Influenza Season Staffing: Interim Report

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.

Project Hypothesis:

The age group with the most reported deaths throughout the country over the dataset year was 65+. Further, children under five years of age are also a vulnerable population. If we allocate more resources to counties with large numbers of senior citizens and small children, we can reduce deaths in those populations from influenza.

Data Set Summaries:

Census Data:

- This is an external data source owned by the U.S. Census Bureau. The U.S. Government has many controls in place to help ensure the most complete data capture possible for the census.
    - The controls include regulations making nonparticipation or falsification a crime ($5,000 fine for failure to participate, $10,000 fine for falsification), multiple methods for survey participation (paper, telephone, electronic) all in multiple languages, and assertive data collection attempts.
    - This is the most trustworthy dataset of this type available.
- The data is collected by survey every 10 years.
    - The annual data in this set is extrapolated and estimated by the U.S. Census Bureau.
- The dataset includes:
    - A breakdown of data by state and county.
    - A breakdown of data by binary gender.
    - A breakdown of data by age group.
    - A data element to identify year.

CDC Influenza Deaths:

- This dataset was extracted from the Centers for Disease Control and Prevention (CDC) Underlying Cause of Death dataset. The CDC has many controls in place to maintain the overall quality and completeness of the data reported.
- The data is collected by survey monthly, with providers self-reporting.
- The extracted dataset we have includes:
    - A breakdown of data by state and month**.**
    - A breakdown of data by 10-yr age groups.
    - The number of reported deaths each month in each of those age groups.

Data Limitations:

Census Data:

- This data becomes less dependable for exact numbers for each year it is removed from the most recent census. However, for our purposes and projections this is the most reliable data available.
    - The numbers for male and female appear to be taken from the census surveys and then extrapolated to whole numbers for each non-census year.
    - The numbers for each age group appear to be estimates, as evidenced by fractional numbers.
- The survey data is entered manually:
    - Although aggregated by computer, the survey forms are completed either by the participant or by a (temporary) government employee.
    - There are multiple opportunities for error, omission, and/or falsification.

CDC Influenza Deaths:

- Counts of less than 10 deaths for any age group in each month are suppressed.
    - This has the potential to introduce bias, which may be significant, into our results.
    - We have no reported death data for anyone under 5yrs old or younger, which is one of the populations considered vulnerable.
- Reporting is not mandatory.
    - Unreported data could result in lower than optimal allocations at the completion of our project. It is, however, in the best interest of each entity to report this data. The expectation is that the impact of unreported data will be de minimis.

Descriptive Analysis of Core Variables:

After cleaning and integrating the datasets, I created fields to directly address the vulnerable population for which we had sufficient data: persons aged 65 years or older. The most relevant descriptive analysis of the resulting created fields follows:

Data Spread		
Variables	65+ Population	65+ Deaths
Dataset Name	Integrated Dataset	
Sample or Population?	Sample	Sample
Normal Distribution?	Yes	Yes
Variance	786,799,499,984	1,028,484
Standard Deviation	887,017	1,014
Mean	806,989	826
Outlier Percentage	8.06%	6.54%

I created a field for 65+ Deaths per 1,000 population. The goal for this field is to normalize the deaths in the vulnerable population to remove the bias of large population centers. If the correlation is very strong, then the field would not be useful to help determine priorities for staffing for our project. If the correlation is moderate to just barely in the strong category, the field could be useful.

Correlation		
Variables	65+ Deaths	65+ Deaths per 1000 Population
Proposed Relationship	Moderate to Strong	
Correlation Coefficient	0.51	
Strength of Correlation	Strong	
Usefulness / Interpretation	Useful for prioritizing staffing	

Results and Insights:

Much of our project hinges on the assumption that people 65 years of age and older is a vulnerable population with respect to influenza mortality. We need to ensure that mortality in the population we have identified as vulnerable is at significantly greater risk than the population as a whole to properly allocate staffing based on this assumption.

Research Testing	
Research Hypothesis	The population of individuals 65 years of age and older is a vulnerable population.
Independent Variable	65+ Deaths
Dependent Variable	5-64 Deaths

I chose the following testing for this purpose:

Statistical Testing	
Null Hypothesis	The number of deaths in the 65+ population will be the same or less than the number of Deaths in the 5-64 age range
Alternative Hypothesis	The number of deaths in the 65+ population will be greater than the number of Deaths in the 5-64 age range
T-Test Type	One-Tailed - For this test we're only interested in "greater than"
Significance level	The significance level will be 0.05 to return a 95% confidence level in the result.
P-value	6.9601E-45
Assessment	The average number of deaths in the 65+ age group is significantly greater than the average number of deaths between ages 5 and 64. The level of confidence in the result is greater than 95% (actually greater than 99%). This sufficiently disproves the Null Hypothesis with respect to our project.

The results of the t-test follow:

t-Test: Two-Sample Assuming Unequal Variances	*65+ Deaths*	*5-64 Deaths*
Mean	826.56	78.82
Variance	1,030,699	22,953
Observations	458	458
Hypothesized Mean Difference	-	
df	477.00	
t Stat	15.5897	
P(T<=t) one-tail	6.9601E-45	
t Critical one-tail	1.6481	

The results for the P-values are given in scientific notation. For those unfamiliar with scientific notation, the actual number is a decimal point followed by 44 zeroes and then the 69601. The key point is that the P-value is less than the level of significance established. The extent to which the number is less than the level of significance is not important. Our level of confidence in our result is greater than 95%.

From the testing, it appears that people 65 years of age and older are a vulnerable population. Allocating based on this assumption appears to be valid. The project should continue under this assumption.

Next Steps:

The next step in our project plan is to proceed with building visualizations. I will be using Tableau for this portion of the project. This will allow additional insights into the most effective and efficient forecasting and, ultimately, allocation of staff for the coming influenza season.

We are still on target for our presentation to management.

Appendix

Clarifying Questions (unknown in green text, answers in red text):

- Which serious complications are related to the flu?
  - There are several, they are determined by doctors, and this will have no impact on this project.
- How is the relationship between influenza and death determined?
  - The relationship between influenza and death is based on official cause of death listed on a death certificate. The first five characters of the ICD-10 code are used for specification.
  - How are comorbidities addressed in this determination?
    - Comorbidities are not addressed. Only one code for cause of death is used.
- When is flu season?
  - Is flu season the same in all states?
    - Flu season is not identical in all states. However, the difference in timing is not significant and does not allow us to gain any efficiencies in allocation.

Ethics and Privacy Questions:

- Are there legal constraints beyond HIPAA that must be considered for gathering and use of data?
  - All datasets used for our project are aggregated. There is no personal or individual data included. The data is publicly available and may be accessed, used, copied, distributed and/or published freely per the respective websites. As our use is internal and not for outside publication, citation is not required.

Glossary of Terms:

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 Centers 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.

ICD-10: The Centers for Medicare and Medicaid Services (CMS) use the International Classification of Diseases (ICD) for claims. The version currently used by CMS is ICD-10. This system is used to "promote international comparability in the collection, processing, classification, and presentation of mortality statistics."