CS24420 Scientific Python --- Semester 2 --- Practical Worksheet 5

Hypothesis Testing

The tasks this week are about conducting hypothesis tests on a real world dataset. The dataset is the "brain size" dataset that we saw in Lecture 3. This dataset summarises a survey of 40 right-handed psychology students at an American university.

Information about gender and body size (height and weight) are included in the dataset, and it is these that you will analyse, by testing three null hypotheses:

1. The average female student's height is the same as the average female American height.
2. The average male student's height is the same as the average male American height.
3. The average female student's height is the same as the average male student's height.

Assuming the following population statistics:

- average female American height = 162.9 cm
- average male American height = 176.4 cm

write a Python program to implement the follow tasks. Your program should print out key results and also the interpretation of any statistical test.

---

1. Download a copy of the brain size dataset as a csv file (`brain.csv`) from Blackboard. and read it into a DataFrame object (here denoted by `df`).

2. Extract the female height data from the DataFrame.

[Hint: Recall from Lecture 3 how to extract rows from a DataFrame using a boolean test condition]

Note: The height and weight columns are in imperial units - inches for height and pounds for weight. You will need to take this into account when testing the data against your population statistics. Either convert the DataFrame columns to metric units, or convert the population statistics to imperial units.

3. Print a statistics summary of the female height data *[Hint: can use describe()]*

4. Conduct a **one sample t-test** to test the hypothesis: *"The average female student's height in the sample differs from the average female American height"*. Remember to state the null hypothesis. Set the confidence level to 95%.

Q: Is the null hypothesis rejected or not?

5. Compute and print out the confidence interval for this test and sample.

---

6. Extract the male height data from the original DataFrame, denoted by m_height.

7. Print a statistics summary of the male height data.

8. Conduct a **one sample t-test** to test the hypothesis: *"The average male student's height in the sample differs from the average male American height"*. Remember to state the null hypothesis. Set the confidence level to 95%.

Important Note: The hypothesis testing routines cannot deal with **missing data**. Since the male height data include some `NaN` items, you must drop these items from the data using `DataFrame.dropna()` before applying the one sample t-test.

e.g: `data = data.dropna()  # data stands for a DataFrame`
or: `data.dropna(inplace=True)`

Q: Is the null hypothesis rejected or not?

9. Compute and print out the confidence interval for this test and sample.

---

10. Conduct a **t-test for two independent samples** to test the hypothesis: "The average female student's height differs from the average mail student's height". Remember to state the null hypothesis. Set a confidence level of 95%.

Q: Is the null hypothesis rejected or not?

2023-03-06 LGT