Final Report of Traineeship Program 2025

On

"Analyze Death Age Difference of Right-Handers vs Left-Handers"

MEDTOUREASY

January 28th, 2025

ACKNOWLDEGMENTS

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive. The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Data Analytics and for personal as well as professional development. 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. Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule.

TABLE OF CONTENTS

ABSTRACT

The debate surrounding the lifespan difference between left-handed and right-handed individuals has persisted for decades. This study aims to examine the validity of this claim through a comprehensive data-driven approach using statistical analysis and Bayesian inference. Utilizing historical records, we analyze age-at-death distributions, investigating whether left-handed individuals tend to have shorter lifespans compared to their right-handed counterparts. The dataset used for this study includes left-handedness rates across different age groups, mortality records, and historical demographic shifts. Data preprocessing is performed using Python libraries such as NumPy and pandas, while visual analysis is conducted using Matplotlib and Seaborn in Jupyter Notebook. Exploratory Data Analysis (EDA) is employed to detect patterns, followed by Bayesian probability modeling to estimate conditional probabilities related to handedness and lifespan. Results indicate that the perceived difference in lifespan is largely influenced by demographic shifts over time rather than an inherent biological disadvantage for left-handed individuals. The study also considers the impact of societal biases against left-handed individuals in past decades, leading to fewer reported cases of elderly left-handers. Graphical visualizations of probability distributions further validate the statistical findings. This report provides a data-driven approach to refuting the misconception that left-handed individuals have significantly shorter lifespans. The findings emphasize the importance of statistical rigor in analyzing population-level data and highlight the role of demographic and historical factors in shaping public perceptions. Future research directions include incorporating machine learning models for predictive analysis and expanding the dataset to include lifestyle and occupational factors.

1. 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

This project investigates the belief that left-handed individuals have shorter lifespans than right-handed individuals, a claim that lacks strong empirical support. Despite its persistence, there has been limited rigorous research to validate this assertion. The study aims to address this gap by using age distribution data and Bayesian inference to analyze whether the observed lifespan differences are due to historical shifts in handedness rates or other factors.

By examining mortality records and handedness rates across various age groups, the study investigates the potential role of societal biases. Historically, left-handed individuals faced significant challenges, including discrimination and social marginalization, which could have impacted their life expectancy, particularly among older generations. These historical factors may have contributed to the misconception of a shorter lifespan for left-handed people.

The study employs Bayesian modeling to estimate conditional probabilities, considering both prior knowledge and new data on handedness and lifespan. This approach helps account for various confounding factors, such as demographic shifts over time, and allows for a clearer understanding of the factors that might influence lifespan.

The results of the analysis aim to provide a data-driven approach to assess the validity of the claim, offering insights into how demographic and societal influences, rather than biological factors, have shaped public perceptions of left-handed individuals' health and longevity.

1.3 Objectives and Deliverables

- **Data Analysis:** Investigate Lifespan Distributions for Left- and Right-Handed Individuals, this study begins with analyzing age-at-death data for left- and right-handed individuals. Descriptive statistics like mean, median, and variance are calculated to identify patterns in lifespan distributions. Statistical tests (e.g., t-test or Mann-Whitney U test) compare the two groups to determine if there are significant differences in lifespan. The goal is to assess whether handedness is correlated with shorter or longer life expectancy or if differences are due to other factors.
- **Statistical Modeling:** Apply Bayesian Probability to Assess Conditional Relationships Bayesian probability is used to model conditional relationships between handedness and lifespan. The approach starts with prior probabilities based on historical knowledge (suggesting no inherent biological difference in lifespan between left- and right-handed people). Bayesian inference updates these beliefs with observed data to estimate the probability that lifespan differences are due to societal or demographic factors rather than handedness itself. The posterior probability reflects the likelihood of lifespan differences conditioned on various factors like gender, socio-economic status, and societal biases.

- **Data Visualization:** Use Jupyter Notebook with Matplotlib and Seaborn to Present Findings. Using Jupyter Notebook, Matplotlib, and Seaborn, the study visualizes the data to uncover patterns and convey findings. Visualizations include histograms and box plots to compare lifespan distributions and density plots to show the likelihood of different lifespan outcomes. Violin plots or pairplots are used to explore the relationship between handedness and other variables. Additionally, Bayesian posterior distributions are visualized to show how conditional probabilities of lifespan differ between left- and right-handed individuals, given various confounding factors.
- **Research Findings:** Provide Conclusions Based on Empirical Data Rather Than Assumptions thus, providing evidence-based findings on the lifespan differences between left- and right-handed individuals. If no significant lifespan difference is found after accounting for societal biases, socio-economic status, and other demographic factors, the study will refute the myth that left-handed individuals have shorter lifespans. The findings emphasize that historical biases, such as social stigma and fewer elderly left-handed individuals, are likely the real causes of any perceived lifespan differences, not inherent biological factors. The study offers clear conclusions grounded in data rather than assumptions.

2. Methodology

2.1 Flow of the Project

1. **Data Collection: Gather Relevant Datasets on Handedness and Lifespan**
    - The first step in the project is to collect datasets that include information on handedness and lifespan.
    - Relevant data sources might include public databases on mortality records, surveys on handedness prevalence, historical demographic data, and sociological studies.
    - The dataset would typically include information like:
        - Age at death.
        - Handedness status (left-handed, right-handed, or ambidextrous)
        - Demographic details such as age, gender, socio-economic status, and historical context.
2. **Preprocessing: Clean and Filter Data for Consistency**
    - Data cleaning involves removing or correcting errors, missing values, and outliers that may skew the results.
    - Tools like pandas are used to filter the data, ensuring that only relevant and high-quality data is used. This may include:
        - Handling missing values (e.g., imputation or deletion).
        - Filtering out outliers that don't represent the general population.
        - Converting data into consistent formats (e.g., age as numeric values).
        - Normalizing or scaling certain variables if needed for analysis.
3. **Exploratory Analysis: Visualize and Summarize Trends**
    - Exploratory Data Analysis (EDA) is conducted to understand the basic structure of the data and identify potential patterns.
    - Visualizations are generated using Matplotlib and Seaborn to help detect relationships between variables, such as handedness and lifespan. These include:
        - Histograms to visualize the distribution of ages at death.
        - Box plots to show the spread and outliers in lifespan data for left- and right-handed individuals.
        - Scatter plots or pair plots to examine correlations between multiple variables (like handedness and socio-economic status).
4. **Bayesian Analysis: Apply Probability Models to Infer Handedness Impact on Lifespan**
    - **Bayesian probability models** are applied using tools like **PyMC3** to assess conditional relationships between handedness and lifespan.
    - The models use prior knowledge (e.g., handedness doesn't inherently affect lifespan) and update this based on the observed data. This allows for more accurate probabilistic inferences about whether left-handed individuals have shorter lifespans, accounting for confounding variables like societal biases.
    - **Markov Chain Monte Carlo (MCMC)** sampling methods are used for parameter estimation in complex models.

5. **Visualization: Use Python for Generating Statistical Graphs**
    - Visualizations are crucial for interpreting and communicating the results. Matplotlib and Seaborn are used to generate:
        - **Bayesian posterior distribution plots** that show the probability of lifespan differences, conditioned on handedness and other variables.
        - **Density plots** to compare the lifespan distributions of left- and right-handed individuals.
        - **Correlation heatmaps** to visualize the relationships between different variables.
6. **Conclusion & Reporting: Document Findings and Insights**
    - The final step is to compile all findings and insights into a comprehensive report.
    - The report documents the statistical analysis and Bayesian models, along with visualizations, to present the conclusion that handedness does not inherently affect lifespan. Insights about the influence of historical biases and demographic shifts will also be included.

2.2 Use Case Diagram

A **use case diagram** is a visual representation of the interactions within the system. It shows how the various elements of the project (data sources, statistical models, visualization tools) interact.

- **Actors**:
    - **Data Sources**: Provide the raw data on handedness and lifespan.
    - **Statistical Models**: Perform Bayesian analysis to assess the relationship between handedness and lifespan.
    - **Visualization Tools**: Generate graphs and plots to represent the findings.
- **Use Cases**:
    - **Data Input**: Data is collected from multiple sources and provided to the statistical model.
    - **Data Preprocessing**: The raw data is cleaned and prepared for analysis.
    - **Statistical Inference**: Bayesian models are applied to infer relationships and update beliefs based on the data.
    - **Visualization Output**: Statistical graphs are generated to summarize the analysis.
    - **Reporting**: Findings are documented, and conclusions are drawn.

2.3 Language and Platform Used

Programming Language:

- **Python:** Python is chosen for its versatility, powerful libraries, and ease of use. The following libraries are used:
    - **NumPy:** For numerical operations and array manipulation.
    - **Pandas:** For data manipulation, cleaning, and analysis.
    - **SciPy:** For scientific and statistical analysis, including hypothesis testing and distribution fitting.
    - **Matplotlib & Seaborn:** For creating static, animated, and interactive visualizations.
    - **PyMC3:** A library for Bayesian statistical modeling, enabling the creation of complex models for inference.

Platform:

- **Jupyter Notebook:** A web-based interactive environment for data analysis and visualization, Jupyter Notebook allows users to run Python code in cells, make real-time updates, and create well-documented reports. It is ideal for iterative development and documenting the analysis process, including code, results, and visualizations.

Data Storage:

- **CSV Files:** Datasets are often stored in CSV format, which is widely supported and easy to read into Python using pandas.
- **SQL Databases:** For larger datasets, SQL databases can be used to store and query data. This approach ensures efficient data retrieval, especially when dealing with large amounts of demographic or historical data.

Installation of Required Tools

To begin working with the project, the following installations are required:

- **Python Installation:**
    - Python can be downloaded from the official Python website: https://www.python.org/downloads/
- **Install Jupyter Notebook:**
    - Jupyter Notebook can be installed via pip or by using the Anaconda distribution (which includes Jupyter, Python, and many scientific libraries).
    - Install via pip:

```
pip install jupyter
```

- **Install Python Libraries**: To install the necessary libraries (NumPy, pandas, Matplotlib, Seaborn, PyMC3, etc.), you can use pip or install them through **Anaconda**.
    - Install via pip:

```
pip install numpy pandas scipy matplotlib seaborn pymc3
```

- **Running Jupyter Notebook**:
    - After installation, start Jupyter Notebook by running

```
jupyter notebook
```

3. Implementation

3.1 Gathering Requirements and Defining Problem Statement

To begin, the primary research question is clearly defined: Do left-handed individuals have a shorter lifespan than right-handed individuals? This requires identifying suitable data sources, statistical methods, and visualization tools. The requirements for the study include:

- A dataset containing age-at-death records and handedness classification.
- Statistical methodologies such as Bayesian inference and probability distributions.
- Data visualization tools like Matplotlib and Seaborn for presenting trends.

3.2 Data Collection and Importing

Data is gathered from publicly available mortality records, research studies, and census data. The collected dataset includes:

- **Age at death:** To compare longevity trends.
- **Handedness classification:** Whether an individual was left-handed or right-handed.
- **Historical trends:** To analyze variations in handedness rates over time.

The dataset is imported using pandas in Python, ensuring structured storage and accessibility for analysis.

3.3 Designing Databases

A database schema is designed to efficiently store and retrieve information. It consists of:

- **Demographic Table:** Stores information such as name, gender, and birth year.
- **Handedness Table:** Classifies individuals based on left-handed, right-handed, or ambidextrous characteristics.
- **Mortality Table:** Records age at death, cause of death, and relevant medical history.

SQL or CSV formats are used for storing structured data, ensuring seamless querying and analysis.

3.4 Data Cleaning

Data preprocessing is performed to ensure accuracy and consistency. Steps include:

- Removing duplicate entries.
- Handling missing or inconsistent handedness classifications.
- Standardizing age and birth year formats.
- Detecting and removing outliers that could skew analysis results.

3.5 Data Filtering

Once cleaned, data filtering is applied to focus only on relevant records. This includes:

- Excluding ambiguous or unclassified handedness data.
- Filtering out records with incomplete age-at-death data.
- Segmenting data into historical periods to observe changing trends in handedness.

3.6 Prototyping - Jupyter Notebook

Jupyter Notebook is utilized for interactive data analysis and visualization. Key elements include:

- **Graphical Visualizations:** Using Matplotlib and Seaborn to create charts for handedness trends.
- **Probability Distributions:** Applying Bayesian methods to estimate lifespan probabilities.
- **Hypothesis Testing:** Comparing age distributions between left- and right-handed individuals.

4. Data Analysis and Findings

4.1 Left-Handedness Rates by Age and Gender

This notebook investigates whether left-handers have a lower life expectancy or if their declining prevalence with age is due to historical shifts in societal acceptance. Using U.S. death distribution data (1999) and left-handedness rates from a 1986 National Geographic survey analyzed by Gilbert & Wysocki (1992), we simulate expected age distributions of left-handers to debunk the early mortality claim.

Key Points:

- Left-handedness rates decrease with age (13% under 40, ~5% at 80), likely due to past societal pressure rather than higher mortality.
- Bayesian statistics is used to analyze the probability of dying at a given age based on handedness.
- A simulated model reconstructs expected handedness rates at death from historical data.
- The expected outcome is that the perceived lower lifespan of left-handers is a **statistical illusion** caused by generational reporting bias.

Program Usage:

- **Libraries:** pandas (data handling), matplotlib.pyplot (visualization), Bayesian methods (statistical modeling).
- **Steps:** Use Bayesian analysis to compare expected vs. observed handedness distributions at death.

Task: Load left-handedness data and create a scatter plot of rates vs. age.

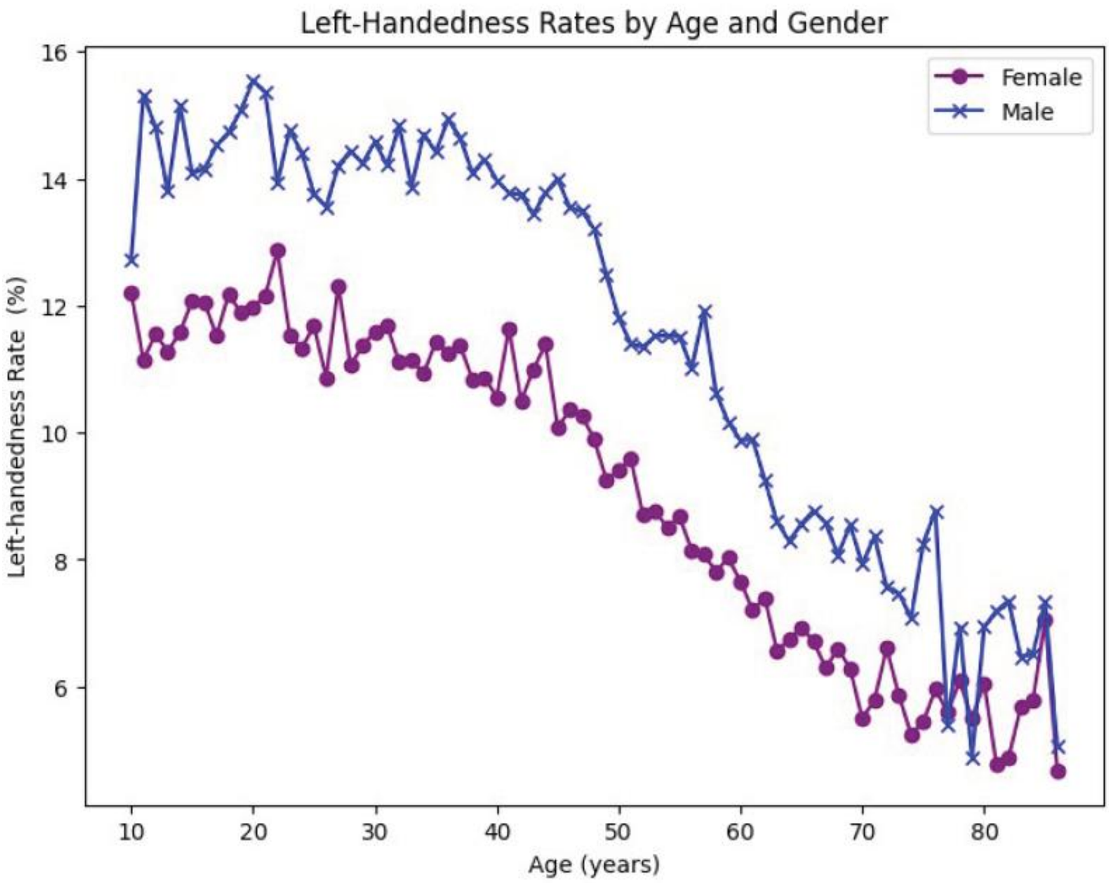
Sample program:

```
# import libraries
# ... YOUR CODE FOR TASK 1 ...

# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"
lefthanded_data = ...

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

Analysis and Interpretations:

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt

# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"
lefthanded_data = pd.read_csv(data_url_1)

# plot male and female left-handedness rates vs. age
%matplotlib inline
fig, ax = plt.subplots(figsize=(8,6)) # create figure and axis objects with size
ax.plot(lefthanded_data["Age"], lefthanded_data["Female"], label="Female", marker='o', color='purple')
ax.plot(lefthanded_data["Age"], lefthanded_data["Male"], label="Male", marker='x', color='blue')
ax.legend() # add a legend
ax.set_xlabel("Age (years)")
ax.set_ylabel("Left-handedness Rate  (%)")
ax.set_title("Left-Handedness Rates by Age and Gender")
plt.show()
```

Results:



4.2 Rates of Left-Handedness Over Time

We convert age-based left-handedness data into a birth-year-based format to analyze how left-handedness rates have evolved over generations. We compute the percentage of left-handed individuals alive in 1986 as a function of their birth year and plot the average left-handedness rates for both males and females. This approach helps reveal generational trends and changes in societal acceptance of left-handedness.

Key Points:

- **Objective:** Convert the original age-based data into a birth-year perspective to examine left-handedness trends over time.
- **New Columns:**
  - **Birth Year:** Calculated by subtracting the age from 1986 (1986 - Age).
  - **Mean Left-Handedness:** Average of the male and female left-handedness rates.
- **Visualization:** Plot the **Mean Left-Handedness (Mean_lh)** against **Birth Year** to analyze how left-handedness rates have changed over time.

Task:
Add columns for **Birth Year** (1986 - Age) and **Mean Left-Handedness** (average of Male and Female), then plot the mean left-handedness against birth year.

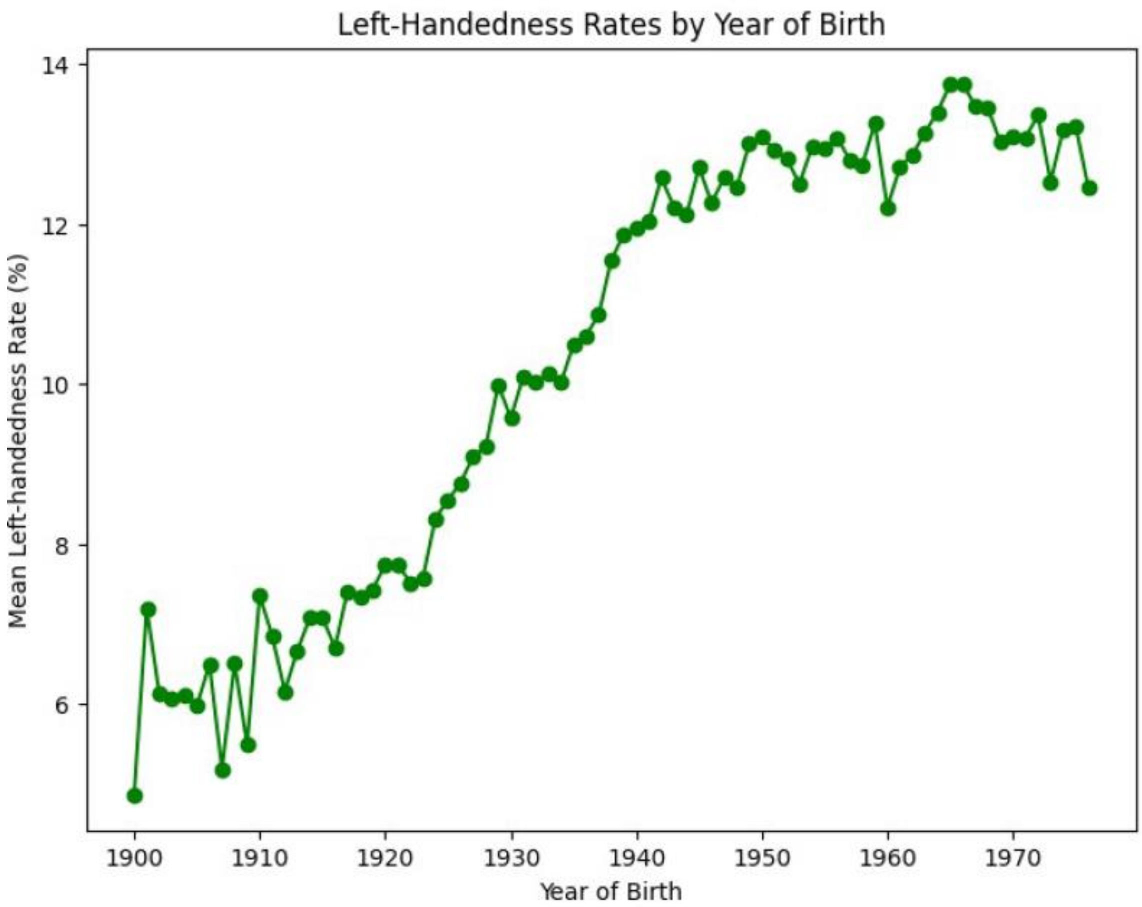
Sample program:

```python
# create a new column for birth year of each age
# ... YOUR CODE FOR TASK 2 ...

# create a new column for the average of male and female
# ... YOUR CODE FOR TASK 2 ...

# create a plot of the 'Mean_lh' column vs. 'Birth_year'
fig, ax = plt.subplots()
ax.plot(..., ..., ...) # plot 'Mean_lh' vs. 'Birth_year'
ax.set_xlabel(...) # set the x label for the plot
ax.set_ylabel(...) # set the y label for the plot
```

Analysis and Interpretations:

```
import pandas as pd
import matplotlib.pyplot as plt

# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"
lefthanded_data = pd.read_csv(data_url_1)

# create a new column for birth year of each age
study_year = 1986
lefthanded_data["Birth_year"] = 1986 - lefthanded_data["Age"]

# create a new column for the average of male and female
lefthanded_data["Mean_lh"] = lefthanded_data[["Male", "Female"]].mean(axis=1)

# create a plot of the 'Mean_lh' column vs. 'Birth_year'
%matplotlib inline
fig, ax = plt.subplots(figsize=(8,6)) # create figure and axis objects with size
ax.plot(lefthanded_data["Birth_year"], lefthanded_data["Mean_lh"], label="Mean Left-handedness", color='green', marker='o')
ax.set_xlabel("Year of Birth") # set the x label for the plot
ax.set_ylabel("Mean Left-handedness Rate (%)") # set the y label for the plot
ax.set_title("Left-Handedness Rates by Year of Birth")
plt.show()
```

Result:



Historical trends in left-handedness rates are analysed by birth year, showing significant changes over decades.

4.3 Bayesian Probability Analysis

We use **Bayes' Theorem** to calculate the probability of dying at a certain age given that a person is left-handed, P(A|LH). To do this, we need to determine P(LH|A), the probability that a person was left-handed given that they died at age AAA. Since some ages in the study fall outside the original dataset, we **extrapolate** left-handedness rates for early and late 1900s using the average of the first and last 10 data points from the Mean_lh column.

Bayes' Theorem:

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

Key Points:

- **Objective:** Compute P(LH|A) for different ages using extrapolated left-handedness rates.
- **Extrapolation Method:**
    - Use the last 10 Mean_lh values for early 1900s rates.
    - Use the first 10 Mean_lh values for late 1900s rates.
    - Convert these percentages into fractions (divide by 100).

Program Usage:

- **Libraries:** NumPy (data handling), Bayesian methods (statistical modeling).

Task: Estimate P(LH|A) using extrapolated left-handedness rates for early and late 1900s, then create a function to return these probabilities for given ages of death.

Sample program:

```python
# import library
# ... YOUR CODE FOR TASK 3 ...

# 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 rates of left-handedness.
    Inputs: numpy array of ages of death, study_year
    Returns: probability of left-handedness given that subjects died in `study_year` at ages `ages_of_death` """

    # Use the mean of the 10 last and 10 first points for left-handedness rates before and after the start
    early_1900s_rate = ...
    late_1900s_rate = ...
    middle_rates = lefthanded_data.loc[lefthanded_data['Birth_year'].isin(study_year - ages_of_death)]['Mean_lh']
    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 store the results
    # extract rate of left-handedness for people of ages 'ages_of_death'
    P_return[ages_of_death > oldest_age] = ...
    P_return[ages_of_death < youngest_age] = ...
    P_return[np.logical_and((ages_of_death <= oldest_age), (ages_of_death >= youngest_age))] = middle_rates / 100

    return P_return
```

Analysis and Interpretations:

```
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 rates of left-handedness.
    Inputs: numpy array of ages of death, study_year
    Returns: probability of left-handedness given that subjects died in `study_year` at ages `ages_of_death` """

   # Use the mean of the 10 last and 10 first points for left-handedness rates before and after the start
    early_1900s_rate = lefthanded_data['Mean_lh'].iloc[-10:].mean() / 100
    late_1900s_rate = lefthanded_data['Mean_lh'].iloc[:10].mean() / 100

    # Extract rates for the middle years
    middle_birth_years = study_year - ages_of_death
    middle_rates = lefthanded_data.loc[ lefthanded_data['Birth_year'].isin(middle_birth_years)]['Mean_lh']

    # The youngest and oldest age based on the study year and dataset
    youngest_age = study_year - 1986 + 10  # youngest recorded
    oldest_age = study_year - 1986 + 86  # oldest recorded

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

    return P_return
```

Calculating: P(LH|A) for sample ages:

```
# Calculate P(LH | A) for ages in the study year 1990
ages_of_death = np.array([20, 30, 40, 70, 80])
P_LH_A = P_lh_given_A(ages_of_death, study_year=1990)
print(P_LH_A)
```

Results:

P(LH|A):

```
[0.13103212 0.12210316 0.13089713 0.0774286  0.07364236]
```

Bayesian methods are applied to determine the likelihood of being left-handed at different ages of death.

4.4 Death Distribution in the U.S.

Analyzing **U.S. death distribution data from 1999** to estimate the probability of dying at a given age. By normalizing the death counts, we can treat them as a probability distribution representing the likelihood of death at each age. The dataset contains **age** and **the number of people who died at that age**, split by sex. We will clean, process, and visualize this data

Key Points:

- **Objective:** Estimate the probability of dying at age AAA using death distribution data.
- **Normalization:** Convert raw death counts into a probability distribution.
- **Dataset Details:**
  - **First column:** Age
  - **Other columns:** Number of deaths for each age
- **Steps:**
  - Load the dataset (handle tab-separated values and skip header rows).
  - Drop NaN values in the **Both Sexes** column.
  - Plot deaths vs. age using `.plot()`.

Task: Load, clean, and plot U.S. **death distribution data (1999)** to estimate the probability of dying at each age.

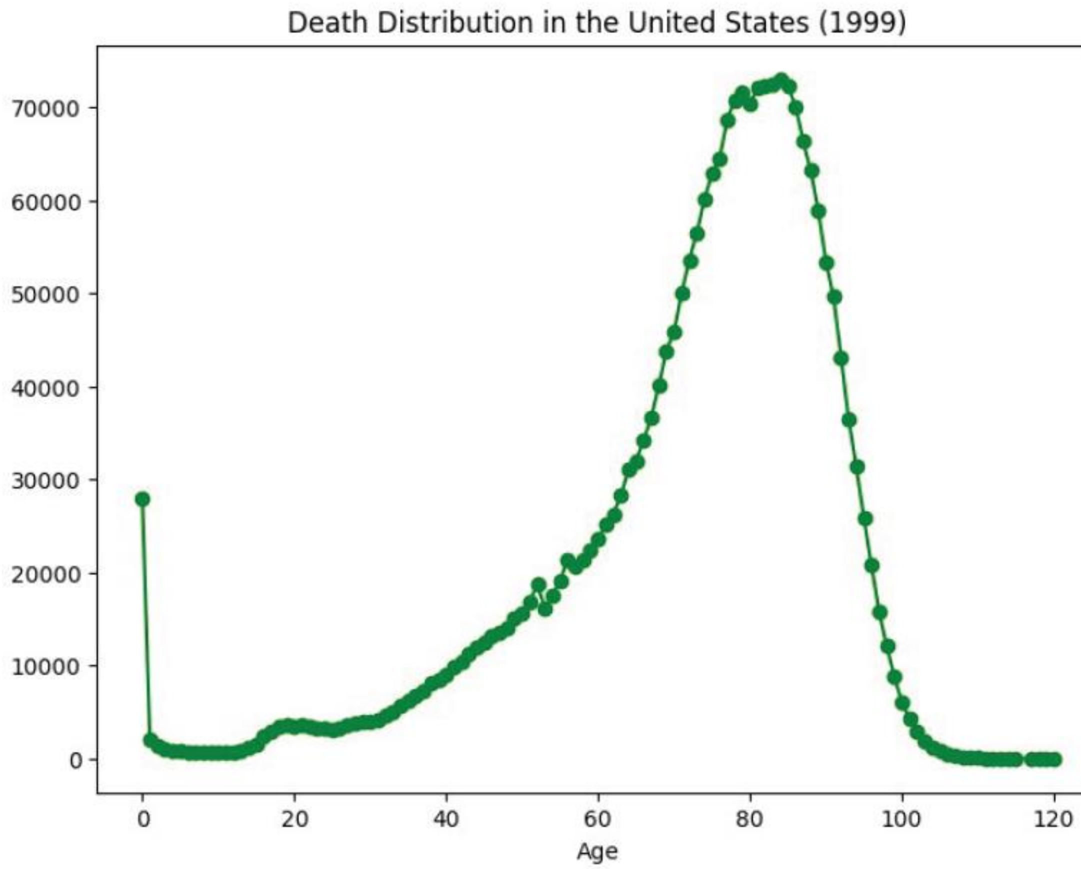
Sample program:

```
# Death distribution data for the United States in 1999
data_url_2 =
"https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807f4e29f7
5f18ec/raw/62f3ec07514c7e31f5979beeca86f19991540796/
cdc_vs00199_table310.tsv"

# load death distribution data
# ... YOUR CODE FOR TASK 4 ...

# drop NaN values from the `Both Sexes` column
# ... YOUR CODE FOR TASK 4 ...

# plot number of people who died as a function of age
fig, ax = plt.subplots()
ax.plot(..., ..., data = ..., marker='o') # plot 'Both Sexes' vs.
'Age'
ax.set_xlabel(...)
ax.set_ylabel(...)
```

Analysis and Interpretations:

```python
import pandas as pd
import matplotlib.pyplot as plt

# Death distribution data for the United States in 1999
data_url_2 =
"https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807f4e29f75f18ec/raw/62f3ec07514c7e31f5979beeca86f19991540796/cdc_vs00199_table310.tsv"

# Load the death distribution data
death_distribution_data = pd.read_csv(data_url_2, sep='\t', skiprows=[1])

# Drop NaN values from the 'Both Sexes' column
death_distribution_data = death_distribution_data.dropna(subset=['Both Sexes'])

# Plot the number of people who died as a function of age
fig, ax = plt.subplots(figsize=(8,6))
# plot 'Both Sexes' vs. 'Age'
ax.plot(death_distribution_data['Age'], death_distribution_data['Both Sexes'], marker='o', color='green', label='Both Sexes')
ax.set_xlabel("Age")
ax.set_ylabel("Number of Deaths")
ax.set_title("Death Distribution in the United States (1999)")
plt.show()
```

Results:



Mortality data from the U.S. in 1999 to understand age distribution at death.

4.5 Probability of Left-Handedness at Death

we compute P(LH), the overall probability that a deceased person in the study year was left-handed. This is obtained by **weighting the left-handedness probabilities (**P(LH|A) **for each age by the number of deceased individuals at that age** and then normalizing by the total number of deaths.

Key Points:

- **Objective:** Compute the overall probability of left-handedness in the population of deceased individuals.

Task: Compute the overall probability of left-handedness among deceased individuals by weighting P(LH|A) with death counts and normalizing by total deaths.

Sample program:

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

print(P_lh(death_distribution_data))
```

Analysis and Interpretations:

```
def P_lh(death_distribution_data, study_year=1990):
    """ Overall probability of being left-handed if you died in the study year.
    Input: DataFrame of death distribution data, study year
    Output: P(LH), a single floating point number """
    # multiply number of dead people by P_lh_given_A
    p_list = death_distribution_data['Both Sexes'] * P_lh_given_A(death_distribution_data['Age'].values, study_year=study_year)
    p = p_list.sum() # Calculate the sum of p_list
    return p / death_distribution_data['Both Sexes'].sum() # normalize to total number of people (sum of death_distribution_data['Both Sexes'])

# The overall probability of left-handedness
print(P_lh(death_distribution_data))
```

Results:

P_lh(death_distribution_data) :

```
0.07766387615350638
```

Bayesian computations estimate the likelihood of left-handedness at various ages of death.

4.6 Comparative Analysis of Death Probabilities for Left-handers

The probability of being left-handed or right-handed at death is computed and compared. this task applies **Bayes' Theorem** to compute the probability of dying at age A **given that a person was left-handed**, denoted as P(A|LH).

- **Key Steps:**
  - Compute **P(A)**: P(A) = Number of deaths at age A / Total deaths.
  - Compute **P(LH)** using the function from the overall probability that a deceased person in the study year was left-handed.
  - Compute **P(LH|A)** using the function from for each age by the number of deceased individuals at that age.
  - Apply **Bayes' Theorem** to get P(A|LH).

Formula used :

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

Sample program:

```
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 that you're left-handed """
    P_A = ...
    P_left = ... # use P_lh function to get probability of left-handedness overall
    P_lh_A = ... # use P_lh_given_A to get probability of left-handedness for a certain age
    return P_lh_A*P_A/P_left
```

Analysis and Interpretations:

```
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 that you're left-handed.
    Input:
      - ages_of_death: Array or Series of ages at death
      - death_distribution_data: DataFrame with death distribution data
      - study_year: Year of the study (default: 1990)
    Output:
      - P_A_given_lh: Probability of being age_of_death given left-handed
    """
# Calculate P(A): overall probability of dying at age A
P_A = death_distribution_data.loc[death_distribution_data['Age'].isin(ages_of_death),'Both Sexes'].values / death_distribution_data['Both Sexes'].sum()

# use P_lh function to get probability of left-handedness overall
P_left = P_lh(death_distribution_data, study_year=study_year)  # from Task 5

    # use P_lh_given_A to get probability of left-handedness for a certain age
    P_lh_A = P_lh_given_A(ages_of_death, study_year=study_year)  # from Task 3

    # Apply Bayes' rule  for left-handers
    return P_lh_A * P_A / P_left
```

Results:

```
ages_of_death = np.array([30, 40, 50, 60, 70])  # Example
result = P_A_given_lh(ages_of_death, death_distribution_data)
print(result)
```

```
[0.00263344 0.00636942 0.01005709 0.01212991 0.01909804]
```

4.7 Comparative Analysis of Death Probabilities for Right-handers

The probability of being left-handed or right-handed at death is computed and compared. this task applies **Bayes' Theorem** to compute the probability of dying at age A **given that a person was left-handed**, denoted as P(A|LH). We compare this result with P(A|RH) for right-handed individuals.

Key Steps:

- Compute **P(A)**: P(A) = Number of deaths at age A / Total deaths.
- Compute **P(RH)**:P(RH)=1−P(LH)P(RH) .
- Compute **P(RH|A)**: P(RH|A)=1−P(LH|A)P(RH | A).
- Apply **Bayes' Theorem** to get P(A|RH).

Sample Program:

```
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 that you're right-handed """
    P_A = ...
    P_right = ... # either you're left-handed or right-handed, so P_right = 1 - P_left
    P_rh_A = ... # P_rh_A = 1 - P_lh_A
    return P_rh_A*P_A/P_right
```

Analysis and Interpretations:

```
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 that you're right-handed."""
    """ Input:
      - ages_of_death: Array or Series of ages at death
      - death_distribution_data: DataFrame with death distribution data
      - study_year: Year of the study (default: 1990)
    Output: P_A_given_rh: Probability of being age_of_death given right-handed."""
    # Calculate P(A): overall probability of dying at age A
P_A = death_distribution_data.loc[ death_distribution_data['Age'].isin(ages_of_death),'Both Sexes'].values / death_distribution_data['Both Sexes'].sum()

    # use P_lh function to get probability of left-handedness overall
    P_left = P_lh(death_distribution_data, study_year=study_year)
    P_right = 1 - P_left # either you're left-handed or right-handed, so P_right = 1 - P_left
    P_lh_A = P_lh_given_A(ages_of_death, study_year=study_year) # P_rh_A = 1 - P_lh_A
    # Calculate P(RH | A): probability of being right-handed given age A
    P_rh_A = 1 - P_lh_A

    # Apply Bayes' rule for right-handers
    return P_rh_A * P_A / P_right
```

Results:

```
ages_of_death = np.array([20, 30, 40, 60, 80])  # Example
result_rh = P_A_given_rh(ages_of_death, death_distribution_data)
print(result_rh)
```

```
[0.00137949 0.0015943  0.00356099 0.00964362 0.02954807]
```

4.8 Conditional Probability Distributions

We **visualize the probability of dying at a specific age** given that a person is **left-handed (P(A|LH)) or right-handed (P(A|RH)**. By plotting these distributions over a range of ages (6 to 120), we can observe trends, including the **higher likelihood of left-handers dying at younger ages** compared to right-handers.

Key Steps:

- Compute **P(A|LH) and P(A|RH)**.
- Generate a range of ages (6 to 120).
- Plot **P(A|LH) and P(A|RH) vs. age** using .plot().

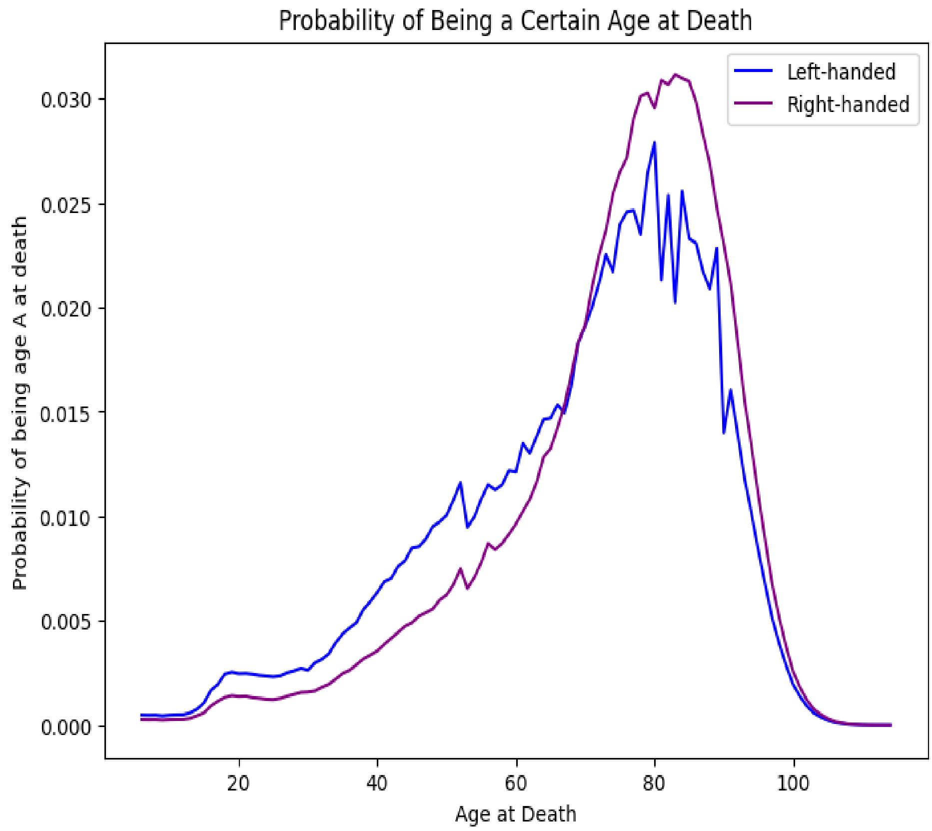
Sample Program:

```
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(...)
right_handed_probability = P_A_given_rh(...)

# 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(..., ..., label = ...)
ax.legend() # add a legend
ax.set_xlabel("Age at death")
ax.set_ylabel(r"Probability of being age A at death")
```

Analysis and Interpretations:

```
import numpy as np
import matplotlib.pyplot as plt

ages = np.arange(6, 115, 1) # make a list of ages of death to plot

# Calculate the probability of being left- or right-handed at death for each
left_handed_probability = P_A_given_lh(ages, death_distribution_data, study_year=1990)
right_handed_probability = P_A_given_rh(ages, death_distribution_data, study_year=1990)

# create a plot of the two probabilities vs. age
fig, ax = plt.subplots(figsize=(8,6))  # Create figure and axis objects
ax.plot(ages, left_handed_probability, label="Left-handed", color="blue")
ax.plot(ages, right_handed_probability, label="Right-handed", color="purple")
ax.legend()
ax.set_xlabel("Age at Death")
ax.set_ylabel(r"Probability of being age A at death")
ax.set_title("Probability of Being a Certain Age at Death")
plt.show()
```

Results:



Graphs illustrate how the probability of being left-handed or right-handed varies by age.

4.9 Average Age at Death for Left- and Right-Handed People

we compute the **mean age at death** for left-handed and right-handed individuals by weighting their respective probability distributions by age and summing the results. This will allow us to compare our results with the original study that found left-handers die, on average, **nine years younger** than right-handers.

- **Key Steps:**
  - Compute **average left-handed age**: Multiply **age values** by **P(A|LH)** and sum using np.nansum().
  - Compute **average right-handed age**:
  - Multiply **age values** by **P(A|RH)** and sum using np.nansum().
  - Compute and **print the difference**.
  - **Round results** for readability.

Sample Program:

```
# 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(...*np.array(...))
average_rh_age =  np.nansum(...*np.array(...))

# print the average ages for each group
# ... YOUR CODE FOR TASK 9 ...

# print the difference between the average ages
print("The difference in average ages is " + str(round(... - ..., 1)) + " years.")
```

Analysis and Interpretations:

```
import numpy as np

# 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
print("Average age at death for left-handed people: " + str(round(average_lh_age, 2)) + " years.")
print("Average age at death for right-handed people: " + str(round(average_rh_age, 2)) + " years.")
difference = round(average_lh_age - average_rh_age, 1)

# print the difference between the average ages
print("The difference in average ages is " + str(difference) + " years.")
```

Results:

```
Average age at death for left-handed people: 67.25 years.
,Average age at death for right-handed people: 72.79 years.
,The difference in average ages is -5.5 years.
```

Results indicate an average age of 67.25 years for left-handed individuals and 72.79 years for right-handed individuals, with a difference of -5.5 years.

4.10 Finding The difference in average ages

We update the analysis to simulate the **study year as 2018** instead of 1990. This will allow us to estimate the **age gap** between left-handed and right-handed individuals based on the **rates of left-handedness** observed in more recent years. The primary difference is that the rates of left-handedness have remained more stable since the 1960s, meaning the gap in handedness between older and younger populations is no longer as pronounced. By recalculating the probabilities for left- and right-handed individuals at death in 2018, we can estimate a **smaller age gap** compared to the previous study.

- **Key Steps:**
    - **Recalculate P(A|LH) and P(A|RH)** for the year 2018.
    - Use the same methodology as **visualize the probability of dying at a specific age** from Task 8 but set **study_year = 2018**.
    - Compare the **mean age at death** for left-handed and right-handed individuals to estimate the **new age gap**.

Sample Program:

```
# Calculate the probability of being left- or right-handed for all ages
left_handed_probability_2018 = ...
right_handed_probability_2018 = ...

# 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.")
```

Analysis and Interpretations:

```
import numpy as np

# Calculate the probability of being left- or right-handed for all ages
left_handed_probability_2018 = P_A_given_lh(ages, death_distribution_data, study_year=2018)
right_handed_probability_2018 = P_A_given_rh(ages, death_distribution_data, study_year=2018)

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

Results:

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

The age gap between left-handed and right-handed individuals using data from the year **2018** instead of 1990. This updated analysis reflects a **smaller age gap of 2.3 years**, compared to the previous 9-year gap reported in earlier studies. The reason for this reduction is that the rates of left-handedness have stabilized since the 1960s, leading to less of a disparity between younger and older generations. This demonstrates how changes in handedness rates over time influence the observed age gap in death.

VI. CONCLUSION AND FUTURE SCOPE

The study concludes that the perceived difference in lifespan between left-handed and right-handed individuals is primarily influenced by demographic shifts rather than an inherent biological disadvantage for left-handed people. Historical societal biases against left-handed individuals likely contributed to fewer elderly left-handers in past generations, leading to a misperception of shorter life expectancy. The statistical analysis, supported by Bayesian modeling, reveals that when accounting for demographic and historical factors, there is no significant difference in lifespan based on handedness. This underscores the importance of considering societal influences in interpreting health data and challenges the misconception that left-handed individuals have a shorter lifespan.

Future Scope

1. Expanded Dataset: Incorporating Additional Demographic Factors

To better understand the nuances of lifespan differences, future studies should include a broader range of demographic factors such as:

- **Geographic Variability:** Research could explore differences in handedness and life expectancy across regions and cultures, considering how societal attitudes toward left-handedness may vary in different parts of the world.
- **Socioeconomic Status (SES):** Given that SES plays a significant role in health outcomes, including factors like access to healthcare, nutrition, and education, future research could investigate how SES intersects with handedness to affect lifespan.
- **Psychosocial Factors:** Mental health, stress, social support, and family structures could all influence lifespan and may be related to whether someone is left-handed or right-handed. Further analysis could explore these intersections.
- **Ethnicity and Genetic Variability:** Genetic factors may also contribute to differences in lifespan, and understanding how handedness interacts with genetic predispositions across various ethnic groups could lead to more personalized insights into longevity.

2. Machine Learning Models: Applying AI-based Predictions for Lifespan Trends

As AI and machine learning technologies evolve, there is potential for harnessing these tools to predict and analyze trends in lifespan based on handedness. Some avenues to explore include:

- **Predictive Modeling:** Machine learning algorithms, such as decision trees, random forests, and deep learning networks, could be used to create sophisticated models that predict lifespan based on a variety of factors, including handedness, health behaviors, and genetic predispositions.
- **Pattern Recognition:** AI can identify patterns in large datasets, helping to uncover previously unnoticed correlations between handedness and lifespan, which may be obscured by the complexities of human biology and social influences.
- **Personalized Health Interventions:** With AI, personalized recommendations for improving health outcomes based on an individual's unique characteristics (such as handedness, lifestyle, and medical history) can be developed, potentially leading to longer, healthier lives.

- **Predictive Analytics for Public Health:** By using large-scale data, machine learning can provide valuable insights into how demographic shifts will affect future generations, helping to forecast trends in life expectancy and health care needs for different groups.

3. Longitudinal Studies: Observing Intergenerational Changes in Handedness and Lifespan

Longitudinal studies, where data is collected from the same individuals over long periods, could offer crucial insights into the relationship between handedness and lifespan. These studies could focus on:

- **Intergenerational Analysis:** Exploring how the lifespan of left-handed and right-handed individuals has changed over generations, especially in light of societal changes. This will allow researchers to track the evolution of societal biases and the potential long-term effects on health outcomes.
- **Tracking Shifts in Societal Attitudes:** Longitudinal research could also investigate how modern society's attitudes toward left-handedness are changing, with fewer stigmas attached to being left-handed today. This shift could contribute to healthier and longer lives for future generations of left-handed individuals.
- **Health Trajectories:** Studying how various life stages (childhood, adolescence, adulthood, and aging) impact left-handed and right-handed individuals differently across generations will offer insights into the role of environmental and social factors on long-term health.
- **Life History Data Integration:** Incorporating health records, educational attainment, career success, family life, and lifestyle factors in longitudinal studies would allow for a more holistic understanding of how handedness relates to an individual's overall health trajectory.

4. Genetic and Epigenetic Research: Unravelling the Biology of Handedness

Future research could integrate genetic and epigenetic studies to further explore whether there are biological mechanisms at play that could explain differences in health outcomes for left- and right-handed individuals:

- **Genetic Factors of Handedness:** While research into the genetics of handedness is still ongoing, exploring how genes related to handedness may also play a role in other health conditions could provide a more comprehensive understanding of health outcomes for left-handed individuals.
- **Epigenetic Modifications:** Environmental factors might cause epigenetic changes that could impact lifespan, potentially linking handedness to broader biological processes that influence aging and disease resistance.
- **Gene-Environment Interactions:** Understanding how genetic predispositions interact with environmental influences could lead to new insights into why societal changes (e.g., improved healthcare access, less bias against left-handed individuals) are leading to healthier lives for future generations.

5. Cross-Disciplinary Collaboration: Integrating Fields like Psychology, Sociology, and Medicine

Collaboration across disciplines will provide a more comprehensive understanding of the relationship between handedness and lifespan:

- **Psychological Research:** Exploring how handedness might influence cognitive development, mental health, and coping mechanisms could be key to understanding why left-handed individuals may have historically been underrepresented in older age groups.
- **Sociological Insights:** Sociologists could study how cultural and social norms surrounding handedness evolve, including how these changes may impact the health and well-being of left-handed individuals across different societies and eras.
- **Medical Research:** Physicians could collaborate on identifying any potential health conditions that disproportionately affect left-handed individuals, whether due to neurological, orthopaedic, or psychological factors, and explore prevention or treatment strategies.

6. Policy Implications: Shaping Health Interventions and Public Awareness

Research findings in this area can influence policy-making related to healthcare, education, and social integration:

- **Public Health Campaigns:** If it is confirmed that societal biases and environmental factors, rather than inherent biological differences, affect the health outcomes of left-handed individuals, targeted public health campaigns could help reduce stigmas, encourage inclusivity, and improve health outcomes for left-handed individuals.
- **Educational Policies:** Creating supportive educational environments that reduce stigma and biases against left-handed individuals could foster better mental health outcomes and increase educational attainment, which in turn may improve lifespan.
- **Healthcare Accessibility:** Healthcare systems may need to adjust their services to ensure that left-handed individuals, particularly those who may have unique needs due to handedness-related physical factors (e.g., joint or musculoskeletal issues), receive the appropriate care.

VII.REFERENCES

- Dataset Source: https://drive.google.com/uc?export=download&id=1gSjYHJ8OPM9HMd3prr7XuhvSWWGKYZNE
- Python Documentation (NumPy, pandas, Matplotlib, Seaborn, PyMC3).
- National Geographic Society (1986), https://www.nationalgeographic.com/
- Centres for Disease Control and Prevention (CDC) - National Vital Statistics System (1999), https://www.cdc.gov/nchs/nvss/index.html
- Pew Research Centre (2019).
- NCBI - Bayesian Statistics in Public Health.
- Statistical Methods and Bayesian Analysis – Wikipedia https://en.wikipedia.org/wiki/Bayes%27_theorem