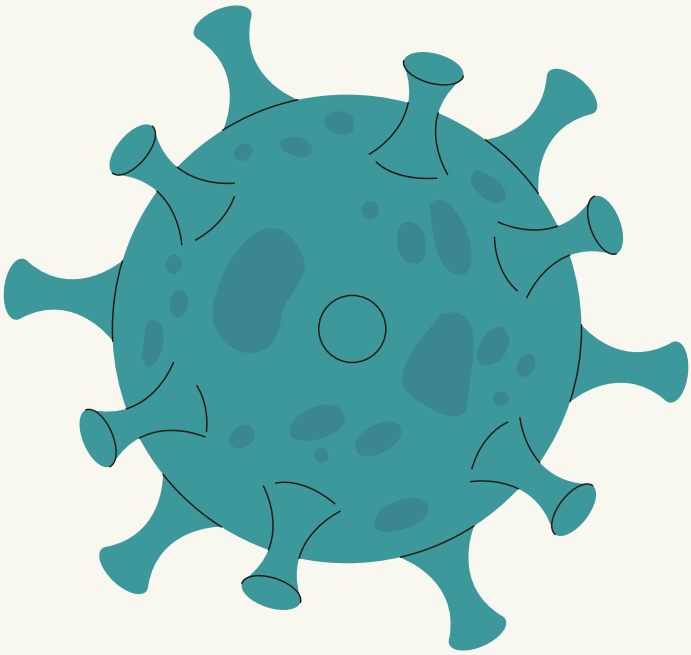

Preparing For The Influenza Season

Interim Report

Schay Kierstin Esparza

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 and Scope

The objective is to determine when to send staff, and how many, to each state.

The scope of the agency is that it 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

Vulnerable populations are patients likely to develop flu complications requiring additional care, as identified by the Centers for Disease Control and Prevention (CDC). Children under the age of 5 and older adults above 65 years old are considered to be vulnerable populations.

"If a vulnerable person contracts the flu, then their mortality rate will be higher than that of a non-vulnerable person."

Data Overview

Population data by Geography

Source: This data is from an external source. The data is provided by the **US Census Bureau** which is a government data source. As government data, the trustworthiness of the data can be verified.

Collection: The data is administrative data collected as part of government data that provides current facts and figures about America's people, places and economy. Data is collected manually. There is a time lag with this data due to the collection done annually.

Contents: The data contains location as county and state, total population and population by gender. Population is further broken down by age groups in increments of 5 years.

Limitation: Due to the data being collected annually, there is a limitation of not being able to see the month-to-month fluctuations in the population. With the data being manually collected, it is prone to error. No biasness in this data set.

Relevancy: Data shows population by state which is helpful in determining how many vulnerable population per state has.

Influenza Deaths by Geography

Source: This data is from an external source. The data is provided by the **Centers for Disease Control and Prevention (CDC)** which is a government data source. As government data, the trustworthiness of the data can be verified.
Collection: The data is administrative data collected as part of the National Vital Statistics Cooperative Program. These are death records in which doctor codes the primary cause of death as "influenza" or "pneumonia"

Contents: The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017 and broken down in to state and age.

Limitation: Death records come from death certificates which only list one cause of death. This could create some discrepancies within vulnerable populations whose decline in health is due to influenza but their cause of death is listed as another reason such as AIDS.

Relevancy: Data shows death from different state and age groups, this is relevant for my hypothesis which looks into vulnerable population deaths.

Descriptive Analysis

Data Integration was performed combining Census Population data and Influenza Deaths data. When integrated, it was noted that one location, Puerto Rico only existed in the Census Population data and not on the other one. Hence, influenza value of Puerto Rico changed to "0".

	Influenza Deaths		*Census Data*	
	Vulnerable Population	**Non-Vulnerable Population**	**Vulnerable Population**	**Non-Vulnerable Population**
Mean	995	407	1,181,712	4,740,329
Standard Deviation	972	133	1,315,689	5,444,845

Vulnerable Population: People under the age of 5 and over the age of 65 years.
Non-Vulnerable Population: People between the ages of 5 to 64 years old.
Calculations were done using the sample version of standard deviation equation.

Data Correlation done on **vulnerable population** in terms of influenza deaths and census data using **Pearson's r** resulted to a **correlation coefficient of 0.94.** This shows a strong relationship and supports the research hypothesis of vulnerable population having a higher mortality rate due to influenza. When performing data correlation as a whole on non-vulnerable population, it resulted to a strong relationship at 0.84. Hence, I wanted to further investigate why by looking at the correlation individually by age group.

The data then reveals that the correlation between influenza death and age increases as the person ages. People aged 5-14 years old had a correlation coefficient of -0.02 which is weak while people aged 55-64 years old resulted to 0.92.

There is a **positive correlation** between age and mortality rate. As the person ages, their mortality rate due to influenza increases as well.

Hypothesis Testing

Null Hypothesis: Mortality rate of vulnerable population is less than or equals to non-vulnerable population.

Alternative Hypothesis: Mortality rate of vulnerable population is greater than or equals to non-vulnerable population.

t-Test: Two-Sample Assuming Unequal Variances

	Influenza Deaths	
	Vulnerable Population	*Non-Vulnerable Population*
Mean	995.0660981	406.6183369
Variance	945519.9678	17519.71941
Observations	469	469
Hypothesized Mean Difference	0	
df	485	
t Stat	12.98590567	
P(T<=t) one-tail	1.32815E-33	
t Critical one-tail	1.648001465	
P(T<=t) two-tail	2.65629E-33	
t Critical two-tail	1.964867287	

One tailed test because the sample mean can only be higher/lower than the population mean.

Alpha: 0.05

P-Value: 1.32815E-33

P-value is significantly less than the alpha, hence we can infer that the null hypothesis can be rejected.

Therefore, rejecting the null hypothesis leads us to the alternative hypothesis that states, ***"Mortality rate of vulnerable population is greater than or equals to non-vulnerable population."***

Remaining Analysis

- Deep dive into current data to understand patterns of influenza deaths by age, geography and time/ season.
- Look into the census data to map out which of the states has the highest concentration of vulnerable population. These are the states that would need more staff members deployed.
- Create data visualizations with Tableau.
- Create final presentation

Appendix

Influenza: a contagious viral infection, often causing fever and aches.

Data Profile for Influenza Deaths by Geography

Variable	Time Variant/ Invariant	Structured/Unstructured	Qualitiative/ Quantitative	Further Characteristics
State	Time Invariant	Structured	Qualitative	Nominal
State Code	Time Invariant	Structured	Qualitative	Nominal
Year	Time Variant	Structured	Quantitative	Discrete
Month	Time Variant	Structured	Qualitiative	Nominal
Month Code	Time Invariant	Structured	Qualitative	Nominal
Ten-Year Age Groups	Time Invariant	Structured	Qualitative	Ordinal
Ten-Year Age Groups Code	Time Invariant	Structured	Qualitative	Ordinal
Deaths	Time Variant	Structured	Quantitative	Discrete

Data Profile for Population Data by Geography (census)

Variable	Time Variant/ Invariant	Structured/Unstructured	Qualitiative/ Quantitative	Further Characteristics
County	Time-Invariant	structured	qualitative	nominal
Year	Time-variant	structured	quantitative	discrete
Population	Time -variant	structured	quantitatve	discrete
Male and female population	Time-variant	structured	quanititative	discrete
Age Groups (under 5 years old - 85 years and over)	Time-Invariant	structured	qualitative	ordinal

Both data sets were integrated. The key variables for both were the State and Year. Age groups were standardized to be grouped in Ten year-age groups.

	Influenza Death	**Census Data**
Data Granularity	State, Month Code and Ten-year age group	State, County and Year

Suppressed Data: Influenza deaths original data contained "suppressed" values which denoted less than 10 deaths. Place holder value used to protect identity of people. This data, was then transformed using random value function on excel since this missing data accounted for 81.7% of the original data set.

Not Stated Data: Influenza deaths original data contained "not stated" values which accounted for 8.3% of the original data. This will still be included in the final report as is. We are unable to use secondary data to prove or to guess what the "not stated" ages are.

Outlier Percentage

For Influenza Data: Vulnerable population is 3% and Non-vulnerable population is 7%.

For Census Data: Vulnerable population is 7% and Non-vulnerable population is 4%