Preparing for Influenza Season: Interim Report

Prepared by: Bryan Lim

Date: July 17, 2022

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.

Hypothesis

If residents in vulnerable populations are administered the flu shot, then hospitals and clinics will see a decrease in the number of patients in this target group affected by the flu.

Data Overview

- **Dataset 1: U.S Census Population**

The data contains the total population of each city and the state it's in from 2009 to 2017. It is broken into categories such as male/female, and as groups ranging from under 5 years old to 85 and older.

- **Dataset 2: CDC Influenza Deaths**

This data indicates the number of deaths for each state by month and age-group from 2009 to 2017. The age-group ranges from 1 years old to 85+ years old.

- **Dataset 3: CDC Influenza Visits**

This data displays weekly data of the total number of patients reporting of influenza-illnesses from 2010 to 2019. It also includes the number of clinics that assisted these patients and are categorized per state.

Data Limitations

Dataset 1: U.S. Census Population

Time lag: The census is conducted every ten years.

Inaccuracies: If the entries are being inputted manually, it will also be prone to human error. In addition, if people are away for work or personal reasons, they will not have answered the questions.

Dataset 2: CDC Influenza Deaths

Inaccuracies: Data is based on death certificates for U.S. residents, based on a single underlying cause of death. Individuals may have multiple health concerns in addition to influenza, but that is not accounted for. For this reason, some of these deaths may be a misrepresentation of influenza being the main cause.

Dataset 3: CDC Influenza Visits

Data is based on each state and the number of clinics that assisted these patients. A missing piece of information is that it does not specify the age group, thus it is difficult to determine if vulnerable populations are mainly affected or not.

Descriptive Analysis

	Variable 1	**Variable 2**
	Total of Vulnerable Population (65 and over)	Vulnerable Population Deaths from Influenza
Mean	268996	299
Standard Deviation	358379.77	388.74591
Variance	1.28436E+11	151123.3812
Correlation Coefficient	0.36716801	

The total population for those 65 and over has a moderately strong relationship with the number of deaths from influenza. If I were to segregate them into age groups, for example 65 to 74, 75 to 84, and 85+, the correlation coefficient would bd **much** stronger. They are each at 0.94 indicating a very strong relationship.

Results & Insights

Null Hypothesis: Vulnerable populations have higher death rates than non-vulnerable populations

Alternative Hypothesis: Vulnerable populations do not have higher death rates than non-vulnerable populations.

One-tailed test: My null hypothesis indicates that we only care about one direction—if vulnerable populations have higher death rates.

- At an alpha of 0.05, or confidence level of 95%, there's no significant difference between vulnerable populations who receive the flu shot, and the number of deaths caused by influenza.

Remaining Analysis and Next Steps

Because this project has an intervention, the reason for the lack of statistical significance must be explored. The next steps will be:

- Holding meetings with project stakeholders.
    - Brainstorming questions and convening medical agency frontline staff, staffing agency administrators, and hospitals and clinics to discuss the effects of flu shots and its prevention.
- Potentially repurposing the research hypothesis
    - Perhaps we are looking at it from the wrong angle. Maybe vulnerable populations are more susceptible to influenza because of other factors such as immunity system, general health, and well-being, etc. Some factors may not be preventable, and we must work around it.
- Repeating the analysis with additional data (if the current data doesn't capture the whole story)

Once all stakeholders have had a chance to discuss the results, and possibly change how we approach this project, the next steps would be:

- Identifying which states have the largest vulnerable populations and comparing their influenza death rates with other states
- Verifying if those states have an adequate number of staff at clinics based on history
- Determining how to best allocate our resources and prioritize states with higher death rates

Appendix

Data Profiles:

U.S. Census Population Data Profile

Variable	Time-Variant/Time-Invariant	Structured/Unstructured	Qualitative/Quantitative	Data Type
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
Under 5 years	Time-Variant	Structured	Quantitative	Discrete
5 to 9 years	Time-Variant	Structured	Quantitative	Discrete
10 to 14 years	Time-Variant	Structured	Quantitative	Discrete
15 to 19 years	Time-Variant	Structured	Quantitative	Discrete
20 to 24 years	Time-Variant	Structured	Quantitative	Discrete
25 to 29 years	Time-Variant	Structured	Quantitative	Discrete
30 to 34 years	Time-Variant	Structured	Quantitative	Discrete
35 to 39 years	Time-Variant	Structured	Quantitative	Discrete
40 to 44 years	Time-Variant	Structured	Quantitative	Discrete
45 to 49 years	Time-Variant	Structured	Quantitative	Discrete
50 to 54 years	Time-Variant	Structured	Quantitative	Discrete
55 to 59 years	Time-Variant	Structured	Quantitative	Discrete
60 to 64 years	Time-Variant	Structured	Quantitative	Discrete
65 to 69 years	Time-Variant	Structured	Quantitative	Discrete
70 to 74 years	Time-Variant	Structured	Quantitative	Discrete
75 to 79 years	Time-Variant	Structured	Quantitative	Discrete
80 to 84 years	Time-Variant	Structured	Quantitative	Discrete
85 years and older	Time-Variant	Structured	Quantitative	Discrete

CDC Influenza Deaths Data Profile

Variable	Time-Variant/Time-Invariant	Structured/Unstructured	Qualitative/Quantitative	Data Type
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

Statistical Testing:

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	0.007541953	0.001336595
Variance	9.67826E-06	1.84511E-06
Observations	459	459
Hypothesized Mean Difference	0	
df	627	
t Stat	39.1636651	
P(T<=t) one-tail	6.5828E-171	
t Critical one-tail	1.647287497	
P(T<=t) two-tail	1.3166E-170	
t Critical two-tail	1.963754701	

A two-sample t-test assuming unequal variances was conducted on two different variables:

- Death rate % of vulnerable populations
- Death rate % of non-vulnerable populations

It produced a t-score of 39.16 which was used to calculate the p-values for both the one-tailed and two-tailed tests. As my null hypothesis is concerned with one direction (if vulnerable populations have higher death rates), I used this value to compare with the statistical significance of 0.05 or a confidence level of 95%. However, because the p-value was higher, I *could not* reject the null hypothesis. If I had stated this was a two-tailed test, then I would reject my null hypothesis. But having worked with the datasets, it is clear that death rates in vulnerable populations are much higher than non-vulnerable.