Medical Staffing for 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

If a there is a high proportion of vulnerable population, then those locations are at higher risk of developing a high influenza-related mortality rate and will require additional staffing relief.

Data Overview

US Census Population: This data set contains US population statistics from 2009-2017 by geographical location. It is further broken down by county & state, year, total population, male total population, female total population, and age group (per 5 years).

CDC Influenza Deaths: The data contains monthly death counts for influenza-related deaths in the United States from 2009-2017. This data set is further broken down by State, month, year, and age groups (per 10 years).

Data Limitations

US Census Population: Since this dataset manually collected, it is possible it is prone to human-bias, typo errors, collection errors, and accuracy errors. Since this dataset looks through the years 2009-2017 it is also important to consider the collection of data has not been consistent in each county through the relevant years. The age group numbers are estimates and may not add up to the total population.

CDC Influenza Deaths: Death certificates only list one cause of death; If the patient had a more serious or chronic ailment influenza would not be listed as cause of death even if it was a contributing factor. The dataset suppressed death counts less than 10 so numbers between 0-9 have been randomly inputted in the suppressed values.

Descriptive Analysis

US_Censuspop	Mean	Standard Deviation	Outlier Percentage
Vulnerable Population	1,193,271	1,327,221	6.97%
Total Population	5,973,848	6,806,790	4.79%
CDC_fludeaths	**Mean**	**Standard Deviation**	**Outlier Percentage**
Vulnerable Population	997	976	3.92%
Total Population	1,434	1,089	4.14%

*Both datasets are sample sizes with a normal distribution, and each has 459 records.
*Vulnerable population is age groups Under 5 and Over 65 years combined.

The hypothesis targets vulnerable populations (Under 5 and Over 65 years) and influenza mortality rates. A statistical analysis focusing on populations totals and flu mortality looked at key variables to find the mean, standard deviation, and outlier percentage. The results are listed in the table above.

Every year, each State has an average of 1,434 flu deaths and an average vulnerable population of 1,193,271. The standard deviation expresses how much the values can vary from the mean, and values typically fall within 2 standard deviations of the mean. If a value is outside of those values, it is determined an outlier. For example, 2 standard deviations away of flu deaths total population would be between 3,614 and -1,453 (since this dataset pertains to human beings, it is impossible to have a negative, so this value is adjusted to 0). There was a total of 19 outliers (values outside of 2 standard deviations away) which equivalates to 4.14%.

The hypothesis also indicates that the higher the vulnerable population, there will be a higher rate of mortality. A correlation test was done between total number of deaths and vulnerable population to check this relationship.

Variables	US_Censuspop Vulnerable Population & CDC_fludeaths Total Deaths
Proposed Relationship	The higher the Vulnerable Population the higher the Mortality
Correlation Coefficient	0.951
Correlation Strength	Strong
Interpretation	The relationship between these two variables seems to strongly support our hypothesis and demonstrates that those considered 'Vulnerable Population' are at higher risk of dying from influenza.

Results & Insights

After confirming the correlation between vulnerable populations and flu deaths, the strength of the relationship was tested. A two-sample t-test (assuming unequal variances) of the number of general population deaths (ages 5 – 64) and vulnerable population deaths (under 5 and over 65 years) was conducted to reject the null hypothesis. See results below.

Null Hypothesis: The mortality rate of Vulnerable Population (under 5 and over 65 years) is less than or equal to the mortality rate of age groups 5 - 64 years.
Alternative Hypothesis: The mortality rate of the vulnerable population (under 5 and over 65 years) is higher than the death rate of age groups 5 - 64 years.

	General Population Deaths	***Vulnerable Population Deaths***
Mean	437.2962963	997.2200436
Variance	15535.64127	953509.3379
Observations	459	459
P(T<=t) one-tail	3.36566E-30	
Significance Level	alpha = 0.05	

As shown in the results above, the mean of vulnerable population deaths is higher than those considered general population. Since the p-value is significantly higher than the significance level, we can confidently reject the null hypothesis.

Remaining Analysis & Next Steps

- Conduct further statistical testing on each state
- Create data visualizations
- Create time forecast for a variable
- Map a variable
- Create word cloud using qualitative data
- Create a narrative to communicate findings and insight
- Record/Present video presentation for stakeholders

Appendix

Hypothesis Development

The following questions were asked when developing the current working hypothesis.

Clarifying & Funneling Questions:

- **Which states have the highest influenza infection rates?**
  - Which states have the lowest vaccination rates? Highest?
  - Which states have highest influenza related death rates? Lowest?
- **Which states have the highest vulnerable populations?**
  - What are vaccination rates within the vulnerable population?
  - What percentage of vulnerable population end up in hospital?
    - Which states have higher hospitalizations of vulnerable population?
- **How many hospitals are in each state?**
  - What are current hospital staffing rates?
    - Which one's are currently understaffed?
    - What are ideal staffing levels at each hospital?
- **What are the most common complications with influenza people seek medical care for?**
  - Which complications require the most resources/medical care?
    - Which type of complications have the highest death rates?
    - Are these complications more prevalent in a certain location or population?
  - How can we better prepare staff to deal with these complications?
  - Are there any preventative treatments available?

Privacy & Ethical Questions:

- Do we need to follow any special security measures while handling this data?
- Is our data compliant with HIPAA laws?
- Was our data ethically sourced?
- What's the procedure in case of a data breach?

Other Proposed Hypotheses

- If a hospital is under-staffed, then it is less likely they will be able to provide proper care for the patients leading to higher influenza mortality rates.
- If an area has a low percentage of population vaccinated, then hospitals will need more staffing assistance in those areas.
- If an individual is under the age of 5, over the age of 65, or suffers from a chronic medical condition, then they are at a higher risk of developing severe complications from the flu, potentially resulting in death.

Data Profiles

CDC Influenza Deaths

Data Counts:

State	1296
State Code	1296
Year	7344
Month	612
Month Code	612
Ten Year Age Groups	5508
Ten Year Age Groups Code	5508
Deaths	66096 (total)

Data Types:

State	time-invariant	structured	qualitative	nominal
State-code	time-invariant	structured	qualitative	ordinal
Year	time-invariant	structured	qualitative	ordinal
Month	time-invariant	structured	qualitative	ordinal
Month code	time-invariant	structured	qualitative	ordinal
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 Completeness:

Deaths	Ten-Year Age Groups
*54013 records are supressed = 81.72% of data set	*5508 records do not state age group = 8.33% of data set
How can we deal with this missing data?	How can we deal with this missing data?
We can do nothing or impute values. Since this is 82% of our data set, I propose we impute values. The CDC suppresses data with less than 10 deaths. We can randomly impute numeric values 0-9 in the supressed cells.	We can do nothing or impute values. In this case since the data is only 8% of our total data set and there is no way to make an educated guess and imputing random numbers would not be accurate, I propose we do nothing and leave the data as is.

Data Uniqueness:
・Data Grain: State - Month Code - Ten-Year Age Group Code – Deaths
・No duplicate records found.

Data Cleaning:
・Imputed random values 0-9 for 54013 suppressed records

Data Timeliness:
・This data set contains records from 2009-2017 and is within a suitable timeline for the project. An updated version is not necessary.

US Census Population

Data Counts:

County	28985
Year	28985
Total Population	28985
Male Total Population	28985
Female Total Population	28985

Data Types:

County	time-invariant	structured	qualitative	nominal
Year	time-invariant	structured	qualitative	ordinal
Total Population	time-variant	structured	quantitative	discrete
Male Total Population	time-variant	structured	quantitative	discrete
Female Total Population	time-variant	structured	quantitative	discrete
Age Ranges	time-variant	structured	quantitative	discrete

Data Completeness:
・There were 3318 blank cells which is 11% of the total data set. In this case it is best to do nothing, as imputing random values would not be accurate and there is not enough information to make an educated guess. Finding a supplemental data source could provide it's own challenges with privacy laws.

Data Uniqueness:
・Data Grain: County - State – Year
・There was a total of 3278 counties that were recorded more than once and 25707 unique records.

Data Cleaning:

- I removed the 3278 duplicated records
- Split County and State into two columns

Data Timeliness:

- This data set contains records from 2009-2017 and is within a suitable timeline for the project. An updated version is not necessary.