# Module 4 Practice Assignment

**Submitted by: Sourabh D Khot**

NUID: 002754952

**Submitted to: Professor Behzad Ahmadi**

Date: March 25, 2022

## Introduction

In this report, we will perform appropriate two-sample t-tests on two different sets of data as given in the assignment [1]. As I am using Python, I will import equivalent libraries and use equivalent functions available in Python.

In [16]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import pingouin as pg
import statistics
```

## Part 1. Two-sample t-test with unequal variance

In this part, we will use the 'cats' dataset to answer whether male and female cat samples have the same bodyweight (Bwt). We will use the following steps:

1. Understand the variable of interest
2. Identify the type of test
3. Define the hypothesis
4. Test the hypothesis
5. Analyze the p-value, interpret the result

### A. Understanding the data

The original 'cats' dataset from R's MASS library is downloaded from [2] as CSV and imported into Python as a pandas dataframe. Let us first understand the data.

In [2]:
```
cats = pd.read_csv('Datasets/MASS_cats.csv')
```

In [34]:
```
print('\033[1m' + 'Table 1: Summary of Data' +'\033[0m')
cats.info()
```

```
Table 1: Summary of Data
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144 entries, 0 to 143
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
```

```
 0   Sex     144 non-null    category
 1   Bwt     144 non-null    float64
 2   Hwt     144 non-null    float64
dtypes: category(1), float64(2)
memory usage: 2.6 KB
```

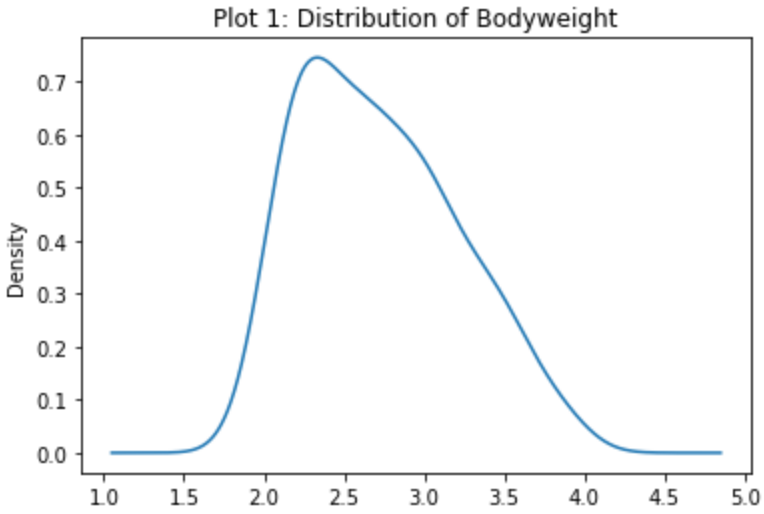
In [4]:
```
cats.Sex = cats.Sex.astype('category') #changing to category variable type
```

Let us understand values of 'Sex' and plot distribution of 'Bwt' [3], as we have to work with these two variables.

In [5]:
```
cats.Sex.unique() #finding unique in categorical attributes
```

Out[5]:
```
['F', 'M']
Categories (2, object): ['F', 'M']
```

In [6]:
```
cats.Bwt.plot.density(title='Plot 1: Distribution of Bodyweight');
```




Plot 1 seems somewhat like a normal distribution.

To compare bodyweight of male and female cats, we will create two separate samples as per the 'Sex' variable.

In [7]:
```
female = cats[cats.Sex == 'F']
male = cats[cats.Sex == 'M']
print('\033[1m' + 'Table 2: Sample Sizes' +'\033[0m')
print("n.female =", female.shape[0])
print("n.male =", male.shape[0])
```

**Table 2: Sample Sizes**
n.female = 47
n.male = 97

## B. Identifying the type of test

We are comparing two random samples independent of each other, and we do not know the population variances. The sample sizes are greater than 30. Assuming the variances are unequal, **we will use the independent two-sample t-test** (Welch's t-test).

## C. Hypothesis statement

We have to test whether male and female cats have the same bodyweight. We can define the hypothesis as below.

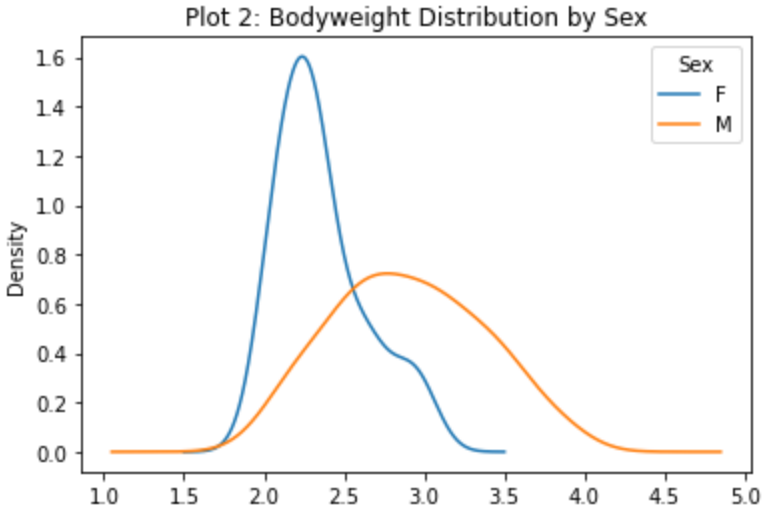
- Null hypothesis, **H0: μ.male = μ.female**
- Alternate hypothesis, **H1: μ.male ≠ μ.female**

This is a two-sided test. Let us consider a confidence level of 0.95 (α = 0.05). From table 2, n.female = 47 and n.male = 97.

## D. Test Output

Let us plot the sample distributions to be compared [4].

In [8]:
```
cats.pivot(columns='Sex',values='Bwt').plot.density(title='Plot 2: Bodyweight Distributio
```




We will use the ttest() function from pingouin library for this test, in which correction = True for Welch's t-test [5].

In [38]:
```
print('\033[1m' + 'Table 3: Two independent sample t-test' +'\033[0m')
pg.ttest(x=male.Bwt, y=female.Bwt, alternative='two-sided', correction=True, confidence=0.
```

**Table 3: Two independent sample t-test**

Out[38]:

| | T | dof | alternative | p-val | CI95% | cohen-d | BF10 | power |
|---|---|---|---|---|---|---|---|---|
| **T-test** | 8.709488 | 136.837883 | two-sided | 8.831034e-15 | [0.42, 0.66] | 1.302837 | 6.959e+11 | 1.0 |

## E. Interpretation

From table 3, the p-value is much less than the considered α (0.05), i.e. there is more than sufficient evidence to reject the null hypothesis. Hence, statistically, from the sample, **we can comfortably say that male and female cats do NOT have the same bodyweight.**

# Part 2. Effect of meditation on sleep quality

Using the data given in the assignment [1], we have to determine whether the meditation workshop improves the sleeping quality score based on a sample of 10 students. We will use the following steps:

1. Understand the variable of interest
2. Identify the type of test
3. Define the hypothesis
4. Test the hypothesis

5. Analyze the p-value, interpret the result

## A. Understanding the data

In [10]:
```
data = {'before': [4.6, 7.8, 9.1, 5.6, 6.9, 8.5, 5.3, 7.1, 3.2, 4.4],
        'after': [6.6, 7.7, 9.0, 6.2, 7.8, 8.3, 5.9, 6.5, 5.8, 4.9]}
score = pd.DataFrame(data)
```

As the same 10 recruited students measured the data before and before, we can assume that data in one row of before and after belongs to the same student. To understand the improvement in score for each student, we will introduce another variable 'D' indicating the difference from before to after.

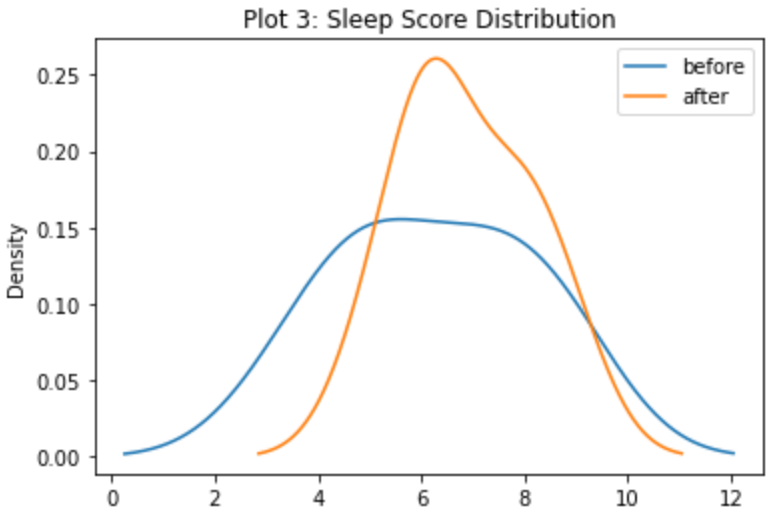
In [11]:
```
score['D'] = score.after - score.before
```

## B. Identifying the type of test

The before and after samples are dependent since corresponding scores belong to the same student. It is reasonable to assume that the students were recruited randomly for the workshop.

As the sample size is 10, we need to further evaluate if the samples are approximately normally distributed. Let us plot their distribution [3].

In [12]:
```
score.iloc[:,[0,1]].plot.density(title='Plot 3: Sleep Score Distribution');
```




Both samples are approximately normally distributed. Hence, **we will use the paired two samples t-test.**

## C. Hypothesis statement

We have to test whether meditation improves the sleeping score, i.e. if there is a positive difference from before to after score. We will define the hypothesis as below.

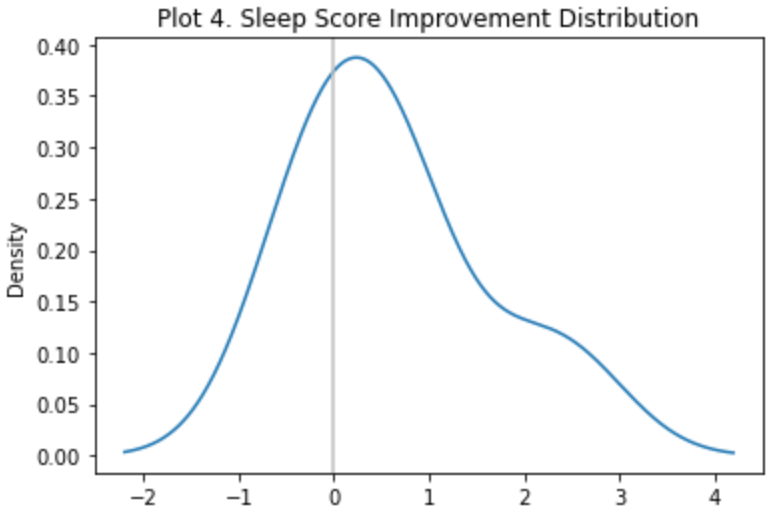
- Null hypothesis, **H0: μ.D = 0**
- Alternate hypothesis, **H1: μ.D > 0**

This is a one-sided right-tail test. As per the given data, α = 0.05 (confidence level of 0.95), n = 10 and degrees of freedom = 9 (n-1).

## D. Test Output

Let us plot the distribution of the difference 'D' variable. Since we are testing whether the difference is statistically greater than 0, I have plotted a vertical line at zero. [3][7][8].

In [13]:
```
score.D.plot.density(title='Plot 4. Sleep Score Improvement Distribution')
plt.axvline(x=0,color='silver');
```




When we perform this test manually, we need to calculate the differences of the sample, mean of differences, standard deviation of the differences and then proceed with the test.

Since we are using Python, specifically pingouin library's ttest() function, we will only need to input the before and after samples. The alternative hypothesis is whether x (after) is 'greater' than y (before) [5].

In [39]:
```
print('\033[1m' + 'Table 4: Two dependent sample t-test' +'\033[0m')
pg.ttest(x=score.after, y=score.before, paired=True, alternative='greater', confidence=0.9
```

**Table 4: Two dependent sample t-test**

Out[39]:

| | T | dof | alternative | p-val | CI95% | cohen-d | BF10 | power |
|---|---|---|---|---|---|---|---|---|
| **T-test** | 1.948098 | 9 | greater | 0.04161 | [0.04, inf] | 0.378623 | 2.412 | 0.295605 |

## E. Interpretation

From table 4, the p-value (0.04161) is less than α (0.05), i.e. there is sufficient evidence to reject the null hupothesis. Hence, statistically, from the sample, **we can say that the meditation workshop improves the sleeping quality score of an individual.**

- This should be a paired two samples t-test since the same recruited students test before and after the meditation workshop.
- The conclusion will not change if the level of significance is increased to 0.10 since the p-value (0.04161) is still less.
- As the before and after samples are paired, we cannot use independent tests. Since we do not know the population variance of the differences, we cannot use the z-test. There are no proportions in the data to use the proportion test. Hence, the usage of the paired two samples t-test is justified.

# Biliography

[1] Canvas Module 4 Practice Assignment: northeastern.instructure.com/courses/105627/assignments/1206558
[2] Cats dataset from R MASS package: r-data.pmagunia.com/dataset/r-dataset-package-mass-cats

[3] Plotting Density Plot: geeksforgeeks.org/density-plots-with-pandas-in-python/
[4] Density Plots: geeksforgeeks.org/multiple-density-plots-with-pandas-in-python/
[5] Two sample t-test: pingouin-stats.org/generated/pingouin.ttest.html
[6] Input data into new dataframe: towardsdatascience.com/15-ways-to-create-a-pandas-dataframe-754ecc082c17
[7] Vertical line stackoverflow.com/questions/19213789/how-do-you-plot-a-vertical-line-on-a-time-series-plot-in-pandas
[8] Color Names: matplotlib.org/stable/gallery/color/named_colors.html