**Influenza season planning: Mid-Project Report**

**Chris Arnold**

**Project Overview**

**Motivation:** The flu season can vary year to year depending on many factors. Hospitals and staffing agencies need a staffing plan utilizing the current pool of employees to optimize allocation of employees to best cover individual state needs during flu season.

**Objective:** This project will review which locations have the highest vulnerable populations against deaths for the flu. With this data a plan will be devised for allocating staff to each state for the upcoming flu season.

**Scope: The staffing agency works with all hospitals in the 50 states of the USA. This project will devise a plan of action for the flu season ahead.**

Hypothesis: If a state has a high count of the vulnerable population, then these regions will have higher flu attributed deaths.

**Data overview:** There were two data sets that were combined, in-order to start the analysis of the flu data from 2009-2017. The first data set was from the CDC and was titled “Influenza deaths by geography, time, age, and gender.”The title thoroughly describes the data , which shows the death by influenza data broken up into specific year with state, state and month data of age groups going from 0-5 through 85+.

The second data set used was “Population data by geography,”and this came from the US census data webpage.This provided population data of all counties within each state for each year broken up into male/female, and age groups ranging also from 0-5 through 85+.

These two sets were combined to show the deaths (due to flu) of each age group in each state, for each year. Through this it will be able to be shown which states have high counts of vulnerable populations by comparing flu deaths to population.

**Data Limitations: Influenza death data + Population data by geography**

**Source 1:** CDC should be highly accurate and representative of the information it contains. The issues that were addressed were manual input error, and some gap filling using other data within the set to estimate some missing data.

**Source 2:** The US census is a manual form that people must participate in by hand and choose to send back to the government for processing. The reporting itself should not be biased, however we need to consider that certain groups may be opposed to giving out their information (likely small). Also since this is a manual collection method, it is prone to having some input errors. Also, it states in the description in the project brief that the numbers are estimates, so there is most likely some extrapolation within the set.

**Descriptive Analysis:** At first, the number of individuals that were in the 85+ age category were compared to the number of deaths in a state due to the flu. The correlation coefficient between these two variables was .9487 (strong). This conclusion reveals that a state with a higher population of 85+ means that it will be more susceptible to flu death.

After further consideration, it was decided that comparing the number of deaths from the flu in the 65+ age category to the number of deaths in the 0-64 Age category was best for standard statistical hypothesis testing. The average amount of deaths in each state from the years of 2009-2017 among the 65+ category was 898 and the average deaths for the same years for the 0-64 age group was 537. So, for a smaller spread of ages the 65+ age group accounted for 63% of the deaths. Furthermore, the standard deviation of the 65+ age group deaths is 973 while the standard deviation of deaths in the 0-67 age group is 119.

**Results and insight:** A student T-test was chosen because this is a relatively large sample size for analysis. The test concluded a 99.99% confidence that deaths among the 65+ age group are not equal to deaths among the 0-64 age group. Thus, we can conclude that the 65+ age group is more susceptible to flu related death than the 0-64 age group.

**Remaining analysis and next steps:** There will be a deeper dive to compare state flu death data over the years of 2009-2017 to identify . The flu death data for each state will help further conclude exactly how to allocate resources according to the population of 65+ age individuals residing in each state

**Appendix**

**Links to data sets**

[Influenza deaths by geography, time, age, and gender](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fcoach-courses-us.s3.amazonaws.com%2Fpublic%2Fcourses%2Fda_program%2FCDC_Influenza_Deaths_edited.xlsx&wdOrigin=BROWSELINK)

[Population by Geography](https://coach-courses-us.s3.amazonaws.com/public/courses/data-immersion/A1-A2_Influenza_Project/Census_Population_transformed_202101.csv)

**Information on the correlation coefficient calculation**

Table

Description automatically generated

Table

Description automatically generatedChart, scatter chart

Description automatically generated

In the preliminary analysis, the two variables 85+ population and deaths of state by year were investigated to see if a correlation could be found between the two. We calculated the correlation coefficient which is a number between -1 and +1 calculated to represent how linearly dependent the two variables are. After performing this calculation, the analysis found that the correlation coefficient between 85+ population and deaths of state by year was +.94 (a very strong relationship). This piece of calculated data can then conclude for us as shown above that if there is a higher population of 85+ years in a state, that state will be more susceptible to flu death. This is further supported if you look at the above scatterplot, which shows the general linear nature.

**Further Exploration**

After some further thought and data exploration, it was decided that looking at the flu deaths slightly differently would be better suited find the answers being sought. After looking into which age groups seemed to show more susceptibility to death from the flu, it was found that the 65+ age group was showing to be much higher than the rest of the population (0-64 age group). This can be seen in the below chart where the 0-64 age group flu deaths is represented by orange and the 65+ age group flu deaths is represented in blue.Chart, bar chart

Description automatically generated

Also if you look at all deaths from the flu over these years, as shown in the below pie chart, you will see that the 65+ age group accounts for 63% of the deaths from the flu from the years of 2009-2017.

Chart, pie chart

Description automatically generated

This also lead to two groups more suited to give significant and applicable results for the project at hand.

**T-Test information**

The Student T-Test was chosen because of the larger sample size and because the true variance for the entire data set is unknown. Estimated values were calculated for inputs missing. As shown below the null hypothesis chosen for this test is that there is no difference between the means of the over 65 age group versus the 0-64 age group. The alternative hypothesis is that the mean of the deaths in the over 65 group were higher than the mean of the deaths in the 0-64 group. An alpha level of .05 was chosen.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

From shown T test we reject the null hypothesis because of the result highlighted in yellow above. Using that value, we now know that if there is really no difference between the flu deaths in the 65+ group and the 0-64 group, there is a less than 1% chance that the difference in means between these groups is due to random chance. Referencing back to our alpha level, less than .01<.05 so we can reject our null hypothesis that there is no difference between the means of the over 65 age group versus the 0-64 age group.