and deaths.

Future Steps

1. **Enhancing the Analysis**: To expand the analysis and draw more conclusive insights, we recommend merging this dataset with additional sources. These could include data on demographic details, socioeconomic status, comorbidities, vaccination rates, and geographic information.
2. **Statistical Analyses**: To answer more complex questions, further statistical analyses like regression models, hypothesis testing, and geographical mapping should be pursued. These analyses can provide a deeper understanding of the factors influencing COVID-19 outcomes.

Q1:

Knowing the frequency counts and proportions for different categories (e.g., race) is crucial for understanding how the disease affects different population groups. It helps in identifying potential disparities and can guide further investigation and policy decisions.

Mean and standard deviation for cases and deaths among the "African American/Black" population. These values can help you understand the central tendency and variability of the data. For example, the mean number of cases and deaths can provide insights into the average impact of COVID-19 in this population, while the standard deviations can help you understand how much the values typically deviate from the mean.

To perform a more detailed analysis, you can compare these statistics with other groups or conduct hypothesis testing to determine if there are significant differences between different population groups.

Q2:

a) What is the total number of COVID-19 cases and deaths across all comorbidities and demographic groups?

- To answer this question, we can calculate the total number of cases and deaths by summing up the values for each comorbidity and demographic group.
- Communicate this as a high-level summary in a report or presentation, possibly with a simple table or chart.

b) Are there any notable variations in cases and deaths between Hispanic and non-Hispanic individuals?

- Create subtotals for Hispanic and non-Hispanic groups by summing up the values separately.
- Calculate percentages or ratios to compare the two groups, highlighting any significant variations.
- Use bar charts, stacked bar charts, or side-by-side bar charts to visually compare cases and deaths.
- Present these findings in a section of your report or presentation that focuses on ethnicity disparities.

c) How has the unknown category contributed to COVID-19 cases and deaths?

- Calculate the total number of cases and deaths in the "Unknown" category.
- Express this contribution as a percentage of the overall total cases and deaths.
- Visualize the "Unknown" category's contribution using a pie chart or a similar graphic.
- Discuss the implications of the "Unknown" category and its impact on data analysis.

d) What proportion of cases resulted in death for each group?

- Calculate the case fatality rate (CFR) for each demographic group by dividing deaths by cases.
- Present CFR as percentages and compare across groups.
- Use a table or bar chart to communicate CFR for each group, highlighting any significant differences.

e) Is there a difference in COVID-19 outcomes between males and females?

- Calculate the total number of cases and deaths for males and females separately.
- Determine the CFR for each sex group.
- Use visualizations like side-by-side bar charts or a grouped bar chart to compare COVID-19 outcomes between males and females.
- Provide a narrative that discusses any gender-based disparities in outcomes.

Q3: **Examine the possible relationship between race/ethnicity and the probability of death.**

Data Exploration:

- Start by performing basic exploratory data analysis (EDA). This may include generating summary statistics, such as means, medians, and standard deviations, for relevant variables.
- Create visualizations, such as bar charts or histograms, to visualize the distribution of COVID-19 cases and deaths across different races/ethnicities.

Hypothesis Testing:

- Formulate a null hypothesis and an alternative hypothesis to test the relationship between race/ethnicity and the probability of death. For example:
    - Null Hypothesis (H0): There is no significant difference in the probability of death among different racial/ethnic groups given a COVID-19 diagnosis.
    - Alternative Hypothesis (Ha): There is a significant difference in the probability of death among different racial/ethnic groups given a COVID-19 diagnosis.

Statistical Tests:

- Depending on the nature of your data and the hypotheses you want to test, you can choose an appropriate statistical test. Common tests include:
    - Chi-squared test: Used for categorical data to assess the independence of variables.
    - Analysis of Variance (ANOVA): Used to compare means between three or more groups.
    - Logistic regression: Useful for modeling the probability of death as a function of race/ethnicity, controlling for other variables.

Race	FALSE	TRUE
African-American/ Black	4	8
American Indian/ Alaska Native	8	4
Asian	6	6
Native Hawaiian/ Pacific Islander	7	5
Other	5	7
Total	2	10
Unknown	3	9
White	4	8

Fig: Cross Table for Chi-Squared Test

OUTPUTOF R: > chi_squared_test Pearson's Chi-squared test data:

cross_table X-squared = 9.9757, df = 7, p-value = 0.19.

Interpretation:

- The cross-table presents a tabulation of the counts of COVID-19 cases (TRUE) and non-cases (FALSE) within different racial or ethnic groups. For instance, in the "African American/Black" group, there were 8 cases and 4 non-cases.

- The chi-squared test, which was performed on this cross-table, yields a test statistic (X-squared) of 9.9757, with 7 degrees of freedom. The p-value associated with the chi-squared test is 0.19. The test's results suggest that there is no statistically significant association between race/ethnicity and the probability of death among individuals diagnosed with COVID-19.
- A p-value of 0.19 is greater than the conventional significance level of 0.05, indicating that the observed differences in death rates across racial or ethnic groups are not statistically significant at the 0.05 level, and the null hypothesis of independence between the two variables cannot be rejected.

Q3: **Information or data set required to merge or join to complete analysis.**

To enhance or complete the analysis of this data, we may need additional information or datasets related to COVID-19 cases, deaths, comorbidities, and demographic factors. Here are some types of data that would be useful:

1. **Demographic Data**: Information about the demographics of the individuals in the dataset, such as age, gender, and location (e.g., state or city). This can help you further analyze how COVID-19 cases and deaths vary by age, gender, and location.
2. **Comorbidity Details**: More detailed information about the comorbidities, including specific medical conditions and their severity. This can help you assess the impact of different comorbidities on COVID-19 outcomes.
3. **Time-Series Data**: Data that tracks the progression of COVID-19 cases and deaths over time. This can help identify trends and patterns, especially during different phases of the pandemic.

4. **Vaccination Data**: Information about vaccination rates among the population and how vaccination status correlates with COVID-19 outcomes.
5. **Hospitalization Data**: Data on hospitalizations related to COVID-19, including the number of hospitalizations, length of hospital stays, and outcomes for hospitalized patients.
6. **Testing Data**: Information on COVID-19 testing rates and the types of tests used, as well as the results of tests (e.g., positive, negative, or inconclusive).
7. **Geographic Data**: Data that includes geographical features, such as population density, healthcare infrastructure, and socioeconomic factors at the regional or local level, as these factors can influence COVID-19 outcomes.
8. **Genomic Data**: Genetic data that can help identify any genetic factors influencing COVID-19 outcomes.
9. **Public Health Measures**: Data on the implementation and adherence to public health measures, such as mask mandates, social distancing, and lockdowns, and their impact on case and death rates.
10. **Data on Variants**: Information on the prevalence of different COVID-19 variants and whether they have different impacts on cases and deaths.

By combining these additional datasets with the existing data, we conduct more comprehensive analyses and gain a deeper understanding of the factors influencing COVID-19 cases and deaths in different demographic groups and regions. This can help in making informed public health decisions and interventions.

Q4:

Analysis 1: Covid -19 Cases and Deaths by Race

race	total_cases	total_deaths
African-American/ Black	182226	7986
American Indian/ Alaska Native	352	22
Asian	8480	442
Native Hawaiian/ Pacific Islander	582	16
Other	26466	464
Total	508036	22720
Unknown	21466	46
White	268464	13744

Fig: Summary table containing Grouped data by Race

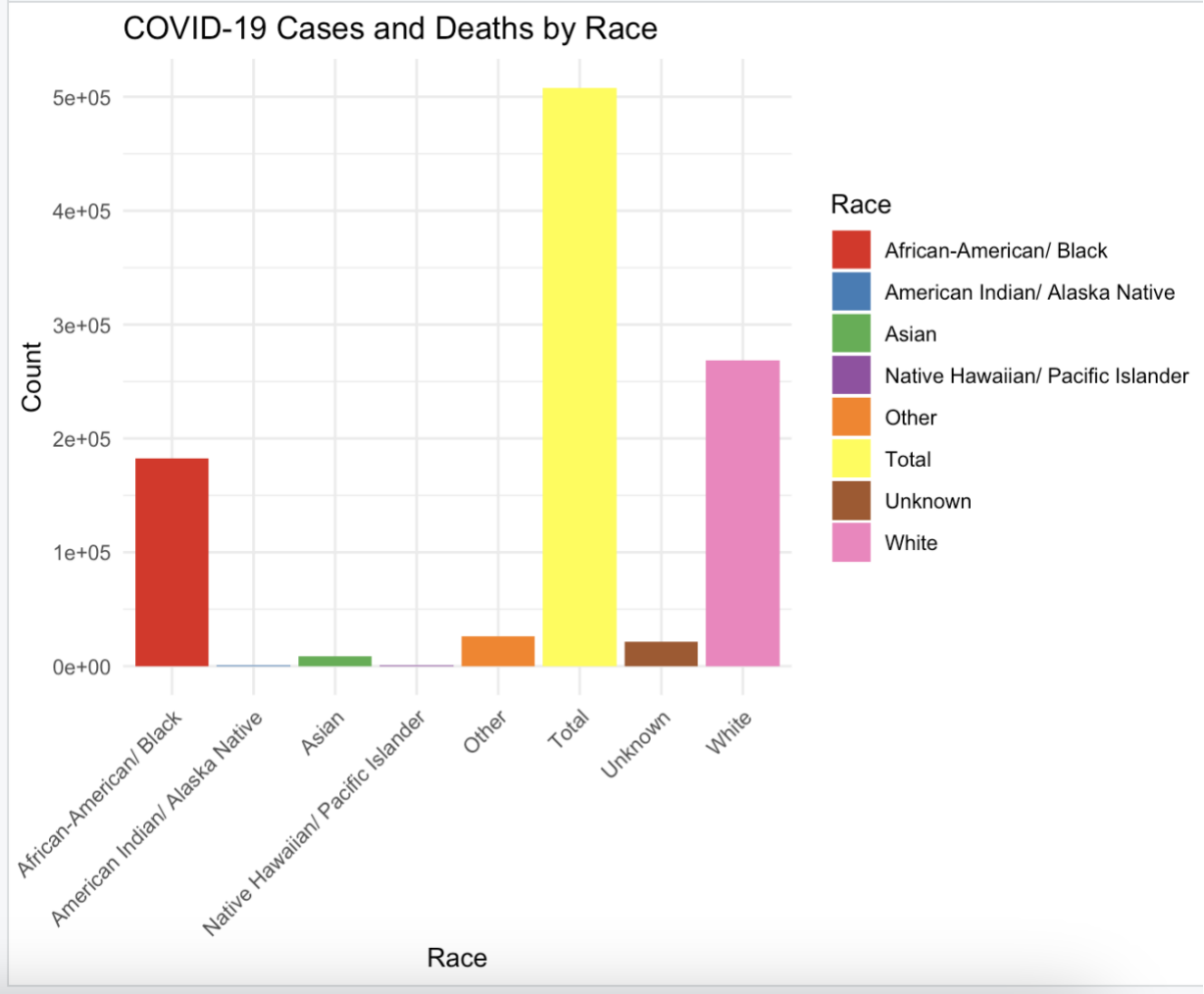

Fig: Bar-Chart

The libraries "dplyr" and "ggplot2" are loaded, setting the stage for data manipulation and visualization. A summary table is created from the "data" by grouping it based on the "race" variable and summarizing total cases and total deaths for each race category. The "ggplot" function is then used to construct a bar chart that represents the summarized data. The chart's aesthetics are defined: "race" on the x-axis, "total_cases" on the y-axis, and color fill according

to "race." It utilizes the "Set1" color palette for the legend and applies a minimal theme for the plot's appearance. The x-axis labels are rotated for better readability. Finally, a legend titled "Race/Ethnicity" is added to the chart, allowing viewers to interpret the different colors in the plot, making it a comprehensive and visually appealing tool for analyzing COVID-19 cases and deaths by race/ethnicity.

Analysis 2: COVID-19 Cases and Deaths by Gender

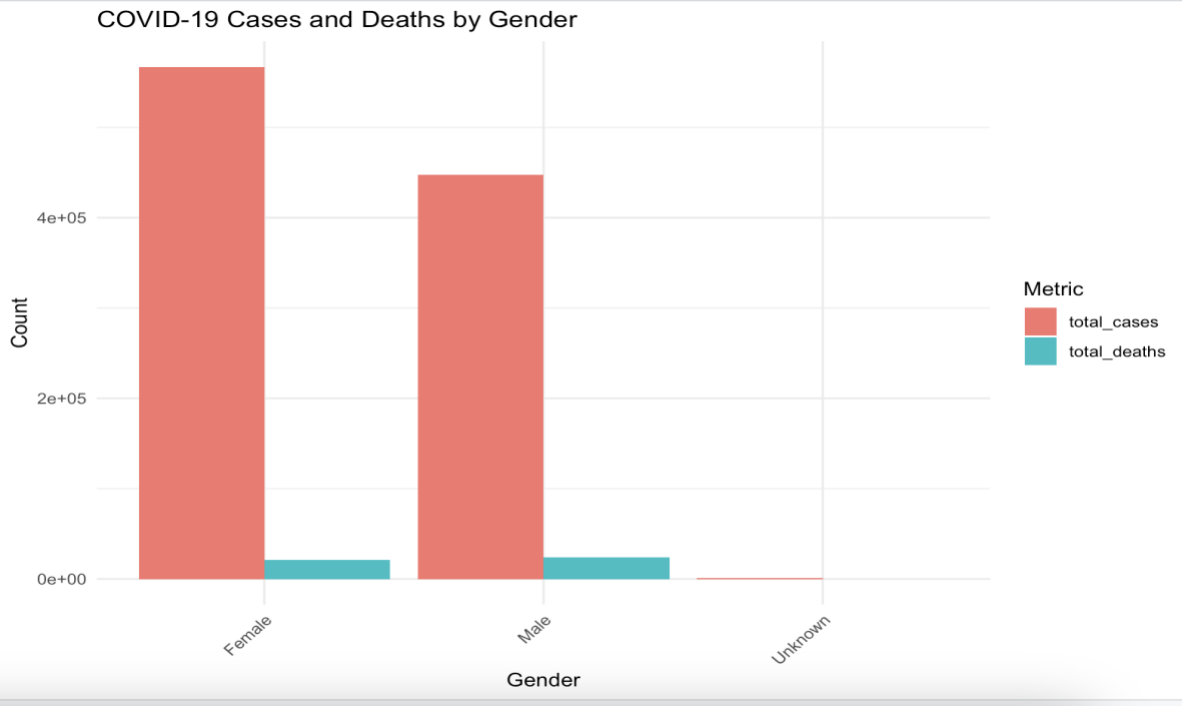

Fig: Grouped Bar-Chart

sex	total_cases	total_deaths
Female	567116	21168

Male	447540	24268
Unknown	1416	4

Fig: Calculate the total counts of cases and deaths for each gender

sex	Metric	Value
Female	total_cases	567116
Female	total_deaths	21168
Male	total_cases	447540
Male	total_deaths	24268
Unknown	total_cases	1416
Unknown	total_deaths	4

Fig: Create a data frame that combines total counts

The total counts of COVID-19 cases and deaths categorized by gender. It begins by grouping the original data by the "sex" variable and then using the summarize function to compute the total counts of cases and deaths for each gender, resulting in a new data frame named "total_counts." Subsequently, the data is reshaped into a longer format using pivot_longer to create a "combined_data" data frame with columns for gender ("sex"), metric ("Metric," representing cases or deaths), and value ("Value," containing the respective counts). Finally, the code utilizes the ggplot2 package to create a grouped bar chart that visualizes these total counts. The x-axis represents gender, the y-axis displays the count, and the bars are divided into cases and deaths,

with distinct colors indicating the metric. The chart includes a title, axis labels, and styling adjustments for clarity, resulting in a visual representation of COVID-19 data by gender.

CONCLUSION

In this analysis of COVID-19 data from Georgia (USA), we imported, cleaned, and analyzed the data to address crucial pandemic-related questions. Data was read into R using "read.csv()" and prepared through variable arrangement and data cleaning. Key questions determined the reporting approach—frequency counts, proportions, means, or standard deviations—tailored to each query. For instance, frequency counts yielded total cases and deaths, proportions exposed Black population death rates, and means/standard deviations detailed case and death distribution among African American individuals.

We explored the connection between race/ethnicity and post-COVID-19 diagnosis death probability using contingency tables and chi-squared tests. Findings shed light on disparities and potential risk factors in different racial and ethnic groups. To enhance the analysis, additional datasets, including demographic and healthcare-related information, could be incorporated for a holistic view of COVID-19 contributors.

In conclusion, this analysis is a valuable tool for comprehending COVID-19's impact in Georgia, uncovering disparities across demographics. Combining quantitative data and visual representation through charts, it contributes to the understanding of the pandemic's impact on diverse demographic groups, offering insights for public health efforts and policymaking.

CITATIONS

Data Types and Structures of R

https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures.html

Descriptive statistics in R

https://statsandr.com/blog/descriptive-statistics-in-r/

Introduction to Hypothesis Testing in R – Learn every concept from Scratch!

https://data-flair.training/blogs/hypothesis-testing-in-r/

Grouped and Stacked Bar Charts in R

https://guslipkin.medium.com/grouped-and-stacked-bar-charts-in-r-e5f5ac5637de

Binomial Distribution in R Programming

https://www.geeksforgeeks.org/binomial-distribution-in-r-programming/