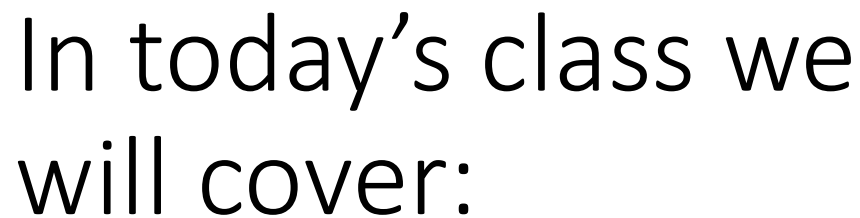


Statistics for Data Analytics

Lecturer: Marina Iantorno

E-mail: miantorno@cct.ie





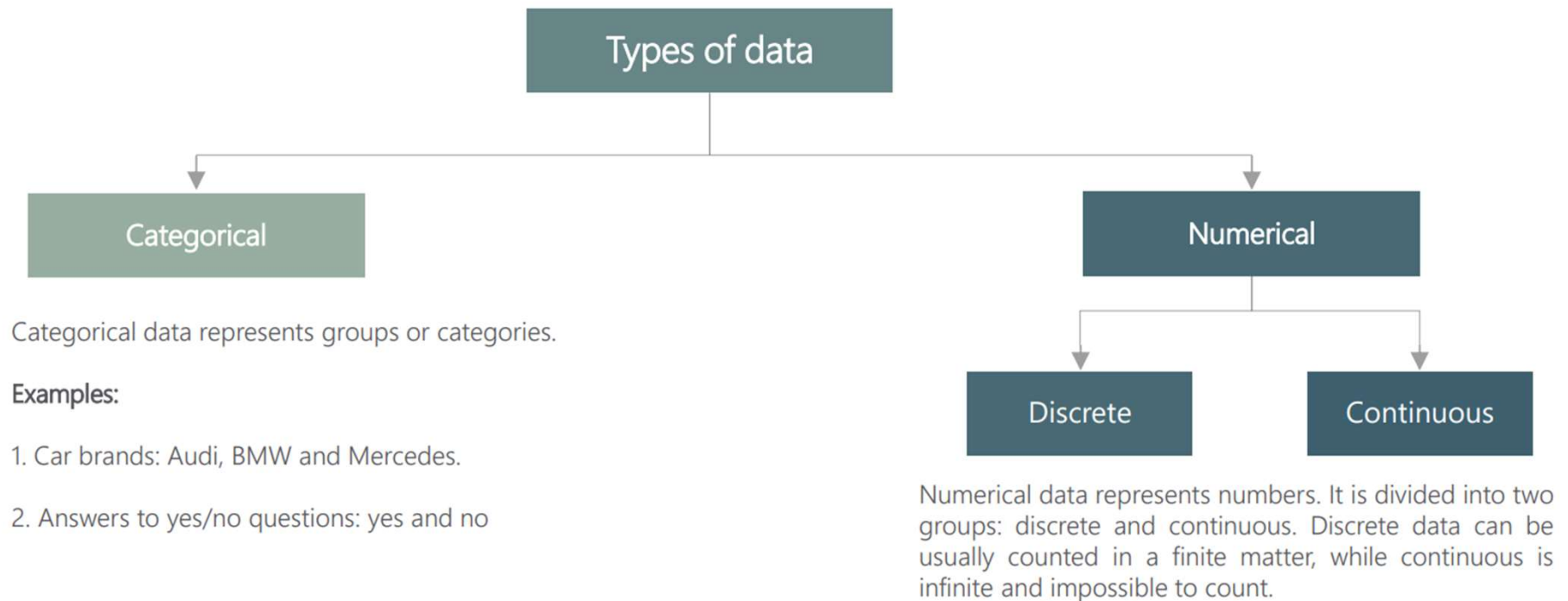
- ❑ Definition of Statistics
- ❑ Central tendency measures and their properties
- ❑ Variation measures and their properties



DESCRIPTIVE STATISTICS

Descriptive Statistics

We use Statistics to describe something about the population. But it is important to understand that



Descriptive Statistics

But what is Descriptive Statistics? As its name indicates, Descriptive Statistics is a group of numeric and graphic techniques that describe and analyse certain population, usually through samples.

The main goal could be:

- Transform the data in information.
- Organise the data
- Obtain concrete information to resume the main characteristics of the variables.
- Offer us an input for Inferential Statistics.



Descriptive Statistics

To be able to get insights about a population we need to take a sample from that population.

- ✓ Random Sample: In a random sample, the probability of ending up in the sample is the same for each element of a population. An example of a simple random sample is the drawing of a ball from a drum or tombola, such as the weekly lottery draw.
- ✓ Stratified: In a stratified random sampling, the elements of a population are divided into groups based on a previously determined attribute. Subsequently, a simple random sample is performed within each of those groups.
- ✓ Cluster sample: Different to simple random selection, individuals are not selected at random, rather entire groups. The cluster analysis is randomly drawn up from all clusters, and all characteristic attributes are examined within any given cluster.

Descriptive Statistics

Careful! Not all the samples are representatives.

A representative sample is a sample that accurately represents the characteristic (or characteristics) of a population.

It is vital to take the sample randomly, otherwise the data would be biased, which means that a sample was collected in such a way that some members of the intended population had a lower or higher probability to be chosen than others. It results in a biased sample of a population, and sometimes the results can be erroneously attributed or influenced by the person in charge of taking the sample.

Let's try an exercise together!

Descriptive Statistics

A laboratory in Dublin wants to gather information about the ages of those who got infected with COVID-19 in Ireland. In order to do so, 10 people were surveyed and these are their answers.

Person	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Age	85	10	90	95	75	20	80	05	60	10







We need to transform these numbers in valuable or useful information.

Our next step, is to define the variable in place. Also, we will analyse this table, but before getting started, we need to sort the data from the smallest to the largest values.

Descriptive Statistics

A laboratory in Dublin wants to gather information about the ages of those who got infected with COVID-19 in Ireland. In order to do so, 10 people were surveyed and these are their answers.

X = age of people who got COVID-19 in Ireland (in years)

Person	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Age (X)	5	10	10	20	60	75	80	85	90	95
										
	X_1	X_2	X_3	X_4	...					X_n

The sample size or number of observations is called “ n ”. In this case $n = 10$.

Descriptive Statistics

➤ Central Tendency Measures

These measures will allow us to get values that resume in one number all the values that the sample provides.

➤ Variation Measures

These measures help us to get values that determine the level of homogeneity within the observations. In other words, we can see through the variation measures how similar/different the values are.

Descriptive Statistics

CENTRAL TENDENCY METRICS

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

Descriptive Statistics

Central Tendency Measures

➤ Mean : \bar{x}

This is the typical average. The calculation is pretty straight forward. We need to sum all the values and divide that result on the total of observations.

$$\bar{x} = \frac{5 + 10 + 10 + 20 + 60 + 75 + 80 + 85 + 90 + 95}{10} = 53 \text{ years old}$$



Formula

$$\bar{x} = \frac{\sum x}{n}$$

Answer: The average age of the people who got infected with covid is 53 years old

Descriptive Statistics

Central Tendency Measures

Advantages of the Mean

- It is easy to calculate it
- It is easy to interpret it
- Existence property: there is **always** a result
- Its result is **always** unique
- Sufficiency property: it is necessary to use all the values to calculate it.

Disadvantages of the Mean

- It is affected by extreme values

Imagine that by a miracle we had a person who is 120 years old , our result would be different and not representative. Those non representative values are known as “outliers”.

Descriptive Statistics

Central Tendency Measures

➤ Mode : Mo

This is the number that presents more repetitions within the sample, regardless of how many times it appears.

We get it by observing our data, there is no a specific formula to calculate it.

5	10	10	20	60	75	80	85	90	95
---	----	----	----	----	----	----	----	----	----

Mo = 10 years old

Answer: The most frequent age of people who got COVID=19 is 10 years old

Descriptive Statistics

Central Tendency Measures

Advantages of the Mode

- It is easy to get it (no calculation involved)
- It is easy to interpret it
- It is not affected by extreme values

Disadvantages of the Mode

- It is not necessary to use all the values to calculate it.
- There could be more than one result or even none.

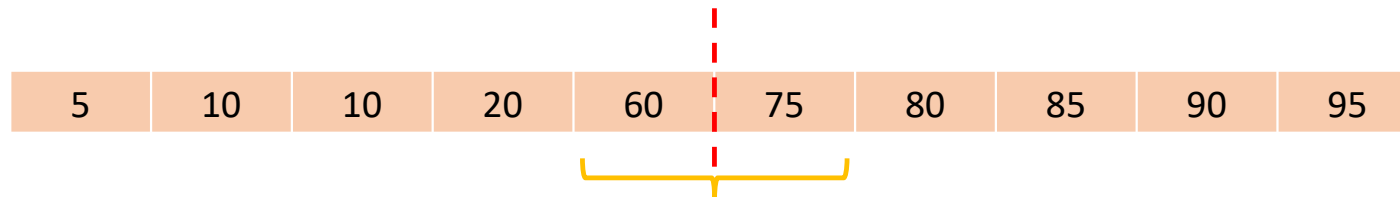
Descriptive Statistics

Central Tendency Measures

➤ Median : Me

This is the value located exactly in the middle of the ordered sequence of values.

We sort the data from the smallest to the largest and we “cut” the sequence in the middle. That value is the Median.

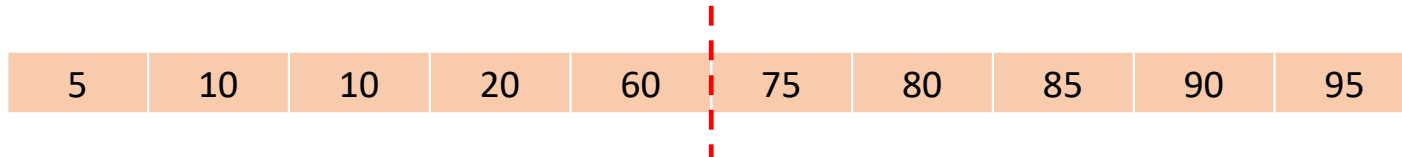


$$Me = \frac{60 + 75}{2} = 67.5$$

Descriptive Statistics

Central Tendency Measures

➤ Median : Me



$$Me = \frac{60 + 75}{2} = 67.5$$

If the number of observations is an even number as in our case, we will get two values, and what we have to do is to calculate an average between them to get the median.

If the number of observations is an odd number, then, we will cut exactly where the median is, and that would be the result.

Answer: The maximum age of the middle of the sample 67.5 years old

Descriptive Statistics

Central Tendency Measures

Advantages of the Median

- It is easy to get it (no real calculation involved)
- Existence property: there is **always** a result
- Its result is **always** unique
- It is not affected by extreme numbers

Disadvantages of the Median

- It is not so easy to interpret. Sometimes the interpretation of the median create confusions.
- There is a debate to decide whether or not the median uses all the values of the sample to be calculated.

Descriptive Statistics

VARIATION METRICS



Descriptive Statistics

Variation Measures

➤ Variance : S^2

The variance by itself does not say much, but through the Standard Deviation, we could know the difference between the values respect to the mean.

This is most commonly used in Finance and Economy. It is useful to measure the volatility (risk) of a variable.

Descriptive Statistics

Variation Measures

➤ Variance : S^2

x	\bar{x}	$(x - \bar{x})$	$(x - \bar{x})^2$
5	53	-48	2304
10	53	-43	1849
10	53	-43	1849
20	53	-33	1089
60	53	7	49
75	53	22	484
80	53	27	729
85	53	32	1024
90	53	37	1369
95	53	42	1764
			12510



Formula

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

$$S^2 = \frac{12510}{9} = 1390 \text{ years}^2$$

Answer: The variance is 1390 years²

Descriptive Statistics

Variation Measures

➤ Standard Deviation : S

This is the square root of the variance. With the standard deviation we go back to unit of measurement, and we will be able to see the difference of the values respect the mean.

$$S = \sqrt{1390 \text{ years}^2} = 37.28 \text{ years}$$

Answer: The average age of the infected people is 53 years old +- 37,28 years.

$$VC = S / \bar{X} * 100 =$$

$$VC = (37.28/53) * 100 = 70.33\%$$

Descriptive Statistics

Now let's try all these calculations in Excel and Python!



THAT'S ALL FOR TODAY

THANK YOU

