BTU ბიზნესისა და ტექნოლოგიების უნივერსიტეტი

python

მონაცემთა ანალიტიკა Python

ლექცია 5: შესავალი სტატისტიკაში. აღწერითი სტატისტიკა. სტატისტიკის ფუნდამენტალური ელემენტები. ჰიპოთეზები და ჰიპოთეზების ტესტირება. ერთ მხრიანი და ორ მხრიანი სტატისტიკური ტესტები.

ლიკა სვანაძე
lika.svanadze@btu.edu.ge

Descriptive vs. Inferential Statistics:

- Statistics is concerned with the describing, interpretation and analyzing of data
- Statistics is often categorized into descriptive and **inferential** statistics.

Descriptive Statistics focuses on summarizing and describing the main features of a dataset. It involves measures like mean, median, mode, variance, and standard deviation.

Inferential Statistics involves making predictions or inferences about a population based on a sample of data. It includes hypothesis testing and estimating parameters.

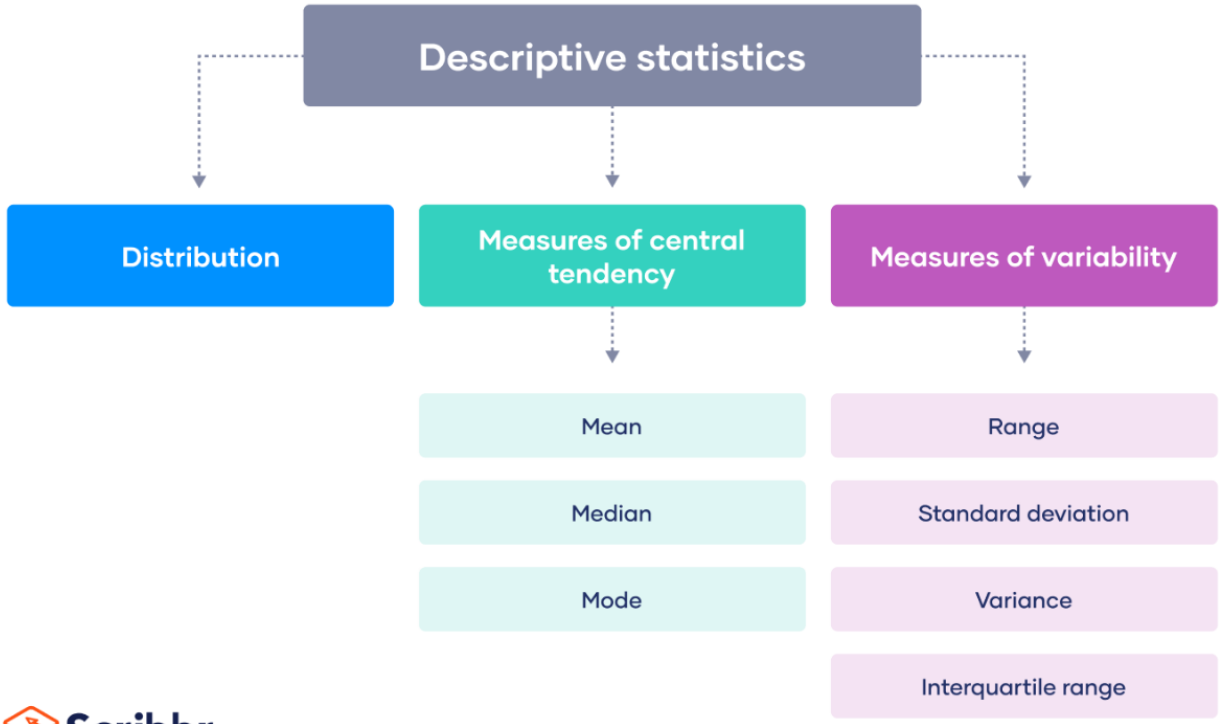
Descriptive Statistics

- Descriptive statistics summarize and organize characteristics of a data set.
- A data set is a collection of responses or observations from a sample or entire population.
- In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).
- Graphical displays are often used along with the quantitative measures to enable clarity of communication

Types of descriptive statistics

There are 3 main types of descriptive statistics:

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** or dispersion concerns how spread out the values are.



Distribution

- **Frequency:** A frequency is the number of times a value of the data occurs. Example: 20 students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3

- **Relative Frequency:** A relative frequency is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes.

- **Cumulative Relative Frequency:** Cumulative relative frequency is the accumulation of the previous relative frequencies.

Frequency, Relative Frequency and CRF

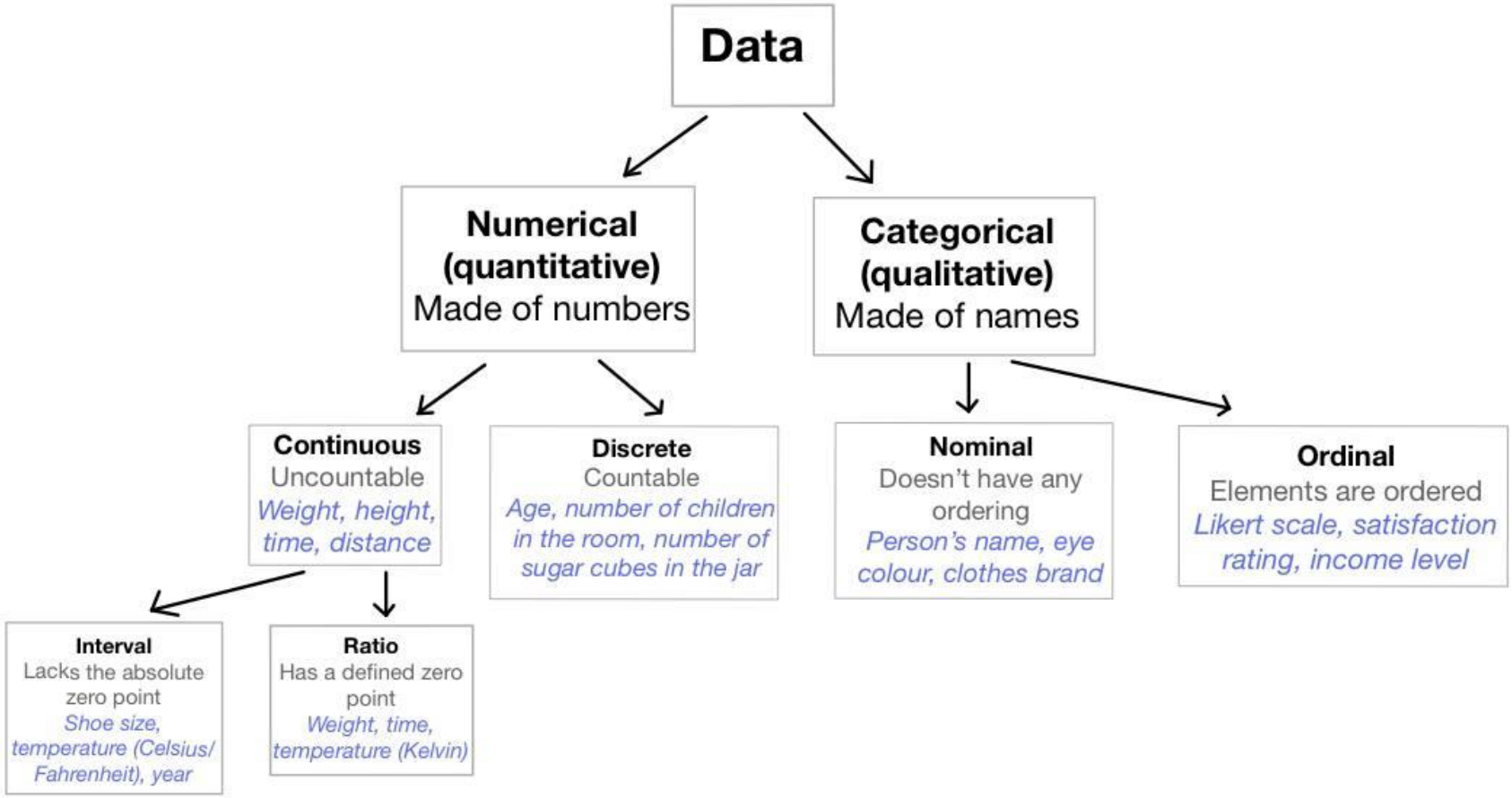
Data Value	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	3/20 = 0.15	0.15
3	5	5/20 = 0.25	0.15 + 0.25 = 0.40
4	3	3/20 = 0.15	0.40 + 0.15 = 0.55
5	6	6/20 = 0.30	0.55 + 0.30 = 0.85
6	2	2/20 = 0.10	0.85 + 0.10 = 0.95
7	1	1/20 = 0.05	0.95 + 0.05 = 1.00

Contingency Table

- A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts.
- To know the relationship between two ordinal or nominal variables then look for contingency table which displays this relationship.

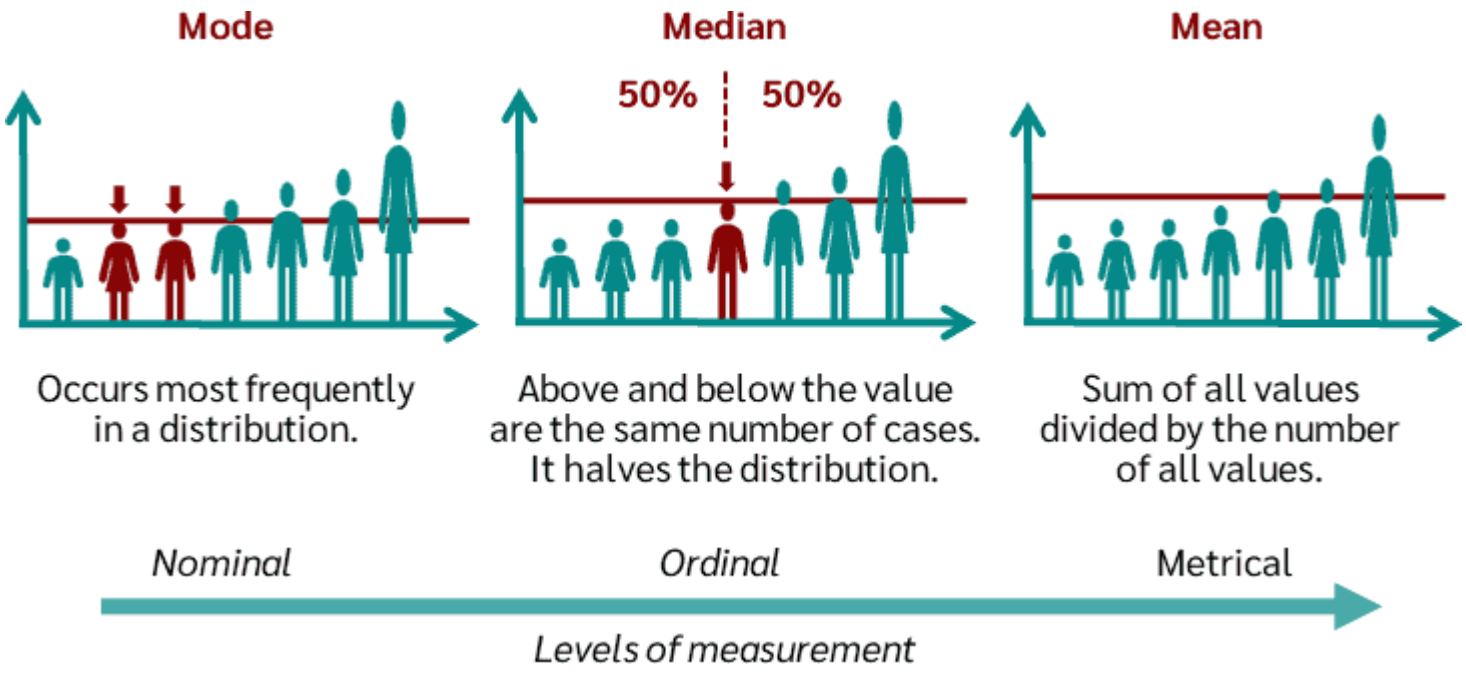
		Sport Preference			
		Archery	Boxing	Cycling	
Gender	Female	35	15	50	**100**
	Male	10	30	60	**100**
		45	**45**	**110**	**200**

Numerical and categorical data



Mean, Median, Mode

In descriptive statistics, the mean, median and mode are location parameter (measures of central tendency). Based on data collected in a sample, the location parameter provide information about where the "center" of the distribution lies



Mean

The mean value can only be calculated for metric variables, i.e. if a metric scale of measurement is given. It indicates where the center of gravity of a distribution can be found. In everyday life it is also called the "average".

The arithmetic mean is the sum of all observations divided by their number n.

The mean value can be calculated by adding up all the values of a variable and then dividing the sum by the number of characteristic values.

The disadvantages of the mean are that it is sensitive to outliers, the value does not have to exist in the data and for the interpretation to be meaningful, the data should have metric scale level.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mean Value — Number of values — Value at the i-th position

A group of 5 statistics students was asked how many cups of coffee they drink per week. The result is 21, 25, 10, 8 and 11 cups. The average is thus 15.

1	2	3	4	5
21	25	10	8	11

$$\frac{21 + 25 + 10 + 8 + 11}{5} = 15$$

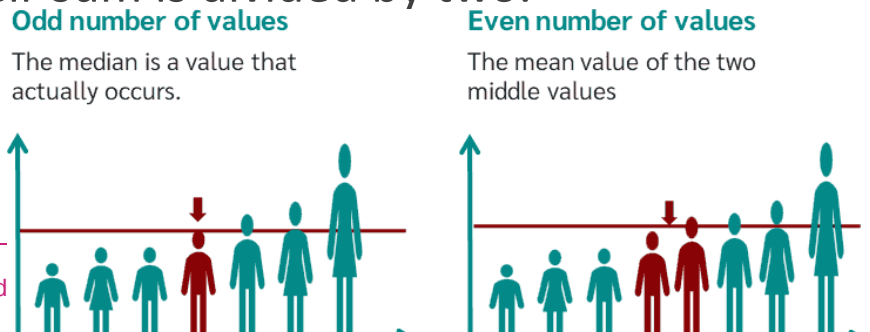
```
In [3]: import numpy as np
        x = [8, 1, 3, 4, 28]
        x = np.array(x)
        print(x)
        mean = np.mean(x)
        print(mean)

        [ 8  1  3  4 28]
        8.8
```

Median

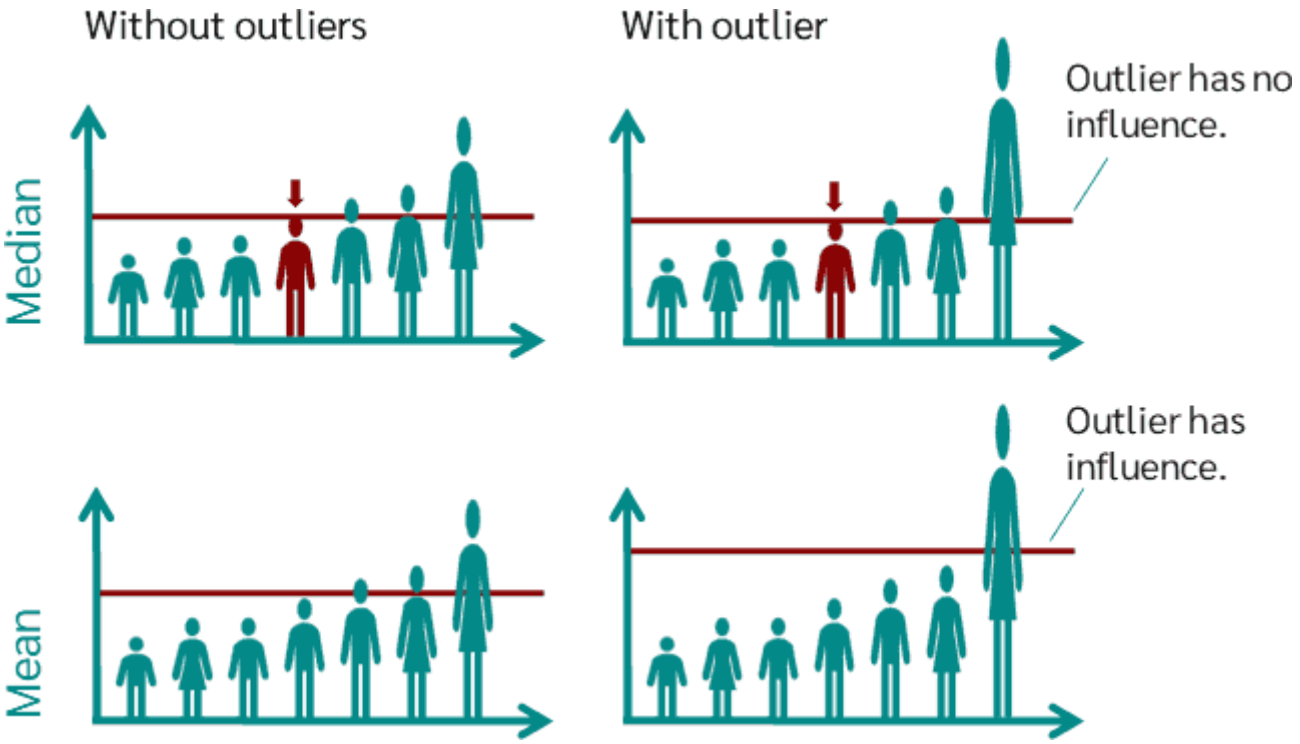
- If the measured values of a variable are ordered by size, the value in the middle is the median. The median is therefore the "middle value" of a distribution. It leads to a division of the series into two parts: one half is smaller and one is larger than the median.
- Since for the calculation of the median the data are ordered, the variables must have ordinal or metric scale level.
- In an ordered series, the median is the value that divides the series into an equal upper and lower range.
- If there is an odd number of characteristic values, then the median is a value that actually occurs.
- If there is an even number of characteristic carriers (persons), the two middle characteristics are added together and their sum is divided by two.

The great advantage of the median is that it is very robust against outliers and that the data only have to be scaled ordinally.



Mean vs Median

Compared to the mean, the median is much more robust against scattering. An outlier usually has no influence on the median, but it has a more or less large influence on the mean.



Implementation

In [9]:
```
import numpy as np
x = [8, 1, 3, 4, 28, 1, 7]
x = np.array(x)
print(x)
mean = np.mean(x)
print(mean)
```

```
[ 8  1  3  4 28  1  7]
7.428571428571429
```

In [10]:
```
median = np.median(x)
median
```

Out[10]:
```
4.0
```

In [11]:
```
from scipy import stats as st
mode = st.mode(x)
mode
```

Out[11]:
```
ModeResult(mode=array([1]), count=array([2]))
```

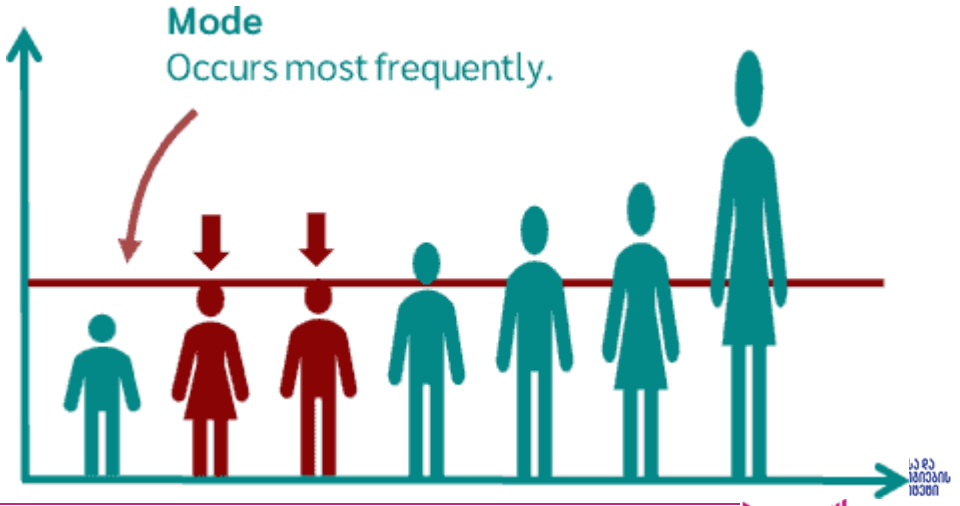
Mode

The mode is the most common value. The mode is therefore the most frequent value in a distribution and corresponds to the highest value in the distribution. It is therefore the value that is "typical" for a distribution.

The mode can be used for both **metric** and **categorical** (nominal or ordinal) variables.

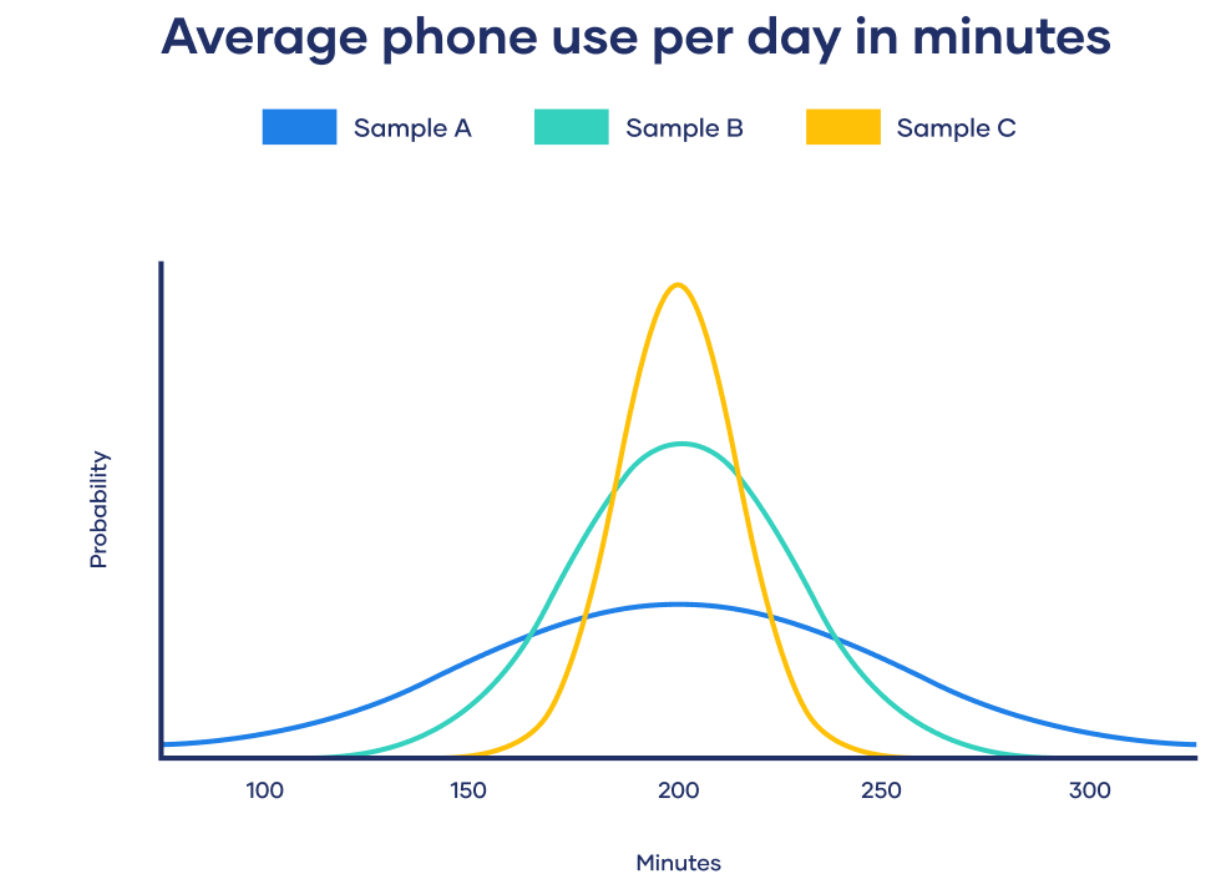
The mode is the value of a distribution that occurs most often.

Example: In a sample of 70 managers from Berlin, 20 drive a Mercedes, 25 a BMW, 10 a VW and 15 an Audi. What is a Mode?



Measures of variability (ცვალებადობა)

- **Measures of variability**
  - Range
  - Standard deviation
  - Variance
  - Interquartile range



Range - np.ptp()

- The range tells you the spread of your data from the lowest to the highest value in the distribution. It's the easiest measure of variability to calculate.
- To find the range, simply subtract the lowest value from the highest value in the data set. Because only 2 numbers are used, the range is influenced by outliers and doesn't give you any information about the distribution of values. It's best used in combination with other measures.
- You can get it with the function **np.ptp():**
- Range example: You have 8 data points from Sample A.
- Disadvantage: It is very **sensitive** to **outliers** and does not use all the observations in a data set.

Data (minutes)	72	110	134	190	238	287	305	324

The highest value (H) is **324** and the lowest (L) is 72.

R = H – L

R = 324 – 72 = **252**

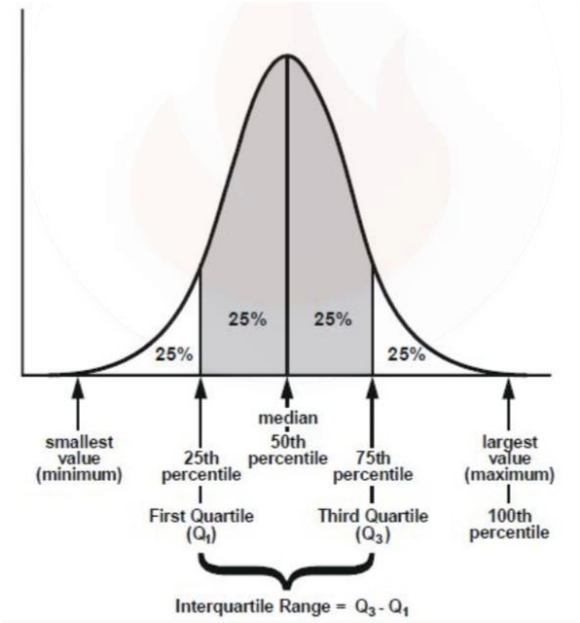
The range of your data is **252** minutes.

```
In [15]: minutes = [72, 110, 134, 190, 238, 287, 305, 324]
         print('range=', max(minutes)-min(minutes))
         minutes = np.array(minutes)
         range_minutes = np.ptp(minutes)
         print("range=", range_minutes)

         range= 252
         range= 252
```

IQR (Interquartile Range)

- It equally divides the distribution into four equal parts called quartiles.
- First 25% is 1st quartile (Q1), last one is 3rd quartile (Q3) and middle one is 2nd quartile (Q2) and it leaves out the extreme values.
- 2nd quartile (Q2) divides the distribution into two equal parts of 50%. So, basically it is same as Median.
- The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1



IQR = Q3- Q1

- The main advantage of the IQR is that it is not affected by outliers because it doesn't take into account observations below Q1 or above Q3.

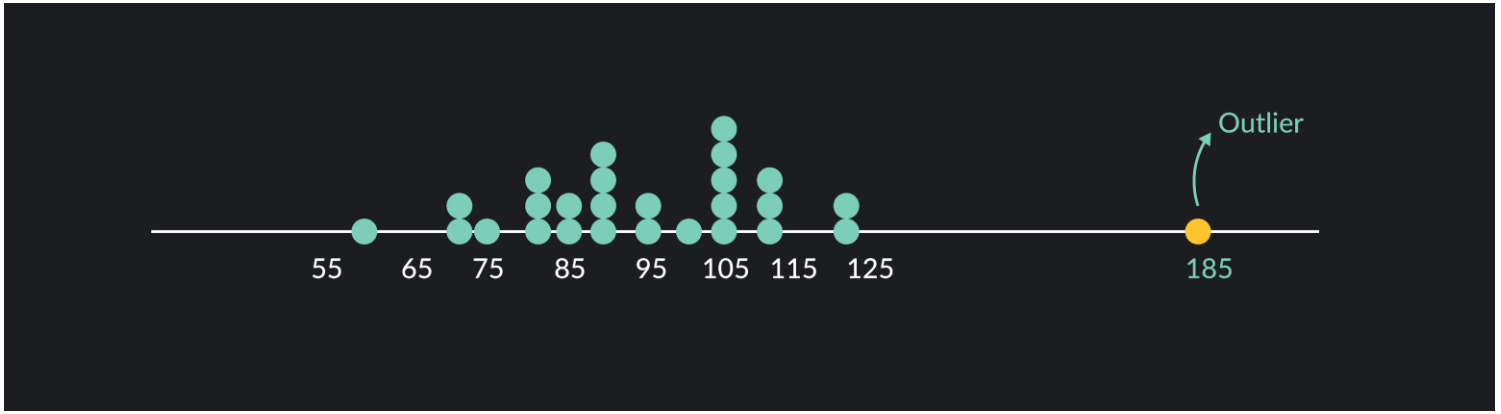
Outliers

- Outliers are values that "lie outside" the other values.
- As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile. Outliers are values that "lie outside“ the other values.

$$\text{Outliers} = Q1 - 1.5 * IQR$$

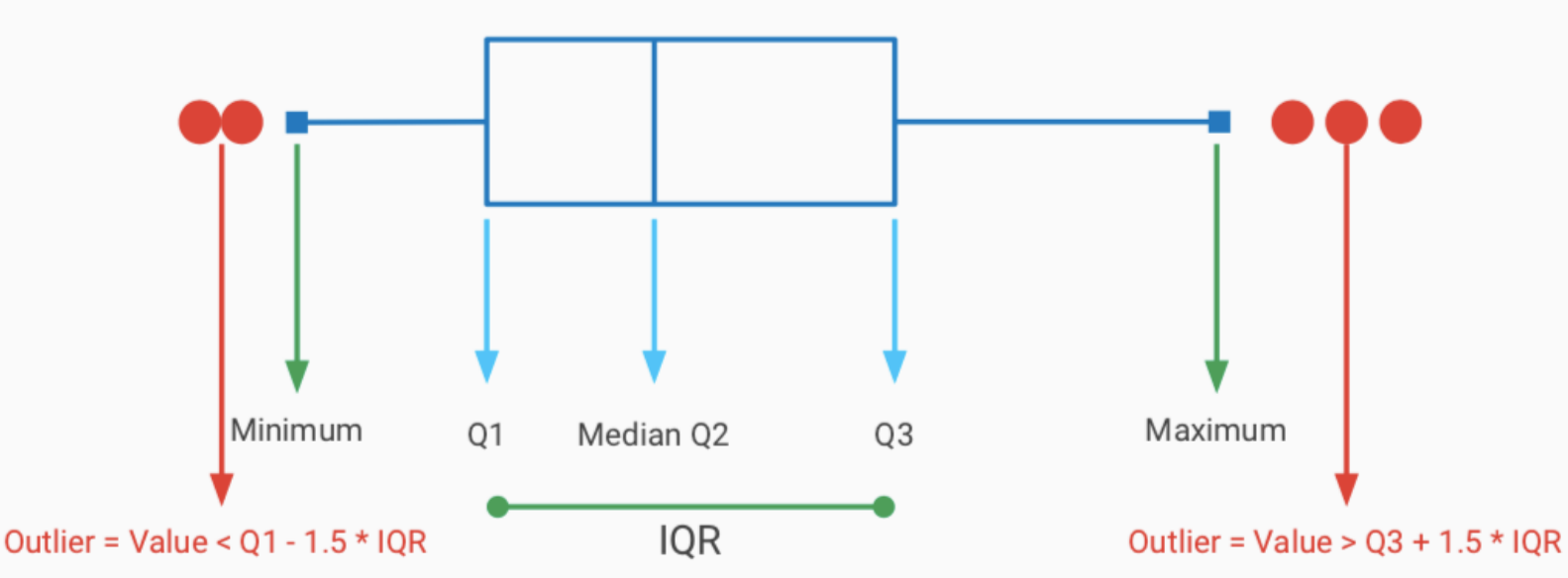
OR

$$\text{Outliers} = Q3 + 1.5 * IQR$$



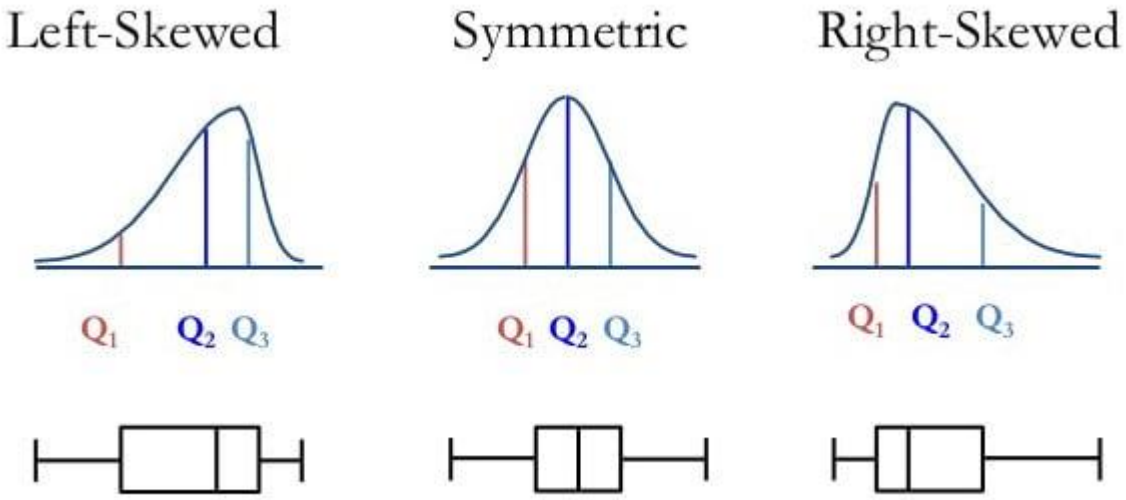
Box Plot

- There is one graph that is mainly used when you are describing centre and variability of your data. It is also useful for detecting outliers in the data.



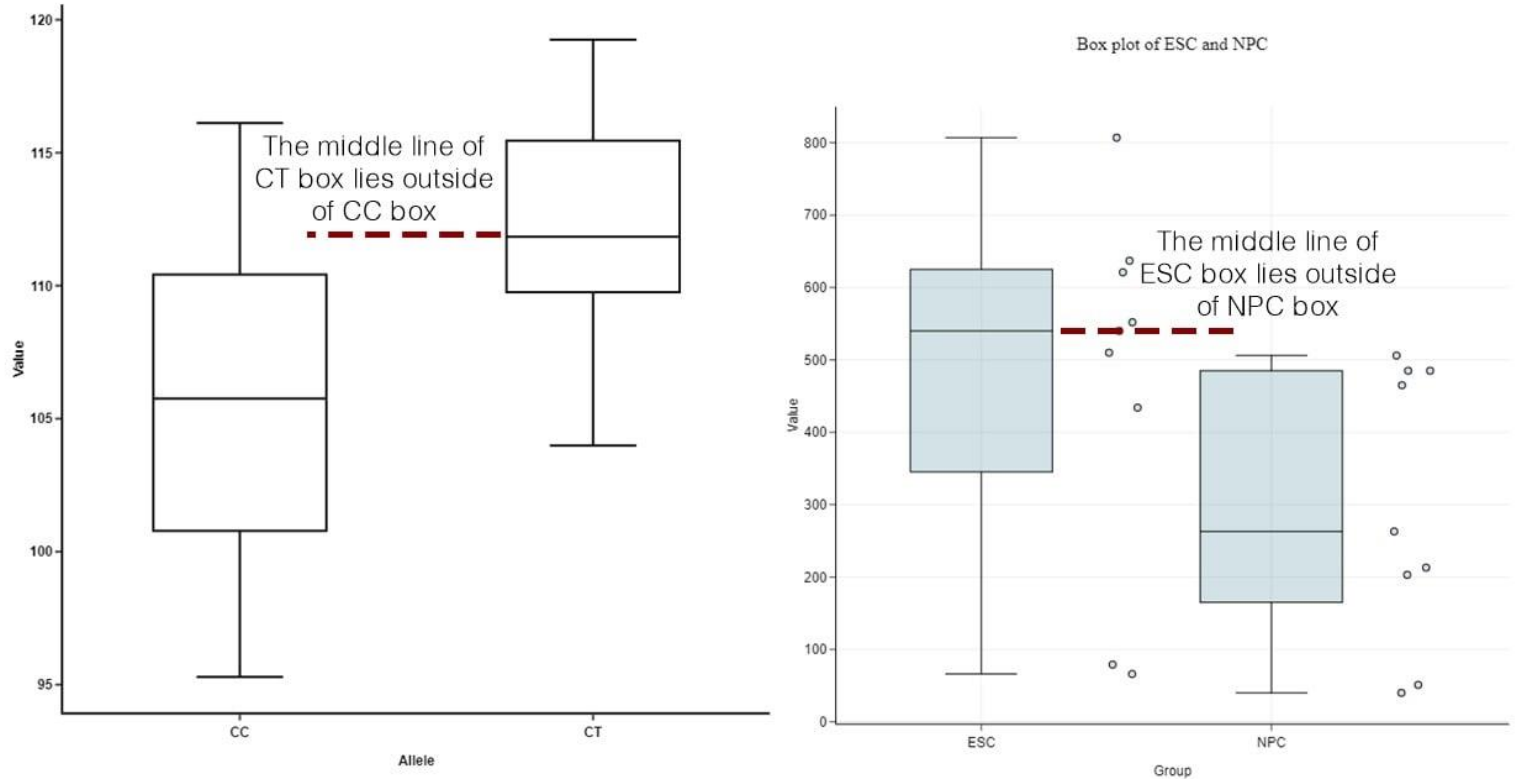
Box Plot

- There is one graph that is mainly used when you are describing centre and variability of your data. It is also useful for detecting outliers in the data.



Box Plot

Compare the respective medians of each box plot. If the median line of a box plot lies outside of the box of a comparison box plot, then there is likely to be a difference between the two groups.

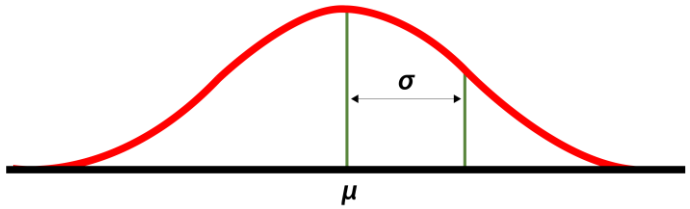


Standard deviation - np.std()

- The **standard deviation** is the average amount of variability in your dataset.
- It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.
- Example: The heights (at the shoulders) of the dogs are:

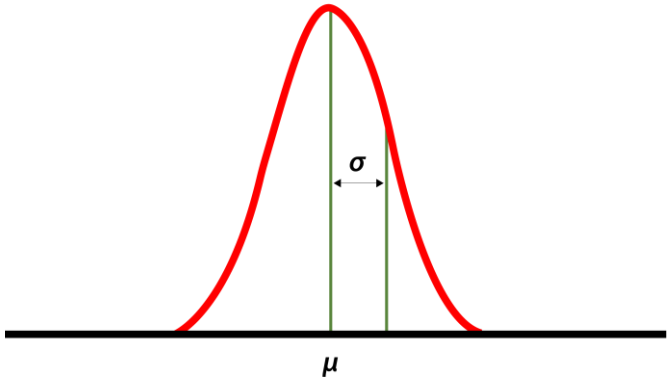
600mm, 470mm, 170mm, 430mm, 300mm

Find out the Mean, the Variance and the Standard Deviation

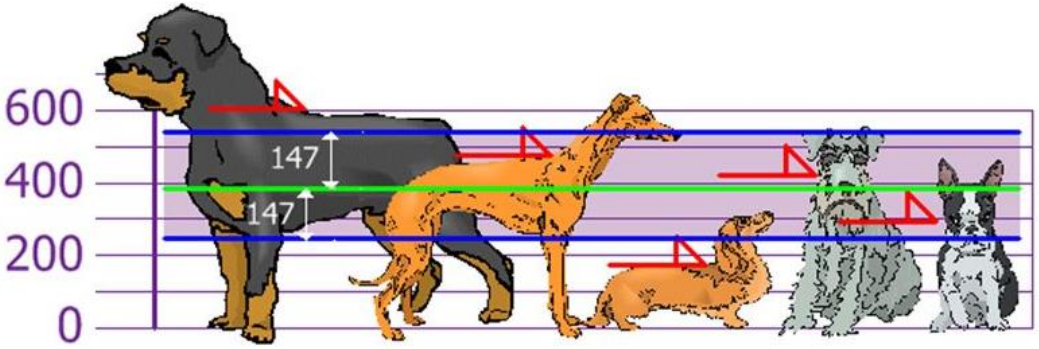


Mean = 394

Std_dev = 147



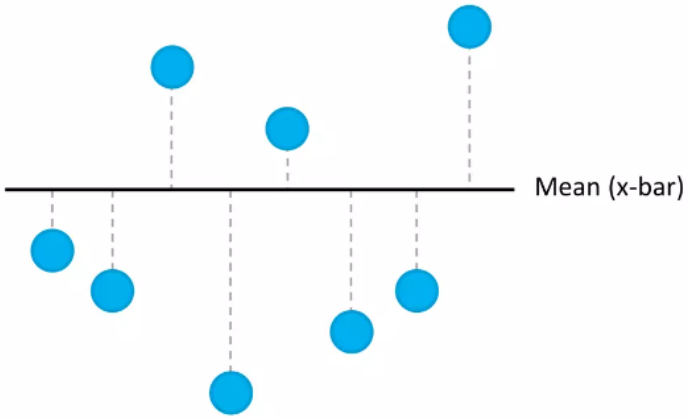
It shows which heights are within one Standard Deviation (147mm) of the Mean



Rottweilers **are** tall dogs and Dachshunds **are** a bit short

Standard deviation - formula

- There are six steps for finding the standard deviation by hand:
  1. List each score and find their **mean**.
  2. **Subtract** the **mean** from each **score** to get the **deviation** from the mean.
  3. **Square** each of these deviations.
  4. Add up all of the squared deviations.
  5. Divide the sum of the squared deviations by N.
  6. Find **the square root** of the number you found.



$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

n is the number of persons

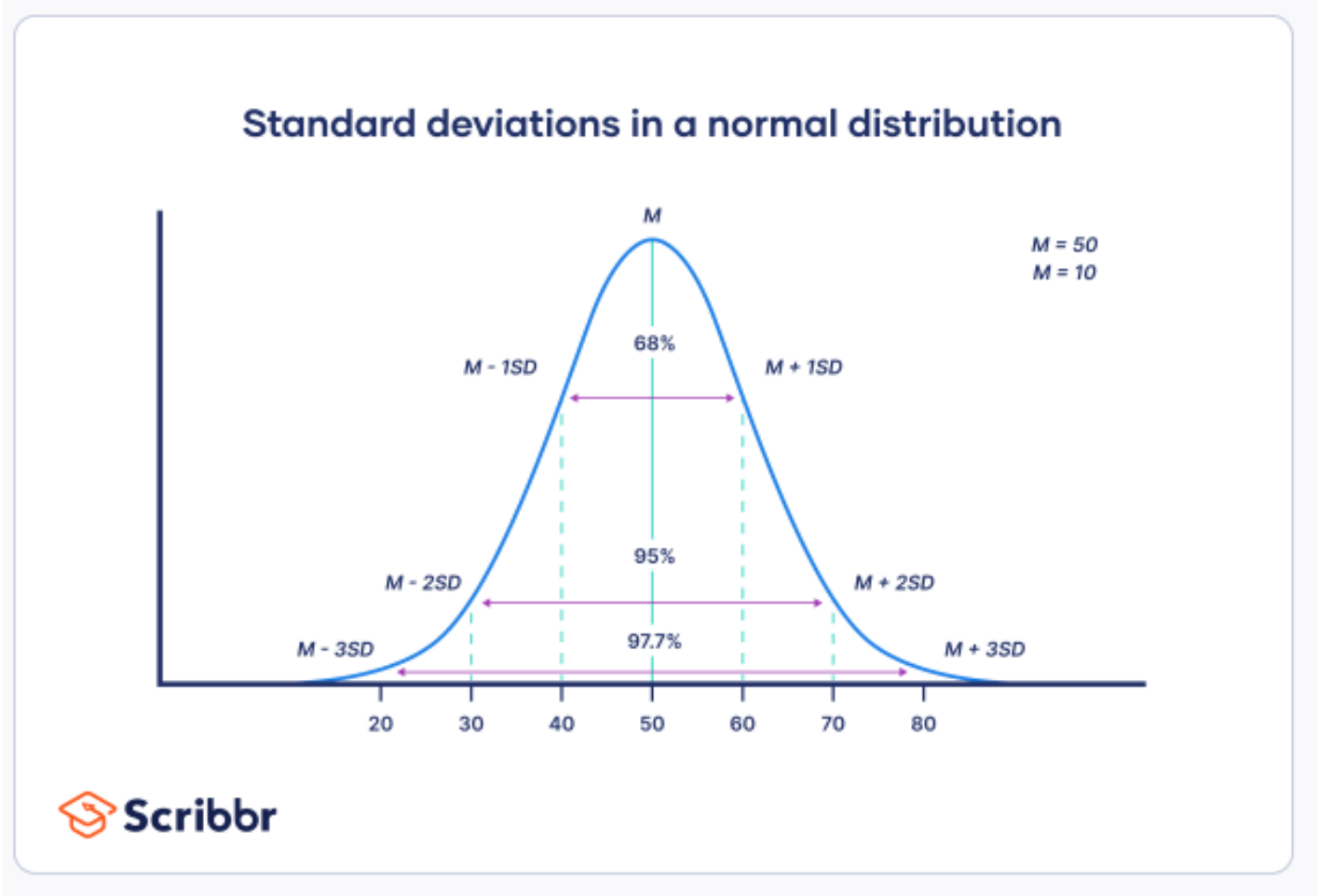
x_i is the size of the individual

$\bar{x}$ is the mean value of all persons

```
In [24]: heights = [600, 470, 170, 430, 300]
         heights = np.array(heights)
         std_heights = np.std(heights)
         std_heights

Out[24]: 147.32277488562318
```

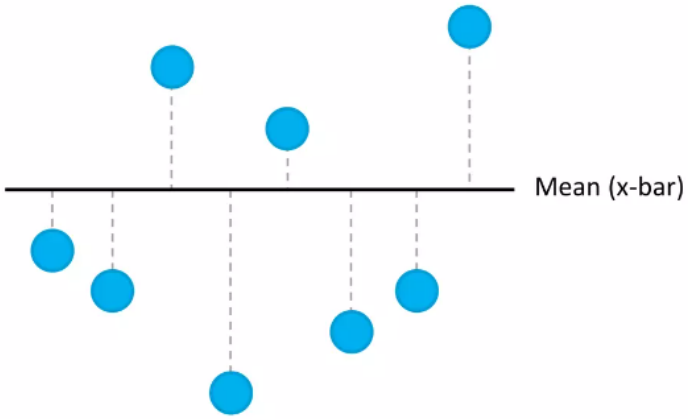
Standard deviation



Variance (დისპერსია) - np.var()

- Standard deviation = Square root of Variance
- **Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.**

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$




```
In [23]: var_heights = np.var(heights)
         var_heights

Out[23]: 21704.0
```

Hypothesis

- A **hypothesis** is an educated guess about something in the world around you.
- It should be testable, either by experiment or observation.
- If you are going to propose a hypothesis, it's customary to write a statement.
- Example:
  - A new medicine you think might work.
  - A way of teaching you think might be better.
- **Hypothesis testing** in statistics is a way for you to test the results of a survey or experiment
- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data
- Example : If I pour less water to the herbs, growth of the plant will increase in size. Size of herb should be measured and technically it should be proven

Types of Hypothesis : H0, Ha

- **Null hypothesis -** the null hypothesis is always the accepted fact. Null hypothesis is denoted by H0: μ1 = μ2, which shows that there is no difference between the two population means.

- **Alternative hypothesis -** Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect. Alternative hypothesis is denoted Ha
- Always try to establish Alternate Hypothesis by rejecting Null Hypothesis.
- The equality sign =, ≤, ≥ should always appear on Null Hypothesis side.

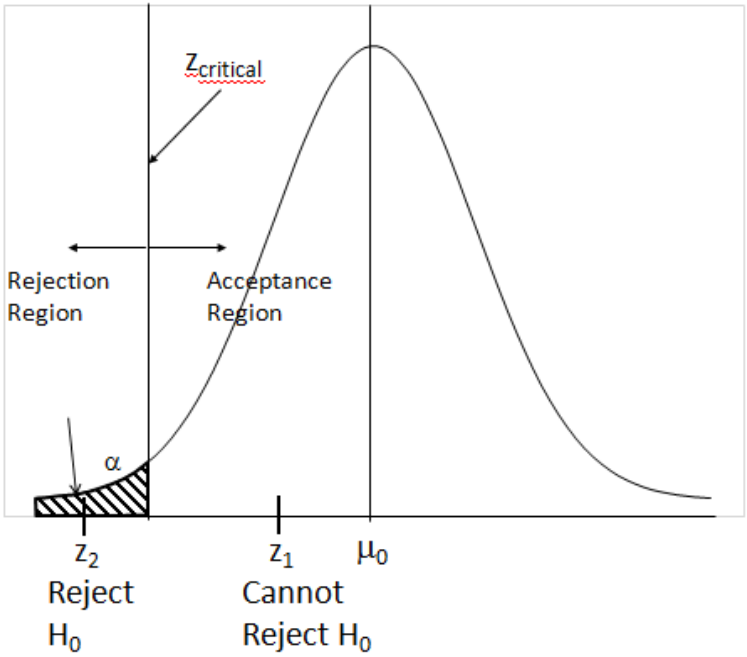
$H_o: \mu \leq \mu_0$ $H_o: \mu \geq \mu_0$ $H_o: \mu = \mu_0$

$H_a: \mu > \mu_0$ $H_a: \mu < \mu_0$ $H_a: \mu \neq \mu_0$

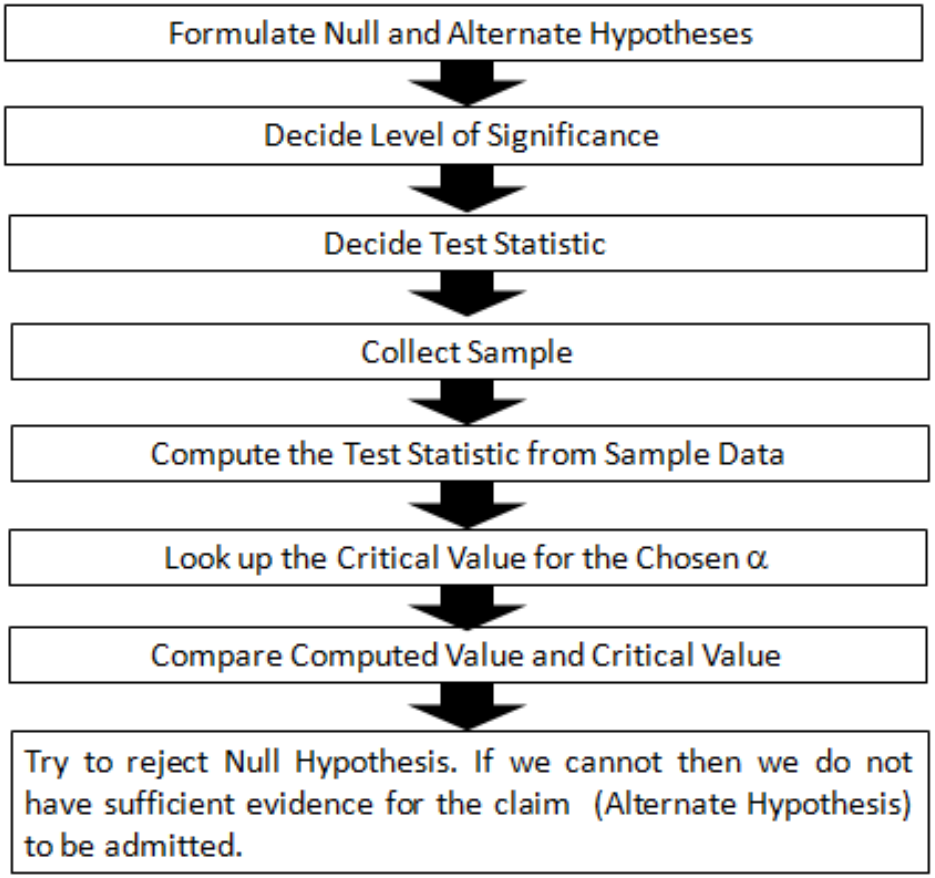
Hypothesis Testing - Example

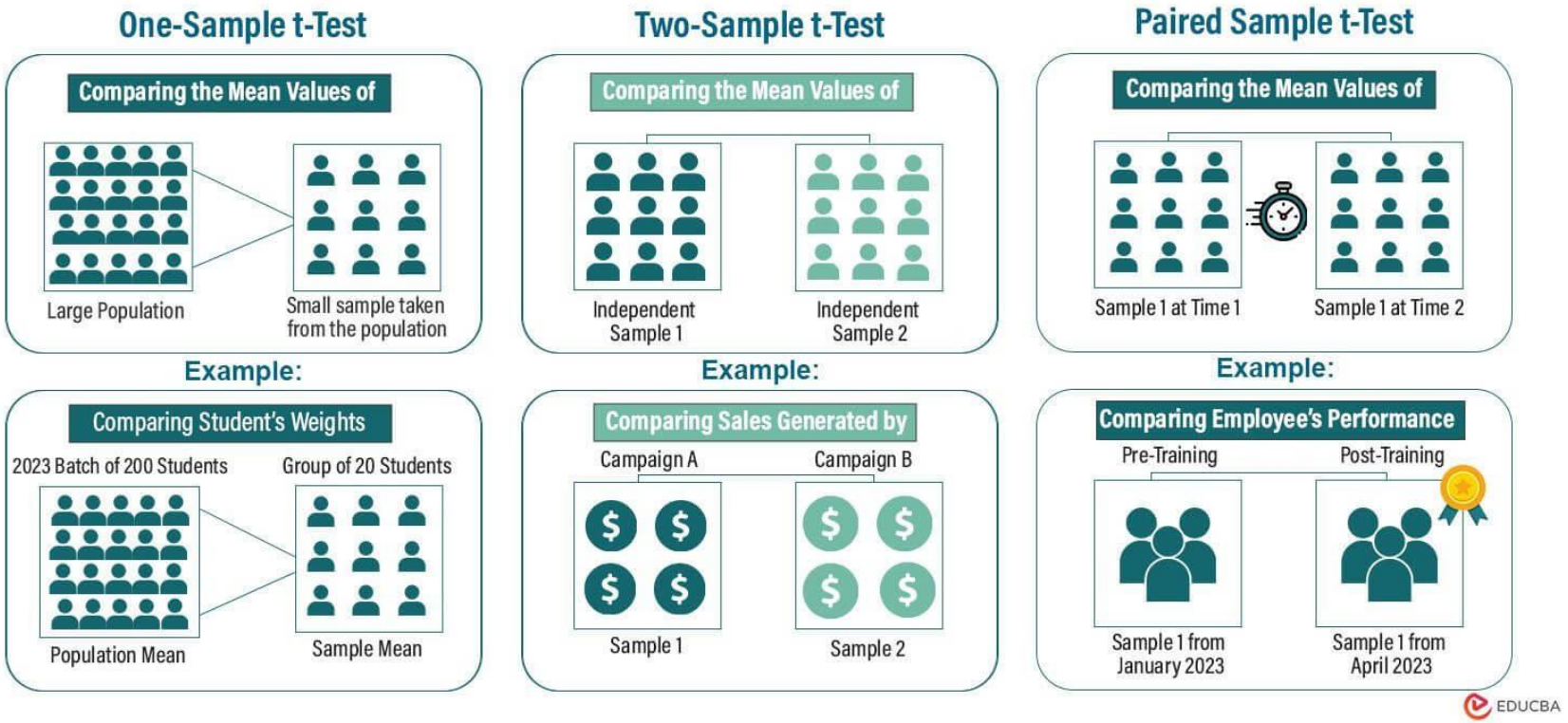
Test Type	Description	Ho / Ha
Testing Research Hypothesis	R&D dept has developed a new battery with higher Ah. The present performance is 80Ah. The new product has a higher performance.	Ho: $\mu \leq 80$
		Ha: $\mu > 80$
Validating a claim	A claim is made that average inflation rate is less than 6.76%	Ho: $\mu \geq 6.76$
		Ha: $\mu < 6.76$
Testing Decision making situations	Crime rate in North Chennai is different from that of South Chennai = 453/year	Ho: $\mu = 453$
		Ha: $\mu \neq 453$

Hypothesis Testing – Identify Rejection area



Hypothesis Testing Procedure




One-Sample t-Test
Comparing the Mean Values of
Large Population
Small sample taken from the population
Example:
Comparing Student's Weights
2023 Batch of 200 Students
Group of 20 Students
Population Mean
Sample Mean
Two-Sample t-Test
Comparing the Mean Values of
Independent Sample 1
Independent Sample 2
Example:
Comparing Sales Generated by
Campaign A
Campaign B
Sample 1
Sample 2
Paired Sample t-Test
Comparing the Mean Values of
Sample 1 at Time 1
Sample 1 at Time 2
Example:
Comparing Employee's Performance
Pre-Training
Post-Training
Sample 1 from January 2023
Sample 1 from April 2023
EDUCBA