Preparing for influenza season: Interim report

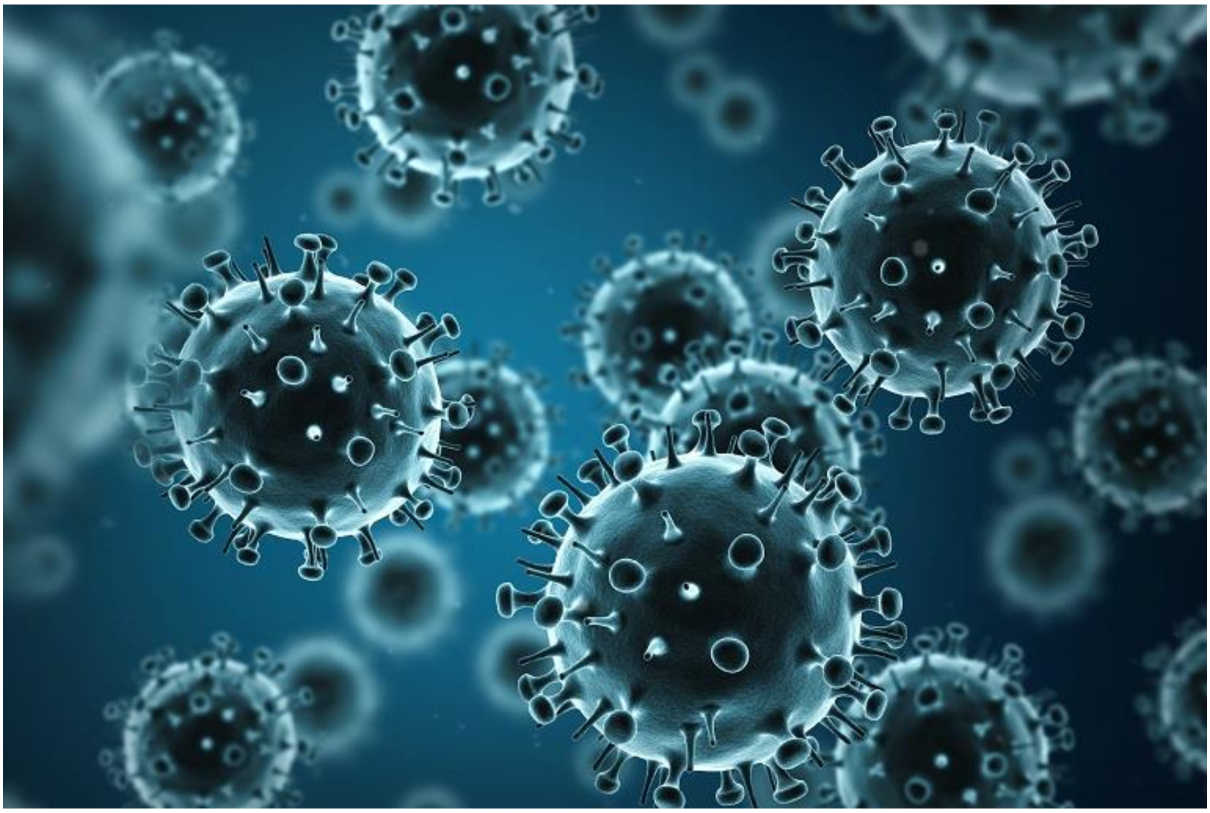

Image retrieved from https://i0.wp.com/www.nfid.org/wp-content/uploads/2019/08/Flu.jpg?ssl=1

By Atika AhmedSudi

1. 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.
- **Goal:** To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. The analysis will help plan for the influenza season, a time when additional staff is in high demand. The final results will

examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

2. Research Hypothesis

If the state has a large number of vulnerable people, then the rate of infection will be high.

3. Data overview

1. Influenza deaths by geography, time, age, and gender.

   Source: Centres for Disease Control and Prevention (CDC).

   Summary: This data provides the number of deaths from the flu by geographic location in the United States (US). The data are recorded monthly and include their age from 2009 to 2017.

2. Population data by geography.

   Source: US Census Bureau

   Summary: This data provides the population counts for each county in the US from 2009-2017. It is reported by the total population, including males and females, and in 10-year age increments.

4. Data Limitations

Influenza data set limitation: As this information is gathered manually, it is prone to typographical error and human error; also, because the data is an estimate rather than exact, there may be some errors in the information precision.

Given that our available data only spans from 2009 to 2017, any estimates we generate for subsequent years would be based on extrapolation and would inherently be estimates of estimates. Therefore, it's important to acknowledge that the accuracy and reliability of these estimates may be further reduced due to the additional layer of uncertainty involved in projecting beyond the available data timeframe.

Population data by geography: Since population data by location is only collected every ten years, approximations may differ. As death certificates are transmitted to the federal government, there may be deaths that go undetected. Estimates are derived from a variety of sources, which may be incomplete.

5. Descriptive Analysis

Data Spread:

Data Set Name	**Population over 65**	**deaths over 65**
Sample or Population?	population	population
Normal Distribution?	normal	normal
Variance	297103482758764	47596711.94
Standard Deviation	17236690.02	13798.07406
Mean	1610465.475	644.3173913
Outlier Percentage	0.217864924	0.217864924

6. Description of one Correlation

Variables	**Population over 65**	**Deaths over 65**
Proposed Relationship	Both of the variables are related so it should be high	
Correlation Coefficient	0.9998	
Strength of Correlation	strong	strong
Usefulness/Interpretation	The strong correlation between a state's population over 65 and its amount of influenza deaths supports the hypothesis of a high rate of influenza mortality for this age group.	

The data shows a strong positive correlation (0.9998) between the population over 65 and deaths over 65 due to influenza, supporting the hypothesis of a high rate of influenza mortality for this age group. Both variables are normally distributed in the population, but the variance and standard deviation are higher for the population over 65 than for deaths over 65. There may be some extreme values (outliers) that could be influencing the correlation coefficient.

7. Summary of Results and Insights

Statistical Hypothesis Testing

Research Hypotheses:	**If the state has a large number of vulnerable people, then the rate of infection will be high**
Null Hypotheses:	There is no significant relationship between the number of vulnerable people in a state and the rate of infection
Alternative Hypotheses:	There is a significant positive relationship between the number of vulnerable people in a state and the rate of infection
Dependent Variable:	the rate of infection
Independent Variable:	age group
Two-tailed or one-tailed	One-Tailed
P-Value set	0.05
P-Value	3.62911E-59
Significant Level	The significance level assessment for a hypothesis test with a one-tailed alternative hypothesis and a p-value of 3.62911E-59 depends on the chosen significance level. If the p-value is less than or equal to 0.05, we can reject the null hypothesis and conclude that there is a significant relationship between the independent and dependent variables.

Next Steps	Vulnerable people have a higher risk of dying from influenza, so states with the highest population over 65 and the highest death rate should be taken into account.

8. Statistical Hypothesis and Interpretation

	Population	Deaths
Mean	5973847.155	265.7102397
Variance	4.63324E+13	82596.18005
Observations	459	459
Hypothesized Mean Difference	0	
df	458	
t Stat	18.80176834	
P(T<=t) one-tail	3.62911E-59	
t Critical one-tail	1.648187415	
P(T<=t) two-tail	7.25821E-59	
t Critical two-tail	1.965157098	

This statistical analysis is testing for a significant difference between the mean number of deaths in a population (with a sample size of 459) and a hypothesized mean difference of zero. The t statistic is 18.80176834, which suggests that there is a significant difference between the sample mean and the hypothesized mean difference. The p-value (P(T<=t)) is very low, indicating that there is strong evidence against the null hypothesis. Therefore, the mean number of deaths in this population is significantly different from the hypothesized mean difference of zero.

9. Remaining Analysis and Next Steps

Additional Analysis:

- The goal of this project is to examine the vaccination history of the vulnerable-aged population and whether or not vaccination status has an effect on the outcome of influenza-related illness.

- If it is discovered that vaccination has reduced the risk of mortality in this demographic, efforts should be directed toward vaccination campaigns to protect these residents from severe complications that could end in hospitalization or death.
- The fundamental motivation for this endeavour is adequate staffing in each state.
- To establish which states, have higher rates of vulnerable-aged citizens, data on current staffing levels in each state and how many patients each staff member can care for should be collected.

Visualization:

- Tableau visualisations of state populations and variable correlations should be generated to indicate the number of vulnerable residents per state and where efforts should be concentrated.

Appendix

Glossary

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

Vulnerable populations: Patients likely to develop flu complications requiring additional care, as identified by the Centres for Disease Control and Prevention (CDC). These include adults over 65 years, children under 5 years, and pregnant women, as well as individuals with HIV/AIDs, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

Additional Context

A count of the historical influenza deaths gives an indication of the severity of flu in an area. Deaths can be prevented with flu shots and adequate medical staff. In the United States, each state has a different population composition, meaning that some states will have more vulnerable populations. In this project, one should pay particular attention to influenza deaths, vulnerable populations, and (optionally) flu-shot rates—particularly in vulnerable populations—to determine medical staffing needs.

Additional Details

Influenza Data set

Data Source: This is external data as this came from CDC, since this is from the government it is reliable information.

Data Collection: The CDC has compiled this administrative collection. This data is collected and grouped annually and represents an estimate of individual births, deaths, race, and migration patterns in the United States; however, because this information is manually entered by government employees, it is subject to human error. This data has a time lag because it is collected and categorized annually.

Data Contents: This data set contains the total population per county from the year 2009 to 2017, it is divided by gender and age (in 5-year intervals) until the age of 85 anything over that age they are grouped.

Data Limitation: As this information is gathered manually, it is prone to misprints and human error; also, because the data is an estimate rather than exact, there may be some errors in the information.

As we only have data from 2009 to 2017, which is 6 years old, and while we can construct estimates for later years, our information is already an estimate, so any estimates we derive will be approximations of estimates.

Data Relevance: This data is applicable to my hypothesis as it is how many vulnerable people live in a certain location.

Influenza Laboratory Tests and Patient Visits data sets

Data Source: These data sets are external as they are owned by the CDC Centre for disease control and Prevention and since this information is from the government it is a reliable source.

Data Collection:

- **Influenza Visits:** Since of the voluntary nature of the data collection and the fact that it is manually recorded, this data was collected as a survey. This data set measures the number of doctor visits and total patients seen by week and state; this information is updated weekly, so there is no time lag.
- **Lab Tests:** As this data was obtained as a survey and was both manual and optional, it is neither complete nor error-free. This data was collected weekly by health providers and clinical labs, so there is no time lag.

Data Contents:

- **Influenza Visits:** From 2010 to 2019, the data includes different age categories, percentage of weighted influenza-like illnesses, percentage of unweighted influenza-like illnesses, the total influenza-like illnesses, number of providers, and total patients.
- **Lab Tests:** From 2010 through 2015, this data set offers information by state and week, the number of samples, and the percentage of positive variables. It also includes data on strains A(H1N1), A(H1), A(H3), A (no subtyping), A (unable to subtype), B, and H2N2v. Each state reports the total number of samples collected as well as the proportion of those that tested positive. It appears to classify positive tests based on the kind of influenza (i.e., A subtype H1N1, H1, H3, A with no subtype, A and unable to subtype, B & H3N2v).

Data Limitations: Both data sets (Influenza visits and Lab Tests) are manually collected by health providers and laboratory clinics and are prone to typing errors and personal biases. Furthermore, both have restrictions in terms of time, with patient visits being four years old and influenza laboratory testing being eight years old. This may require calculating averages for years not included, creating margins of error.

Data Relevance: As this data is unreliable and incomplete and with the limitations provided complete with human error that can’t be fixed with such a large amount of data, it is information that has no relevance to my hypothesis.

<u>Children’s Flu Shots Data Set</u>

Data Source: This is an external data set as the information is owned by the CDC and originates from the University of Chicago, which uses a third party to conduct surveys, the NIS (National immunization survey) as this is owned by the government it is a reliable source.

Data Collection: Survey data is collected over the phone and manually entered from a random sample of parents. Each child's vaccine provider receives a questionnaire. When the quantity of results reaches a certain threshold, they are manually input into a computer and uploaded to a bigger database. Individual responses are used to generate survey data, which is not always standardized or reliable. since all of this is time-consuming there is a time lag.

Data Contents: This data set contains vaccination of children ranging from 6 months to 17 years old, education, number of children, marital status, demographic information, type of insurance, insurance history, and year. It is only valid for 2017, and it includes information about family demographics such as ethnicity, education level, poverty level, and marital status.

Data Limitations: Because the survey is performed over the phone, misprints are possible. If the interviewer is biased, they may change what they type to reflect their ideas. The data cannot be used to produce final recommendations for the number of medical workers needed in each state.

Data Relevance: This data has no significance to my hypothesis as this information is biased and might have human error which cannot be used in the final product.