**Final Report of Traineeship Program 2020**

*On*

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

**MEDTOUREASY**



20th November 2023

Aditya Agrawal



# ACKNOWLEDGMENTS

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 also, 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.

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



# TABLE OF CONTENTS

Acknowledgments i

|  |  |
| --- | --- |
| **Sr. No.** | **Topic** |
| **1** | Introduction |
|  | 1.1 About the Company |
|  | 1.2 About the Project |
| **2** | Methodology |
|  | 2.1 Flow of the Project |
|  | 2.2 Language and Platform Used |
| **3** | Implementation |
|  | 3.1 Ask |
|  | 3.2 Prepare |
|  | 3.3 Process |
|  | 3.4 Analyze |
|  | 3.5 Share and Act |
| **4** | References |



## **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.

## **About the Project**

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



# METHODOLOGY

## **Flow of the Project**

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

Ask

Act

Share

Analyze

Process

Prepare



## **Language and Platform Used**

## **Language:** Python 3

## **Frameworks:** Pandas, Matplotlib, NumPy

## **IDE:** Jupyter Notebook



# IMPLEMENTATION

## **Ask**

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 development of the project.

## **Prepare**

Data collection is a systematic approach for gathering and measuring information from a variety of sources in order 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 COVID-19 has been collected through various GitHub repositories, mentioned as follows:

"https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw /aec88b30af87fad8d45da7e774223f91dad09e88/lh\_data.csv"

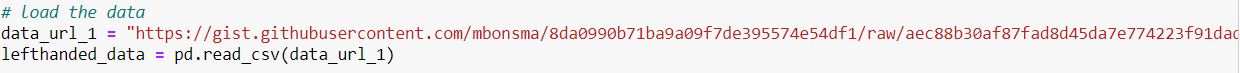
Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, filtered according to the requirements and needs of the project.

Packages Used:

**Pandas**: - Pandas Data Frame is a two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).

**Functions Used:**

**.read\_csv ():** It is a wrapper function for read data frame that mandates a comma as separator and uses the input file's first line as header that specifies the table's column names. Thus, it is an ideal candidate to read CSV files.





## **Process**

***“Quality data beats fancy algorithms”***

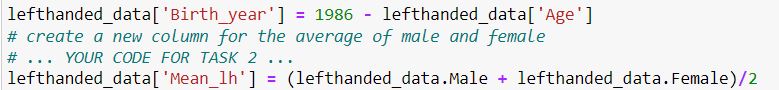
Data is the most imperative aspect of Analytics and Machine Learning. Everywhere in computing or business, data is required. But many a times, the data may be incomplete, inconsistent or may contain missing values when it comes to the real world. If the data is corrupted then the process may be impeded or inaccurate results may be provided. Hence, Data cleaning is considered a foundational element of the basic data science.

Data Cleaning means the process by which the incorrect, incomplete, inaccurate, irrelevant or missing part of the data is identified and then modified, replaced or deleted as needed.

In our data set we have added 2 new columns namely Birth\_year and Mean\_lh.

lefthanded\_data['Birth\_year'] = 1986 - lefthanded\_data ['Age']

lefthanded\_data['Mean\_lh'] = (lefthanded\_data ['Male'] + lefthanded\_data ['Female']) / 2



In the data set death\_distribution Nan values are dropped from `Both Sexes` column Below is the code for the reference:

C:\Users\hp\Desktop\Coursera\Capture.JPG



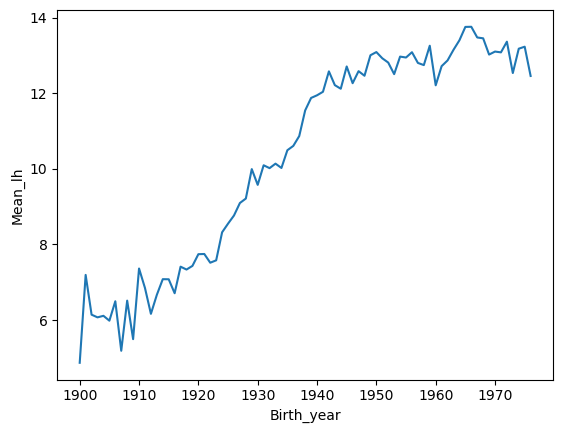
## **Analyze**

## **Exploratory Data analysis (EDA)**

Calculated the mean of left-handed % for Birth year.

lefthanded\_data['Mean\_lh'] = (lefthanded\_data['Male'] + lefthanded\_data['Female']) / 2

Create visualizations to compare the distributions of age at death for both groups.





**Bayesian Analysis**

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

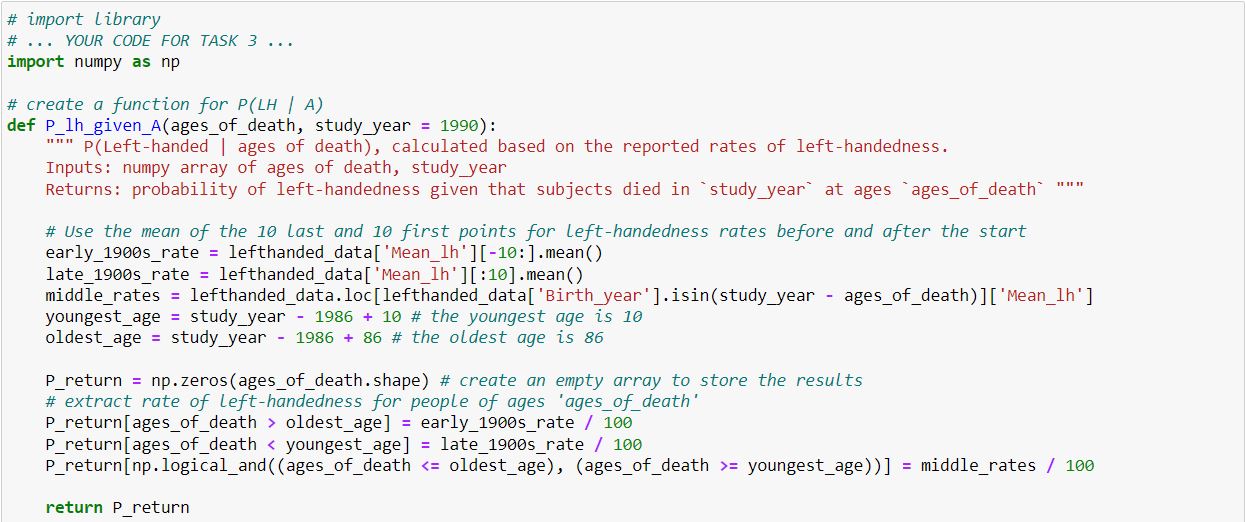
We want to calculate the probability of dying at age A given that you're left-handed. Let's write this in shorthand as P(A | LH). We also want the same quantity for right-handers: P(A | RH).

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

𝑃(𝐴|𝐿𝐻)=𝑃(𝐿𝐻|𝐴)𝑃(𝐴)𝑃(𝐿𝐻)

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

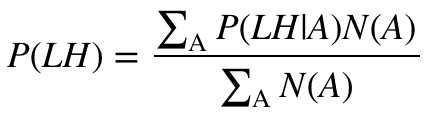
To calculate P(LH | A) for ages that might fall outside the original data, we will need to extrapolate the data to earlier and later years. Since the rates flatten out in the early 1900s and late 1900s, we'll use a few points at each end and take the mean to extrapolate the rates on each end. The number of points used for this is arbitrary, but we'll pick 10 since the data looks flatish until about 1910.

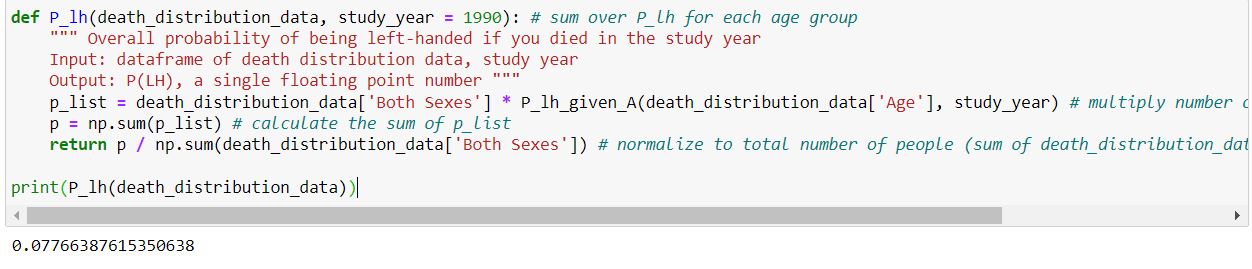




**The Overall probability of left\_handness**

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





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

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

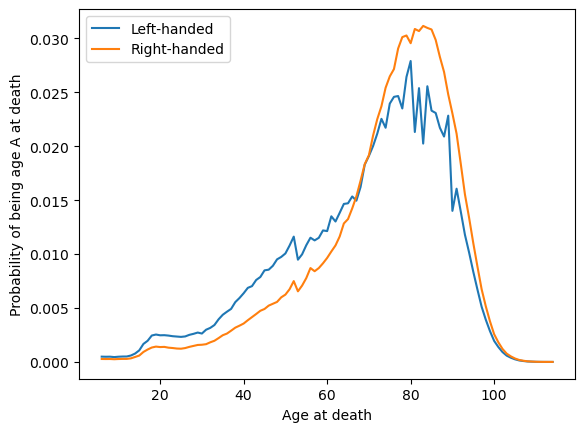
𝑃(𝐴|𝐿𝐻)=𝑃(𝐿𝐻|𝐴)𝑃(𝐴)𝑃(𝐿𝐻)



**Plotting the distributions of conditional probabilities**

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

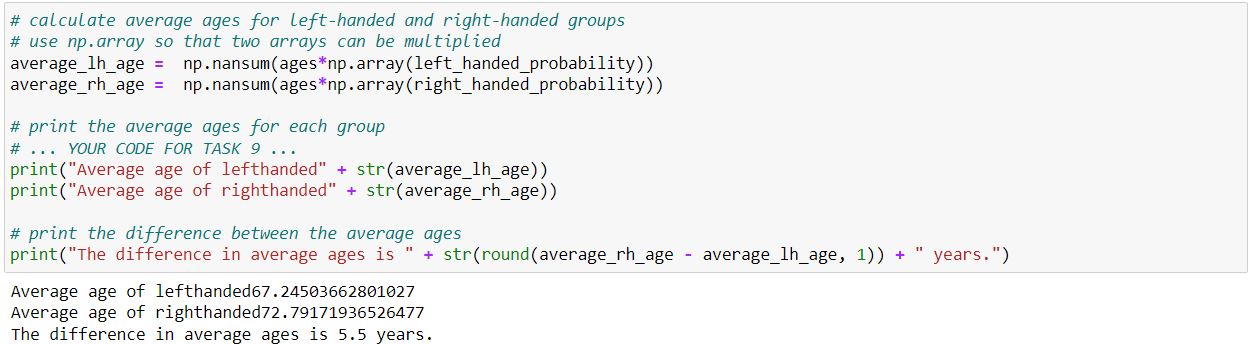
Notice that the left-handed distribution has a bump below age 70: of the pool of deceased people, left-handed people are more likely to be younger.



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.

Average age of left-handed people at death=∑𝐴𝐴(𝐴|𝐿𝐻)

Average age of right-handed people at death=∑𝐴𝐴(𝐴|𝑅𝐻)





## **Share and Act**

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

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

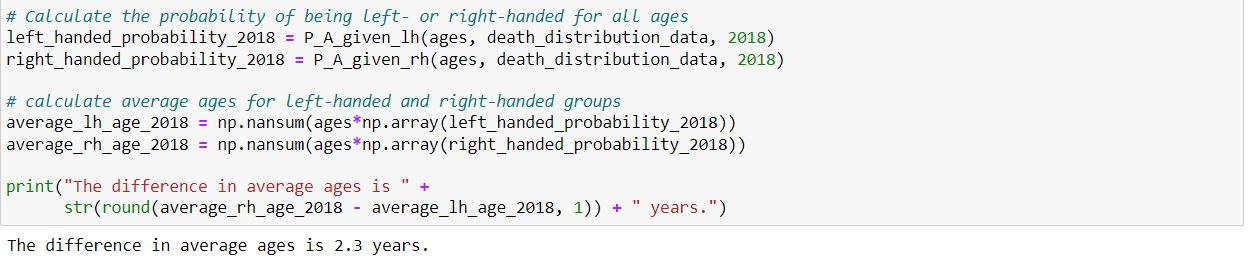
We used death distribution data from almost ten years after the study (1999 instead of 1991), and we used death data from the entire United States instead of California alone (which was the original study).

We extrapolated the left-handedness survey results to older and younger age groups, but it's possible our extrapolation wasn't close enough to the true rates for those ages.

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

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







# REFERENCES

### Data Collection

The following websites have been referred to obtain the input data and statistics:

* + - 1. <https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv>
      2. <https://gist.githubusercontent.com/mbonsma/2f4076aab6820ca1807f4e29f75f18ec/raw/62f3ec07514c7e31f5979beeca86f19991540796/cdc_vs00199_table310.tsv>