DSCI 6607 – Fall 2024

Assignment 5*

Question 1

Consider `Occupation.txt` data set from your directory.

1. Use python and import the data set.
2. Write a python script which computes the median of the age variable for each occupation.
3. Write a python script which computes the male ratio per occupation and returns them from the most to the least.
4. For each occupation, calculate the 90% interval for the mean age.

Hint: To report the 90% interval of a data set, sort the data set and report the 2.5 and 97.5 percentiles of the data. [**20 points**]

Question 2

Consider the α-trimmed standard deviation:

$$S_\alpha = \left(\frac{1}{N_\alpha - 1} \sum_{j=[n\alpha]+1}^{[n(1-\alpha)]} (X_j - \bar{T}_\alpha) \right)^{1/2}$$

where $\bar{T}_\alpha = \frac{1}{N_\alpha - 1} \sum_{j=[n\alpha]+1}^{[n(1-\alpha)]} X_j$.

1. Write a python function which takes a x and α and returns the S_α.
2. Apply you function to the `age` variable in `Occupation.txt` data. [**20 points**]

Question 3

It is well-known that the writing style of J. D. Salinger is highly self-conscious and vernacular.

The data set `Catcher_in_the_rye.txt` is a sample text randomly selected from Salinger' novel `The Catcher in the Rye`.

1. Write a python program which computes the number of times each word occurs in the sample text. The output should be reported as a list of tuples. Each tuple shows two elements where the first one is the word (as a string) and the second element represents the number of times that word occurred. [**20 points**]

*This content is protected and may not be shared, uploaded, or distributed without written permission from Dr. Armin Hatefi.

Question 4

Consider the `tips` data set

```
import seaborn as sns
import matplotlib.pyplot as plt

tips_data = sns.load_dataset("tips")
```

Write a python script using seaborn library to show

- A categorical scatter plot that represents the relationship between the `total_bill` and `tip` columns.
- Differentiate the points by the day of the week using different colors.
- Use different marker styles for lunch (`Lunch`) and dinner (`Dinner`) times.
- Ass labels for the x-axis, y-axis, and title.
- Add a legend that clearly indicates the day of the week.

[**20 points**]

Due on Tuesday, December 3th, by 5 pm

Have fun!