Data Analytics Traineeship Report - March 2024

Analysis of the Death Age Difference Between Right-Handers and Left-Handers

MedTourEasy

Namrata Muralidharan
27.03.2024

ACKNOWLEDGEMENTS

The traineeship opportunity that I had with MedTourEasy was a wonderful chance to learn and understand the intricacies of Bayesian statistics and Python in Data Science for personal as well as professional development.

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.

I would also like to thank the team of MedTourEasy and my colleagues who made the working experience productive and very conducive.

CONTENTS

1. ABSTRACT

Handedness is an individual's preferential use of one hand, known as the dominant hand, due to it being stronger, faster or more dextrous. It is the tendency to be more skilled and comfortable using one hand instead of the other for tasks such as writing and throwing a ball. The other hand, comparatively often the weaker, less dextrous or simply less subjectively preferred, is called the non-dominant hand. Although the percentage varies worldwide, in Western countries, 85 to 90 percent of people are right-handed and 10 to 15 percent of people are left-handed.

Several studies in the 1990s have claimed left-handers die earlier than right-handers due to a variety of factors ranging from elevated accident risk to developmental disabilities. Left-handers were said to die approximately nine years earlier than right-handers on average. However, reported rates of left-handedness have increased from just 3% in the early 1900s to about 11% today.

More recent studies have disputed this claim, as the social stigma of left-handedness would have prevented individuals from reporting their hand preference accurately, thus skewing the data, especially in older populations. Recent studies also indicate that the rates of left-handedness are not a factor of younger death age but rather of the year of birth. If the same study was conducted today, there would be a shifted version of the same distribution as a function of age. This project investigates the effect this changing rate has on the apparent mean age of death of left-handed people.

2. SUMMARY

2.1 About the Company

MedTourEasy, an online medical tourism marketplace, provides you with the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on your specific health needs, and affordable care while meeting the quality standards that you expect to have in healthcare.

2.2 Introduction

In 1991, Coren and Halpern analysed lifespan studies and found that left-handers were underrepresented in the older age groups. They claimed that this was due to the reduced longevity of left-handers, who die around nine years younger than right-handers. They argued that this was due to environmental factors such as elevated accident susceptibility. Several other factors such as left-handedness as a marker for neurological and birth-related stressors, developmental delays and compromised immune systems were also evidenced. However, several studies have provided other reasons for the underrepresentation.

A National Geographic survey in 1986 resulted in over a million responses that included age, sex, and hand preference for throwing and writing. Gilbert and Wysocki analyzed this data in 1992 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. Based on an analysis of a subgroup of people who throw left-handed but write right-handed, they concluded that this age dependence was primarily due to the changing social acceptability of left-handedness.

2.3 Objectives and Deliverables

The project aims to show that if the same study was done today the difference in the death age would be less than nine years.

This project aims to deliver the following:

1. Use Bayesian statistics to show that the death age gap between left-handers and right-handers would be different in the later years than when the study was conducted.
2. A visualisation depicting the shifting age gap across the years based on extrapolated data.

3. METHODOLOGY

1. The problem statement was defined after preliminary research.
2. Appropriate data sources were gathered to solve the problem.
3. Exploratory data analysis was conducted and the datasets were appropriately cleaned and manipulated for better readability and relevance.
4. The data was analysed using Bayesian statistics
5. Appropriate visualisations were created.
6. The process was documented in a report for better reproducibility.

3.1 Language and Platform Used

Python

Python is a high-level, general-purpose programming language. It was created by Guido van Rossum, and released in 1991. Python's design philosophy emphasizes code readability with the use of significant indentation. It supports multiple programming paradigms, including structured, object-oriented and functional programming. Python has been considered the preferred choice among data scientists throughout the globe. Here are some reasons why:

1. **Easy To Learn:** Being an open-source platform, Python has a simple and intuitive syntax that is easy to learn and read. This makes it a great language for beginners to learn data science.
2. **Cross-Platform:** Being a developer, you don't need to worry about the data types. The reason is, that Python allows developers to run the code on Windows, Mac OS X, UNIX, and Linux.
3. **Portable:** Being an easy & beginner's friendly programming language, Python is highly portable in nature which means that a developer can run their code on different machines without making any further changes.
4. **Extensive Library:** Python has several powerful libraries that make data analysis and visualization easy. Pandas is a library for data manipulation and analysis, NumPy is a library for numerical computation, and Matplotlib is a library for data visualization.

5. **Community Support:** Python has a large and active community that supports and contributes to the development of various libraries and tools for data science. This community has created many useful libraries, including Pandas, NumPy, matplotlib, and SciPy, which are widely used in data science.

Visual Studio Code

Visual Studio Code is a code editor. Visual Studio Code is free and is a source-code editor redefined and optimized for building and debugging modern web and cloud applications. It was developed by Microsoft for Windows, Linux and macOS. Visual Studio Code is a source code editor that can be used with a variety of programming languages, including C, C#, C++, Fortran, Go, Java, JavaScript, Node.js, Python, Rust, and Julia. It is built on the Electron framework, which is used to develop Node.js web applications that run on the Blink layout engine

The Python extension makes VS Code an excellent Python editor and works on any operating system with a variety of Python interpreters. It leverages all of VS Code's power to provide auto-complete and IntelliSense, linting, debugging, and unit testing, along with the ability to easily switch between Python environments, including virtual and conda environments.

Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

4. IMPLEMENTATION

4.1 Identifying/defining the problem

This is the first step wherein the requirements are collected from the clients to understand the deliverables and goals to be achieved after which a problem statement is defined which has to be adhered to while developing the project.

Halpert and Cohern (1991) state that left-handers die nine years younger than right-handers. However, since left-hander reporting is affected negatively by social stigma, this project proposes that the death age gap would lessen over time i.e., the mean death age of left-handers would increase over time.

4.2 Data Collection and Importing

Data collection is a systematic approach for gathering and measuring information from a variety of sources to obtain a complete and accurate picture of an interest area. It helps an individual or organization to address specific questions, determine outcomes and forecast future probabilities and patterns.

The data has been collected through various data sources, mentioned as follows:

1. Death distribution data for the United States from the year 1999. (source)
2. Rates of left-handedness digitized from a figure in ‘Hand preference and age in the United States’ by Gilbert and Wysocki (1992). (source)

Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, and filtered according to the requirements and needs of the project.

4.3 Python Libraries Used

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. Other than pre-compiled codes, a library may contain documentation, configuration data, message templates, classes, values, etc. As we don't need to write the same code again and again for different programs. Python libraries play a very vital role in the fields of Machine Learning, Data Science, Data Visualization, etc.

Pandas: Pandas is a fast, powerful, flexible and easy-to-use open-source data analysis and manipulation tool, built on top of the Python programming language. It allows us to analyze big data and make conclusions based on statistical theories and can clean messy data sets, making them readable and relevant. pandas is an open-source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

NumPy: NumPy is Python's fundamental package for scientific computing. It is a library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. NumPy is used at the core of many popular packages in the world of Data Science and machine learning.

Matplotlib: matplotlib.pyplot is a state-based interface to matplotlib, an object-oriented plotting library. It is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. It also opens figures on your screen and acts as the figure GUI manager.

Seaborn: Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

4.4 Python Code Blocks

In this section, we will look at the code blocks created in the Python notebook to understand the process of analytics in the project.

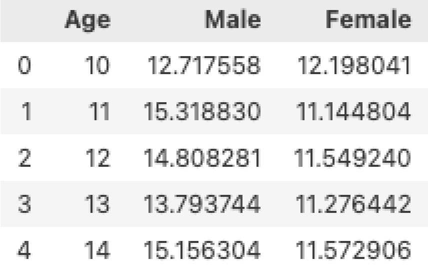
Where are the old left-handed people?

In this introductory code, block, the environment is set up by loading the libraries necessary for the completion of the analysis and importing and exploring the first dataset.

1. First, the aforementioned libraries were loaded and aliased appropriately for more convenient usage. Then the first dataset was loaded in from a URL using the **read_csv()** from pandas. This function imports a CSV file to DataFrame format. This dataset is saved as ‘lefthanded_data’.

2. The data was explored using the following functions to check for data type validity, null values and if any data cleaning/manipulation would be necessary to make it more readable and relevant.

The dataframe contains the following columns:
Age: The age of the individuals during the time of the survey
Male: The average left-handedness rate of males at a certain age
Female: The average left-handedness rate of females at a certain age

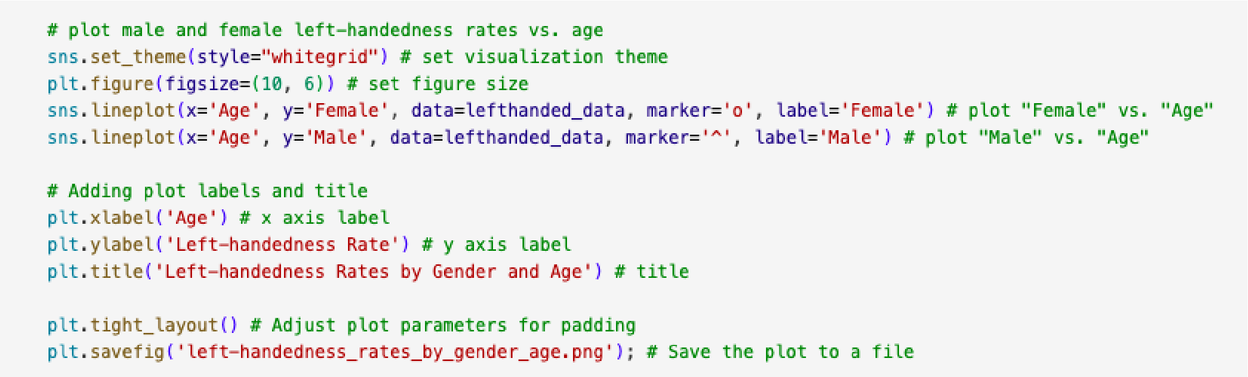


	Age	Male	Female
0	10	12.717558	12.198041
1	11	15.318830	11.144804
2	12	14.808281	11.549240
3	13	13.793744	11.276442
4	14	15.156304	11.572906

3. Since there were no null values or any inconsistencies in the data, no data cleaning is necessary.

4. Functions/methods used:
    1. **read_csv()**: Read a comma-separated values (csv) file into DataFrame.
    2. **.head()**: This function returns the first n rows for the object based on position. By default, it returns five rows.
    3. **.isna()sum()**: Returns the number of missing values in each column.

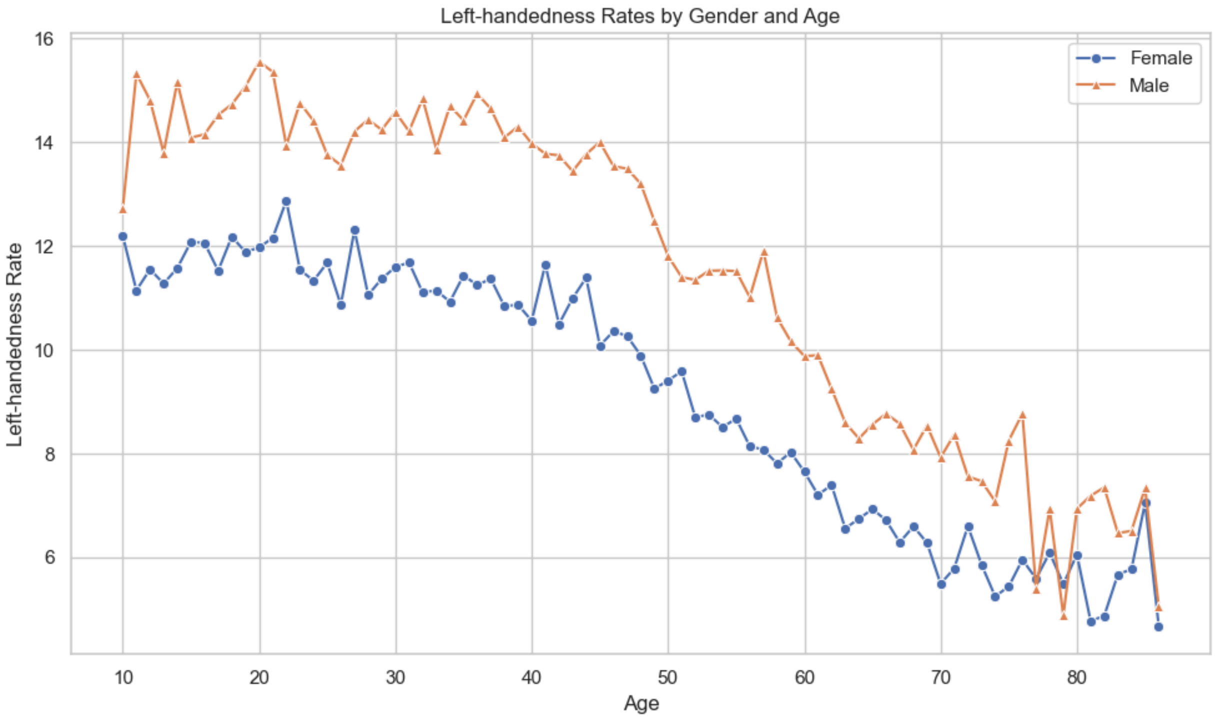
To get a better understanding of left-handedness rates, a graph was plotted using the seaborn package. This graph looks at the left-handedness rates over age and gender. Below is a sample of code used to create visualisations.



```
# plot male and female left-handedness rates vs. age
sns.set_theme(style="whitegrid") # set visualization theme
plt.figure(figsize=(10, 6)) # set figure size
sns.lineplot(x='Age', y='Female', data=lefthanded_data, marker='o', label='Female') # plot "Female" vs. "Age"
sns.lineplot(x='Age', y='Male', data=lefthanded_data, marker='^', label='Male') # plot "Male" vs. "Age"

# Adding plot labels and title
plt.xlabel('Age') # x axis label
plt.ylabel('Left-handedness Rate') # y axis label
plt.title('Left-handedness Rates by Gender and Age') # title

plt.tight_layout() # Adjust plot parameters for padding
plt.savefig('left-handedness_rates_by_gender_age.png'); # Save the plot to a file
```

Note: Instead of using ‘;’ at the end of the code, ‘plt.show()’ can also be used to display the plot.

Implementing the above code results in the following visualisation, which describes left-handedness in males and females over age.



The rates seem to reduce consistently as the age increases across both genders, while males have a higher rate of left-handedness.

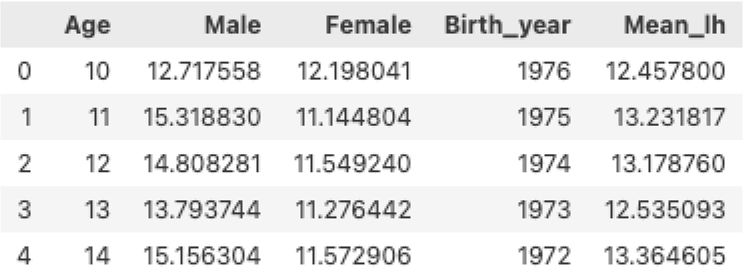
Rates of left-handedness over time

Two more columns are calculated and added to better understand the data:

Birth_year: Since the study was done in 1986, the birth year can be calculated by subtracting the age from 1986.

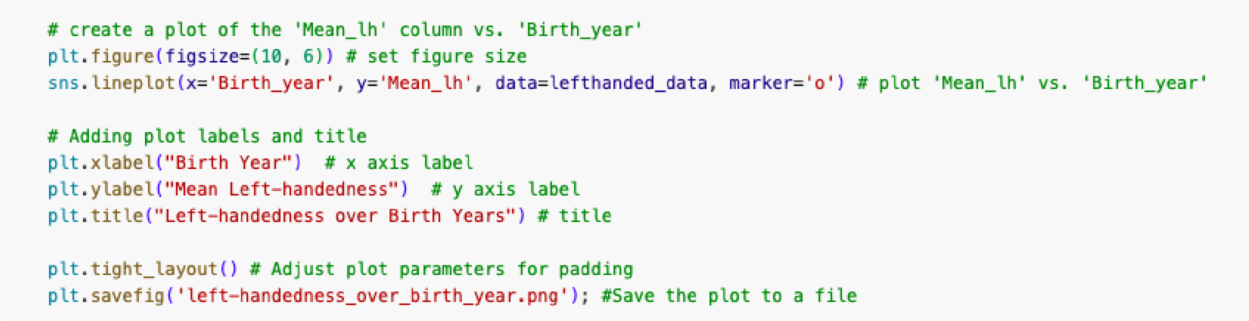
Mean_lh: The mean left-handedness rate for both genders at a certain age. This is calculated by taking the mean of the ‘Male’ and ‘Female’ columns.

Then use .head() to see if the columns have been implemented.



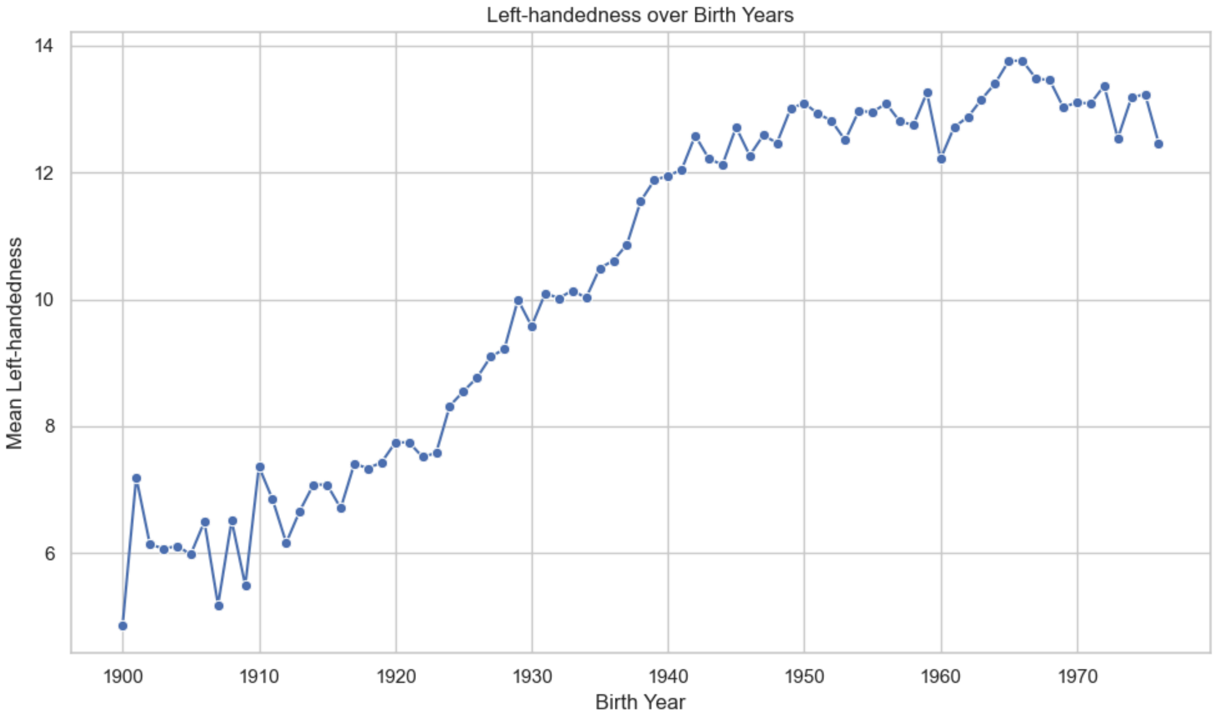
	Age	Male	Female	Birth_year	Mean_lh
0	10	12.717558	12.198041	1976	12.457800
1	11	15.318830	11.144804	1975	13.231817
2	12	14.808281	11.549240	1974	13.178760
3	13	13.793744	11.276442	1973	12.535093
4	14	15.156304	11.572906	1972	13.364605

Using seaborn, plot the two newly created columns against each other i.e., mean left-handedness for both genders over the year of birth. This can be done using the code below.



```
# create a plot of the 'Mean_lh' column vs. 'Birth_year'
plt.figure(figsize=(10, 6)) # set figure size
sns.lineplot(x='Birth_year', y='Mean_lh', data=lefthanded_data, marker='o') # plot 'Mean_lh' vs. 'Birth_year'

# Adding plot labels and title
plt.xlabel("Birth Year")  # x axis label
plt.ylabel("Mean Left-handedness")  # y axis label
plt.title("Left-handedness over Birth Years") # title

plt.tight_layout() # Adjust plot parameters for padding
plt.savefig('left-handedness_over_birth_year.png'); #Save the plot to a file
```

This code will result in the following visualization, which looks at the amount of left-handedness reporting according to the year of birth.



Looking at left-handedness over birth years, it is evident that rates of left-handedness have increased over the years. This could be due to a few reasons, such as:

1. The social stigma around left-handedness decreased over time and more left-handers reported the handedness preference.
2. The stigma is more pronounced in adults than in children.

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 **Bayes' theorem** is needed, a statement about conditional probability which allows us to update our beliefs after seeing the evidence.

The probability of dying at age A given left-handedness can be written as **P(A | LH).**

In the same way, for right-handers, it can be written as **P(A | RH)**.

To find the probability of dying at a certain age A given left-handedness can be written as:

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

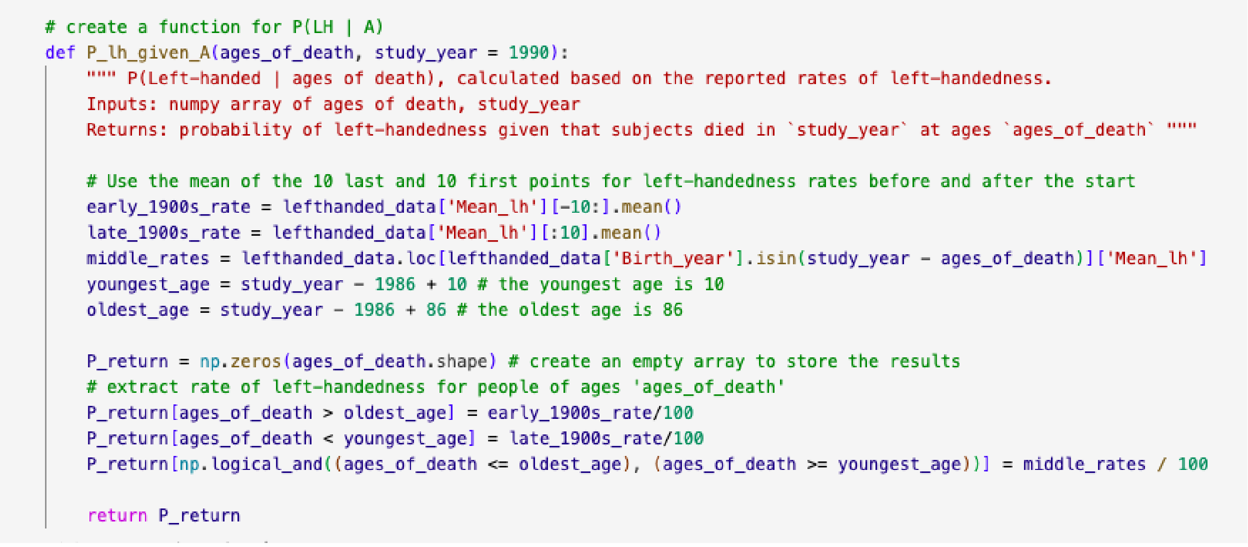
$$P(A \mid LH) = \frac{P(LH|A) * P(A)}{P(LH)}$$

Where:

- **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.
- **P(LH)** is the overall probability of being left-handed.

To calculate P(LH | A) for ages that might fall outside the original data, the data needs to be extrapolated 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 to calculate the mean and extrapolate the rates on each end. The number of points used for this is arbitrary, since the data looks comparatively until about 1910, 10 points are considered.

This is calculated by creating a function ‘P_lh_given_A’ which takes two arguments namely ‘ages_of_death’ (a NumPy array of ages) and ‘study_year’ which we specify to ‘1990’. It returns an output ‘P_return’ which is the probability of left-handedness given that subjects died in ’study_year’ at ages ‘ages_of_death’. This can be created in Python using the following code.



```
# 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'][-10:].mean()
    late_1900s_rate = lefthanded_data['Mean_lh'][:10].mean()
    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] = 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 >= youngest_age))] = middle_rates / 100

    return P_return
```

When do people normally die?

To estimate the probability of living to an age A, a distribution of death can be created. This can be done using data that gives the number of people who died in a given year and their ages at death. If the numbers are normalized to the total number of people who died, this data could be a probability distribution that gives the probability of dying at age A. The death distribution data used in this project is from the entire population of the USA for the year 1999.

1. This dataset is also loaded the same way as the previous one, with some additions as this is a TSV file. This dataset is saved as ‘death_distribution_data’.

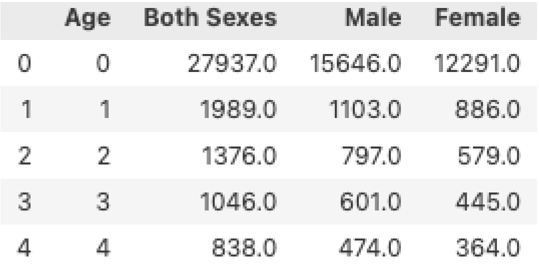
This dataframe contains the following tables:

Age: The age at death

Both Sexes: Total number of deaths at a given age

Male: Male deaths at a given age

Female: Female deaths at a given age



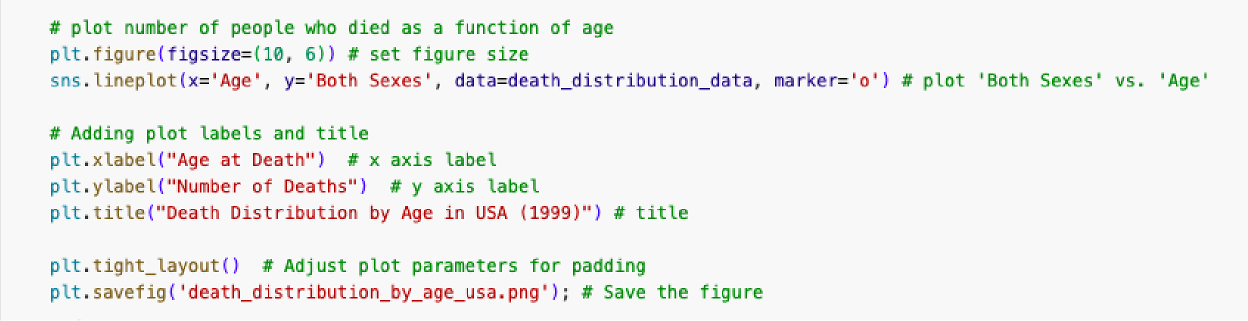
	Age	Both Sexes	Male	Female
0	0	27937.0	15646.0	12291.0
1	1	1989.0	1103.0	886.0
2	2	1376.0	797.0	579.0
3	3	1046.0	601.0	445.0
4	4	838.0	474.0	364.0

2. Then the rows containing NaN values from the ‘Both Sexes’ are removed as part of data cleaning.

Functions Used:

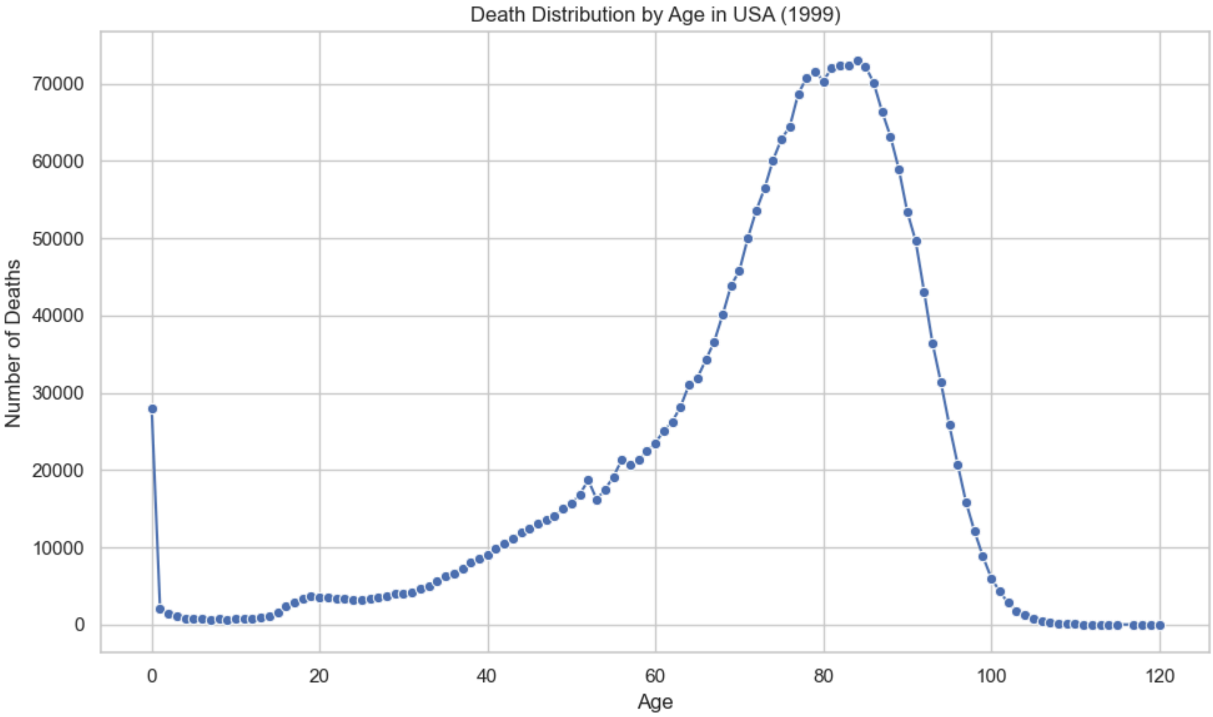
dropna(): removes the rows that contain NULL values. ’Subset’ is an array which limits the dropping process to passed rows/columns through the list. Here we specify (subset = ‘Both Sexes’)

Using seaborn, plot the total deaths against age at death. This can be done using the following code:



```
# plot number of people who died as a function of age
plt.figure(figsize=(10, 6)) # set figure size
sns.lineplot(x='Age', y='Both Sexes', data=death_distribution_data, marker='o') # plot 'Both Sexes' vs. 'Age'

# Adding plot labels and title
plt.xlabel("Age at Death")  # x axis label
plt.ylabel("Number of Deaths")  # y axis label
plt.title("Death Distribution by Age in USA (1999)") # title

plt.tight_layout()  # Adjust plot parameters for padding
plt.savefig('death_distribution_by_age_usa.png'); # Save the figure
```

This results in the following visualisation, which shows the number of deaths recorded at every age ranging from 0 to 120 in the USA in 1999.



Death Distribution by Age in USA (1999)

The trend looks consistent with the usual and conventional distribution of death ages.

The overall probability of left-handedness

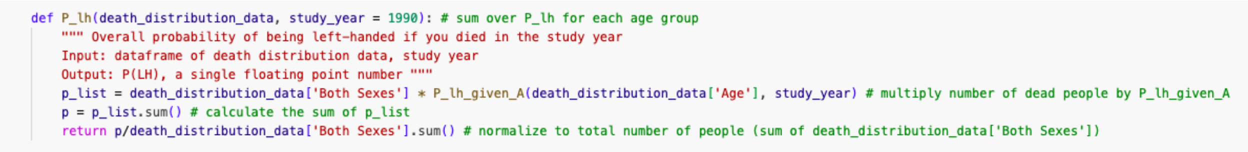
In the previous code block, P(A) was determined, and now P(LH) is next. P(LH) is the probability that a person who died in a particular study year is left-handed, assuming nothing else is known about them. This is the average left-handedness in the population of deceased people, and it can be calculated 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 given as follows:

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

where

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

This is calculated by creating a function ‘P_lh’ which takes two arguments namely ‘death_distribution_data’ dataframe and ‘study_year’ which we specify to ‘1990’ as in the previous function as well. It returns an output of the overall probability of being left-handed if you died in the study year as a single floating point number (**0.07766387615350638**). The code can be expressed as follows:



```
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 = death_distribution_data['Both Sexes'] * P_lh_given_A(death_distribution_data['Age'], study_year) # multiply number of dead people by P_lh_given_A
    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'])
```

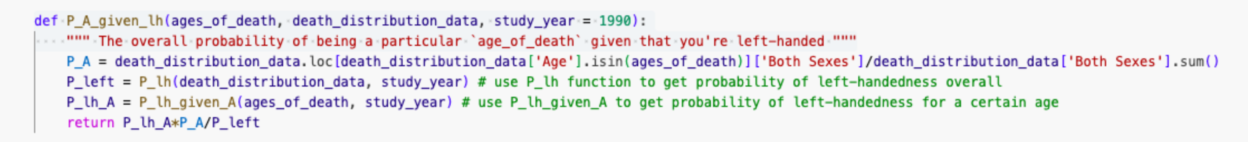
Putting it all together: dying while left-handed (i)

Now all three quantities are calculated: P(A), P(LH), and P(LH | A). Bayes' rule and the three quantities can be used to determine. to get P(A | LH) the probability of being age A at death (in the study year) given that an individual is left-handed. To make this answer meaningful, it must also be compared to P(A | RH), the probability of being age A at death given that an individual is right-handed.

We're calculating the following quantity twice, once for left-handers and once for right-handers. For left-handers, it is calculated using the following equation:

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

This is calculated by creating a function ‘P_A_given_lh’ which takes two arguments namely ‘ages_of_death’ (a NumPy array of ages) and ‘study_year’ which we specify to ‘1990’ as in the previous functions. It returns an output value of the overall probability of being a particular `age_of_death` given that you're left-handed. The code can be expressed as below.



```python
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 = death_distribution_data.loc[death_distribution_data['Age'].isin(ages_of_death)]['Both Sexes']/death_distribution_data['Both Sexes'].sum()
    P_left = P_lh(death_distribution_data, study_year) # use P_lh function to get probability of left-handedness overall
    P_lh_A = P_lh_given_A(ages_of_death, study_year) # use P_lh_given_A to get probability of left-handedness for a certain age
    return P_lh_A*P_A/P_left
```

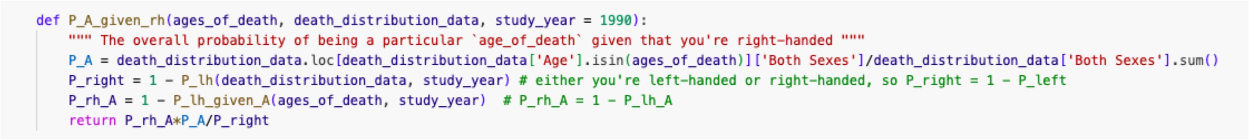
Putting it all together: dying while left-handed

And now for right-handers, it can be calculated using the following formula:

$$\mathbf{P(A \mid RH) = 1 - P(LH \mid A) * \underline{P(A)}}$$

P(LH)

This is calculated by creating a function ‘P_A_given_rh’ which takes two arguments namely ‘ages_of_death’ (a NumPy array of ages) and ‘study_year’ which we specify to ‘1990’ as in the previous functions. It returns an output value of the overall probability of being a particular `age_of_death` given that you're right-handed. The calculations are essentially the same as the previous function except for the first value being ‘1 - P(LH | A)’ instead of simply ‘P(LH | A)’.
This can be written in code as given below.



```
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 = death_distribution_data.loc[death_distribution_data['Age'].isin(ages_of_death)]['Both Sexes']/death_distribution_data['Both Sexes'].sum()
    P_right = 1 - P_lh(death_distribution_data, study_year) # either you're left-handed or right-handed, so P_right = 1 - P_left
    P_rh_A = 1 - P_lh_given_A(ages_of_death, study_year)  # P_rh_A = 1 - P_lh_A
    return P_rh_A*P_A/P_right
```

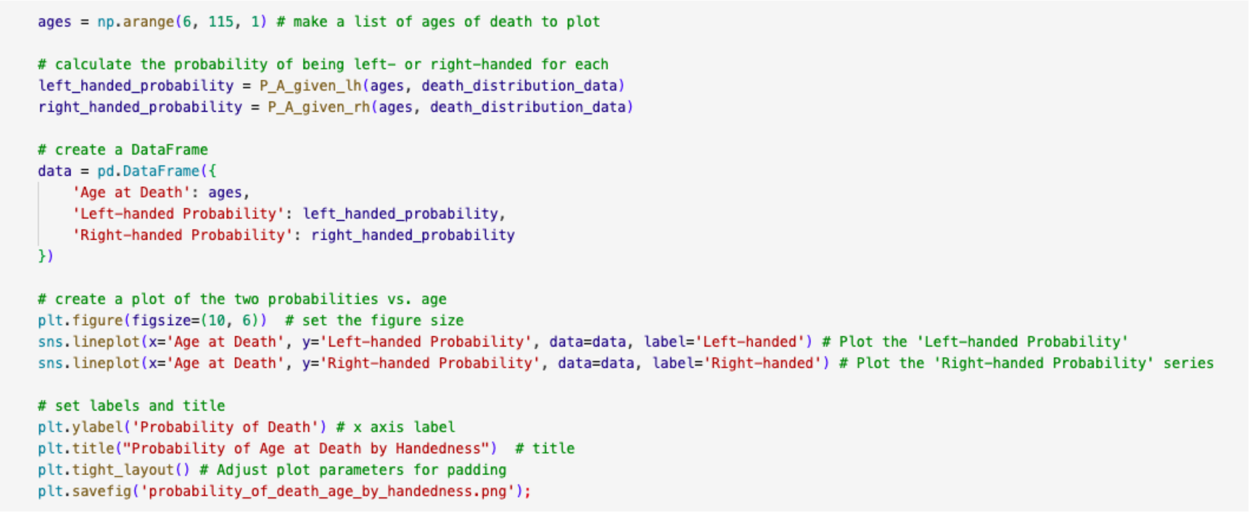
Plotting the distributions of conditional probabilities

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 using seaborn.

1. First, make a list of ages of death to plot. This can be done using the following function:
   np.arange(): Return evenly spaced values within a given interval from the NumPy library. arange can be called with a varying number of positional arguments

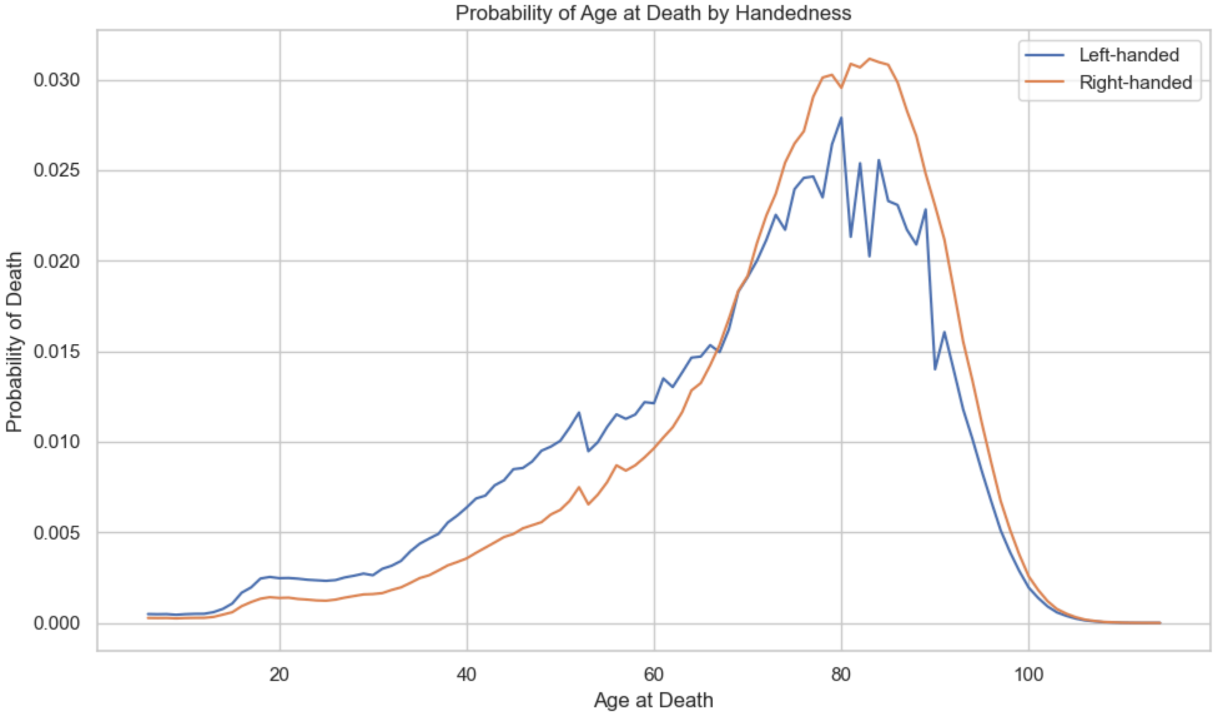
2. Calculate the probability of being left- or right-handed for each using the previous functions ‘**P_A_given_lh**’ and ‘**P_A_given_rh**’
3. Create a new dataframe with the age and left and right-handedness probabilities using the following function:

   pd.dataFrame(): Creates Pandas DataFrame is a structure that contains two-dimensional data and its corresponding labels. It is a part of the Pandas library
4. Finally, plot the dataframe using seaborn. This can be done using the following code.



```
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 DataFrame
data = pd.DataFrame({
    'Age at Death': ages,
    'Left-handed Probability': left_handed_probability,
    'Right-handed Probability': right_handed_probability
})

# create a plot of the two probabilities vs. age
plt.figure(figsize=(10, 6))  # set the figure size
sns.lineplot(x='Age at Death', y='Left-handed Probability', data=data, label='Left-handed') # Plot the 'Left-handed Probability'
sns.lineplot(x='Age at Death', y='Right-handed Probability', data=data, label='Right-handed') # Plot the 'Right-handed Probability' series

# set labels and title
plt.ylabel('Probability of Death') # x axis label
plt.title("Probability of Age at Death by Handedness")  # title
plt.tight_layout() # Adjust plot parameters for padding
plt.savefig('probability_of_death_age_by_handedness.png');
```

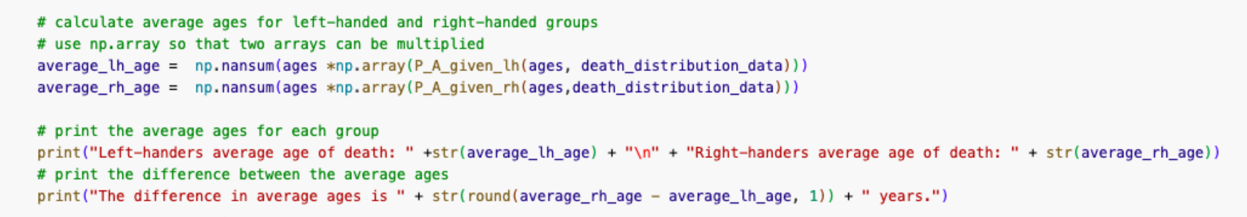
This would result in the visualisation below, which shows the probability of dying at a certain age based on handedness.



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.

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. This can be done with the following code.



```python
# 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(P_A_given_lh(ages, death_distribution_data)))
average_rh_age =  np.nansum(ages *np.array(P_A_given_rh(ages,death_distribution_data)))

# print the average ages for each group
print("Left-handers average age of death: " +str(average_lh_age) + "\n" + "Right-handers average age of death: " + str(average_rh_age))
# print the difference between the average ages
print("The difference in average ages is " + str(round(average_rh_age - average_lh_age, 1)) + " years.")
```

We get the following results for the year 1990:

Left-hander's average age of death: 67.24503662801027

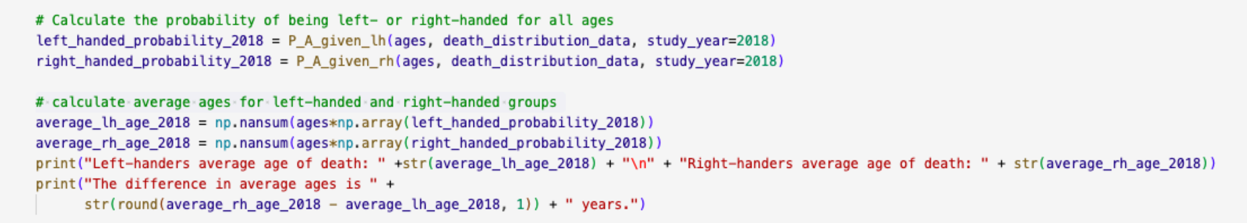
Right-hander’s average age of death: 72.79171936526477

The difference in average ages is 5.5 years.

We get 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. 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.

Final comments

To conclude, let's calculate the age gap we'd expect if we did the study in 2018 instead of in 1990. The same code as above can be used, after calculating the probability for the year 2018.



```python
# 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("Left-handers average age of death: " +str(average_lh_age_2018) + "\n" + "Right-handers average age of death: " + str(average_rh_age_2018))
print("The difference in average ages is " +
      str(round(average_rh_age_2018 - average_lh_age_2018, 1)) + " years.")
```

We get the following results for the year 2018:

Left-handers average age of death: 70.28773299940532

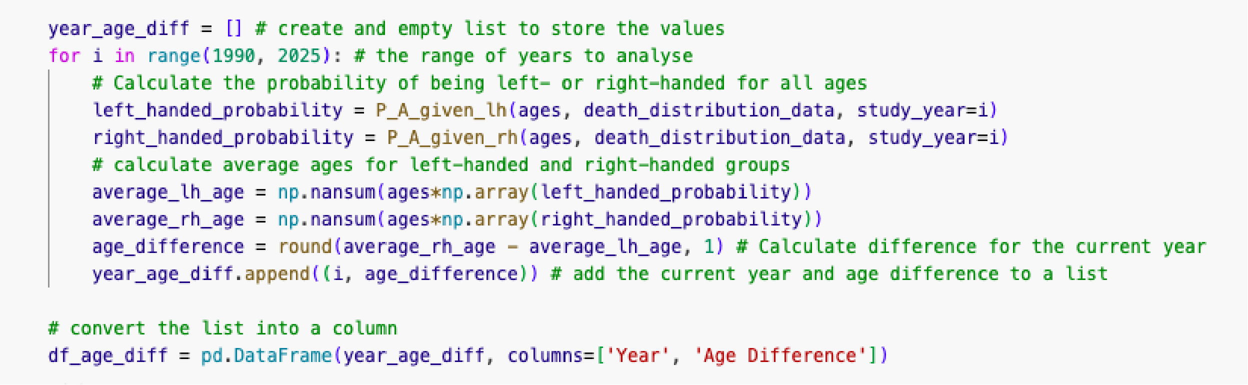
Right-handers average age of death: 72.62899693809848

The difference in average ages is 2.3 years.

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.

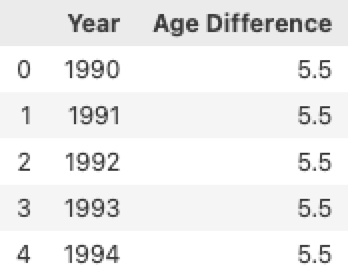
Finally, let's look at an estimation of how the age gap closes over time and see what the death age gap between left-handers and right-handers will be in 2024. We can do this by looking at a line chart depicting the age gap across the years.

First, we need to calculate the values to be plotted. This can be built on the previous code blocks with a few modifications to return a list of values regarding the death age difference.



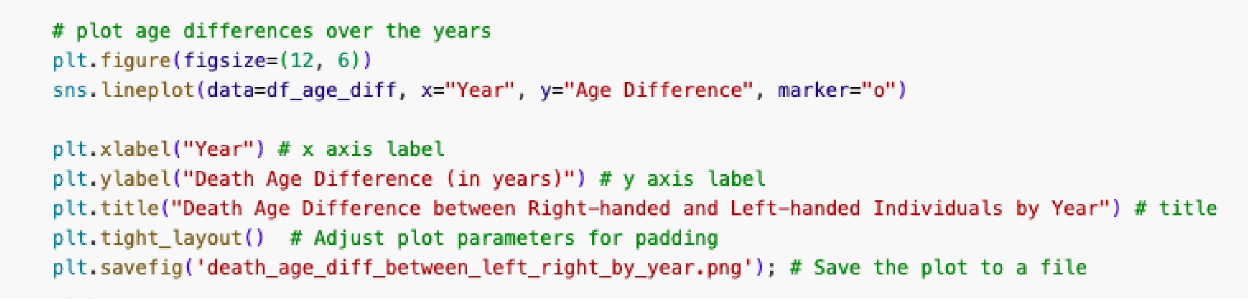
```
year_age_diff = [] # create and empty list to store the values
for i in range(1990, 2025): # the range of years to analyse
    # Calculate the probability of being left- or right-handed for all ages
    left_handed_probability = P_A_given_lh(ages, death_distribution_data, study_year=i)
    right_handed_probability = P_A_given_rh(ages, death_distribution_data, study_year=i)
    # calculate average ages for left-handed and right-handed groups
    average_lh_age = np.nansum(ages*np.array(left_handed_probability))
    average_rh_age = np.nansum(ages*np.array(right_handed_probability))
    age_difference = round(average_rh_age - average_lh_age, 1) # Calculate difference for the current year
    year_age_diff.append((i, age_difference)) # add the current year and age difference to a list

# convert the list into a column
df_age_diff = pd.DataFrame(year_age_diff, columns=['Year', 'Age Difference'])
```

Now we have a dataframe with the year and the death age difference between left-handers and right-handers. The first five rows should look like this.



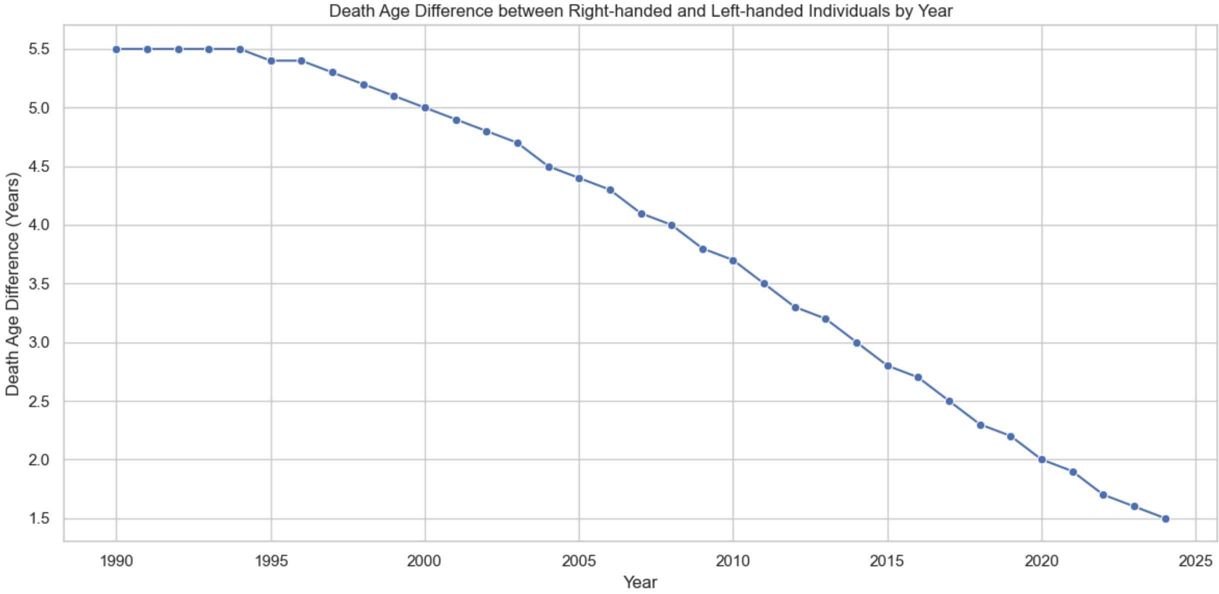
	Year	Age Difference
0	1990	5.5
1	1991	5.5
2	1992	5.5
3	1993	5.5
4	1994	5.5

Now, this data can be plotted using the code below.



```
# plot age differences over the years
plt.figure(figsize=(12, 6))
sns.lineplot(data=df_age_diff, x="Year", y="Age Difference", marker="o")

plt.xlabel("Year") # x axis label
plt.ylabel("Death Age Difference (in years)") # y axis label
plt.title("Death Age Difference between Right-handed and Left-handed Individuals by Year") # title
plt.tight_layout()  # Adjust plot parameters for padding
plt.savefig('death_age_diff_between_left_right_by_year.png'); # Save the plot to a file
```

This would result in the following visualization, which looks at the death age gap between left-handers and right-handers over the years, ranging from 1990 to 2024.



It’s important to note that this is an estimation based on extrapolation and may vary from real-world data.

It is very clear that over the years the gap can get progressively smaller, with the gap in 2024 estimated to only be 1.5 years.

5. CONCLUSION

Left-handers have faced several unfair discriminations and biases over the past decades, but current studies are disproving many of the myths associated with left-handedness. While many things are designed for right-handers such as knives and scissors, it is unlikely that the risk of accidents for left-handers is so great that it affects mortality. Several studies over the past few years have shown this gap to be nothing more than a myth. A recent study by Ferres et al., (2023) also asserts that" cultural pressures may have caused an underestimation of the true rate of left-handedness." As shown in this project, the death age gap of nine years between left-handers and right-handers has most likely reduced over the years due to higher self-reporting of left-handedness. The social stigma associated with left-handedness was much higher in the 1900s compared to the 21st century. This must have led to inconsistencies in the reported instances of left-handedness. If the same survey and study were conducted now, it is most likely that the death age difference would be much less pronounced than it was in the 1990s.

6. LIMITATIONS AND FUTURE SCOPE

6.1 Limitations

The difference in death ages of left and right-handers calculated in this project is less than the 9-year gap measured in the study. Some of the following approximations may be the cause:

1. The death distribution data used was almost ten years after the study (1999 instead of 1991).
2. The dataset analysed in this study was much larger than the original study. The death data used was from the whole country of the United States instead of only California which was used in the original study. The was much larger than the original study.
3. The left-handedness survey results were extrapolated to older and younger age groups, but it is possible our extrapolation was not close enough to the true rates for those ages.
4. The handedness rates are mostly self-reported, which can be affected by various factors and differ from actual hand preferences.

6.2 Future Scope

The variability expected in the age difference purely because of random sampling can be calculated in the future. Several questions regarding variability could be answered, such as:

1. If a smaller sample of recently deceased people was taken and assigned handedness with the probabilities of the survey, what does that distribution look like?
2. How often would an age gap of nine years be encountered using the same data and assumptions?

These answers could be calculated using the tools of random sampling.

7. REFERENCES

1. Coren, Stanley & Halpern, Diane. (1991). Left-Handedness: A marker for decreased survival fitness. Psychological bulletin. 109. 90-106. 10.1037//0033-2909.109.1.90.
2. Gilbert AN, Wysocki CJ. Hand preference and age in the United States. Neuropsychologia. 1992 Jul;30(7):601-8. doi: 10.1016/0028-3932(92)90065-t. PMID: 1528408.
3. Ferres, J.L., Nasir, M., Bijral, A. et al. Modeling to explore and challenge inherent assumptions when cultural norms have changed: a case study on left-handedness and life expectancy. Arch Public Health 81, 137 (2023). https://doi.org/10.1186/s13690-023-01156-6
4. Death distribution data for the United States from the year 1999. (source)
5. Rates of left-handedness digitized from a figure in ‘Hand preference and age in the United States’ by Gilbert and Wysocki (1992). (source)