Final Report of Traineeship Program 2023

On

“Analyze Death Age Difference of Right Handers with Left Handers”

MEDTOUREASY

28th Nov 2023

ACKNOWLEDGMENTS

I express my heartfelt appreciation to the team at MedTourEasy for fostering an enriching and conducive working environment throughout my Data Analytics traineeship.

My gratitude goes to my colleagues whose support and collaboration, despite the distance, were instrumental in fostering productivity and creating a positive learning atmosphere.

The traineeship at MedTourEasy has been pivotal in broadening my understanding of Data Analytics. This experience has significantly contributed to both my personal and professional growth, offering a profound learning curve.

I am sincerely thankful to the professionals who generously shared their guidance and expertise, thereby shaping my journey during the traineeship project.

Special thanks to the Training & Development Team at MedTourEasy for extending the opportunity to undertake this valuable traineeship. Their support and insights into the intricacies of the Data Analytics profile have been invaluable in executing the project with excellence.

TABLE OF CONTENTS

ABSTRACT

This project investigates the potential variance in the age of death between right-handed and left-handed individuals. The study aims to explore whether there exists a statistically significant difference in life expectancy or age of mortality based on hand preference.

Using a dataset encompassing demographic and hand-preference information, this research employs statistical analysis and data visualization techniques to discern any correlations or disparities in the lifespan between right-handed and left-handed subjects. Factors such as gender, geographic location, and socio-economic backgrounds may also be considered in the analysis.

The findings from this investigation seek to contribute insights into the potential relationship between hand preference and lifespan, shedding light on potential implications for healthcare and understanding human longevity.

Keywords: Hand Preference, Mortality, Life Expectancy, Statistical Analysis, Demographic Factors, Data Visualization

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

This project aims to delve into the intriguing relationship between hand preference and lifespan, particularly focusing on the purported claim of early mortality among left-handed individuals. Leveraging age distribution data, this investigation seeks to explore whether observed differences in the average age at death can be attributed solely to varying rates of left-handedness over time.

Utilizing Python's pandas library and employing Bayesian statistics, the analysis centers on assessing the likelihood of mortality at specific ages based on an individual's handedness. By employing statistical modeling techniques, this study aims to reveal insights into the potential correlation between hand preference and life expectancy.

The notebook developed for this project will facilitate the examination of age-related mortality patterns, aiming to discern any discernible disparities in life expectancy between right-handed and left-handed individuals. Through comprehensive data analysis and statistical inference, the objective is to substantiate or refute claims regarding differential mortality rates based on hand preference.

This investigation not only seeks to contribute to the ongoing discourse on handedness and longevity but also underscores the utilization of advanced statistical methodologies to explore and potentially refute long-standing assumptions regarding the mortality rates associated with left-handedness.

1.3 Objectives:

1. **Investigate Handedness and Mortality:** Analyze age distribution data to explore if there's a substantial difference in the average age at death between right-handed and left-handed individuals.
2. **Assess Impact of Changing Left-Handedness Rates:** Determine if variations in mortality rates could be attributed solely to fluctuations in left-handedness rates across different time periods.
3. **Utilize Bayesian Statistics:** Employ Bayesian statistical methods to assess the probability of mortality at specific ages based on an individual's handedness, enabling a nuanced understanding of potential correlations.
4. **Refute Misconceptions:** Aim to challenge or substantiate claims suggesting an early death trend among left-handed individuals by providing empirical evidence derived from robust statistical analysis.

Deliverables:

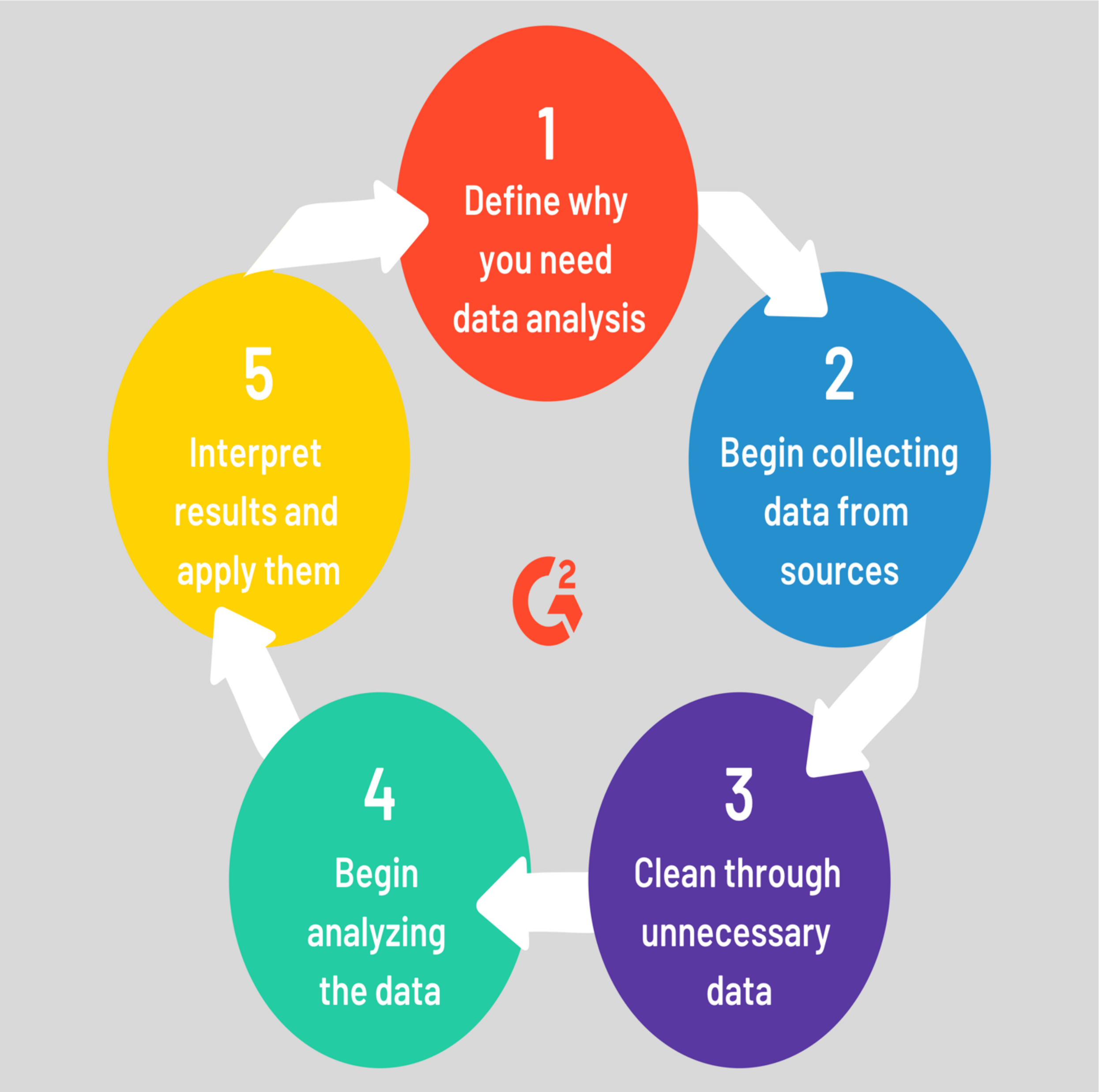
1. **Data Analysis Notebook:** A comprehensive Python notebook utilizing pandas and Bayesian statistics to perform detailed analysis and visualization of age distribution data for right-handed and left-handed individuals.
2. **Statistical Findings Report:** A detailed report summarizing the key findings, statistical insights, and conclusions drawn from the analysis, aiming to address the correlation between handedness and mortality rates.
3. **Visual Representations:** Visual aids such as graphs, charts, and diagrams to illustrate age-related mortality patterns and disparities between right-handed and left-handed groups.
4. **Documentation of Methodology:** Clear documentation outlining the methodologies, statistical models, and data preprocessing techniques employed in the analysis.
5. **Insights for Discussion:** Key insights and implications derived from the analysis, providing a basis for discussions on debunking or supporting prevailing beliefs regarding handedness and longevity.

These deliverables aim to encapsulate the thorough analysis conducted using Bayesian statistics to explore the relationship between handedness and mortality, with the objective of providing concrete evidence to either validate or challenge existing notions.

I. METHODOLOGY

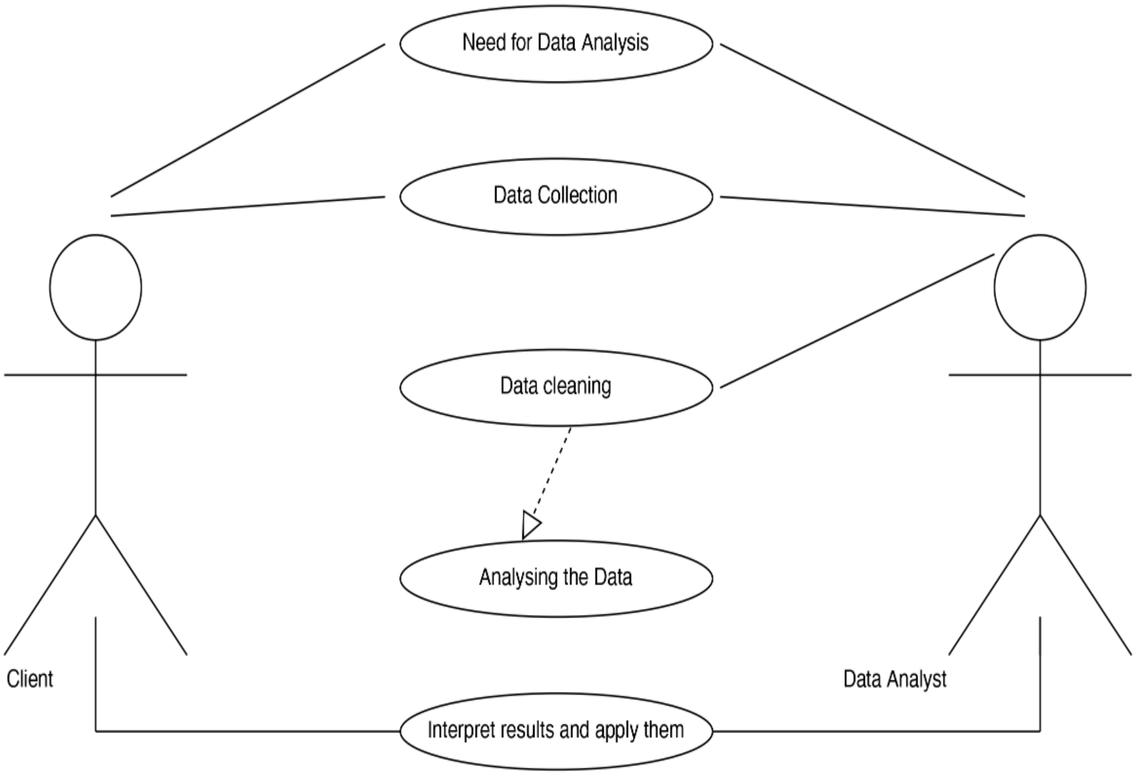
2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives and deliverables. Each step has been explained in detail in the following section.



1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

2.2 Use Case:



In this data analysis project, the analyst aims to investigate the alleged early mortality among left-handed individuals by examining age distribution data. Using Python's pandas library and Bayesian statistics, the project involves retrieving, cleaning, and analyzing age-related datasets to discern differences in average age at death between right-handed and left-handed individuals. The analyst seeks to explore correlations between handedness and life expectancy, intending to substantiate or refute claims of differential mortality rates. Ultimately, this project contributes to understanding the relationship between hand preference and longevity while employing advanced statistical methodologies to challenge prevailing assumptions about mortality rates associated with left-handedness.

2.3 Language and platform used:

The project involves data analysis, specifically exploring age distribution data to investigate the relationship between left-handedness and average age at death using Bayesian statistics. The technologies mentioned in the context suggest the use of a Jupyter Notebook, utilizing the Python programming language along with the Pandas library for data manipulation and Bayesian statistics for analysis.

Key components involved:

1. **Python**: A popular programming language used for data analysis, machine learning, and scientific computing.

2. **Jupyter Notebook**: An interactive web-based computational environment that allows for the creation and sharing of documents containing live code, equations, visualizations, and explanatory text.

3. **Pandas**: A powerful Python library for data manipulation and analysis, particularly offering data structures and operations for manipulating numerical tables and time series data.

4. **Bayesian Statistics**: A branch of statistics that deals with probability inference where probabilities express degrees of belief.

II. IMPLEMENTATION

Here's the comprehensive implementation plan:

1. **Define Project Objective:**
   - Investigate alleged early mortality among left-handed individuals using age distribution data.

2. **Data Gathering and Preparation:**
   - **Data Collection:** Retrieve datasets containing age distribution data and handedness information.
   - **Data Cleaning:** Use Pandas to clean datasets, handling missing values, duplicates, and ensuring consistency.
   - **Data Integration:** Merge or concatenate datasets to align them for analysis.

3. **Exploratory Data Analysis (EDA):**
   - **Descriptive Statistics:** Compute basic statistics (mean, median, std. deviation) for left-handed and right-handed groups.
   - **Visualization:** Utilize histograms, box plots, etc., to explore age distributions and identify patterns.

4. **Bayesian Statistical Analysis:**
   - **Model Building:** Construct Bayesian models for probability distributions of age at death for both groups.
   - **Hypothesis Testing:** Use Bayesian methods to test differences in average age at death between left-handed and right-handed groups.

5. **Correlation Investigation:**
   - **Correlation Analysis:** Assess relationships between handedness and life expectancy using statistical techniques.
   - **Inferential Analysis:** Draw inferences about mortality rates associated with left-handedness.

6. **Conclusion and Insights:**
   - **Summarize Findings:** Detail results, highlighting significant observations and differences.
   - **Insights and Recommendations:** Offer insights into hand preference and longevity, suggesting further research areas.

7. **Documentation and Reporting:**
   - **Report Generation:** Create a detailed report outlining methodology, findings, conclusions, and visualizations.
   - **Presentation:** Prepare a presentation summarizing key insights and discoveries for communication.

8. **Reflection and Iteration:**
   - **Feedback and Improvement:** Gather feedback, review the analysis process, and consider necessary improvements or additional investigations.

This detailed plan facilitates a thorough exploration of the relationship between handedness and mortality rates, employing advanced statistical methodologies to challenge and redefine existing assumptions.

III. SAMPLE SCREENSHOTS AND OBSERVATIONS

1. Where are the old left-handed people?

In this notebook, we will explore this phenomenon using age distribution data to see if we can reproduce a difference in average age at death purely from the changing rates of left-handedness over time, refuting the claim of early death for left-handers. This notebook uses `pandas` and Bayesian statistics to analyze the probability of being a certain age at death given that you are reported as left-handed or right-handed.

A National Geographic survey in 1986 resulted in over a million responses that included age, sex, and hand preference for throwing and writing. Researchers Avery Gilbert and Charles Wysocki analyzed this data and noticed that rates of left-handedness were around 13% for people younger than 40 but decreased with age to about 5% by the age of 80. They concluded based on analysis of a subgroup of people who throw left-handed but write right-handed that this age-dependence was primarily due to changing social acceptability of left-handedness. This means that the rates aren't a factor of *age* specifically but rather of the *year you were born*, and if the same study was done today, we should expect a shifted version of the same distribution as a function of age. Ultimately, we'll see what effect this changing rate has on the apparent mean age of death of left-handed people, but let's start by plotting the rates of left-handedness as a function of age.

This notebook uses two datasets: death distribution data for the United States from the year 1999 (source website here) and rates of left-handedness digitized from a figure in this 1992 paper by Gilbert and Wysocki.

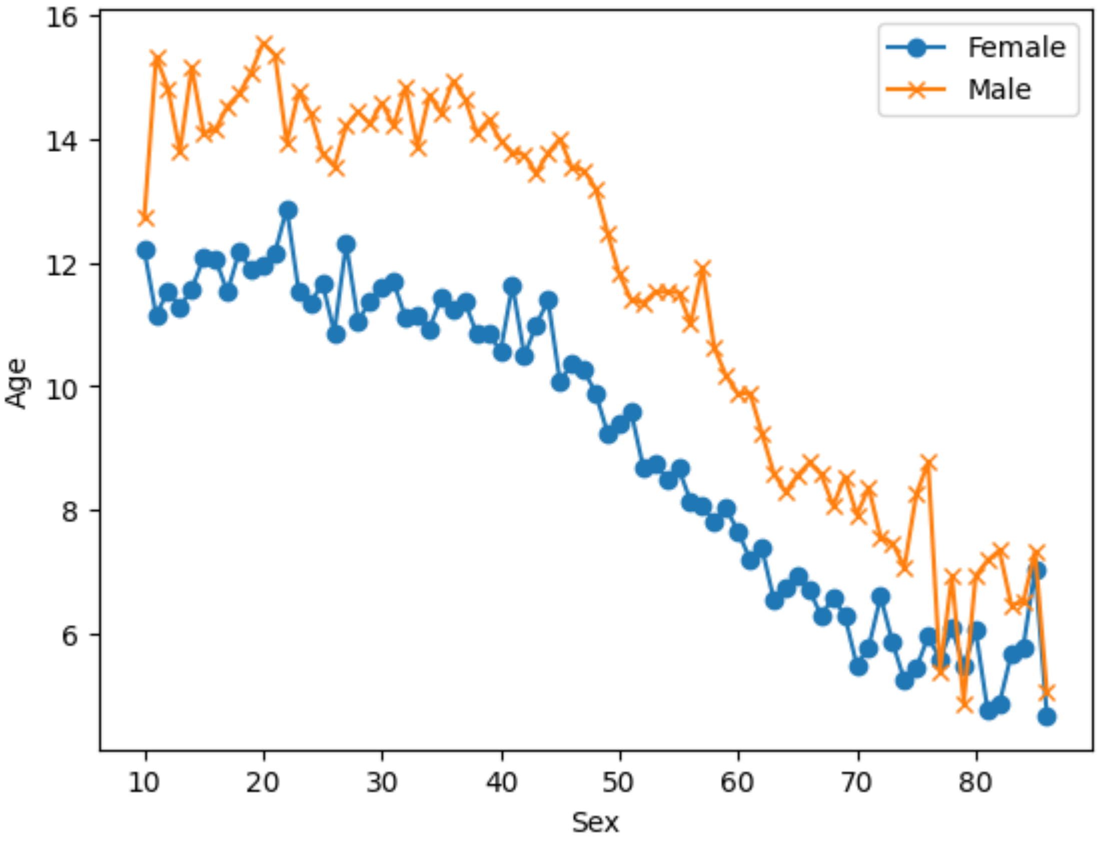
In [ ]:

```
 # import libraries
# ... YOUR CODE FOR TASK 1 ...
import pandas as pd
import matplotlib.pyplot as plt
# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7d
lefthanded_data = pd.read_csv(data_url_1)

# plot male and female left-handedness rates vs. age
%matplotlib inline
fig, ax = plt.subplots() # create figure and axis objects
ax.plot('Age', 'Female', data = lefthanded_data, marker = 'o') # plot "Femal
ax.plot('Age', 'Male', data = lefthanded_data, marker = 'x') # plot "Male" v
ax.legend() # add a legend
ax.set_xlabel('Sex')
ax.set_ylabel('Age')
```

Out[ ]:

```
Text(0, 0.5, 'Age')
```



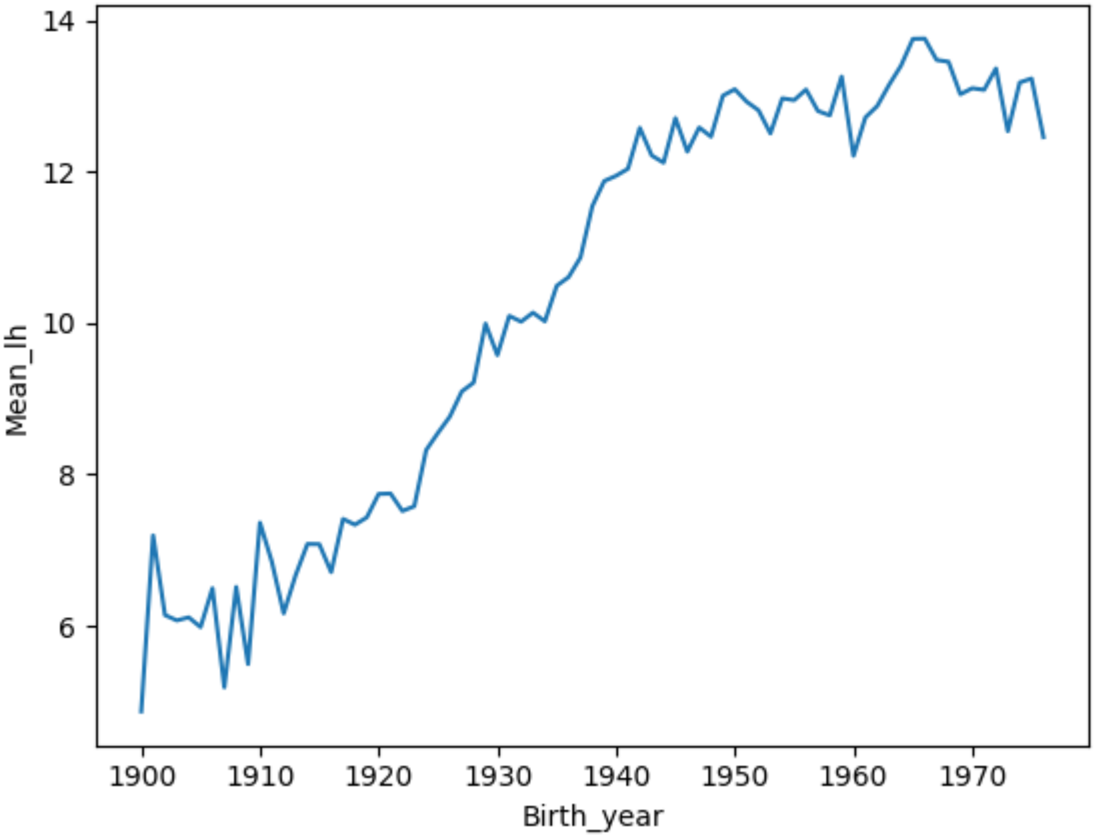
2. Rates of left-handedness over time

Let's convert this data into a plot of the rates of left-handedness as a function of the year of birth, and average over male and female to get a single rate for both sexes.

Since the study was done in 1986, the data after this conversion will be the percentage of people alive in 1986 who are left-handed as a function of the year they were born.

```
In [ ]: # create a new column for birth year of each age
        # ... YOUR CODE FOR TASK 2 ...
        lefthanded_data['Birth_year'] = 1986 - lefthanded_data['Age']
        # create a new column for the average of male and female
        # ... YOUR CODE FOR TASK 2 ...
        lefthanded_data['Mean_lh'] = lefthanded_data[['Male', 'Female']].mean(axis=1
        # create a plot of the 'Mean_lh' column vs. 'Birth_year'
        fig, ax = plt.subplots()
        ax.plot('Birth_year', 'Mean_lh', data = lefthanded_data) # plot 'Mean_lh' vs
        ax.set_xlabel('Birth_year') # set the x label for the plot
        ax.set_ylabel('Mean_lh') # set the y label for the plot
```

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



3. Applying Bayes' rule

The probability of dying at a certain age given that you're left-handed is **not** equal to the probability of being left-handed given that you died at a certain age. This inequality is why we need **Bayes' theorem**, a statement about conditional probability which allows us to update our beliefs after seeing evidence.

We want to calculate the probability of dying at age A given that you're left-handed. Let's write this in shorthand as P(A | LH). We also want the same quantity for right-handers: P(A | RH).

Here's Bayes' theorem for the two events we care about: left-handedness (LH) and dying at age A.

$$P(\quad |LH) = \frac{P(LH| \quad)P(\quad)}{P(LH)}$$

P(LH | A) is the probability that you are left-handed *given that* you died at age A. P(A) is the overall probability of dying at age A, and P(LH) is the overall probability of being left-handed. We will now calculate each of these three quantities, beginning with P(LH | A).

To calculate P(LH | A) for ages that might fall outside the original data, we will need to extrapolate the data to earlier and later years. Since the rates flatten out in the early

1900s and late 1900s, we'll use a few points at each end and take the mean to extrapolate the rates on each end. The number of points used for this is arbitrary, but we'll pick 10 since the data looks flat-ish until about 1910.

In [ ]:

```
# import library
# ... YOUR CODE FOR TASK 3 ...
import numpy as np
# create a function for P(LH | A)
def P_lh_given_A(ages_of_death, study_year = 1990):
    """ P(Left-handed | ages of death), calculated based on the reported rat
    Inputs: numpy array of ages of death, study_year
    Returns: probability of left-handedness given that subjects died in `stu

    # Use the mean of the 10 last and 10 first points for left-handedness ra
    early_1900s_rate = lefthanded_data['Mean_lh'][-10:].mean()
    late_1900s_rate = lefthanded_data['Mean_lh'][:10].mean()
    middle_rates = lefthanded_data.loc[lefthanded_data['Birth_year'].isin(st
    youngest_age = study_year - 1986 + 10 # the youngest age is 10
    oldest_age = study_year - 1986 + 86 # the oldest age is 86

    P_return = np.zeros(ages_of_death.shape) # create an empty array to stor
    # extract rate of left-handedness for people of ages 'ages_of_death'
    P_return[ages_of_death > oldest_age] = early_1900s_rate / 100
    P_return[ages_of_death < youngest_age] = late_1900s_rate / 100
    P_return[np.logical_and((ages_of_death <= oldest_age), (ages_of_death >=

    return P_return
```

4. When do people normally die?

To estimate the probability of living to an age A, we can use data that gives the number of people who died in a given year and how old they were to create a distribution of ages of death. If we normalize the numbers to the total number of people who died, we can think of this data as a probability distribution that gives the probability of dying at age A. The data we'll use for this is from the entire US for the year 1999 - the closest I could find for the time range we're interested in.

In this block, we'll load in the death distribution data and plot it. The first column is the age, and the other columns are the number of people who died at that age.

In [ ]:

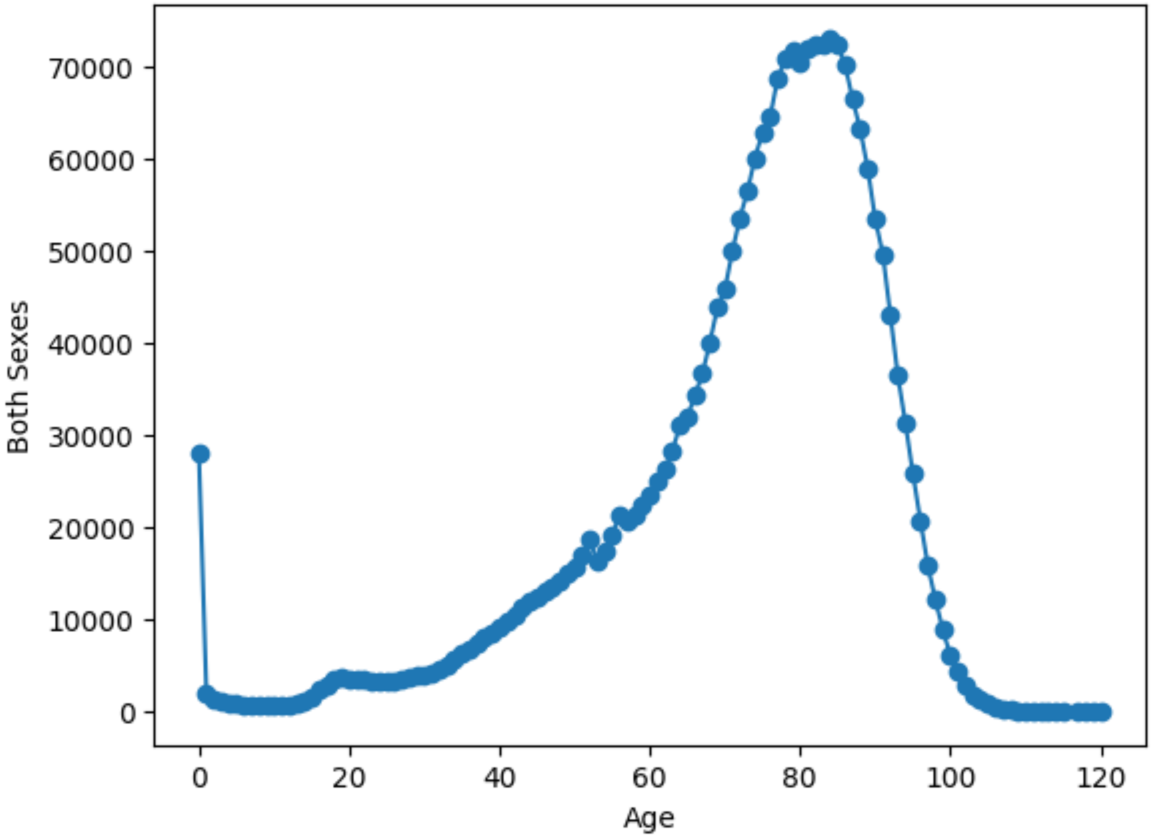
```
# Death distribution data for the United States in 1999
data_url_2 = "https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807

# load death distribution data
# ... YOUR CODE FOR TASK 4 ...
death_distribution_data = pd.read_csv(data_url_2, sep='\t', skiprows=[1])
# drop NaN values from the `Both Sexes` column
# ... YOUR CODE FOR TASK 4 ...
death_distribution_data = death_distribution_data.dropna(subset = ['Both Sex
# plot number of people who died as a function of age
fig, ax = plt.subplots()
```

```
ax.plot('Age', 'Both Sexes', data = death_distribution_data, marker='o') # p
ax.set_xlabel('Age')
ax.set_ylabel('Both Sexes')
```

Out[ ]:

```
Text(0, 0.5, 'Both Sexes')
```



5. The overall probability of left-handedness

In the previous code block we loaded data to give us P(A), and now we need P(LH). P(LH) is the probability that a person who died in our particular study year is left-handed, assuming we know nothing else about them. This is the average left-handedness in the population of deceased people, and we can calculate it by summing up all of the left-handedness probabilities for each age, weighted with the number of deceased people at each age, then divided by the total number of deceased people to get a probability. In equation form, this is what we're calculating, where N(A) is the number of people who died at age A (given by the dataframe `death_distribution_data`):

$$P(LH) = \frac{\sum_{\mathrm{A}} P(LH|A)N(A)}{\sum_{\mathrm{A}} N(A)}$$

In [ ]:

```
def P_lh(death_distribution_data, study_year = 1990): # sum over P_lh for ea
    """ Overall probability of being left-handed if you died in the study ye
    Input: dataframe of death distribution data, study year
    Output: P(LH), a single floating point number """
    p_list = death_distribution_data['Both Sexes'] * P_lh_given_A(death_dist
    p = np.sum(p_list) # calculate the sum of p_list
    return p / np.sum(death_distribution_data['Both Sexes']) # normalize to

print(P_lh(death_distribution_data))
```

```
0.07766387615350638
```

6. Putting it all together: dying while left-handed (i)

Now we have the means of calculating all three quantities we need: P(A), P(LH), and P(LH | A). We can combine all three using Bayes' rule to get P(A | LH), the probability of being age A at death (in the study year) given that you're left-handed. To make this answer meaningful, though, we also want to compare it to P(A | RH), the probability of being age A at death given that you're right-handed.

We're calculating the following quantity twice, once for left-handers and once for right-handers.

$$P(\quad|LH) = \frac{P(LH|\quad)P(\quad)}{P(LH)}$$

First, for left-handers.

In [ ]:

```
def P_A_given_lh(ages_of_death, death_distribution_data, study_year = 1990):
    """ The overall probability of being a particular `age_of_death` given t
    P_A = death_distribution_data['Both Sexes'][ages_of_death] / np.sum(deat
    P_left = P_lh(death_distribution_data, study_year) # use P_lh function t
    P_lh_A = P_lh_given_A(ages_of_death, study_year) # use P_lh_given_A to g
    return P_lh_A*P_A/P_left
```

7. Putting it all together: dying while left-handed (ii)

And now for right-handers.

In [ ]:

```
def P_A_given_rh(ages_of_death, death_distribution_data, study_year = 1990):
    """ The overall probability of being a particular `age_of_death` given t
    P_A = death_distribution_data['Both Sexes'][ages_of_death] / np.sum(deat
    P_right = 1 - P_lh(death_distribution_data, study_year) # either you're
    P_rh_A = 1 - P_lh_given_A(ages_of_death, study_year) # P_rh_A = 1 - P_lh
    return P_rh_A*P_A/P_right
```

8. Plotting the distributions of conditional probabilities

Now that we have functions to calculate the probability of being age A at death given that you're left-handed or right-handed, let's plot these probabilities for a range of ages of death from 6 to 120.

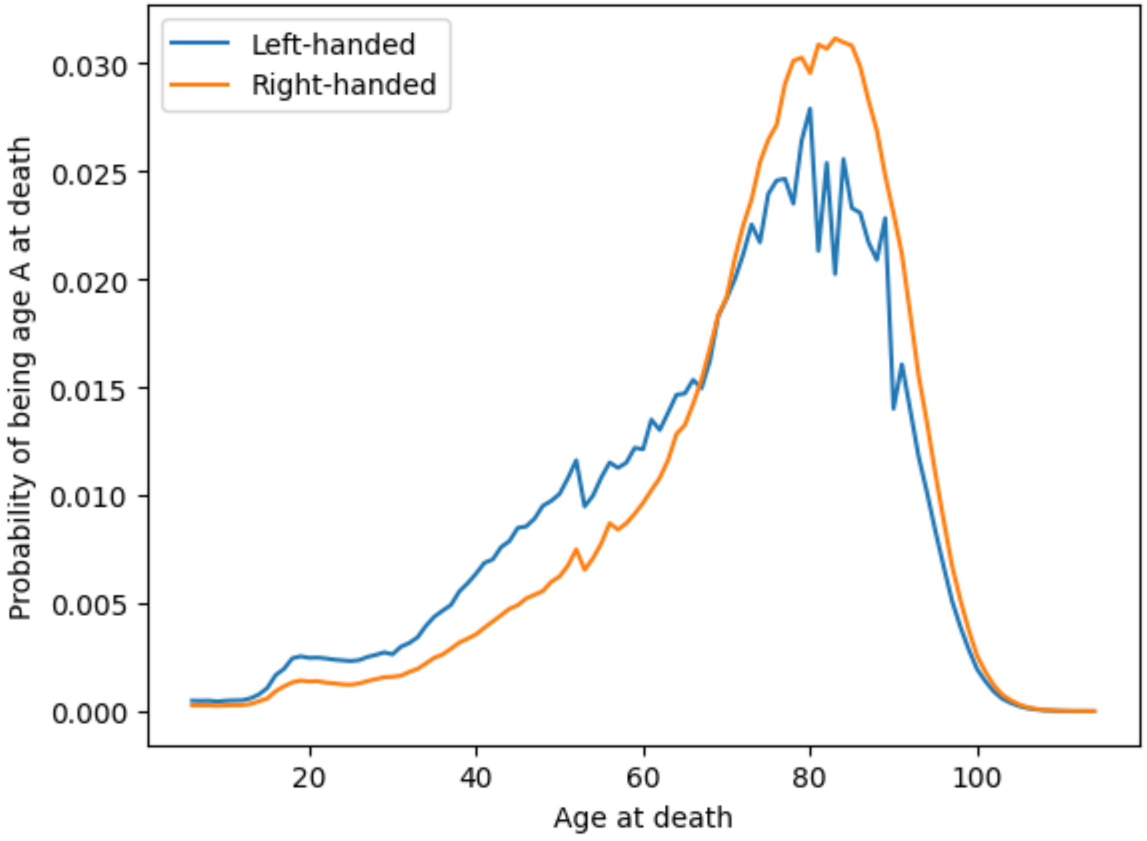
Notice that the left-handed distribution has a bump below age 70: of the pool of deceased people, left-handed people are more likely to be younger.

```
In [ ]: ages = np.arange(6, 115, 1) # make a list of ages of death to plot

        # calculate the probability of being left- or right-handed for each
        left_handed_probability = P_A_given_lh(ages, death_distribution_data)
        right_handed_probability = P_A_given_rh(ages, death_distribution_data)

        # create a plot of the two probabilities vs. age
        fig, ax = plt.subplots() # create figure and axis objects
        ax.plot(ages, left_handed_probability, label = "Left-handed")
        ax.plot(ages, right_handed_probability, label = 'Right-handed')
        ax.legend() # add a legend
        ax.set_xlabel("Age at death")
        ax.set_ylabel(r"Probability of being age A at death")
```

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



9. Moment of truth: age of left and right-handers at death

Finally, let's compare our results with the original study that found that left-handed people were nine years younger at death on average. We can do this by calculating the mean of these probability distributions in the same way we calculated P(LH) earlier, weighting the probability distribution by age and summing over the result.

$$\text{Average age of left-handed people at death} = \qquad P(\quad |LH)$$

$$\text{Average age of right-handed people at death} = \qquad P(\quad |RH)$$

In [ ]:

```
# calculate average ages for left-handed and right-handed groups
# use np.array so that two arrays can be multiplied
average_lh_age =  np.nansum(ages*np.array(left_handed_probability))
average_rh_age =  np.nansum(ages*np.array(right_handed_probability))

# print the average ages for each group
# ... YOUR CODE FOR TASK 9 ...
print("Average age of lefthanded" + str(average_lh_age))
print("Average age of righthanded" + str(average_rh_age))

# print the difference between the average ages
print("The difference in average ages is " + str(round(average_rh_age - aver
```

```
Average age of lefthanded67.24503662801027
Average age of righthanded72.79171936526477
The difference in average ages is 5.5 years.
```

10. Final comments

We got a pretty big age gap between left-handed and right-handed people purely as a result of the changing rates of left-handedness in the population, which is good news for left-handers: you probably won't die young because of your sinisterness. The reported rates of left-handedness have increased from just 3% in the early 1900s to about 11% today, which means that older people are much more likely to be reported as right-handed than left-handed, and so looking at a sample of recently deceased people will have more old right-handers.

Our number is still less than the 9-year gap measured in the study. It's possible that some of the approximations we made are the cause:

1. We used death distribution data from almost ten years after the study (1999 instead of 1991), and we used death data from the entire United States instead of California alone (which was the original study).
2. We extrapolated the left-handedness survey results to older and younger age groups, but it's possible our extrapolation wasn't close enough to the true rates for those ages.

One thing we could do next is figure out how much variability we would expect to encounter in the age difference purely because of random sampling: if you take a smaller sample of recently deceased people and assign handedness with the probabilities of the survey, what does that distribution look like? How often would we encounter an age gap of nine years using the same data and assumptions? We won't do that here, but it's possible with this data and the tools of random sampling.

To finish off, let's calculate the age gap we'd expect if we did the study in 2018 instead of in 1990. The gap turns out to be much smaller since rates of left-handedness haven't increased for people born after about 1960. Both the National Geographic study and the 1990 study happened at a unique time - the rates of left-handedness had been changing across the lifetimes of most people alive, and the difference in handedness between old and young was at its most striking.

In [ ]:

```
# Calculate the probability of being left- or right-handed for all ages
left_handed_probability_2018 = P_A_given_lh(ages, death_distribution_data, 2
right_handed_probability_2018 = P_A_given_rh(ages, death_distribution_data,

# calculate average ages for left-handed and right-handed groups
average_lh_age_2018 = np.nansum(ages*np.array(left_handed_probability_2018))
average_rh_age_2018 = np.nansum(ages*np.array(right_handed_probability_2018)

print("The difference in average ages is " +
      str(round(average_rh_age_2018 - average_lh_age_2018, 1)) + " years.")
```

```
The difference in average ages is 2.3 years.
```