NOVEMBER 5, 2022

INFLUENZA SEASON

INTERIM REPORT

BY NGAWANG DHUNDUP

Ngawang Dhundup

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 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 we send more resources to states with high vulnerable population then mortality rate will decrease.

DATA OVERVIEW

US CENSUS DATA:

This data source is external, from the US Governmental Bureau, and therefore it is trustworthy. The data is administrative, and represents an actual count of the population, by age group in the different counties and states. Data is collected through online questionnaires, once a year using a survey form produced by the Bureau, which local governmental bodies fill out and return. The data contains census data collected between 2009 and 2017, including the total population per county/state, then also broken down into sex (total female, total male) and age groups from under 5 years to 85 years and over (in increments of 5 years).

INFLUENZA DEATHS DATA:

This data source is external, provided by the Centers for Disease Control and Prevention (CDC) through their National Center for Health Statistics. This is a trustworthy data source, collected as part of the National Vital Statistics Cooperative Program. Each of the U.S. states and territories is required to record all births, deaths, marriages, and divorces within their jurisdiction. Death records come from death certificates, in which a doctor codes the primary cause of death as “Influenza” or “Pneumonia”.

The data 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 groups from under 5 years to 85 years (in increments of 10 years).

DATA LIMITATIONS

A key limitation for both datasets is timeliness, in both datasets the data is collected once a year and there would be delay in data relevancy. There is also a concern for data accuracy because the census data is survey based and dependent on human input, so there could be missing/inaccurate data for certain areas. In the case of Influenza deaths dataset there is also a case for bias, due to the fact that the doctor who codes the cause of death, may have attributed the death to influenza, even though there may have been other co-morbidities.

DESCRIPTIVE ANALYSIS

The table below is a descriptive summary for the two variables, Vulnerable population and Not vulnerable population concerned with the research hypothesis.

DESCRIPTIVE FOR VULNERABLE VS NOT VULNERABLE		
	Sum not vulnerable population deaths	**Sum Vulnerable population deaths**
Mean	416.63	1016.80
Variance	14183.16	944307.02
Standard Deviation	119.09	971.75
P(T<=t) two-tail	8.37872E-34	

There was a strong relationship found between the vulnerable population (sum of people over 65, and under 5 years) and influenza deaths, supported by a correlation coefficient of 95%. The outlier percentage for both variables were close or less than 5% so it is within the norm and negligible.

DESCRIPTIVE FOR VULNERABLE POPULATION		
	Sum USA census Vulnerable population	**Sum Influenza deaths Vulnerable Population**
Mean	1193271.58	1016.80
Variance	1761516451463.14	944307.02
Standard Deviation	1327221.33	971.75
Outlier Percentage	6.97%	3.92%
P(T<=t) two-tail	6.40968E-61	
Proposed Relationship:	A positive relationship for vulnerability and death related to the flu	
Correlation Coefficient	95%	
Strength of Correlation	Strong Relationship	

RESULTS AND INSIGHTS

The results were as expected and the more vulnerable population did in fact have significantly more deaths compared to the rest of the population.

In order to arrive at this conclusion I conducted tests on two different statistical hypothesis:

- Null hypothesis: The vulnerable group (population aged 65+ years and under 5 years) have equal deaths than the rest of the population

- Alternative hypothesis: The vulnerable group (population aged 65+ years and under 5 years) have more deaths than the rest of the population

The average number of deaths for vulnerable population is 1,016.80, while the rest of the population the average is only 416.63, a major **difference of 600.17**.

DEATHS OF VULNERABLE		
t-Test: Two-Sample Assuming Unequal Variances		
	Sum of Vulnerable deaths	*Sum of Vulnerable pop*
Mean	1016.79956	1193271.58
Variance	944307.021	1.7615E+12
Observations	459	459
Hypothesized Mean Difference	0	
df	458	
t Stat	-19.245622	
P(T<=t) one-tail	3.2048E-61	
t Critical one-tail	1.64818741	
P(T<=t) two-tail	6.4097E-61	
t Critical two-tail	1.9651571	

I conducted a (two-tailed) t-test with an **alpha of 0.05** which resulted in a p-value of **6.4097E-61**, meaning we can say with a 95% confidence level that the higher numbers of deaths in vulnerable populations is statistically significant. Therefore, we can **reject** the null hypothesis and state that there is evidence for the alternative hypothesis.

DEATH RATE PERCENTAGE VS		
t-Test: Two-Sample Assuming Unequal Variances		
	Sum of Not vulnerable	*Sum of Vulnerable*
Mean	0.02%	0.12%
Variance	0.00%	0.00%
Observations	459	459
Hypothesized Mean Difference	0	
df	575	
t Stat	-31.150585	
P(T<=t) one-tail	7.587E-126	
t Critical one-tail	1.64750796	
P(T<=t) two-tail	1.517E-125	
t Critical two-tail	1.96409822	

I conducted another (2 tailed) t test to see what the death rate percentage was for the two groups, vulnerable and not vulnerable. With a significance value of 0.05 and with a p-value of 1.517E-125, I can see with 95% confidence that the sum of vulnerable population has a higher death rate of 0.12% compared to the low 0.02% of the not vulnerable group.

REMAINING ANALYSIS AND NEXT STEPS

Based on the results of the analysis, the next steps would be to identify areas where there are large vulnerable populations and ensure that we send more staff and supplies in order to reduce death rates. The data has a significant level of confidence to use prescriptive analysis and distribute staff to the most affected communities. The project overview also asks us to understand whether the influenza season differs from state to state, which will require further analysis for us. Additional analysis that could be done would be to investigate vaccine rates across states, as it has been shown that vaccines deter the impact of influenza, as well as whether weather plays a part in influenza infections and ultimately death.

Therefore, states with lower vaccine rates would potentially need more medical staff to help cater to patients.

Further deliverables will include a Tableau storyboard with temporal, statistical, and spatial visualizations. The final research will be presented as a video, including visualizations, conclusions, and recommendations

APPENDIX

PROJECT 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 influenza season, a time when additional staff are 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.

STAKEHOLDER IDENTIFICATION:

- Medical agency frontline staff (nurses, physician assistants, and doctors)
- Hospitals and clinics using the staffing agency's services
- Influenza patients
- Staffing agency administrators

SUCCESS FACTORS:

The project's success will be based on:

- A staffing plan that utilizes all available agency staff per state requirements, without necessitating additional resources
- Minimal instances of understaffing and overstaffing across states (a state can be considered understaffed if the staff-to-patient ratio is lower than 90% of the required ratio and overstaffed if greater than 110%)

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 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

DATA SET SOURCES:

1. Influenza deaths by geography, time, age, and gender Source: CDC
   Download Data Set

2. Population data by geography
   Source: US Census Bureau
   Download Data Set
   NOTE: keep in mind that these numbers are estimates, hence the sum of the numbers from the different age groups may not sum up to the total in the first columns, but that's totally okay.

ADDITIONAL STATISTICAL TESTS

DESCRIPTIVE ANALYSIS FOR 2 VARIABLES

DESCRIPTIVE FOR VULNERABLE POPULATION		
	Sum USA census Vulnerable population	**Sum Influenza deaths Vulnerable Population**
Mean	1193271.58	1016.80
Variance	1761516451463.14	944307.02
Standard Deviation	1327221.33	971.75
Outlier Percentage	6.97%	3.92%
P(T<=t) two-tail	6.40968E-61	
Proposed Relationship:	A positive relationship for vulnerability and death related to the flu	
Correlation Coefficient	95%	
Strength of Correlation	Strong Relationship	

INFLUENZA DEATHS

DESCRIPTIVE FOR VULNERABLE VS NOT VULNERABLE		
	Sum not vulnerable population deaths	**Sum Vulnerable population deaths**
Mean	416.63	1016.80
Variance	14183.16	944307.02
Standard Deviation	119.09	971.75
P(T<=t) two-tail	8.37872E-34	

T TEST 1

DEATHS OF VULNERABLE		
t-Test: Two-Sample Assuming Unequal Variances		

	Sum of Vulnerable deaths	*Sum of Vulnerable pop*
Mean	1016.79956	1193271.58
Variance	944307.021	1.7615E+12
Observations	459	459
Hypothesized Mean Difference	0	
df	458	
t Stat	-19.245622	
P(T<=t) one-tail	3.2048E-61	
t Critical one-tail	1.64818741	
P(T<=t) two-tail	6.4097E-61	
t Critical two-tail	1.9651571	

T TEST 2

DEATH RATE PERCENTAGE VS		
t-Test: Two-Sample Assuming Unequal Variances		
	Sum of Not vulnerable	*Sum of Vulnerable*
Mean	0.02%	0.12%
Variance	0.00%	0.00%
Observations	459	459
Hypothesized Mean Difference	0	
df	575	
t Stat	-31.150585	
P(T<=t) one-tail	7.587E-126	
t Critical one-tail	1.64750796	
P(T<=t) two-tail	1.517E-125	
t Critical two-tail	1.96409822	

DATA SPREAD

	variable 1	Variable 2
Dataset Name	USA census population Vulnerable	Influenza deaths vulnerable
Sample or Population?	Sample	Sample
Normal Distribution?	Normal Distribution	Left Skew
Variance	1.76152E+12	944307.0209
Standard Deviation	1327221.327	971.7546094
Mean	1193271.579	1016.799564
Deviations		
upper		
1 standard deviation	2520492.906	1988.554174
2 standard deviation	3847714.234	2960.308783
lower		
1 standard deviation	-133949.748	45.04495488
2 standard deviation	-1461171.075	-926.7096545
Outlier Count	32	18
Outlier Percentage	6.97%	3.92%
Correlation		
Variables:	Flu deaths and Populations for vulnerable pop	
Proposed Relationship:	A positive relationship for age and death related to the flu	
Correlation Coefficient	0.945984438	
Strength of Correlation	Strong Relationship	