

**2018 STAFFING STRATEGIES FOR INFLUENZA SEASON**

**Interim Report**

**August 18, 2023**

**Haley Bustle Carrer Foundry Data Immersion Project**

**PROJECT OVERVIEW & HYPOTHESIS**

**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***

Influenza is a seasonal epidemic across the United States that sees an increase in the need for care in both illness and critical illness patients, particularly for those in at risk populations (65 years and older). In certain states with larger populations, and consequently larger elder populations, will have a higher number of influenza-related deaths and therefore are predicted to require more staffing.

***Variables***

Our analysis will study the population of citizens over the age of 65 [years of age] as an independent variable, and the death rates of this same population as its depended variable. Further, we will study the relationship of these deaths to deaths of the population under 65 [years of age] to verify if our hypothesis is correct.

**DATA OVERVIEW**

***Data Set One: US Census Report***

This is an external data source collected by the U.S. Government and is provided by the U.S. Census Bureau containing information on a population at a specific time and location. The data set in question is from the 2009 to 2017 collection period and is broken down into specific categories: county, state, gender, and age. Furthermore, each age group is broken down to age categories, starting at under 5 years and moving upwards in increments of five years (5 to 14 years, 15 to 19 years, etc.).

***Data Set Two: Influenza Deaths Report***

This is an external data source owned and provided by the CDC (Centers for Disease Control and Prevention) containing information on deaths due to influenza. This information is categorized based upon state, year, age, and deaths from 2009 through 2017. This information is further categorized into county, state, year, age group [in 10-year increments, such as 15-24, 25-34, etc.], and deaths related to each age group.

**DATA LIMITATIONS**

***Data Set One: US Census Report***

**Collection:** As indicated from the US Census website, the information and data collected by them is collected both through surveys and administrative data ([Source](https://www.census.gov/about/what/admin-data.html#:~:text=The%20Census%20Bureau%20uses%20data%20from%20a%20variety%20of%20sources.&text=Some%20data%20are%20collected%20from,additional%20data%20from%20other%20sources.)) and is a trusted government resource. However, this collection only happens even ten years ([Source](https://www.census.gov/programs-surveys/censuses.html#:~:text=The%20U.S.%20census%20counts%20every,takes%20place%20every%2010%20years.&text=Learn%20about%20all%20the%20decennials%20from%201790%20to%20present.)), which can cause massive discrepancies in a data set. In addition to a time lag, this data has further room for error, as it is collected both through survey (where human error can occur) and administratively (where computer error can occur). Surveys are also at the mercy of people themselves, as not everyone will fill out or respond to them, which can also skew data collection for one or more populations.

**Contents:** This data set includes population data from all fifty states, District of Columbia, and Puerto Rico for gender, age categories, and year data was collected. However, this data set is from 2009 – 2017 and the most recent census was in 2020, which will cause a data discrepancy.

**Missing Data:** As a raw data set, this report contains missing data, duplicate entries, mis-represented counties & states (for example, M?xico instead of México), and decimal points in population counts. It is also important to note that this is a survey of a population and may not be completely accurate, accounting for non-participation or human error. This information must be used with the mindset of an approximate population count for each state, year, and age group.

***Data Set 2: Influenza Deaths Report***

**Collection:** This data set is provided & owned by the CDC (Centers for Disease Control and Prevention) which is a trusted government source. This information is collected with a variety of usage data collection methods, such as: hundreds of public health and clinical laboratories reporting influenza tests/results to W.H.O. Collaborating Laboratories Systems and NREVSS (National Respiratory and Enteric Virus Surveillance System and outpatient facilities report to ILINet (U.S. Outpatient Influenza-like Illness Surveillance Network) ([Source](https://www.cdc.gov/flu/weekly/overview.htm)). This is a collection system that may contain both human and computer errors (such as duplicate or error entries) since the CDC is analyzing so many variables regarding positive results.

**Contents:** This data set includes death reports for each state in a specified month & year in age groups of 10 years (for example, 15-24, 25-34, etc.) for the years 2009 to 2017.

**Missing Data:** As a raw data set, this report contains missing data points and death suppression, which can skew results or create a data bias. As noted in the CDC’s website, any death count under nine [9] is suppressed, and any death count under “20” is considered unreliable ([Source](https://wonder.cdc.gov/wonder/help/ucd.html)). This will need to be accounted for when conducting our analysis and testing our hypothesis.

**DESCRIPTIVE ANALYSIS**

Our in-depth descriptive analysis aims to show the trend of the at-risk population [65+ years of age] versus the population as a whole to better understand if this group should be our target group to provide additional staffing for in the upcoming flu season. To do this first we can study the relationship between total population size versus deaths in this age group.

By normalizing the data, we can study and determine the relationship, if any, between our independent variable (population of 65+ years) and dependent variable (deaths of 65+ years), as shown in the table below.

|  |  |  |
| --- | --- | --- |
|  | Variable 1, Independent | Variable 2, Dependent |
| Dataset Name | Total Population 65+ Years | Influenza Related Deaths 65+ Years |
| Normal Distribution? | Yes | Yes |
| Variance | 7.73132E+11 | 948302.5304 |
| Standard Deviate | 879279.0969 | 973.8082616 |
| Mean | 801596.1409 | 873.2029915 |
| Number of Outliers | 34 | 20 |
| Outlier Percentage | 7% | 4% |

Finally, we can study the correlation coefficient, which was determined to be 0.937744049 indicating a strong positive linear correlation. This suggests that there is correlation between the number of deaths reported in the Influenza Deaths Data Sheet and the US Census Report Data sheet; as one would increase, so would the other ([Source](https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php)).

**RESULTS AND INSIGHTS**

***Null Hypothesis***: People **under** the age of 65 years have a higher death toll from influenza that those over 65 years of age.

***Alternative Hypothesis****:* People **over** 65 years have a higher death toll from influenza than those under 65 years of age.

|  |
| --- |
| t-Test: Two-Sample Assuming Unequal Variances |

|  |  |  |
| --- | --- | --- |
|  | *Deaths Under 65+ Years* | *Deaths Over 65+ years* |
| Mean | 480.5576923 | 873.2029915 |
| Variance | 19816.94527 | 948302.5304 |
| Observations | 468 | 468 |
| Hypothesized Mean Difference | 0 |  |
| df | 487 |  |
| t Stat | -8.632942542 |  |
| P(T<=t) one-tail | 4.27806E-17 |  |

The null hypothesis states that people under the age of 65 years will have a higher death toll from influenza than those over 65 years, while our alternative hypothesis claims the opposite. Using a one-tail t-test and a 0.05 alpha rate, we can determine if the null should be accepted or rejected. In this case, our p-value was indicated at 4.27806E-17, which is significantly smaller than our alpha value (or significance value of 0.05). This means we can conclude with 95% confidence that the null hypothesis can be rejected and that the alternative hypothesis to be true. Our null hypothesis is that people **over** 65 years have a higher death toll from influenza than those **under** 65 years of age.

**REMAINING ANALYSIS AND NEXT STEPS**

With the information gathered from our descriptive and statistical analysis from these reports gathered between 2009 and 2017, we can comfortably conclude that populations with a larger population, and therefore larger elder population, will have a higher death toll and require more staffing. With consideration to funding and the limited number of staff on hand (nurses, physician assistants, and doctors), the following steps are recommended for consideration before final project presentation:

* Use predictive analysis to anticipate future populations of elderly to accommodate staffing needs.
* Prioritize states with the highest population & death count to appropriately assign existing staff members.
* Create simplified visualizations and reports for current staff to be distributed prior to the influenza season.
* Determine when, if any, the influenza season is most active and when staff distribution is most critical to incur the least number of deaths amongst patients [of all ages].

**APPENDIX**

Project Management Report, [Source](file:///C:\Users\boudl\OneDrive\Documents\Data%20Analysis%20Bootcamp\Achievement%201\Bustle%20A1-3.pdf)

Data Sheet 1, U.S. Census Report, [Excel File](file:///C:\Users\boudl\OneDrive\Documents\Data%20Analysis%20Bootcamp\Achievement%201\1.5%20CDC%20Influenza%20Deaths%20BASE%20CLEAN.xlsx)

Data Sheet 2, Influenza Deaths Report, [Excel File](file:///C:\Users\boudl\OneDrive\Documents\Data%20Analysis%20Bootcamp\Achievement%201\1.5%20CDC%20Influenza%20Deaths%20BASE%20CLEAN.xlsx)

Integrated Data Sheet, including statistical & descriptive analysis, [Excel File](file:///C:\Users\boudl\OneDrive\Documents\Data%20Analysis%20Bootcamp\Achievement%201\Bustle%201.10%20Final%20Report,%20Excel%20Report.xlsx)