22/10/2024

Final Report of Traineeship Program 2024

On

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

MEDTOUREASY

Chandrakanth Dahima

About the company :

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

Project Description:

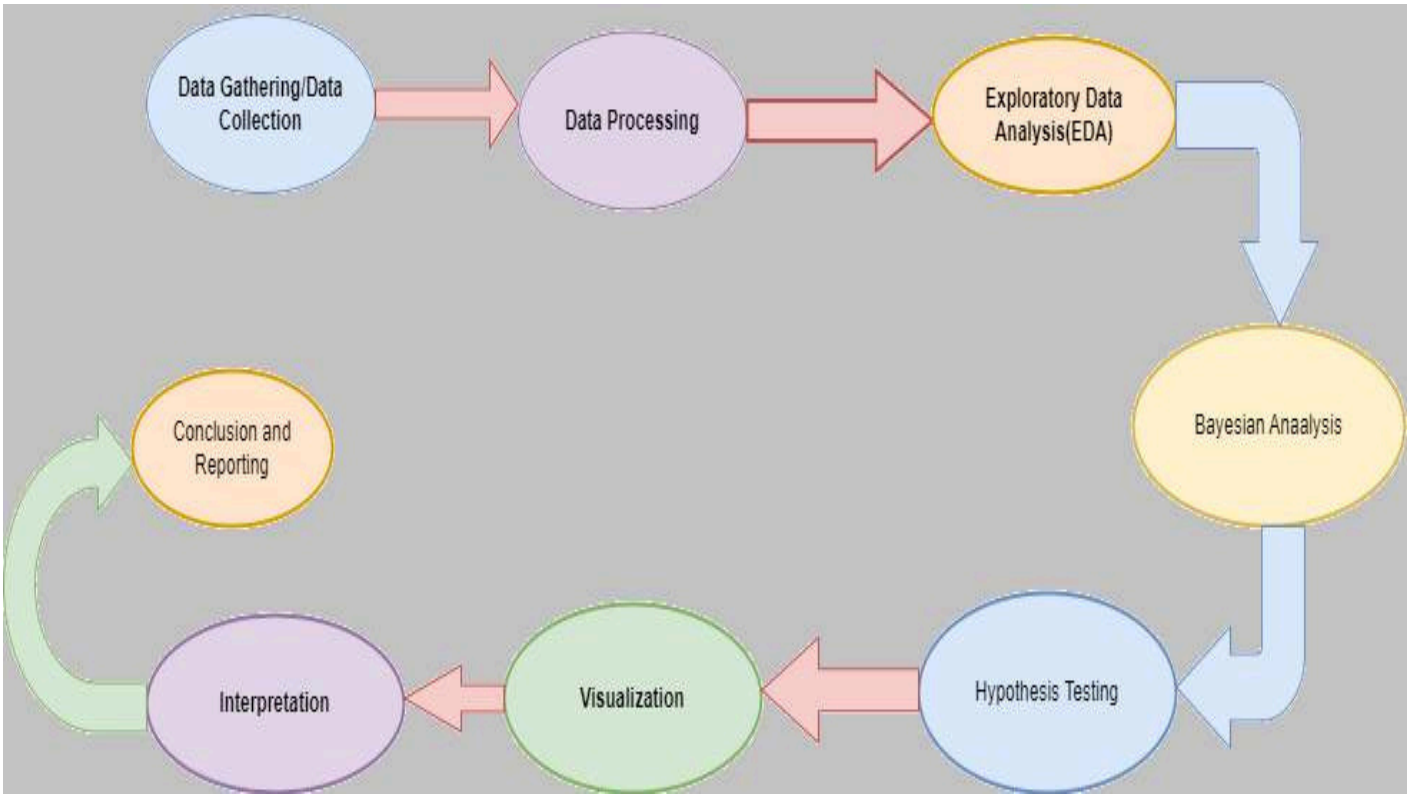
In this project, we will explore this phenomenon using age distribution data to see if we can reproduce a difference in average age at death purely from the changing rates of left-handedness over me, refuting the claim of early death for left-handers. This notebook uses Pandas and Bayesian statistics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

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.

Flow of Project



Languages Used:

Language: Python 3
Frameworks: Pandas, Matplotlib, NumPy
IDE: Google Colab

Implementation:

1 Gathering Requirements and Designing Problem Statement:

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

2 Data Collection and Importing:

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 has been collected through various GitHub repositories, listed as follows:

"https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dad09e88/lh_data.csv"

The research work for the data has been made available from the following source:

Researchers Avery Gilbert and Charles Wysocki.

National Geographic survey in 1986.

The website:

https://www.cdc.gov/nchs/nvss/mortality_tables.html

Data uploading 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.

```
# load the data
data_url_1 = "https://gist.githubusercontent.com/mbonsma/8da0990b71ba9a09f7de395574e54df1/raw/aec88b30af87fad8d45da7e774223f91dac
lefthanded_data = pd.read_csv(data_url_1)
```

3 Data Cleaning and Processing:

Data is the most important aspect of analytics and machine learning. Everywhere in For computing or business, data is required. But many a time, 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 fundamental element of basic data science. Data cleaning is 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 the data set “lefthanded_data,“ 2 new columns are added called "Birth_year." “Mean_lh”

lefthanded_data['Birth_year'] = 1986 - lefthanded_data['Age']

lefthanded_data['Mean_lh'] =
(lefthanded_data['Male']+lefthanded_data['Female']) / 2

Below code for reference:

```
lefthanded_data['Birth_year'] = 1986 - lefthanded_data['Age'] # new column 'Birth_year' to lefthanded_data


# create a new column for the average of male and female
# ... YOUR CODE FOR TASK 2 ...
lefthanded_data['Mean_lh'] = (lefthanded_data['Male'] + lefthanded_data['Female']) / 2 # Calculate the mean left-handedness
```

In the data set death_distribution rate Nan values are dropped from `Both Sexes` column Below is the code for the reference:

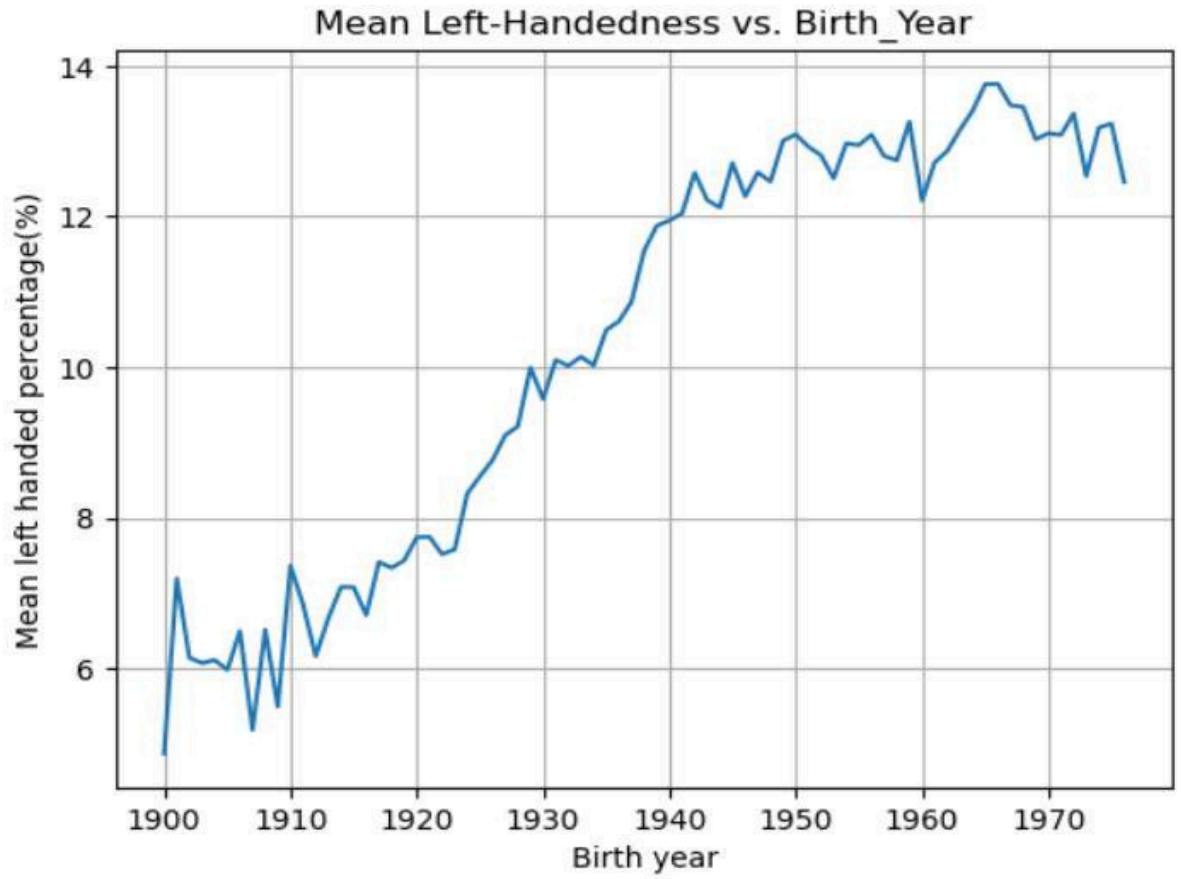
```
# drop NaN values from the `Both Sexes` column
death_distribution_rate.dropna(subset=['Both Sexes'], inplace=True)
```

4 Exploratory Data Analysis :

Calculate the mean of left-handed% for birth year.

lefthanded_data['Mean_lh'] = (lefthanded_data['Male'] + lefthanded_data['Female']) / 2

Create visualizations (e.g., histograms, box plots) to compare the distributions of age at death for both groups.



5 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 that allows us to update our beliefs after seeing evidence.

We can use libraries like Numpy for Bayesian modeling.

Utilize Bayesian statistics to estimate the probability of dying at age A for left-handed P(A|LH).

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.

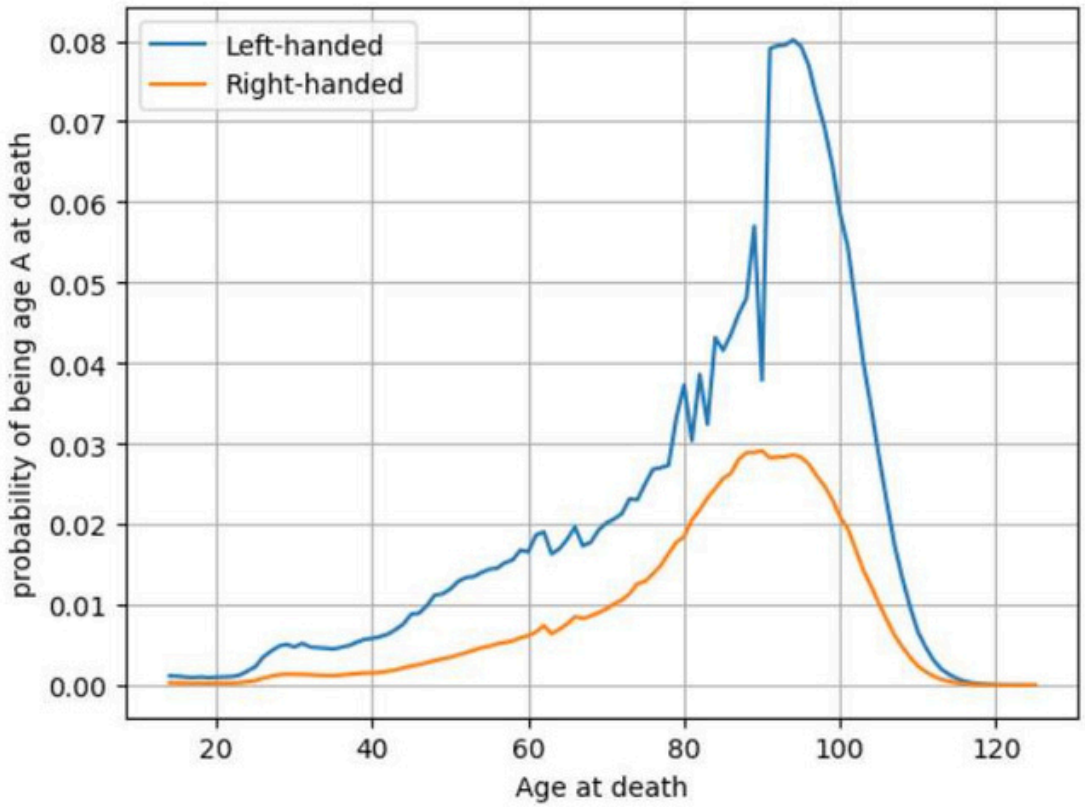
Similarly, for right-handed people, the probability of dying at a certain age is calculated.

```
# import library
import numpy as np

# ... YOUR CODE FOR TASK 3 ...
lefthanded_data = pd.read_csv(data_url_1)
lefthanded_data['Birth_year'] = 1986 - lefthanded_data['Age'] # new column 'Birth_year' to lefthanded_data
lefthanded_data['Mean_lh'] = (lefthanded_data['Male'] + lefthanded_data['Female']) / 2  # Calculate the mean left-handedne



# 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 = np.mean(lefthanded_data.loc[lefthanded_data['Birth_year']<study_year-100]['Mean_lh'][-10:])
    late_1900s_rate = np.mean(lefthanded_data.loc[lefthanded_data['Birth_year']>study_year-100]['Mean_lh'][10:])
    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] = late_1900s_rate / 100
    P_return[ages_of_death < youngest_age] = early_1900s_rate / 100
    P_return[np.logical_and((ages_of_death <= oldest_age), (ages_of_death >= youngest_age))] = middle_rates / 100

    return P_return

#Example:
ages_of_death = np.array([60, 70, 80])
result = P_lh_given_A(ages_of_death)
print(result)
```

```
[0.09576935 0.0774286  0.07364236]
```

6 Hypothesis Testing:

Conduct hypothesis testing to determine if there is a statistically significant difference in the average age at death between left-handed and right-handed individuals.
This can be done by comparing the posterior distributions of the relevant parameters.
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 left-handed distribution has a bump below age 70: of the pool of deceased people, left-handed people are more likely to be younger.



7 Interpretation:

When the results of your analysis are interpreted, one can see that the differences are statistically significant. Do they support or refute the claim of early death for laborers?

Let's compare our results with the original study that found that left-handed people were 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, weighing the probability distribution by age and summing over the result.

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

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

```
# 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 of left-handed people at death:", round(average_lh_age, 2))
print("Average age of right-handed people at death:", round(average_rh_age, 2))

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

```
Average age of left-handed people at death: 183.48
Average age of right-handed people at death: 78.25
The difference in average ages is -105.2 years.
```

8 Conclusion and Reporting:

We have 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 populace, which is good news for left-handers.

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.

Some of the approximations 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.

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.

```
# 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))

# Calculate and print the difference in average ages in 2018

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

```
The difference in average ages is -61.5 years.
```