Nirav Bariya

03/03/2024

Influenza Season: Interim Report

Project Overview

- **Motivation:** The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.
- **Objective:** Determine when to send staff, and how many, to each state.
- **Scope:** The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.

Research Hypothesis

1. If your age is smaller than 5 years, then you have higher risk of dying from flu.
2. If your age is more than 65 years, then you have higher risk of dying from flu.

Data Overview

Census Population: The data contains county-wise total population and population of different age groups for every year from 2009-2017. The age groups differ by a margin of 5 years.

Column Name	**Definition**	**Example**
County	Name of County along with state	Acadia Parish, Louisiana
Year	Year of the Census	2009
Total population	Total Population Count of the County	59616
Male Total Population	Total Male Population of the County	28963
Female Total population	Total Male Population of the County	30653
Under 5 years	Total Population under age 5	4590
5 to 9 years	Total population between age 5 to 9	4471.2
...	...	...
85 years and over	Total Population over 85 years	894.24

CDC Influenza Deaths: This data set contains information about state-wise death mortality between the year 2009-2017. This data set also tells us about the age group of number of deaths.

Column Name	Definition	Example
State	US State	Wyoming
Year	Year of record	2013
State Code	Code for State between 1 and 56	3
Month	Month of the record	April

Month Code	Code of the month between 1 and 12	04
Ten-Year Age Groups	Age group in which the death was recorded	1-4 years, 5-14 years, 15-24 years etc.
Ten-Year Age Groups Code	Code for age groups	1-4, 15-24, 35-44 etc.
Deaths	Number of deaths recorded in this age group	An integer or text string “Suppressed”

Data Limitations

1. **Census Population:** Due to the various methods employed to collect the data, it is likely that the dataset is representative of the population. However, there are some years for which the population data is missing for some county. This will limit the aggregation process and comparing yearly population state-wise. Further, the data was collected every year, so there was a considerable gap between when the survey was conducted and when the results were published creating a scope of error. Some of the survey methods include feeding manual data in the system which may involve error. I believe because this data is published by a government agency it is likely to have minimum errors and it is the best data source available for this analysis.

2. **CDC Influenza Deaths:** Data for Puerto Rico is not available for this data set. This means we will not be able to accurately describe the situation related to deaths in Puerto Rico. Further, some data is suppressed due to privacy concerns. Therefore, not all the deaths due to influenza are accounted for in this data set. Also, the results of the analysis will only be relevant for the year 2018 as the data available is for 2009-2017. Therefore, the results may not be applicable for present year However, this data set is the most accurate data set we can get for this analysis.

Descriptive Analysis

There were several variables in both the data sets, out of which following variables were key to our hypothesis.

Variable	Population (65-74 age group)	Mortality rate (65-74 age group)	Population (75-84 age group)	Mortality rate (75-84 age group)	Population (85+ age group)	Mortality rate (85+ age group)
Mean	440464	0.03%	254032	0.12%	107100	0.61%
Standard Deviation	479173	0.05%	280736	0.014%	121222	0.72%

We were informed that vulnerable population, which comprises of age groups 65+ years and children below age 5, are at the highest risk of dying from influenza.

We figured that:

- population for age group 65-74 has a **strong positive correlation**, **0.94**, with mortality rate for the same age group.
- population for age group 75-84 has a **strong positive correlation**, **0.94,** with the mortality rate for the same age group.
- population for age group 85+ has a **strong positive correlation**, **0.94,** with the mortality rate for the same age group.

This means that as the population for each of these age groups increases the mortality rate also increases.

Variable	Population Under 5 years	Mortality under 5 years
Mean	382641	0.00 %
Standard Deviation	452782	0.00%

For population under 5 years of age, the mortality was 0% across all states. Therefore, we were **not able to calculate correlation coefficient** in this case as doing so would mean dividing by 0. Further, it is safe to assume that there is no strong positive correlation as any increase or decrease in population doesn't change mortality rate.

Results and Insights

There were two hypothesis which were tested. They are as follows.

First Hypothesis

Null Hypothesis: Average mortality rate for age less than 5 years is equal to the average mortality rate of entire population.

Alternate Hypothesis: Average mortality rate for age less than 5 years is not equal to average mortality rate of total population.

Result: At an alpha level of 0.05, or 95% confidence level, **there was a significant difference** between the two groups of population.

Second Hypothesis

Null Hypothesis: Average mortality rate of the population is equal to average mortality for 65+ years age group

Alternate Hypothesis: Average Mortality rate of the population is not equal to average mortality for 65+ years age group.

Result: At an alpha level of 0.05, or 95% confidence level, **there was a significant difference** between the two groups. Therefore, we can reject null hypothesis with a 95% confidence level.

Remaining Analysis Work and Next Steps

After the results we have discovered during hypothesis testing, we will follow the following steps to complete our project:

- We will prioritize states with large vulnerable population and categorize each state as high-need, medium-need, or low-need based on their vulnerable population
- We will identify the months of the year during which influenza cases are at highest, this will help us in creating a staffing plan
- We will find if the influenza-months for each state are same or different
- We will create various types of visualizations, including spatial visualization of the spread of the virus, to communicate the results more effectively with our stakeholders
- Finally, we will prepare a video presentation for our stakeholders communicating the results of analysis and staffing plan.

Appendix

1. **Project Overview:** Link to Business Requirement Document This link provides a complete project overview of the project.
2. **Hypothesis:** The business requirement document highlighted vulnerable population as 65+ years age group and less than 5 years age group. Based on this information, we hypothesized that mortality rate for vulnerable group must be higher than non-vulnerable population. We considered an alpha of 0.05, or 95% confidence level for this testing. Further, both of our hypothesis involved looking in one direction, greater than or equal to. Therefore, we considered one tailed student's t-test for checking the hypothesis.

3. **Data Overview:** The following dataset were used for the analysis.
   (i) Influenza deaths by geography
   Source: CDC
   Dataset
   (ii) Population data by geography, time, age, and gender
   Source: US Census Bureau
   Dataset