Northeastern University

ALY-6010

PROBABILITY THEORY AND INTRODUCTORY STATISTICS

BY INSTRUCTOR

ROY WADA

MODULE-2 PROJECT REPORT

SUBMITTED BY

ITI ROHILLA

On

Nov-14-2023

SUMMARY

The juxtaposition of datasets from the previous and current weeks underscores a paradigmatic shift in data granularity and organization. In the ancient week, data aggregated at the group level provided a macroscopic view, while the current week's individual-level data delves into intricate details.

This evolution enables a comprehensive examination of COVID-19 deaths in Georgia, revealing not just broad demographic trends but also unveiling individual nuances. The 'proportion_BlackHispanic_Deaths' metric, a lighthouse in this analysis, accentuates the pronounced disparities within Black and Hispanic populations, shedding light on critical disparities that warrant targeted public health interventions.

The meticulous construction of a detailed summary table further enriches the analysis, offering a multifaceted perspective on age-related characteristics and demographic patterns that shape the landscape of COVID-19 fatalities in the region.

Moreover, the exploration extends beyond demographic considerations, delving into individual health indicators. The calculation and interpretation of the 'mean_indicator' provide a nuanced understanding of the prevalence of chronic conditions, indicating that approximately 52.33% of individuals in the dataset exhibit such conditions.

This fine-grained insight into health status adds a layer of complexity, emphasizing the need for tailored healthcare strategies that account for the prevalence of chronic conditions among the affected population. The inclusion of this metric not only enriches the analysis but also underscores the intricate interplay between individual health factors and the severity of COVID-19 outcomes.

Q: Data set organization from last week v/s this week

The data sets from the last week and this week differ in their organization and granularity. In the previous week's data, each row represented aggregated information for specific groups, such as comorbidity, sex, ethnicity, and race, providing counts and proportions of cases and deaths. The dataset was structured to showcase overall patterns and proportions within various demographic categories.

In contrast, the current week's data is organized at the individual level, with each row representing a person's information, including age, ethnicity, race, sex, county, and chronic condition. This allows for a more detailed examination of each person's characteristics and circumstances related to COVID-19 deaths in Georgia.

The shift from aggregated to individual-level data provides a more granular perspective, enabling a closer analysis of specific cases and contributing factors at the individual level.

Q: Report Count, Proportion, Means, SD, and CI in the Same Table:

ethnicity	count	mean	sd	ci_lower	ci_upper
Hispanic/ Latino	1101	61.9130435	15.6023796	60.9904224	62.8356646
Non-Hispanic/ Latino	20491	70.4690879	12.8465655	70.2931825	70.6449933
Unknown	117	59.7982456	17.9533219	56.5108319	63.0856593

Table: Proportion of deaths among Black or Hispanic population

The 'proportion_BlackHispanic_Deaths' variable is computed as the sum of Black and Hispanic deaths divided by the total number of deaths. This proportion is a key metric for understanding the distribution of deaths among these specific ethnic groups.

The subsequent analysis involves creating a summary table, broken down by ethnicity, that includes the count of deaths, mean age, standard deviation, and confidence interval for age.

This detailed summary provides insights into the age-related characteristics of Black and Hispanic populations affected by COVID-19 deaths, allowing for a nuanced understanding of the demographic patterns within these groups.

Q: Understanding the mean of an indicator variable?

Variable	Value
Indicator Chronic Condition (Yes)	11360
Mean Indicator	0.523285273388917

Table: Summary Indicator

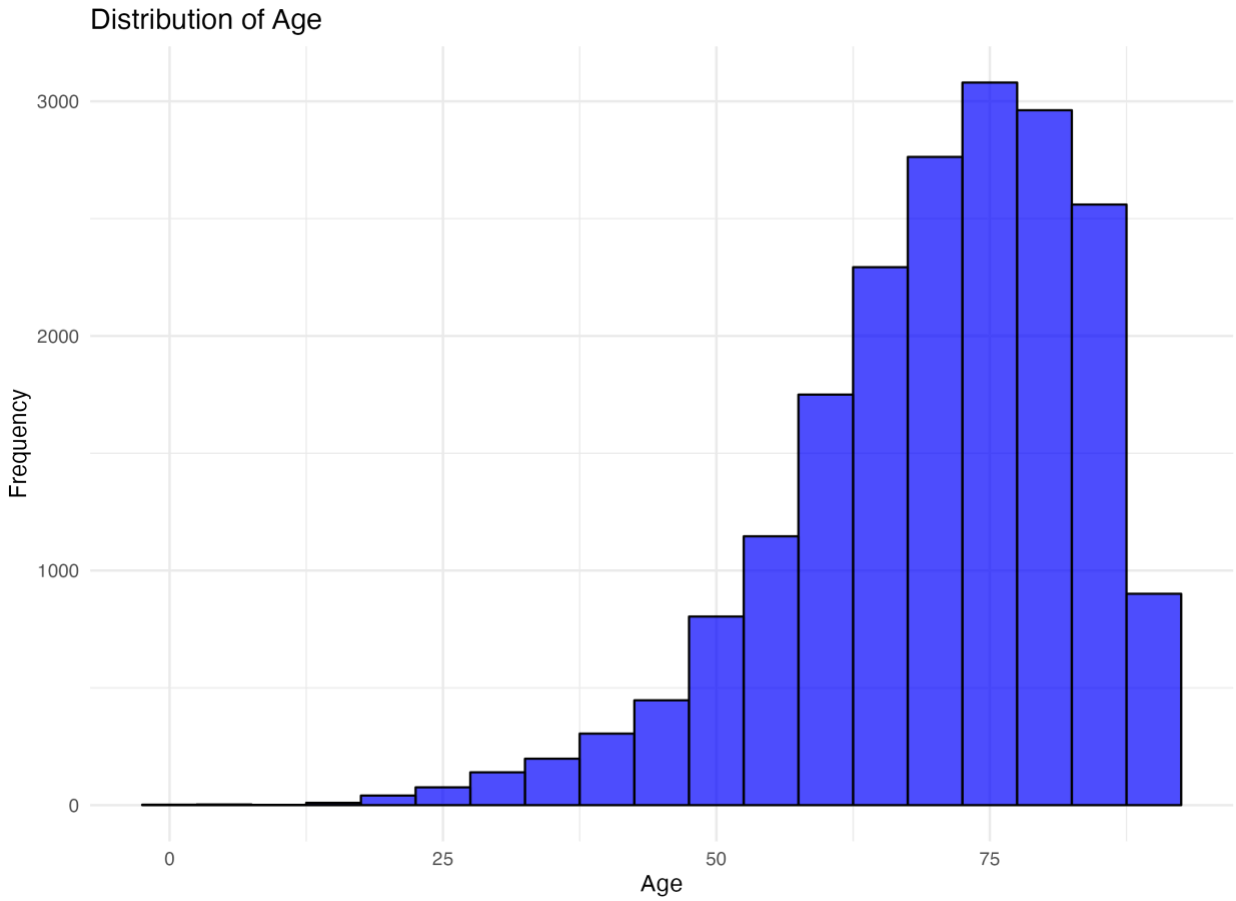
When we calculate the mean of an indicator variable, we are essentially calculating the proportion of cases in which the indicator is equal to 1 (or "Yes" in this case). An indicator variable is a binary variable that takes on the values of 0 or 1, where 1 typically represents the presence of a certain condition or characteristic, and 0 represents the absence.

In this specific case:

- If the indicator variable is 1 for individuals with a chronic condition, and 0 for those unknown or without, the mean of the indicator will give the proportion of individuals in the dataset who have a chronic condition.

So, the **mean_indicator** is 0.5232853, which suggests that approximately 52.33% of the cases in the dataset have a chronic condition.

Q: **Visualization 1:**



The histogram visually represents the distribution of ages in the dataset, offering insights into key characteristics. The peak of the histogram provides a central tendency, indicating the most common age group. The width of the distribution reflects the range of ages, with a wider spread suggesting greater variability.

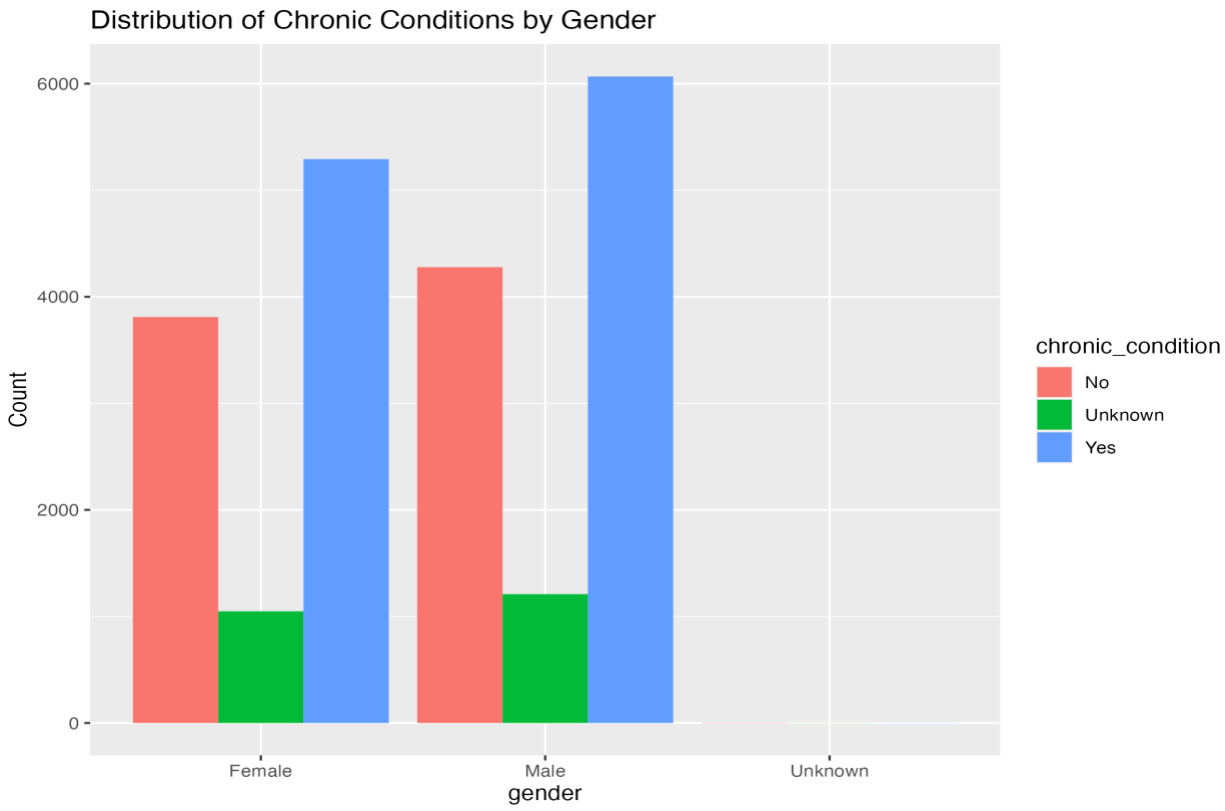
Skewness helps identify if the age distribution is symmetric or skewed, providing clues about the prevalence of younger or older individuals. Peaks or clusters highlight common age groups, while isolated bars may indicate outliers.

Sparse areas or gaps in the histogram can signify missing data or underrepresented age ranges. This visualization serves as a quick overview, offering valuable insights into the age demographics and patterns within the dataset.

Q: Visualization 2:

sex	chronic_condition	Count
Female	No	3810
Female	Unknown	1048
Female	Yes	5292
Male	No	4278
Male	Unknown	1209
Male	Yes	6067
Unknown	No	3
Unknown	Unknown	1
Unknown	Yes	1

Table: Summary_Table_Gender



The chart provides a clear comparison of the count of individuals with and without chronic conditions across different genders. The title "Distribution of Chronic Conditions by Gender" appropriately describes the chart's content. By examining the chart, one can gain insights into how chronic conditions are distributed among different genders in the dataset.

Healthcare professionals can utilize the insights to understand how chronic conditions are distributed among different genders within a population, enabling them to tailor interventions and healthcare strategies accordingly. Public health researchers may find this analysis valuable for identifying gender-specific patterns in chronic conditions, leading to a better understanding of health disparities.

Q: When to use a Grouped Bar Chart:

Grouped bar charts are commonly used when you want to compare the distribution or frequency of a categorical variable across two or more levels of another categorical variable. Each bar in the chart represents a specific category, and within each category, bars are grouped by the levels of the second categorical variable.

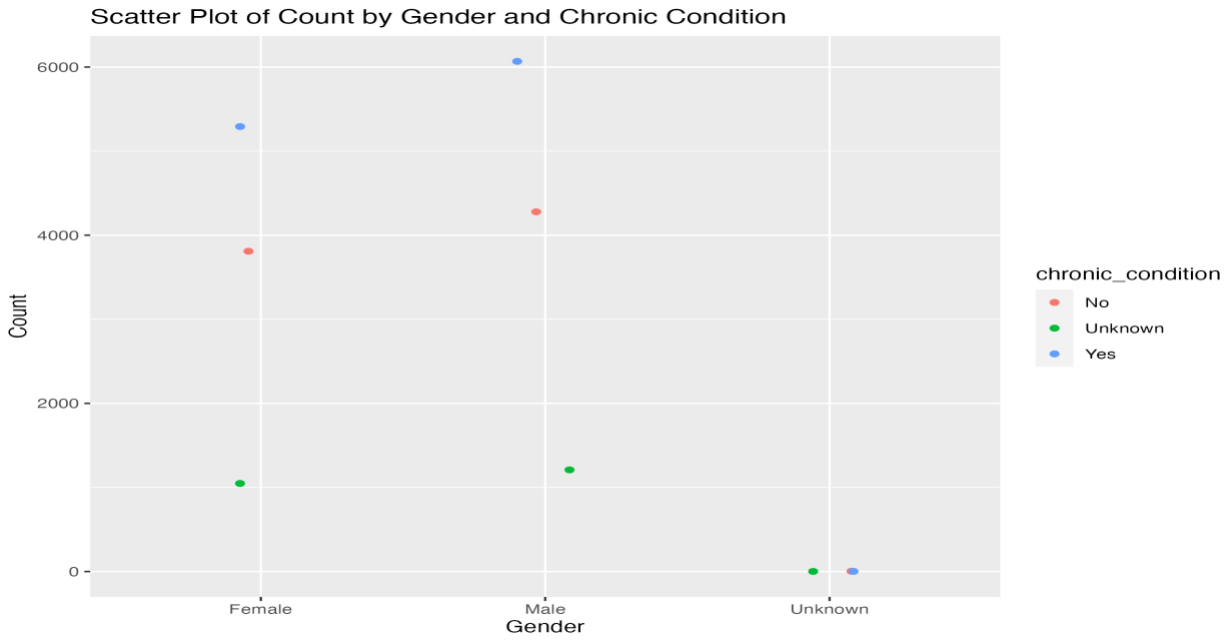
This type of chart is effective for visualizing the relationship between two categorical variables and understanding how the distribution of one variable varies across different groups of the other variable.

Grouped bar charts are useful for presenting and comparing counts or frequencies of different categories within each group, providing a clear visual representation of the data's structure and patterns.

Q: Grouped bar charts v/s Scatter Plot:

Grouped bar charts excel in scenarios where you aim to compare the distribution or frequency of a categorical variable across different levels of another categorical variable. This visual representation is effective for conveying insights about the count or proportion of occurrences within each category and subgroup. Each bar in the chart represents a specific category, and these bars are grouped according to the levels of the second categorical variable. This makes it easy to discern variations and patterns in the data.

In contrast, a scatter chart is better suited for showcasing the relationship between two continuous variables. It excels in illustrating correlations and trends within numerical data. While scatter plots are valuable for exploring connections between variables on a continuous scale, grouped bar charts shine when emphasizing categorical comparisons and the frequency distribution within distinct groups. Therefore, the choice between the two depends on the nature of the data and the analytical focus – whether it's about relationships and correlations or categorical distributions and comparisons.



CONCLUSION

In conclusion, the transition from aggregate to individual-level data serves as a cornerstone in unraveling the intricate tapestry of COVID-19 deaths in Georgia. The rich summary table, augmented by the exploration of the 'mean_indicator,' and bolstered by insightful visualizations, paints a vivid picture of the multifaceted nature of this public health challenge.

However, the journey does not conclude here. The exploration of gender-specific patterns, as exemplified in the summary table detailing chronic condition distribution, adds yet another layer of complexity to our understanding. This nuanced approach lays the foundation for informed decision-making, empowering healthcare professionals and policymakers to design interventions that address the unique characteristics within various demographic groups.

Moreover, as we delve into the age distribution through the compelling histogram visualization, we gain further insights into the prevalence and impact of COVID-19 across different age groups. Peaks and clusters within the histogram highlight common age groups, aiding in the identification of vulnerable populations.

Sparse areas or gaps in the histogram may signify underrepresented age ranges, prompting a closer examination of potential disparities in healthcare access or reporting. This visual exploration enhances our grasp of the age-related dynamics, facilitating targeted interventions and resource allocation tailored to the specific needs of distinct age cohorts.

Together, these multifaceted analyses and visualizations form a comprehensive guide for healthcare professionals, researchers, and policymakers navigating the complex landscape of COVID-19, fostering a more resilient and adaptive public health response.

CITATIONS

Getting both column counts and proportions in the same table in R

- https://stackoverflow.com/questions/9438193/getting-both-column-counts-and-proportions-in-the-same-table-in-r

The Basics of Indicator Variables

- https://online.stat.psu.edu/stat462/node/161/

How to Create a Grouped Barplot in R?

- https://www.geeksforgeeks.org/how-to-create-a-grouped-barplot-in-r/

Understanding scatter plots in R

- https://www.dataanalytics.org.uk/graphs-data-visualization-using-r/#scatter_plots

How to Read and Use Histograms in R

- https://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/