Unpaired t-test

Explanation

- In general: t-tests are used to check whether the *difference in the means of two samples* (or one sample vs a general population mean) are more likely *due to actual differences* (for example, due to a difference in treatment) or simply *due to statistical fluctuations*.

- Simplified, the t-test is calculated as: $t = \frac{variance\ between\ groups}{variance\ within\ groups}$

- Therefore, a high t-value implies that the differences between groups are significant and a low t-value implies that the differences are merely statistical fluctuations due to sampling error or random chance.

- The unpaired t-test is then used to estimate the significance of the differences between two independent / unpaired samples that have (roughly) the same variance.

- Assumptions to use an unpaired t-test:
  - The observations are sampled independently
  - The dependent variable is normally distributed
  - The variance within the groups is (roughly) the same

Unpaired t-test: Real life example

- Compare the average height of individuals grouped by gender: male and female groups, which are two independent groups.

- Data set extracted from: https://www.kaggle.com/mustafaali96/weight-height

- The dataset contains weight and height data by gender.

- The assumptions done on the datasets are:

1)Both height distributions are normal distributions
2)There are no big amount of outliers
3)The variances are assumed equal in order to use the Unpaired t-test

Unpaired t-test: Real life example

1)As shown on Figure 1 both variables fit well into a normal distribution.

2)In Figure 2 the outliers are shown, considering the dataset contains 10k entries is assumed to be good enough.

3)In Figure 3 the variance of both variables is shown. Unpaired t-test will be used.

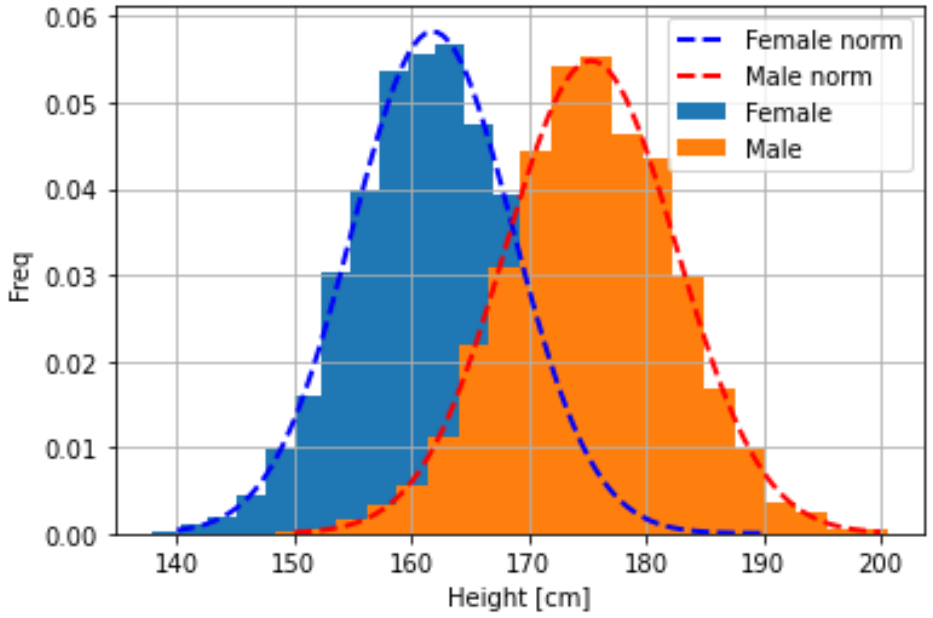


Figure 1

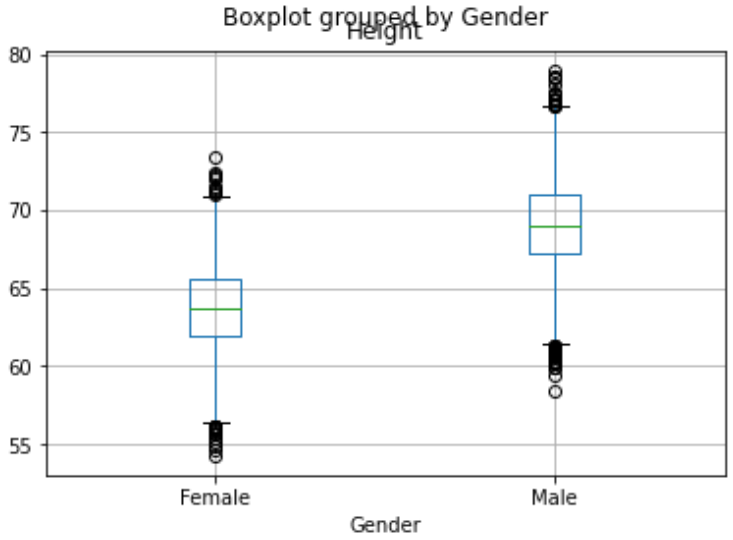


Figure 2

```
Male var: 52.9
Female var: 46.9
```

Figure 3

Unpaired t-test: Real life example

Considering the hypothesis:

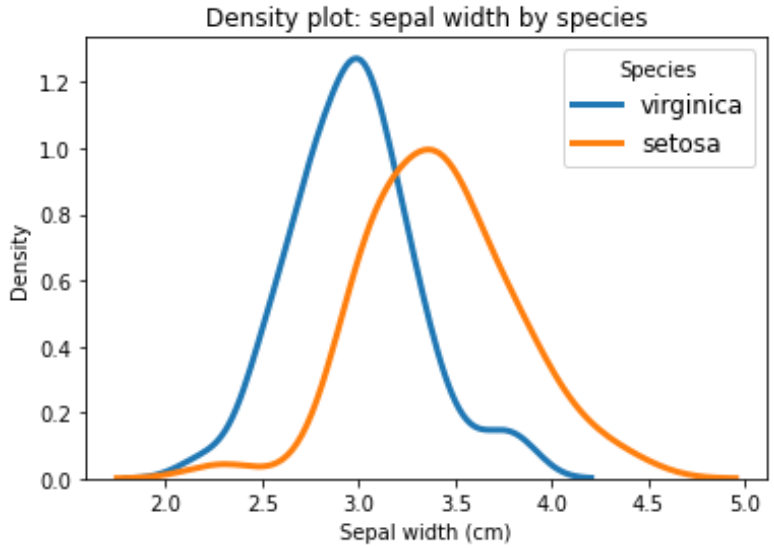
$$H_0: m_{Male} = m_{Female}$$

```
Ttest_indResult(statistic=95.60271449148863, pvalue=0.0)
```

Since p-value is 0 we reject the null hypothesis concluding that the two mean values for the two variables are different.

Unpaired t-test (aka. independent t-test)

- Comparison of **sepal width** of **virginica** and **setosa** species.



- D'Agostino and Pearson's tests (p=0.28 and p=0.39) as well as Shapiro-Wilk tests (p=0.18 and p=0.20) suggest normal distribution for both variables.

```
stats.ttest_ind(df_setosa['swidth'],df_virginica['swidth'], equal_var = True)
# equal_var = True -> assume equal variances
```

```
Ttest_indResult(statistic=6.289384996672061, pvalue=8.916634067006443e-09)
```

- Unpaired t-test with equal variance assumption indicates that we can **reject** the null hypothesis of equal width with $p < 0.01$. Most likely, these samples **do not** come from a population with same mean and variance.