Medical Staff Planning
Influenza Season

Interim Report

Simone van der Graaf

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 you are 65 years and older, then the chances of dying from the flu are higher.

Data Overview

Data source: population data by geography, time, age and gender

The data is a complete count of the US population per county and state, from 2009 until 2017. It includes the total, male and female population and age groups per 5 years, until 85 years and over. The data source comes from US Census Bureau.

Data source: influenza deaths by geography

The data comes from CDC and contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken into two categories: state and age.

Data Limitations

Data source: population data by geography, time, age and gender

Time lag: Every 10 years, the US Census Bureau conducts a census counting every resident in the United States. Because it's only done every 10 years it has a risk that it is not up to date and therefore has limitations. The data shows that for some counties one or more years are missing. Many duplicated have been found and removed, because the data is collected manually there is always a risk of duplicates and/or errors.

Data source: influenza deaths by geography

Inaccuracies: When there are 1 to 9 deaths counted per state and year in different age groups, it's described as "Suppressed" instead of actual numbers given. Therefore, some of the data is not accurate and has been replaced with random numbers.

Descriptive Analysis

	Deaths 65 years and older	Total Deaths
Mean	897	1434
Standard Deviation	972	1082

The number of deaths of people 65 years and older has a strong correlation (0,9988) with the total number of deaths. This means that both variables increase almost identically. These numbers are from all states in the years 2009-2017. Which indicates that the relationship between the two variables is very similar throughout the US in those years.

Results and Insights

Null Hypothesis: people of 65 years and older have lower chances of dying from the flu.

Alternative Hypothesis: people of 65 years and older have higher chances of dying from the flu.

T-test results (based on a significance level of 0.05):

	% of deaths 0-64 years	*% of deaths 65 years and older*
Mean	0,000269646	0,001313872
Variance	7,66014E-08	2,75655E-07
P-value	**4,61452E-170**	

After including the percentage of deaths of people younger than 65 years and people of 65 years and older in the statistical hypothesis test, I calculated a p-value of less than 0.01. Based on the significance level of 0.05, the p-value is significantly smaller. I can therefore reject the null hypothesis and claim that indeed the chances of dying from the flu are higher when you are 65 years and older.

Remaining Analysis and Next Steps

Now that we have certainty that people of 65 years and older are indeed more vulnerable, it is important to compare the number of deaths of this vulnerable group per state. With this information the medical staffing agency will know how to divide more staff over the most vulnerable states. Another crucial factor in the remaining analysis is to determine whether influenza occurs seasonally or throughout the entire year. Composition, statistical, spatial, and temporal visualizations will be part of further analysis.

The project will be completed by presenting the final results in a video recording that will be sent to the stakeholders. The analysis will also be published as a Tableau Storyboard.

Appendix

Business Requirements

- Provide information to support a staffing plan, detailing what data can help inform the timing and spatial distribution of medical personnel throughout the United States.
- Determine whether influenza occurs seasonally or throughout the entire year. If seasonal, does it start and end at the same time (month) in every state?
- Prioritize states with large vulnerable populations. Consider categorizing each state as low-, medium-, or high-need based on its vulnerable population count.
- Assess data limitations that may prevent you from conducting your desired analyses.

Hypothesis Development

Project: create an analysis to help a medical staffing agency with planning resources and distributing staff for an upcoming influenza season.

Clarifying questions:

1. Which people are most vulnerable to getting the flu?
2. How long is the flu season?
3. Why does the medical staffing agency need help?

Funneling questions:

1. Is every year the same group of people the most vulnerable to getting the flu?
2. Is the duration of the flu season the same in every state?
3. Does the medical staffing agency distribute their staff the same way every year?

Privacy and ethical questions:

1. Are there privacy laws we need to adhere to related to collecting, storing, and analyzing data from medical staff?
2. Do we need to ask permission to the patients, to be able to use their data?

Data Profiles

Data source: population data by geography, time, age and gender

County and Year: The duplicates have been removed.
All Age Groups: The population numbers have been rounded up by removing decimals.

Data Types

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
County	Time-invariant	Structured	Qualitive	Nominal
Year	Time-invariant	Structured	Qualitive	Ordinal
Total population	Time-invariant	Structured	Quantitative	Discrete
Male Total population	Time-invariant	Structured	Quantitative	Discrete
Female Total population	Time-invariant	Structured	Quantitative	Discrete
Under 5 years	Time-variant	Structured	Quantitative	Continuous
5 to 9 years	Time-variant	Structured	Quantitative	Continuous
10 to 14 years	Time-variant	Structured	Quantitative	Continuous
15 to 19 years	Time-variant	Structured	Quantitative	Continuous
20 to 24 years	Time-variant	Structured	Quantitative	Continuous
25 to 29 years	Time-variant	Structured	Quantitative	Continuous
30 to 34 years	Time-variant	Structured	Quantitative	Continuous
35 to 39 years	Time-variant	Structured	Quantitative	Continuous
40 to 44 years	Time-variant	Structured	Quantitative	Continuous
45 to 49 years	Time-variant	Structured	Quantitative	Continuous
50 to 54 years	Time-variant	Structured	Quantitative	Continuous
55 to 59 years	Time-variant	Structured	Quantitative	Continuous
60 to 64 years	Time-variant	Structured	Quantitative	Continuous
65 to 69 years	Time-variant	Structured	Quantitative	Continuous
70 to 74 years	Time-variant	Structured	Quantitative	Continuous
75 to 79 years	Time-variant	Structured	Quantitative	Continuous
80 to 84 years	Time-variant	Structured	Quantitative	Continuous
85 years and over	Time-variant	Structured	Quantitative	Continuous

Data source: influenza deaths by geography

State: All N/A values have been changed to the correct state.
Year: The mistakes in "Year" were easily solved by checking the column "Month", which also shows the year.
Deaths: The suppressed values that represent any number between 1 and 9, have been replaced by random numbers between 1 and 9.
Ten-Year Age Groups: All "Not Stated" values have been removed.

Data Types

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
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-variant	Structured	Qualitative	Ordinal
Ten-Year Age Groups Code	Time-variant	Structured	Qualitative	Ordinal
Deaths	Time-variant	Structured	Quantitative	Discrete

Additional Information

The report informs that the p-value is less than 0.01, lower that the significance level of 0.05.
The table on the following page shows all results from the t-test used in the statistical analysis:

T-test: Two-Sample Assuming Unequal Variances

	% of deaths 0-64 years	*% of deaths 65 years and older*
Mean	0,000269646	0,001313872
Variance	7,66014E-08	2,75655E-07
Observations	458	458
Hypothesized Mean Difference	0	
df	693	
t Stat	-37,65285167	
P(T<=t) one-tail	4,61452E-170	
t Critical one-tail	1,647055388	
P(T<=t) two-tail	9,229E-170	
t Critical two-tail	1,96339306	