Learning Objectives

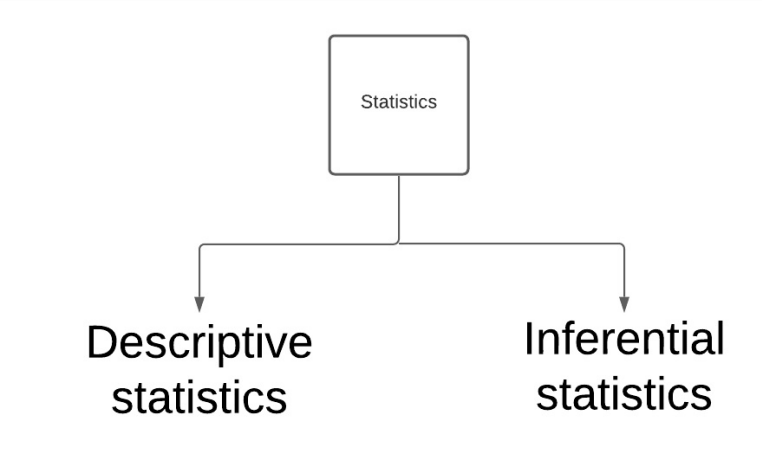
In the third module of this course, you will learn how to handle different scenarios in order to better manipulate your data.

Learning Objectives:

1. Learn some comparison test
2. Visualizing data frame base on comparison
3. Statistical categories

Comparison Test

The term “statistics” refers to the study of extracting meaningful information using a set of mathematical procedures for organizing, summarizing, analyzing, and interpreting information.



images/stat

- **Descriptive Statistics**: is used to organize and summarize facts
  For example, the average SAT score for the class of 2020 was 1051.
- **Inferential Statistics:** is used to analyze and interpret samples of data.

As scientists, we need standardized techniques to make sense of information in an accurate way.
In the first few chapters we went over some descriptive statistics. In order to tackle some comparison tests we first need to go over some terminology.

	Population	***Sample***
Definition	The entire set of individuals of interest in a particular	A set of individuals selected from a population, usually intended to be representative of the population
Examples	College students	90 randomly chosen students on campus

Naturally occurring discrepancy or error that exist between a sample and the corresponding population is referred to as **Sampling Error**.

Standard Deviation

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Standard deviation is important because it helps us better understand indicative of is our descriptive statistic is in connection to our sampling data. A standard deviation close to zero indicates that data points are close to the mean, whereas a high or low standard deviation indicates data points are respectively above or below the mean.

There are scenarios where the mean (average) is not indicative of the general population or sample, so standard deviation highlights those scenarios.

> To calculate the standard deviation of those numbers:
>
> 1. Calculate the mean
> 2. Then for each number: subtract the Mean and square the result
> 3. Find the mean of the squared results
> 4. Return the square root of the mean of square results

In order to calculate our standard deviation

```
import statistics

print((statistics.stdev(sample)))
```

> **Compare the standard deviation of the following samples**
>
> - [1,23,4,25,67,8]
> - [1,2,3,4,5]
> - [1,2,3,999]

Comparison tests:

These tests look for the difference between the means of variables, the comparison of Means.
* T-tests are used when comparing the means of precisely two groups (e.g. the average heights of men and women).
* Independent t-test: Tests the difference between the same variable from different populations (e.g., comparing dogs to cats)
* ANOVA and MANOVA tests are used to compare the means of more than two groups or more(e.g. the average weights of children, teenagers, and adults).

Null Hypothesis, Alternative Hypothesis & P-Value

There are three important **key terms** that you should know and they are: **null hypothesis**, **alternative hypothesis**, and **p-value**. Any time a statistic test is conducted, the goal is always to determine if the test shows a **significant** difference between two sets of data. The null hypothesis is rejected in favor of the alternative hypothesis if the result shows **significant** difference between the data. On the other hand, we fail to reject the null hypothesis if the result shows **no significant** difference in the data. To determine if there is significance or not, we look at the **p-value**.

- A p-value of **0.05 or below** means there is a **significant difference** between the data, in which case we will **reject the null hypothesis** in favor of the alternative hypothesis.
- A p-value of **greater than 0.05** means there is **no significant difference** between the data, in which case we will **fail to reject the null hypothesis**.
- When it comes to these hypotheses, we always refer to either **rejecting** the null hypothesis (in other words, the alternative hypothesis is supported) or **failing to reject** the null hypothesis (null hypothesis is supported). The **null hypothesis** is always the hypothesis being tested.

For example, if The p-value is calculated to be `6.2e-09` which is **significantly less than 0.05**. Therefore, we will reject the null hypothesis in favor of the alternative hypothesis that there **is** a significant difference between the groups.

T-test

t-test allows us to compare the average values of the two data sets and determine if they came from the same population

Two Sample t-test

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not.

```
from scipy.stats import ttest_ind

class1= [324,433,555,6234,4353]
class2= [320,502,589,6100,4200]

print(ttest_ind(class1,class2,equal_var=True))
```

Every t-value has a p-value to go with it. The p-value helps you determine how significant your results are.

- A Large t-score tells you the groups are different.
- A Large p-value(>0.05) tells you there is small evidence against the null hypothesis(no statistical relationship or significance in the group being looks at).

Researchers tend to work to reject the null hypothesis. In other words, a low p-value is what we are looking for so we can better explain or reject.
.
Paired t-test
—-
A paired t-test is used when we are interested in the difference between two variables for the same subject.

Anova

Like the t-test, ANOVA helps you find out whether the differences between groups of data are statistically significant. It works by analyzing the levels of variance within the groups through samples taken from each of them.

Variance measures how far each number in the set is from the mean and thus from every other number in the set.