Preparing for Influenza Season: Interim Report

Prepared by: Balachandar Kaliappan

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.

Research Hypothesis:

H_1: Children under age 5 have a higher chance of dying due to Influenza compared to age groups between more than 5 years and less than 65 years.

H_2: Individuals aged more than 65 have a higher chance of dying due to Influenza compared to age groups between more than 5 years and less than 65 years.

Data Overview:

Two data sets were used for this project:

1. Influenza deaths by geography dataset provided by CDC.
   The influenza data set has the influenza death cases of each state in the US, along with information on the months, years, and age groups of participants. The main limitation is that they don't report death values below 10, which puts us in a situation where we create random numbers to impute the data.
2. Population data provided by the US Census Bureau.
   The population data provide estimates of each county within the US states. The data includes males, females, and the total population. Further, it is split to inform the population by age group. Since the dataset does not have accurate population information, it is a significant limitation of the study.

Additionally, the influenza dataset contains information for the states of Washington, D.C., and Washington. However, only Washington was present in the population data set. Hence, the data for Washington, D.C., is not counted for the data analysis.

Both datasets have been merged for analysis after performing data profiling, cleaning, and wrangling.

Data Summary:

Table 1 shows the summary statistics of the two variables taken for the study.

	Under 5 - Death Rate	**Under 5 - Population**	**Above 65 - Death Rate**	**Above 65 - Population**
Mean	109.68	393689.79	907.95	820269.64
Variance	435.34	210723641756.06	959218.10	790647644378.49
Standard Deviation	20.86	459046.45	979.40	889183.70

Note: For the analysis purposes, only the normalized dataset is used.

Correlation results:

A correlation test was performed to assess whether the increase in the vulnerable group's population is associated with the death rate of the vulnerable group. It was found that there was a very weak positive correlation between the age group 5 and the population of age group 5 (*Pearson r = 0.05*), and a similar observation was found for the age group more than 65 years (*Pearson r = 0.01*)

Results and Insights:

The interim report took two hypotheses; it assessed whether the death rate (in percentage) among the vulnerable group is higher than that of the non-vulnerable group.

In hypothesis 1, I tested whether the age group 5 has a higher death rate than the non-vulnerable group aged above 5 and below 65. The one-tailed t-test results indicated that there is a significant difference between the death rates of vulnerable and non-vulnerable age groups. However, the results showed that the death rate of the age group below 5 is significantly lower than that of the non-vulnerable age group—above 5 and below 65 (see Appendix 1; Table 2).

The mean differences indicate that the identified age group 5 had a significantly lower death rate compared to the non-vulnerable groups.

Table 2: Hypothesis 1 test results

Variables/ Parameters	**Values**
Mean death rate – below 5 years:	0.1069
Mean death rate – above 5 years and below 65 years:	0.347
t-statistic:	-37.726
One tail p-value:	< 0.001

In hypothesis 2, I tested whether the age group 65+ has a higher death rate compared to the non-vulnerable group.

The results showed that the death rate for individuals above 65 is significantly higher than that of the non-vulnerable age group—those aged 5 to 65. The mean differences indicate that individuals above 65 had a substantially higher death rate compared to the non-vulnerable group (see Appendix 2; Table 3).

Table 3: Hypothesis 2 test results

Variables/ Parameters	Values
Mean death rate – above 5 years:	0.544
Mean death rate – above 5 years and below 65 years:	0.347
t-statistic:	19.417
One tail p-value:	< 0.001

Conclusion:

Out of the two vulnerable groups – those below 5 years and those above 65 years - the medical team identified them based on age category. Only individuals above 65+ years are more vulnerable and have a higher chance of death compared to the non-vulnerable age group – above 5 years to less than 65 years.

Next steps:

The rest of the report, including spatial visualization, temporal trend analysis, detailed age group analysis, and socioeconomic correlations, will be completed in the coming days. Additionally, I will explore how age is linked to the death rate and examine how the death rate is different across states and counties.

Final presentation:

In the final presentation, I will present the key findings, supported by visuals and recommendations to address the project's goals. Notably, the stakeholders will be able to understand the influenza occurrence patterns to allocate manpower to the most vulnerable counties/ states.

Appendix

Appendix 1:

One-tailed t-test results: <5 years vs. 5+ to <65 years

t-Test: Two-Sample Assuming Unequal Variances

	< 5 years	*>5 to 65 years*
Mean	0.106919602	0.347481088
Variance	0.002443432	0.01585284
Observations	450	450
Hypothesized Mean Difference	0	
df	584	
t Stat	-37.72689389	
P(T<=t) one-tail	5.2537E-159	
t Critical one-tail	1.64746699	
P(T<=t) two-tail	1.0507E-158	
t Critical two-tail	1.964034381	

Appendix 2:

One-tailed t-test results: 65+ years vs. 5+ to <65 years

t-Test: Two-Sample Assuming Unequal Variances

	> 65 years	*>5 to 65 years*
Mean	0.544235673	0.347481088
Variance	0.030350053	0.01585284
Observations	450	450
Hypothesized Mean Difference	0	
df	818	
t Stat	19.41764393	
P(T<=t) one-tail	1.14739E-69	
t Critical one-tail	1.64671855	
P(T<=t) two-tail	2.29477E-69	
t Critical two-tail	1.962868294	