Interactive Modeling of Human Immunodeficiency Virus in the United States

Tyler Audino, Tikiri Ekanayake, Arlen Gyden, and Quinn Yuan

University of Florida

MAD 2502: Intro to Computational Math

Dr. Jason Harrington

December 8th, 2022

Introduction

The purpose of this project was to create a tool that would help to identify at-risk communities in need of more health and wellness education and resources, specifically in regards to Human Immunodeficiency Virus, also known as HIV. This would be achieved through the analysis of HIV data records obtained from AIDSVu, another interactive mapping tool that helps to visualize HIV impacts across the nation (Emory University's Rollins School of Public Health, 2022).

With this data, we hoped to make an interactive mapping tool, along with statistical analysis tools similar to what was found on the AIDSVu website; however, we find it important to note that the idea of making an interactive choropleth map formulated before seeing the one provided by AIDSVu, and no attempts to find the exact code AIDSVu used for their map was made. Our map is completely original code and has the ability to display different data compared to the AIDSVu map.

Additionally, although our program is specifically designed to work with the HIV dataset from AIDSVu, our program could be transferred for use for similar geographically-focused datasets of other diseases by changing the strings and file names used in the code to align with different datasets. We chose to focus on HIV because of the ease of availability of public geographically-focused data through AIDSVu.

This tool was created to aid individuals who may not be scientifically literate enough to dissect datasets that have been given directly to them. We aimed to create a program which makes data visualization and understanding more accessible, especially to those who do not have a scientific background. Our tool hopes to make it easier for the general public to interpret this data and come to their own conclusions.

Our final product consists of a custom choropleth map generator, a T-Test calculator, along with an ANOVA calculator. All of these tools are capable of taking in user input to compare data between races, genders, transmission types, and ages, depending on which tool was selected and the availability of data.

HIV/AIDS

Human immunodeficiency virus, more commonly known as HIV, is a sexually transmitted infection (STI), that attacks the body's immune system, therefore eroding away the body's defense against other diseases and viruses. Eventually, the infected's condition worsens and reaches the final stage: acquired immunodeficiency syndrome, also known as AIDS. There is no cure for HIV yet and the human body is not able to get rid of the infection, which means that once someone has been infected, it will be there for the remainder of their life. However, this does not mean that there is no treatment for it, and the one that does exist is quite effective in reducing the amount or viral load of HIV in the blood (What are HIV and AIDS?, 2022).

After being infected, there are three different stages of HIV, with the last being AIDS. The first stage is acute HIV infection, starting at around 2 to 4 weeks after being infected. Symptoms during this stage include fever, headache, rash, and general flu-like symptoms; however, it has been noted that in some cases the infected have experienced no symptoms. This is the stage where the viral load in the blood is the highest. Essentially, this means that the amount of HIV RNA copies that are reported per millimeter of blood is the highest then. This also means that it is significantly easier to transmit the infection to someone else. On the brighter end, if the infected were to start treatment now, they would experience the most benefits from it.

The second stage is chronic HIV infection, which is also sometimes called the asymptomatic HIV infection or clinical latency. At this stage, the infected may not experience

any symptoms related to HIV, which often leads people to believe that the 'fever' they previously had was only a fever, and not something else. Receiving treatment during this time can significantly decrease the chances of the infection developing into AIDS and can reduce the risk of transmission to essentially none. Without receiving treatment, the infection can advance into AIDS within 10 years or more, but in some cases it may be quicker.

The final stage of HIV is acquired immunodeficiency syndrome or AIDS. At this stage, the body's immune system has taken significant damage and can no longer hold itself against infections that could be potentially severe to them in this condition. The infected viral load is incredibly high, which means they are able to transmit it easily to others. The average person who does not receive treatment at this stage typically lives for around three more years (U.S. Department of Health and Human Services, 2021).

HIV can be transmitted in various ways, most of them involving exchanging of body fluids from infected people into the bloodstream of HIV-negative people, such as blood, semen, rectal fluids, vaginal fluids, and breast milk. HIV will immediately lose its activity and infectivity when exposed to the air, and will die soon. In addition, HIV will also be killed immediately in alcohol, disinfectant, high temperature, and in the lack of a host, but will survive for hours in fresh blood. Therefore, normal physical contact with an HIV-infected person is not easily contagious. HIV infections are not spread by routine human contact such as kissing, hugging, handshakes, or sharing of personal belongings, food, or liquids. It is difficult to transmit HIV because the amount of HIV in tears, sweat, saliva, and urine from daily contact is very small, the concentration is really low, and when the amount does not exceed a particular scale, the immune system of the human body can play a role and kill it. AIDS can spread through injecting drugs, sexual activity, and mother-to-child transmission since only blood, semen,

vaginal fluid, breast milk, and wound exudate with a high viral concentration can cause transmission. It should be noted that individuals with HIV who are virally suppressed and using antiretroviral medications do not transfer the virus to their sexual partners (What are HIV and AIDS?, 2022).

HIV transmission is primarily tied to human social conduct, and it is totally possible to prevent and minimize infection by regulating human social behavior due to the unique requirements of HIV on its transmission pathway and the fragile nature of HIV. As mentioned above, HIV is transmitted mainly through infected blood, semen or vaginal secretions entering the body and causing infection. Therefore, the primary approach to prevent HIV is to effectively avoid body fluid interchange in daily life with people who have HIV. The best strategies to prevent contracting HIV are to use condoms appropriately, refrain from sharing needles with others, and avoid coming into contact with anybody else's bodily fluids or blood. Moreover, HIV prevention medicines such as pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) can also be effective in preventing people from contracting HIV (What are HIV and AIDS?, 2022).

Data Frame Organization

To begin our project we first had to research and identify possible public datasets that could be used to model HIV statistics in the United States. After some time, we discovered the AIDSVu website which is a public database that contains numerous public health resources that track and monitor the presence of HIV nationally. Originally, we planned to visualize HIV on a local level by utilizing the new diagnoses data for each year in each county. However, we soon realized that these spreadsheets were riddled with missing or incomplete data for many counties nationwide. Ultimately, the datasets we chose to model were new diagnoses by state from 2008

to 2020, as these were widely available and had relatively consistent data by state. These datasets were all excel files separated by year, so we concatenated them into one excel file that could be converted into a pandas dataframe for future manipulation.

Although there are much less data discrepancies in the HIV statistics by state, we still needed a method to remove unusable data types from the spreadsheet to generate models without error. Due to the set up of our program, we could not simply clean the data by removing states (rows) with unusable data types because not all of their columns were expendable. Therefore, we created the dataskip function to iterate through the columns of data being used based on the user's parameters. This function effectively removes all unusable data types from the user-specified columns and stores them in a list for modeling without permanently altering the original dataframe.

For the ANOVA table, the data frame had to be manipulated a different way in order to be read properly by the code. The ANOVA code correlates two variables, by observing the cases where more than one trait was observed and calculating if there is enough chance for this correlation to be random. For example, if the user wanted to see whether gender and race impacted the number of HIV cases, then the data inputted would not be data for only HIV cases where only gender was identified and data for HIV cases where only race was identified, but instead data for HIV cases where race and gender were specified.

The data found already provided cases for the interaction between each of the four variables observed: race, sex, age, and transmission type (e.g. data for White females, Black people aged 13-24, etc.). What was left to do was to reorganize this into six different columns of cases, for each combination of two for the variables observed, along with two columns of labels for each case column describing those cases, for a total of 18 new columns. It should be noted

however, that for many transmission and other variable cases, negative numbers were recorded for cases as information for that category was unavailable. So, those instances were removed from the data frame.

Choropleth Map

Choropleth maps—also known as heat maps—are maps that use color gradients for data visualization. They are commonly used to display how a variable varies across a geographic region, with each color or shade indicating the intensity or prevalence of the variable. We chose to develop a function to create custom choropleth maps because it is a commonly used model for health data, and is visually easy to understand from a user standpoint.

Our main function is the mapselect function, which first gathers the data parameters the user would like for their map, and then develops the interactive map based on these inputted data parameters. The user can choose up to two parameters at a time, in the categories of sex, race, age, and transmission type, and then the user decides if they would like to look at the data as cases or rates. An example of data that can be visualized is “New Diagnoses Female and Asian Rate” which has the parameters of ‘Female’ for sex and ‘Asian’ for race. If no parameters are chosen, all data is displayed. Once the user selects their parameters, the map plot is created and displays the data with a map for each year of available data, from 2008-2020, and this can be animated or paused to look at one specific year. Additionally, the user can move their cursor over a state of interest and get the exact data value per state.

To code the map, the package Plotly.express was imported as px, and the Pandas package as pd. A class called maps was created, and uses the px.choropleth function to generate the maps. This px.choropleth function requires input of a Pandas data frame, a string that corresponds to the name of a column header within that dataframe, and a geoJSON or JSON file that encodes

the geographic structure to be mapped. The maps generated within this program all utilize data from the United States, so the JSON file used encodes the geography of the United States including state borders (Gallo, 2020).

The main function to generate these choropleth maps is the mapselect function. Mapselect calls the function branchmenu, which calls the selection function, which then calls upon sexmenu, racemenu, agemenu, or transmissionmenu if needed. Selection asks the user what parameter categories—if any—they would like to use for their map through input functions. If a user wants to define a category of parameter, selection will call the specific menu function for that category, such as sexmenu or agemenu, to identify the specific parameter the user wants, such as Male or Female. These specific menus use input statements to gather the parameters, and while loops to ensure the user enters a correct string. Once the specific parameters have been identified, selectionmenu returns a list of the parameters the user wants.

When branchmenu calls selection, the list returned from selection is used to find the corresponding header name within the dataframe named data, through if statements and lists. These if statements search for the key words of each parameter as strings, such as “Black” in all of the headers of the dataframe. When searching for just one parameter, any headers with the word “and” are omitted to exclude data with two parameters, such as “New Diagnoses IDU and White Cases.” When two parameters are being searched for, all strings including the first parameter are found with an if statement and saved to a list, and then another if statement searches for the second parameter in this list. Finally, branchmenu identifies one specific string which corresponds to the user’s parameters, and returns this string.

To account for errors, when a user provides overly strict parameters, in which there is no data available and no string is found, branchmenu will ask the user to select new parameters.

Additionally, when a user provides parameters which correspond to multiple data types, the function will ask the user to pick between the two possible dataframe headers. A common occurrence of this is when the user selects any one parameter of sex, race, or age, along with IDU transmission and cases, which will often return two possibilities, an example of which is: 'New Diagnoses IDU and Age 35-44 Cases' or 'New Diagnoses MSM/IDU and Age 35-44 Cases'. A list of these possibilities will be returned and the user will be asked to select one.

Finally, after branchmenu has returned a string of the header name, the function mapselect will first use this string to clean the data through the dataskip function, and will also call the range_finder function to define a consistent range for the map scale based on the data. The dataskip function cleans the data selected by iterating through the data and skipping over unusable data points, which were indicated in the dataset by negative numbers (-1, -4, or -9) or as "N/A". Dataskip then returns the cleaned data as a new dataframe. Mapselect will then use the cleaned dataframe, as well as the range provided from range_finder to generate a new object using the maps class, and then will display this map object to the user.

The final product from running the mapselect function is an interactive choropleth map, which displays data from 2008-2020, based on the parameters provided from the user, as previously described.

Data Analysis

In addition to the display of the choropleth map, the program also provides different statistical testing methods, aiming to provide users with more comprehensive and detailed descriptive data analysis. Descriptive statistical analysis makes a statistical description of the relevant data of all variables in the sample population, mainly including measures of central

tendency, variability, distribution, and some statistical graphics (Hayes, n.d.). In this program, there are two statistical tests, ANOVA and T-test, used to analyze data with statistical aids.

One of the tools made for this project was an ANOVA calculator which takes in input for what variables to test for correlation. ANOVA stands for analysis of variance, and it is a test "used to compare differences of means among more than two groups," (Edanz-Learning-Team, 2021). These differences are then checked to see if it is likely to have been caused by chance or not. Should the P-value be less than 0.05, the null hypothesis stating that the means are all equal can be rejected, and it can be stated there is indeed correlation between the populations.

For the ANOVA tool, say the user was to input 'gender_age.' This would mean that the user wishes to see if being a certain gender and age correlate with the amount of HIV cases in the nation during that year or all the years that data is available for. If the P-Value calculated for the interaction between gender and age (represented in the program by: C(gender_gender_age):C(age_gender_age)) is smaller than 0.05, than the user can interpret this to mean that there is an effect between the two variables on HIV cases across the United States.

The code for this tool begins with importing pandas, statsmodels.api, and statsmodels.formula.api. The pandas package was used for reading in the data frame made for this tool. Statsmodels.api was used to gain access to the actual ANOVA function calculator which outputs a table with the degrees of freedom, sums squared, means squared, F-Value, and the P-Value (PR(>F)), the latter of which is used for the interpretation of the statistics here. Statsmodels.formula.api was used to arrange the data in an appropriate manner that could be properly read by the anova function code. The packages and code reading the data frame are kept in a separate cell as once that has happened, it does not need to happen again.

Next, a series of questions/inputs are given to the user in order to identify what column of cases in the data frame would be used. The first asks which two variables to compare and type them in as presented. The next asks if the user would like to observe this data within a certain year, or use the data set as a whole. Should the user not give a valid input for either of those, then a while loop is initiated until the user finally types in a valid input.

With a valid year selection, the code then narrows down the dataset and stores it for later as a separate variable: observed_year. Using the user input for variable comparisons, the string is manipulated multiple times in order to match the headers for their columns in the data frame. For example, if the user chose gender_age, then the column of cases it needs to be matched to is gender_age_cases. The columns of cases for this data frame are what is being used as the dependent variable for the function. The other two columns matched (first_iv and second_iv for the independent variables) contain information for each of those cases regarding what its characteristics are and which specific part of it. The first independent variable noted for this example is gender_gender_age, which refers to the column of the same heading in the data frame. That column contains the genders documented for all the cases containing information for only gender and age. The second independent variable is age_gender_age, whose column contains the age ranges for the cases where gender and age were documented together.

The next part of the code creates an F-string that is later used in the first argument of the ols.fit() function. This string is formatting the dependent variables and independent variables in the needed way, along with creating the formatting for the interaction between the two variables. This string is then inputted into ols.fit() along with the section/year within the data that was specified by the user and is assigned to the variable ‘model.’ ‘Model’ is then inputted into the ANOVA function accessed by the statsmodels.api package, and the type of ANOVA test is

specified (type two in this case, as there are two independent variables). The ANOVA table produced is then stored as variable ‘result’ and printed. Attempts were made to try and isolate just the P-value from this table; however, the appropriate code for doing this was not found. The code used for creating the ANOVA table is an adapted version from the one found on GeeksforGeeks (2022) explaining how to perform a two-way ANOVA.

The other statistical modeling tool made in the program is the T-test, also known as the Student’s T test. T-test is a parametric statistical test often used in hypothesis testing to compare the difference between means of two groups, analyzing whether two groups are different from one another (Bevans, 2020). The premise of the T-test is that the sample is required to obey a normal distribution or an approximate normal distribution. Otherwise, some transformations (logarithm, square root, reciprocal, etc.) and some non-parametric test methods can be used to try to convert it into data that obeys a normal distribution. However, when the sample size is greater than 30, the set of data can be considered to be approximately normally distributed (Bevans, 2020). The result from the T-test function in the project generates the relationship between male and female new diagnoses cases in different states, indicating if there is a significant difference between them.

The use of T-test in Python requires some open-source package installations, including pandas, scipy.stats, and matplotlib.pyplot since the T_test function also creates boxplots based on user inputs. The T_test() function begins by requesting the user's input regarding the state in which the male and female new diagnostic cases should be compared. After the user selects a state, the program organizes the data in accordance with the particular state that was chosen by the user and ensures the user’s input is accurate with the use of a while function. Through simple

data organization, the chosen state's new diagnoses male and female cases are assigned as sample1 and sample2 for the operation of T-test.

The P-value displayed from running the ttest_ind() method quantifies the probability of an outcome that is more extreme than the observed sample observations, when the null hypothesis is true. If the P-value is small, the probability of the occurrence of the null hypothesis is very small. In this case, the null hypothesis would be there is not a statistically significant difference between sample1 and sample2 from the chosen state. Therefore, according to the principle of probability, it is reasonable to reject the null hypothesis, which means there is a statistical difference between samples (Hayes, n.d.). In general, the smaller the P-value, the more significant the result. In order to compare the p-value with the threshold (α=0.05), the program extracts the P-value from the ttest_ind() result display and prints out the analysis of the t-test to inform the user if the male and female new diagnoses cases in the chosen state are different from each other.

Furthermore, the T_test() function also includes the use of boxplot to better illustrate the relationship between male and female cases in one state. Box plots are a standard way to describe the distribution of data through 5 numeral indicators. These 5 numbers include: minimum value, the first quantile, median, the third quantile, the maximum value. Additionally, boxplot is able to clearly display the information of outliers, and let users understand the nature of symmetry, relationship, and kurtosis of the data samples.

The display of boxplot effectively helps users understand the differences between male and female new diagnosis cases in the selected states. The upper and lower bounds of the box are the upper and lower quartiles of the data respectively. Therefore, the size of the box reflects the

volatility of the data to some extent. The orange line in the middle of the box is the median of the data, which represents the average level of the sample data.

Conclusion

Over the course of developing our program we uncovered a few considerable limitations of our datasets from AIDSVu. Firstly, because the HIV statistics were not collected on a national level, it is apparent that there are inconsistencies between state data collection methods and reporting processes. Additionally, as cases are not reported per capita based on each state's overall population, the data when modeled appears to consistently visualize places like California as outliers rather than populous states. This is important to note as the goal of our modeling is to identify areas in need of HIV education in the United States. As for the ANOVA testing, the datasets did not include enough information to compare the statistical differences between individual parameters (i.e compare differences between the age groups 13-24 and 25-34, or compare differences specifically between females and Asian).

One of the major roadblocks that we ran into while scripting our program was ensuring that the branchmenu, selection, and mapselect functions worked seamlessly. The debugging process for these interactions was one of the larger efforts that took up a considerable amount of time. Ultimately, it was through the dataskip function and maps class that we were able to achieve this goal. Since last presenting in class, we have fixed the errors in the user input menu for transmission, specifically the IDU selection. We added code to branchmenu that identifies when there is no data returned for the parameters provided, and asks the user to choose new parameters. Further, a new function was created to ensure that the scales of the maps are consistent throughout all years of data, called range_finder. As for the T-test function, we added F-strings to ensure that each graph is properly titled and displayed with labeled axes.

In essence, accomplishing the goal of modeling HIV in the United States took immense teamwork and communication on behalf of the group as a whole. To detail, below is a brief description of each group member's primary area of work; note that this list is not entirely reflective of each member's efforts as each member participated in the development and debugging of all sections.

Coding Credits:

Data Concatenation and Cleaning - Arlen

Choropleth Map - Tyler/Arlen

ANOVA - Tikiri

T-Test - Quinn

Paper Credits:

Introduction: Tikiri/Tyler

HIV/AIDS - Quinn/Tikiri

Data Frame Organization - Arlen/Tikiri

Choropleth Map - Tyler/Arlen

Data Analysis - Quinn/Tikiri

Conclusion - Arlen

References

Bevans, R. (2020, January 31). *An introduction to T-tests | definitions, formulas and examples*. Scribbr. Retrieved December 12, 2022, from https://www.scribbr.com/statistics/t-test/

Box plot in python using Matplotlib. (n.d.). Geeksforgeeks. Retrieved December 12, 2022, from https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/#:~:text=A%20Box%20Plot%20is%20also,median%2C%20third%20quartile%20and%20maximum.

Centers for Disease Control and Prevention. (2020, October 28). *HIV transmission*. Centers for Disease Control and Prevention. Retrieved December 12, 2022, from https://www.cdc.gov/hiv/basics/transmission.html

Edanz-Learning-Team. (2021, December 30). *ANOVA explained: How to compare differences of means*. Edanz Learning Lab. Retrieved December 11, 2022, from https://learning.edanz.com/anova-explained/

Emory University's Rollins School of Public Health. (2022, November 3). Tools & Resources. AIDSVu. Retrieved December 12, 2022, from https://aidsvu.org/resources/#/datasets

Gallo, K. (2020, August 18). *USA states GeoJSON*. Kaggle. Retrieved November 10, 2022, from https://www.kaggle.com/datasets/pompelmo/usa-states-geojson?resource=download

Hayes, A. (n.d.). *Descriptive statistics: Definition, overview, types, examples*. Investopedia. Retrieved December 12, 2022, from https://www.investopedia.com/terms/d/descriptive_statistics.asp

Hayes, A. (n.d.). *T-Test: What it is with multiple formulas and when to use them*. Investopedia. Retrieved December 12, 2022, from https://www.investopedia.com/terms/t/t-test.asp

How to merge multiple Excel files into a single file with python. GeeksforGeeks. (2022, March 7). Retrieved December 12, 2022, from https://www.geeksforgeeks.org/how-to-merge-multiple-excel-files-into-a-single-files-with-python/

How to perform a two-way ANOVA in python. GeeksforGeeks. (2022, February 28). Retrieved December 11, 2022, from https://www.geeksforgeeks.org/how-to-perform-a-two-way-anova-in-python/

spicy.stats.ttest_ind. (n.d.). SciPy documentation. Retrieved December 12, 2022, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

U.S. Department of Health and Human Services. (2021, August 20). *The stages of HIV infection*. National Institutes of Health. Retrieved December 12, 2022, from https://hivinfo.nih.gov/understanding-hiv/fact-sheets/stages-hiv-infection

What are HIV and AIDS? HIV.gov. (2022, June 15). Retrieved December 12, 2022, from https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids

Note: references used for both the paper and coding