Final Report on the Traineeship Program 2023

On

“ANALYZE DEATH AGE DIFFERENCE OF RIGHT HANDERS WITH LEFT HANDERS”

MED TOUR EASY

28 July 2023

ACKNOWLDEGMENTS:

I would also like to thank the team of MedTourEasy for providing me with an opportunity to be a part of Data Analytics Traineeship. The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Analytics and all the steps to be done while working on a project. I am very obliged for having a chance to interact with so many professionals who guided me throughout the traineeship project and made it a great learning curve for me. The training covered basics of Data Analytics to developing an entire project. Time allotted for learning and Executing each module was very sufficient. Thank you

ABSTRACT:

Handedness refers to the preferred or dominant hand an individual use while performing manual tasks. There are three type of handedness namely right handedness, left handedness, ambidexterity. Right handedness is the preference of using right hand for manual tasks. Approximately 90 percent of the population are right handed. Left handedness is the preference of using left hand for manual tasks. Approximately 10 percent of the population are right handed. Ambidexterity is a rare occurrence of using both hands. The individuals don’t have a strong preference for both hands. Does these handedness impact the death age of an individual. One popular claim is that left handed individuals die at early age than right handed individuals. This claim may or may not be true. Based on the average age at death data and the individuals handedness our objective is to refute the claim of early death for left handers. The goal is to utilize Bayesian statistics to calculate the probability of an individual being a specific age at death, considering whether they are reported as left-handed or right-handed. By incorporating the reported handedness information into the analysis, the project aims to determine if there are any differences in the age distribution of deaths between left-handers and right-handers.

I. INTRODUCTION

1.1 ABOUT THE COMPANY:

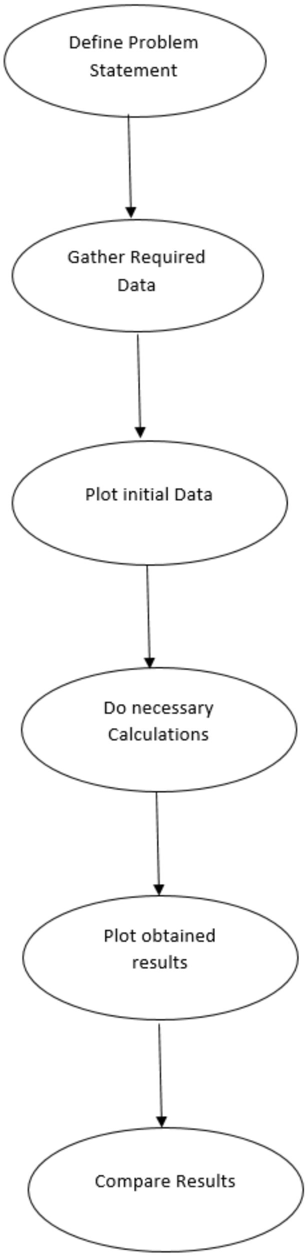
MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

1.2 ABOUT THE PROJECT:

The project Analyze Death Age Difference of Right Handers with Left Handers project delves into the relationship between handedness and age at death. The study aims to investigate the widely debated claim that left-handed individuals are more likely to experience early death. By analyzing age distribution data and handedness rates, the project seeks to determine if there is a significant difference in the age at death between left-handed and right-handed individuals, and whether this difference can be attributed to changing rates of left-handedness over time. Using pandas and Bayesian statistics, the project calculates the probabilities of being a particular age at death given left-handedness or right-handedness. By incorporating the reported handedness information, the analysis allows for a comparison of age distributions between the two groups. The study reveals that left-handed individuals are not likely to die at a younger age solely due to their handedness, contrary to some earlier studies claims. The impact of changing rates of left-handedness over time is explored, highlighting how the representation of left-handed individuals in different age groups has evolved. While the project acknowledges certain limitations, such as using death distribution data from a different year and the extrapolation of left-handedness rates, it underscores the significance of considering population dynamics and historical context in interpreting the age gap.

II. METHODOLOGY:

2.1 FLOW OF PROJECT:



2.2 DATA USED:

The data is obtained from secondary data sources. The death distribution data for United States for 1999 was obtained from Centers for Disease Control and Prevention. The death distribution data for the year 1999 contains information about the total number of deaths, number of dead male , number of dead female in each age with ages ranging from 0 to 124. The handedness preference data was obtained from National Library of Medicine. This data was published by A N Gilbert and C J Wysocki in 1992. The data is present as a csv file and is loaded into pandas data frame. The data is read using red_csv() function.

2.3 LANGUAGES USED:

The analysis and visualization id done in Jupyter Notebook .It is a popular web-based interactive computing environment for creating and sharing documents that contain live code, visualizations, explanatory text, and more. Jupyter Notebook supports various programming languages, including Python, R, Julia, etc. It allows to write and execute code in cells, making it ideal for data analysis, prototyping, and sharing code and visualizations. The primary language used for this handedness analysis is python. Various python inbuilt libraries are used to examine the data also for visualizing the data namely

pandas: It is a powerful data manipulation and analysis library in Python. It provides data structures and functions for efficiently working with structured data, such as CSV files, Excel sheets, and SQL tables. Importing pandas as pd will import the pandas library and assigns it the alias pd, making it accessible as pd.

matplotlib.pyplot: It is a plotting library in Python that provides a MATLAB-like interface for creating a variety of plots and visualizations. It is typically imported as import matplotlib.pyplot as plt to provide a shorthand alias. The plt.subplots() function and the ax.plot(), ax.legend(), ax.set_xlabel(), and ax.set_ylabel() methods are part of this library and are used for creating plots and customizing their appearance.

plt.subplots: It is a function in the matplotlib.pyplot library that creates a figure and one or more subplots (axes) within that figure. It returns both the figure object (fig) and the axes object (ax).

ax.plot: It is a method of the axes object (ax) that is returned by plt.subplots(). It is used to plot data on the specified axes. It takes the x-coordinates, y-coordinates, and other optional parameters to specify the type of plot (e.g., line plot, scatter plot).

ax.legend: It is a method of the axes object (ax) that adds a legend to the plot. The legend displays labels for different elements (such as lines or markers) in the plot, allowing the viewer to identify what each element represents.

ax.set_xlabel: It is a method of the axes object (ax) that sets the x-axis label for the plot. It takes a string argument representing the label text

ax.set_ylabel: It is a method of the axes object (ax) that sets the y-axis label for the plot. It also takes a string argument representing the label text.

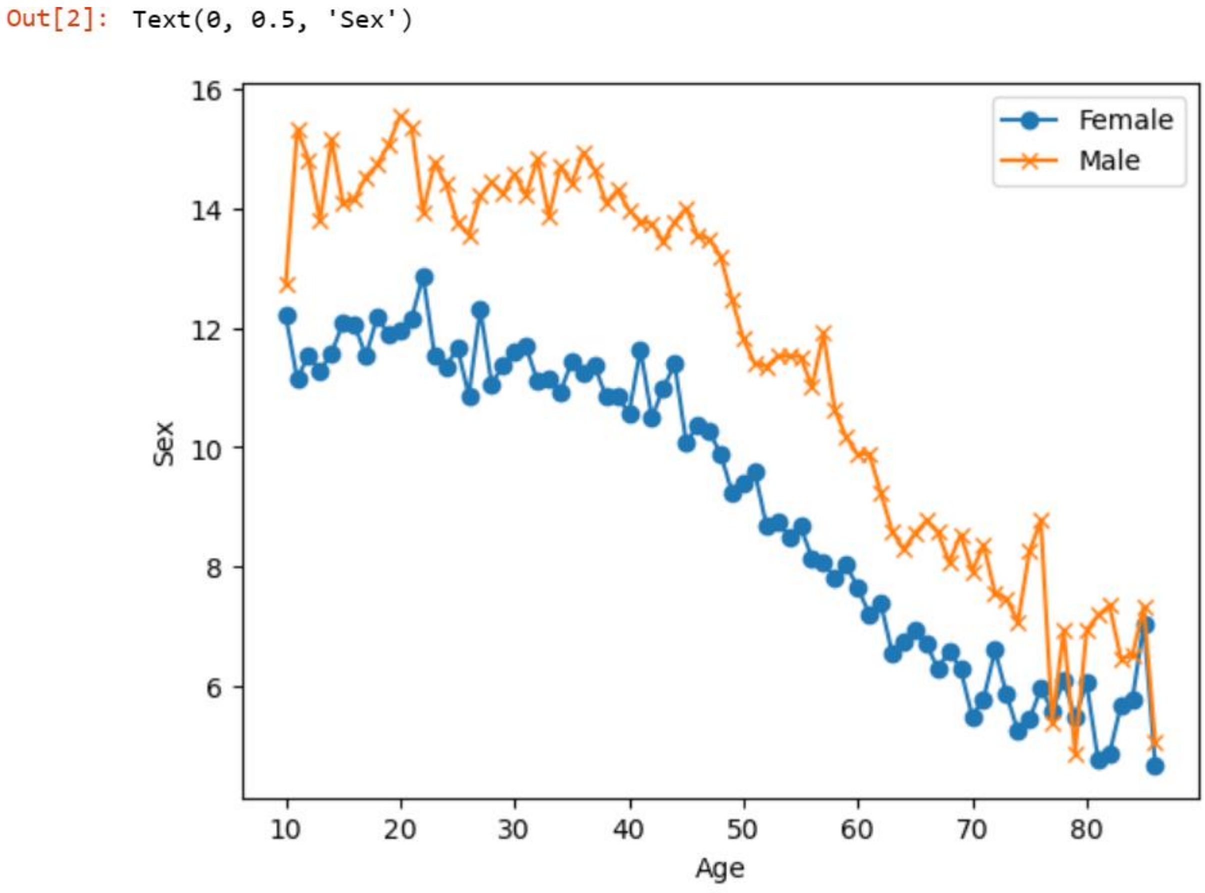
numpy: A fundamental library for numerical computing in Python. It provides multidimensional arrays, mathematical functions, and tools for array manipulation. Importing it as import numpy as np allows you to use numpy functions and classes with the shorthand alias np.

np.arange: It is a function from the NumPy library (import numpy as np) that returns an array with evenly spaced values within a specified range. It is commonly used to create arrays for indexing, looping, or plotting purposes.

III. IMPLEMENTATION :

3.1 CREATING PLOTS WITH ORIGINAL DATA :

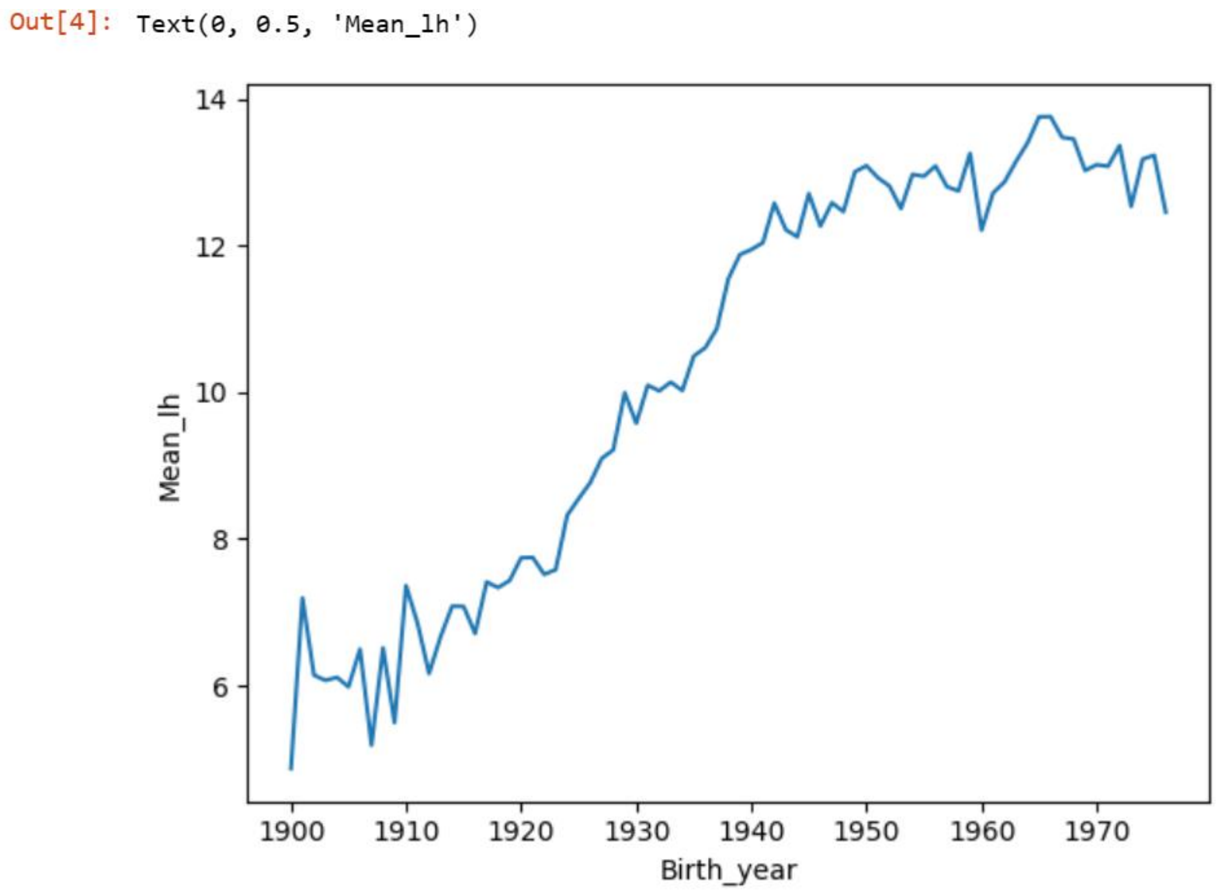
Based on the left handed male and female data and their ages a graph is plotted as Age vs Sex. The sex includes both male and female.



```
Out[2]: Text(0, 0.5, 'Sex')
```

The graph contains two lines. Each lines contains multiple points representing the age of left handed male or female. The orange plot indicates left handed males and the blue plot indicates left handed females.The CSV file is loaded into the lefthanded_data variable. Using the ax.plot the females are marked using ‘o’ and the males are marked using ‘x’ . Using ax.set_xlabel the x axis is named as Age and using ax.set_ylabel the y axis is named as Sex.The age data starts from 10 years upto 86 years .The resulting visualization will help analyze and understand the left-handedness trends based on age and gender

In the same lefthanded data a new column is formed named as Birth_year which stores birth year of each individual. The birth year is calculated from age and current year. Age subtracted from current year gives the birth year. Another column is created in the same data named as Mean_lh. This column stores the average of male and female columns for each row. The mean() function is applied along axis=1 to calculate the row wise mean of the male and female columns.



```
Out[4]: Text(0, 0.5, 'Mean_lh')
```

A graph is plotted with the Birth_year and the Mean_lh values. Birth_year along the x axis and Mean_lh along the y axis. The plot will display a line graph where each data point represents the average left-handedness rate for a particular birth year. By plotting the average left-handedness rates against birth years this visualization can help analyze any trends or patterns in left-handedness over time.

3.2 BAYES RULE:

3.2.1 EXPLANATION:

Probability of Being Left-Handed Given That You Died at a Certain Age (P(LH | A)) represents the likelihood of an individual being left-handed, given that they died at a specific age. It is calculated by considering the number of individuals who died at age A and were left-handed relative to the total number of individuals who died at that age. Probability of Dying at a Certain Age Given That You're Left-Handed (P(A | LH)) represents the likelihood of an individual dying at a specific age, given that they are left-handed. It is calculated by considering the number of left-handed individuals who died at that particular age relative to the total number of left-handed individuals in the population. Both these probabilities are not the same. The difference between these probabilities arises because they are based on different sets of data. P(A | LH) focuses on left-handed individuals and calculates the likelihood of dying at a specific age among that group. P(LH | A) focuses on individuals who died at a specific age and calculates the likelihood of being left-handed among that group. This inequality is why Bayes' theorem is used. Bayes theorem is a statement about conditional probability which allows us to update our beliefs after seeing evidence. (P(A | LH)) and P(LH | A) can be calculated as

P(A | LH) = Number of left-handed individuals who died at age A / Total number of left-handed individuals

P(LH | A) = Number of left-handed individuals who died at age A / Total number of individuals who died at age A

The Bayes theorem for P(A | LH) is

$$P(A \mid LH) = \frac{P(LH \mid A) * P(A)}{P(LH)}$$

where P(A) s the overall probability of dying at age A and P(LH) is the overall probability of being left-handed .

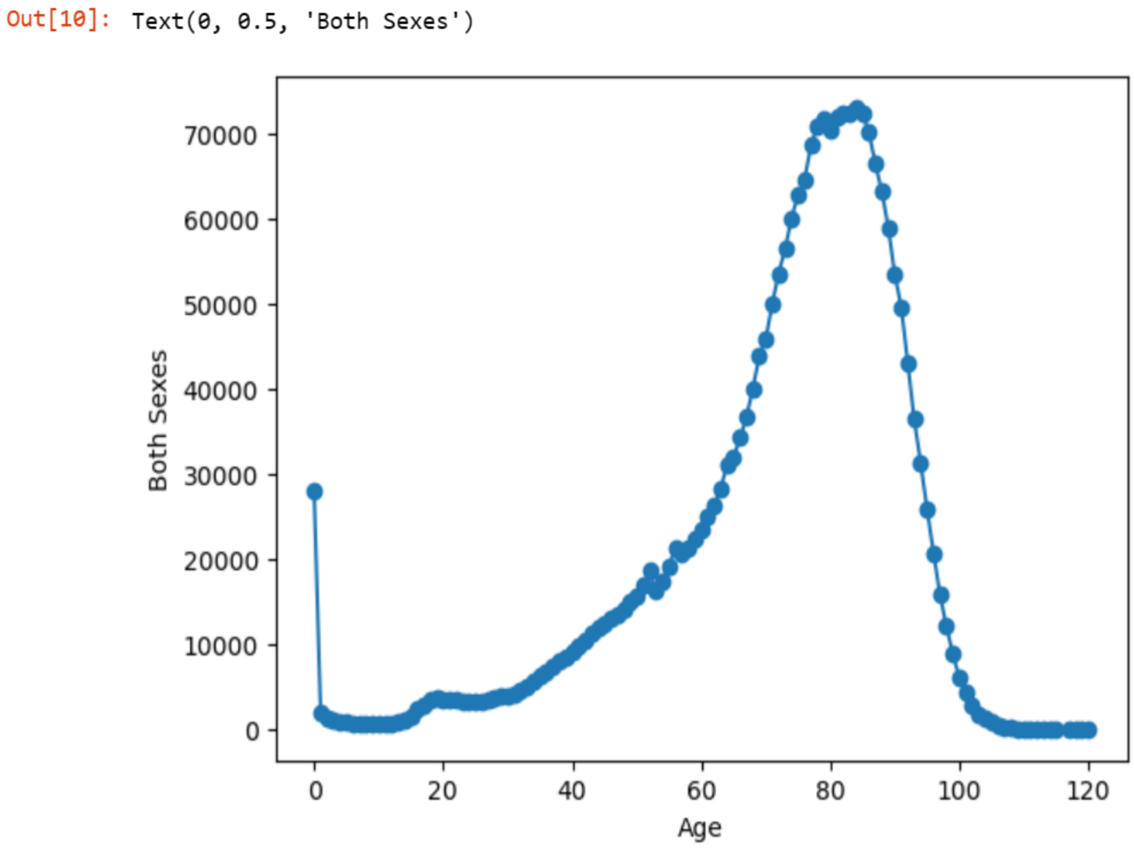
Since the original data might not cover all ages (ages that might fall outside the data range), the rates may need to be extrapolated. In this case, the rates are assumed to flatten out in the early 1900s and late 1900s. To extrapolate the rates on each end, a few points from each end of the data will be used, and the mean will be taken to estimate the rates. The number of points used for extrapolation is arbitrary, but we use 10 points since the data looks almost flat until about 1910.

3.2.2 EXECUTION:

A function named P_lh_given_A(ages_of_death, study_year = 1990) that calculates the probability of being left-handed given that subjects died in a specific study_year at ages specified in the ages_of_death is defined. The function calculates the mean of the last 10 points (early_1900s_rate) and the first 10 points (late_1900s_rate) from the 'Mean_lh' column in lefthanded_data. This is used to estimate the left-handedness rates in the early 1900s and late 1900s, respectively. The function extracts the left-handedness rates (middle_rates) from lefthanded_data for people with ages specified in ages_of_death for the study_year. It uses the 'Birth_year' column in lefthanded_data to find the rows corresponding to the ages specified in study_year - ages_of_death. The function calculates the youngest and oldest ages (youngest_age and oldest_age) based on the study_year. The youngest age is set to 10 (since the minimum age in the data is 10 years), and the oldest age is set to 86 (since the maximum age in the data is 86 years). An empty array P_return is created to store the results of the probabilities of being left-handed for the specified ages_of_death. The function calculates the probabilities of being left handed and assigns them to the corresponding elements of the P_return array based on the ages_of_death and the estimated left-handedness rates from the early 1900s, late 1900s, and middle years. The function returns the P_return array, which contains the probabilities of being left-handed for each age of death specified in the ages_of_death array.

3.3 CREATING PLOTS WITH DEATH DATA:

The death data source includes the number of people who died at different ages in the US for the year 1999. This data will be used to create a distribution of ages of death and estimate the probability of living to a specific age. The first column of the data represents the age, and the other columns represent the number of people who died at each age. The data is loaded into the variable death_distribution_data. Any NaN values present in any rows in the ‘Both Sexes’ columns will be dropped using dropna() function. Plots are created to visualize the number of people who died as a function of age. It uses the 'Age' column on the x-axis and the 'Both Sexes' column (number of deaths) on the y-axis. The ax.plot() function is used to plot the data, with the 'o' marker representing data point. The plot will help visualize how many people died at each age, giving insights into the age distribution of deaths for that year.



```
Out[10]: Text(0, 0.5, 'Both Sexes')
```

3.4 OVERALL PROBABILITY OF LEFT HANDEDNESS:

P(LH | A) is the probability of being left-handed given that an individual died at age A (calculated using the function P_lh_given_A() in the previous code block).

N(A) is the number of people who died at age A (given by the dataframe death_distribution_data).

With these data we can calculate P(LH), the probability that a person who died in a particular study year is left-handed, assuming we know nothing else about them as :

$$P(LH) = \frac{\Sigma_A P(LH \mid A) N(A)}{\Sigma_A N(A)}$$

This calculation is performed by defining a function P_lh(death_distribution_data, study_year) which takes two inputs namely death_distribution_data which is a dataframe containing the death distribution data, including the age and the number of people who died at each age, study_year indicating the year in which the subjects died. By default, it is set to 1990. p_list is calculated by multiplying the number of people who died at each age (death_distribution_data['Both Sexes']) with the conditional probability of being left-handed at that age, given by the function P_lh_given_A(death_distribution_data['Age'], study_year). This is done element-wise, so p_list will be a series with the same length as the dataframe. The sum of p_list using np.sum(p_list) is found. died in the specified study_year, the sum of p_list is divided by the total number of people who died, given by np.sum(death_distribution_data['Both Sexes']). This normalization ensures that the probabilities are relative to the total number of deceased individuals.Finally, the P_lh() function is called with the death_distribution_data dataframe to obtain the overall probability of being left-handed(a single floating-point number) if an individual died in the specified year. After execution value **0.07766387615350638** is returned.

3.5 AGE AT DEATH LEFT HANDEDNESS VS RIGHT HANDEDNESS:

The probability of being age A at death given that you're left-handed P(A | LH) is calculated using P(A), P(LH), and P(LH | A). Similarly the probability of being age A at death given that you're right-handed P(A | RH) is calculated using P(A), P(RH), and P(RH | A). This calculation is done as:

$$P(A|LH) = \frac{P(LH|A) * P(A)}{P(LH)}$$

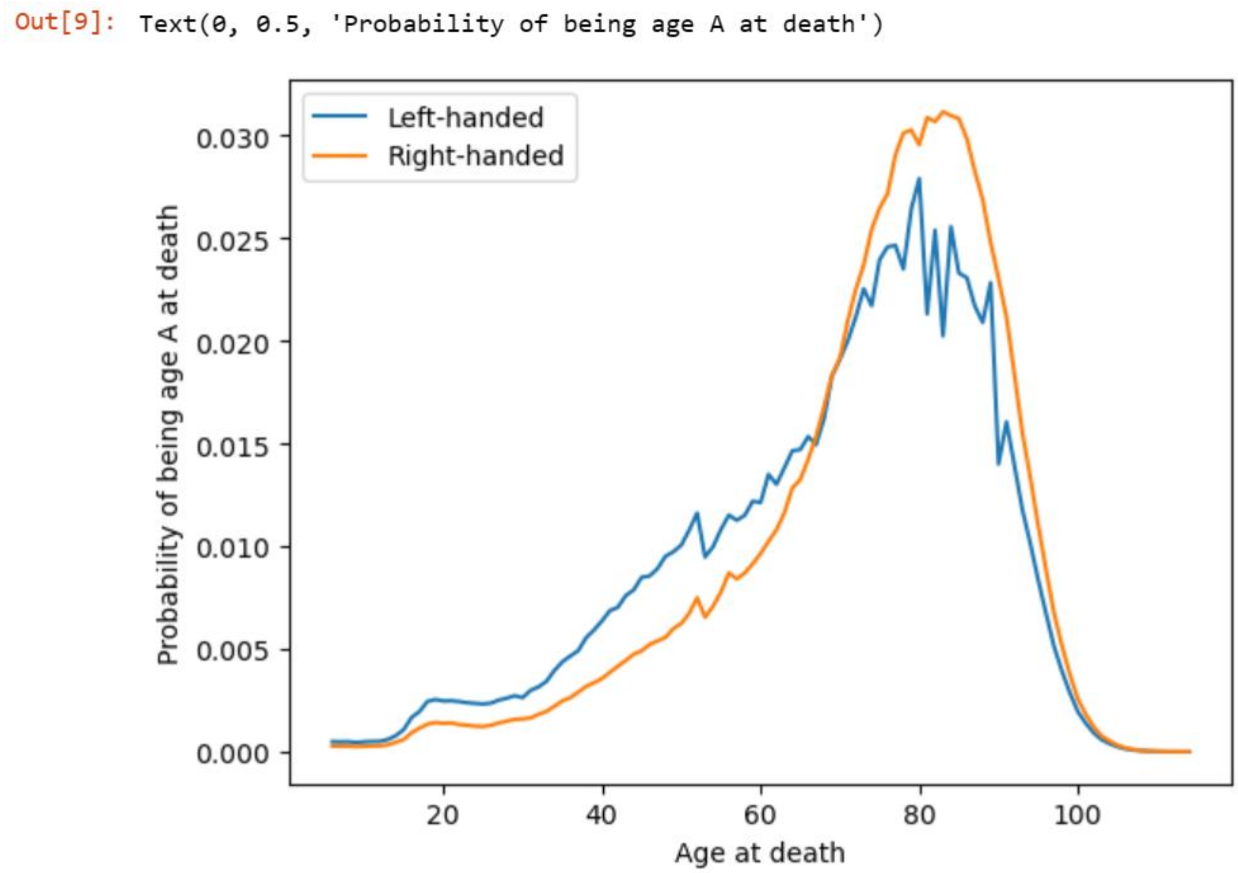
$$P(A|RH) = \frac{P(RH \mid A) * P(A)}{P(RH)}$$

Use the calculated P(LH|A) ,P(LH),P(A) values.

And P(RH|A) =1-P(LH|A), P(RH)=1-P(LH)

3.6 PLOTTING THE CONDITIONAL PROBABILITIES:

The calculated probability of being age A at death given that you're left-handed or right-handed is plotted for a range of ages of death from 6 to 120.The left-handed distribution has a bump below age 70: of the pool of deceased people, left-handed people are more likely to be younger.Create a plot of the two probabilities vs. age using plt.subplots() to create figure and axis objects. Then,plot the obtained probabilities (left_handed_probability and right_handed_probability) against the corresponding ages. The plot includes two curves: one for left-handed individuals labeled "Left-handed" and another for right-handed individuals labeled "Right-handed." The x-axis represents the age at death, and the y-axis represents the probability of being age A at death.



```
Out[9]: Text(0, 0.5, 'Probability of being age A at death')
```

3.7 COMPARING RESULTS:

The original study that found that left-handed people were nine years younger at death on average. This result is compared with our results. To do this the average age of left-handed people at death and the average age of right-handed people at death is calculated using the probability distributions obtained earlier.

Average age of left-handed people at death = Σ_A AP(A| LH)

Average age of right-handed people at death= Σ_A AP(A| RH)

A - represents the age of death.

P(A | LH) - represents the probability of being age A at death given that an individual is left-handed, which is calculated using the P_A_given_lh() function.

P(A | RH) - represents the probability of being age A at death given that an individual is right-handed, which is calculated using the P_A_given_rh() function.

The average age of left handed people at death is obtained as **67.24503662801027.** The Average age of right handed people at death is obtained as **72.79171936526477**. The difference in average ages after rounding up is **5.5 years**.

3.8 ADDITIONAL RESULTS:

The analysis using Bayesian statistics revealed a significant age gap between left-handed and right-handed individuals, purely as a result of the changing rates of left-handedness in the population. The conclusion was that left-handed individuals are unlikely to die younger due to their handedness. However, the calculated age gap was still less than the 9-year gap measured in the original study. Several factors could have contributed to the difference between the calculated age gap and the original study's result such as use of Different Death Distribution Data: The analysis used death distribution data from 1999, which was almost ten years after the original study conducted in 1990.

Additionally, the data included the entire United States rather than focusing solely on California, as in the original study. Extrapolating the left-handedness rates to older and younger age groups may have introduced some approximation errors, affecting the overall results. To gain further insights into the age gap, it would be valuable to explore the variability we would expect to encounter in the age difference purely due to random sampling. By taking smaller

samples of recently deceased people and assigning handedness with the probabilities from the survey, it is possible to determine the distribution of age gaps and how often an age gap of nine years would occur using the same data and assumptions.

The age gap expected if the study were conducted in 2018 was found to be much smaller than in 1990. This is because rates of left-handedness had not increased significantly for people born after around 1960, resulting in a less striking difference in handedness between older and younger individuals. The difference in average ages between right-handed and left-handed individuals in 2018 after rounding up is **2.3 years**

IV. CONCLUSION:

The analysis explored the age gap between left-handed and right-handed individuals at the time of the original study and further investigated how this gap might have changed over time. Using Bayesian statistics, the study aimed to determine if the changing rates of left-handedness in the population could explain any differences in the age distribution of deaths between the two groups.The analysis revealed that, based on the probabilities calculated using the provided data, left-handed individuals are not likely to die younger purely because of their handedness. The calculated age gap between left-handed and right-handed individuals was found to be smaller than the age gap reported in the original study, but still significant. Several factors could have contributed to the difference between the calculated age gap and the original study's result, including the use of different death distribution data and the extrapolation of left-handedness rates.

Additionally, the study highlighted the importance of considering changing rates of left-handedness over time and their impact on the age gap. It provided insights into the variability and random sampling involved in estimating the age difference between the two groups. It also projected the age gap that might be expected if the study were conducted in 2018, demonstrating a much smaller gap due to stable rates of left-handedness after approximately 1960.It also showcased the complexities of interpreting age gaps between left-handed and right-handed individuals, emphasizing the significance of historical context and changes in handedness rates. It highlights the uniqueness of the time when the original study and the National Geographic study took place, with changing rates of left-handedness influencing the differences observed in handedness among different age groups.

In summary, the analysis offers valuable insights into the age gap between left-handed and right-handed individuals, providing a zmore detailed understanding of how handedness rates and age distribution interact. It underscores the importance of considering population dynamics and historical trends when exploring such phenomena and encourages further research to understand the interplay between handedness and age at death.

V. REFERENCES

a) NVSS - Mortality Tables (cdc.gov)

b) Deaths By Single Years of Age, Race, and Sex: United States, 1999 (cdc.gov)

c) Hand preference and age in the United States - PubMed (nih.gov)

d) https://towardsdatascience.com/hands-on-bayesian-statistics-with-python-pymc3-arviz-499db9a59501

e) https://www.activestate.com/resources/quick-reads/how-to-display-a-plot-in-python

f) Google Data Analytics Professional Certificate- Coursera

g) Google Advanced Data Analytics Certificate- Coursera