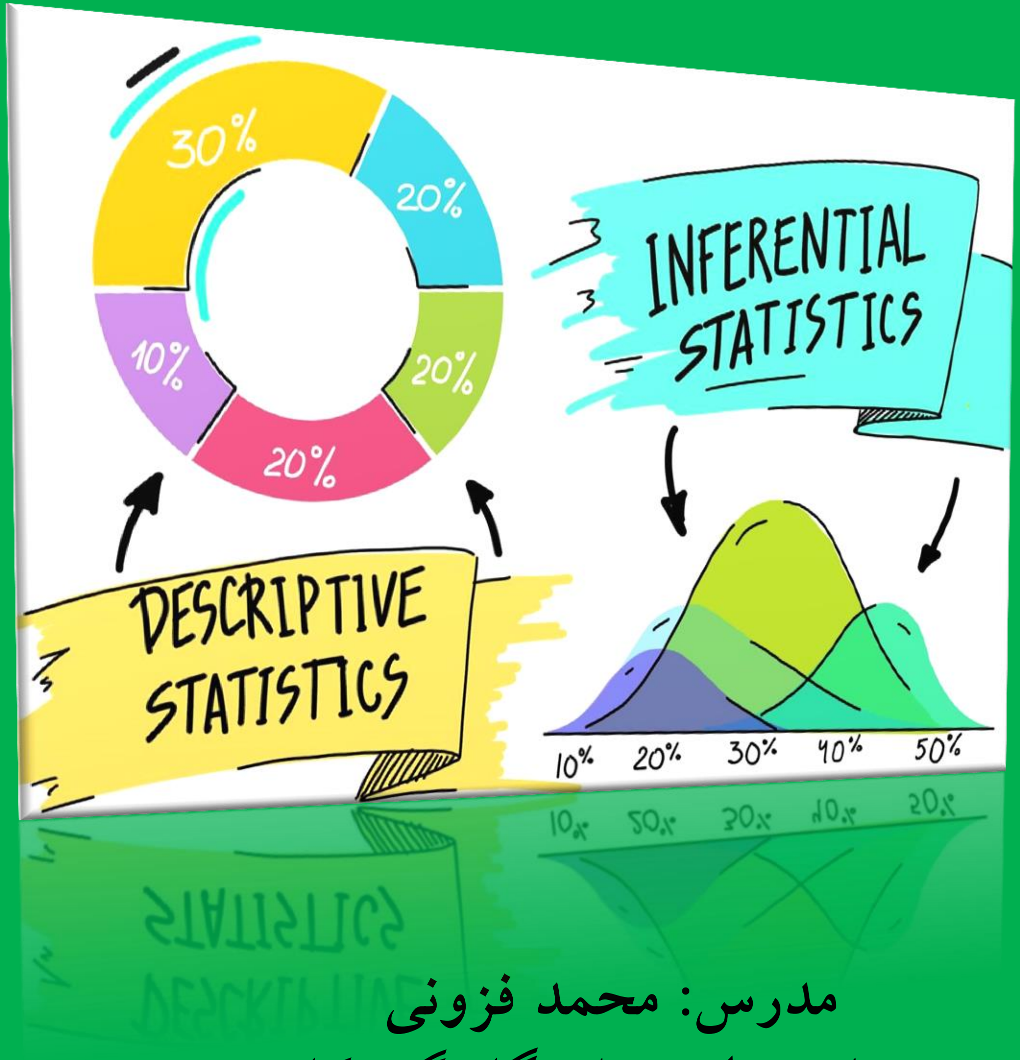
دورهٔ آموزشی «علم داده»

Data Science Course

جلسهٔ هشتم:
آمار توصیفی
و تحلیل داده‌ها



مدرس: محمد فزونی
عضو هیات علمی دانشگاه گنبدکاووس

پائیز ۱۳۹۹

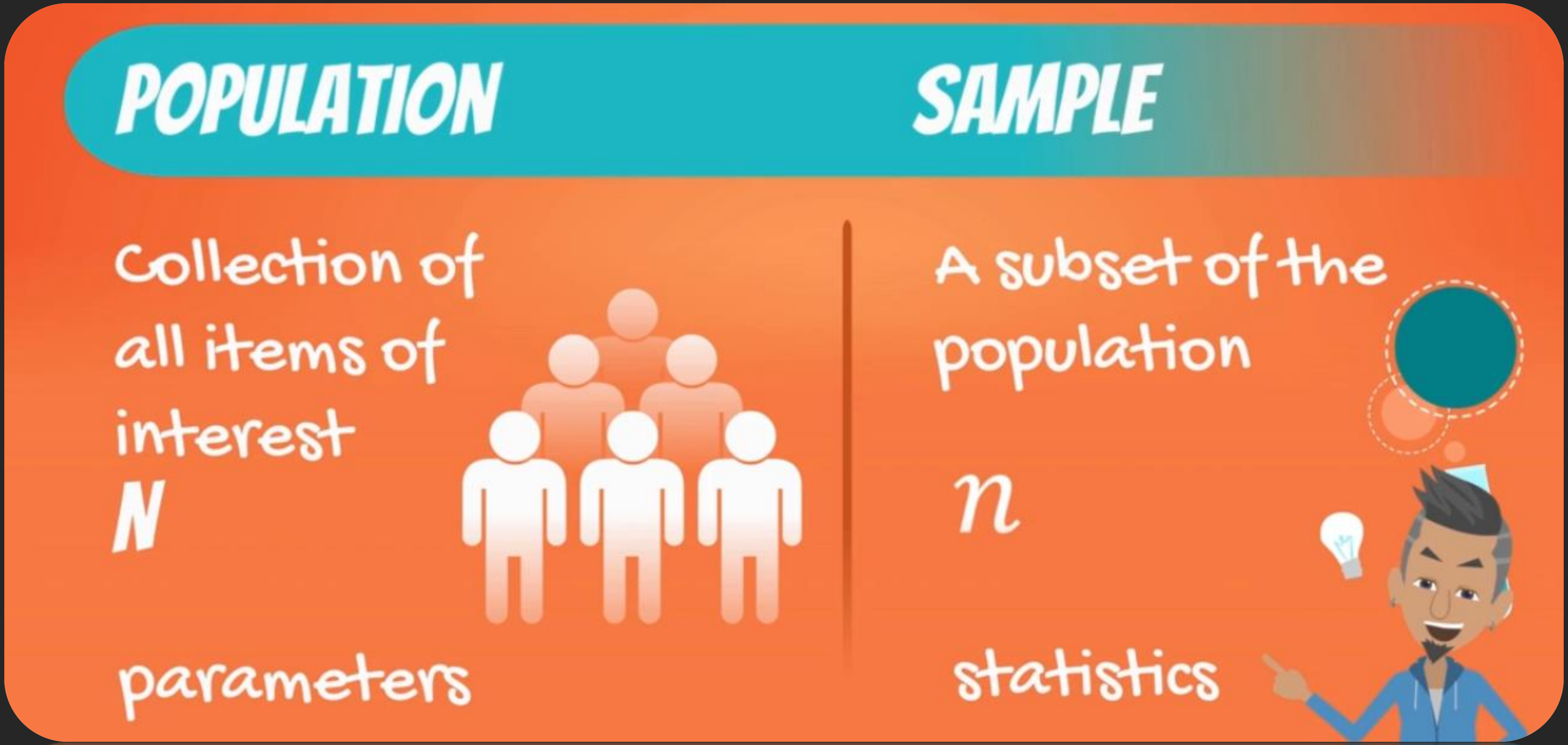
About me…

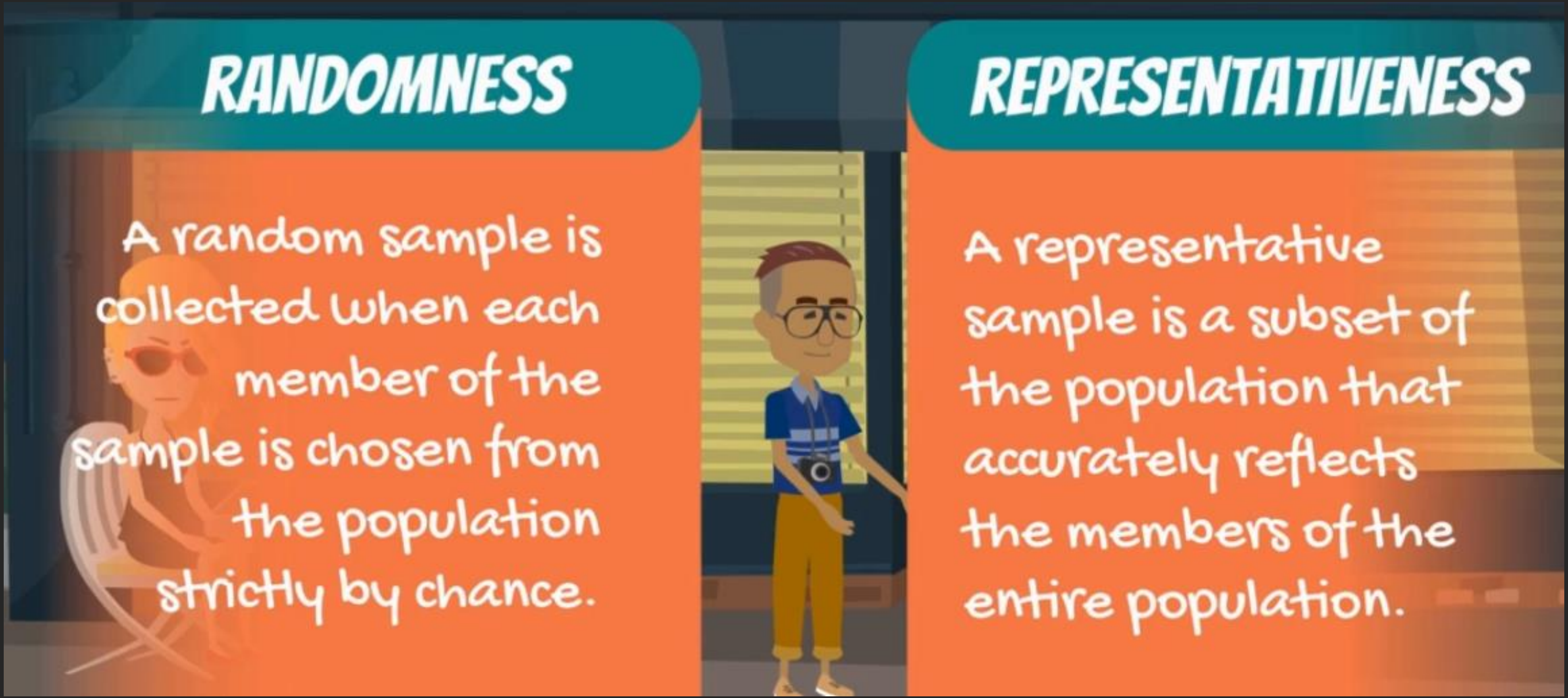
Mohammad Fozouni (Ph.D.)
Dep. of Math & Statistics
Gonbad Kavous University

- fozouni@hotmail.com
- https://m-fozouni.ir
- http://profs.gonbad.ac.ir/fozouni/en

#data_science_fozouni

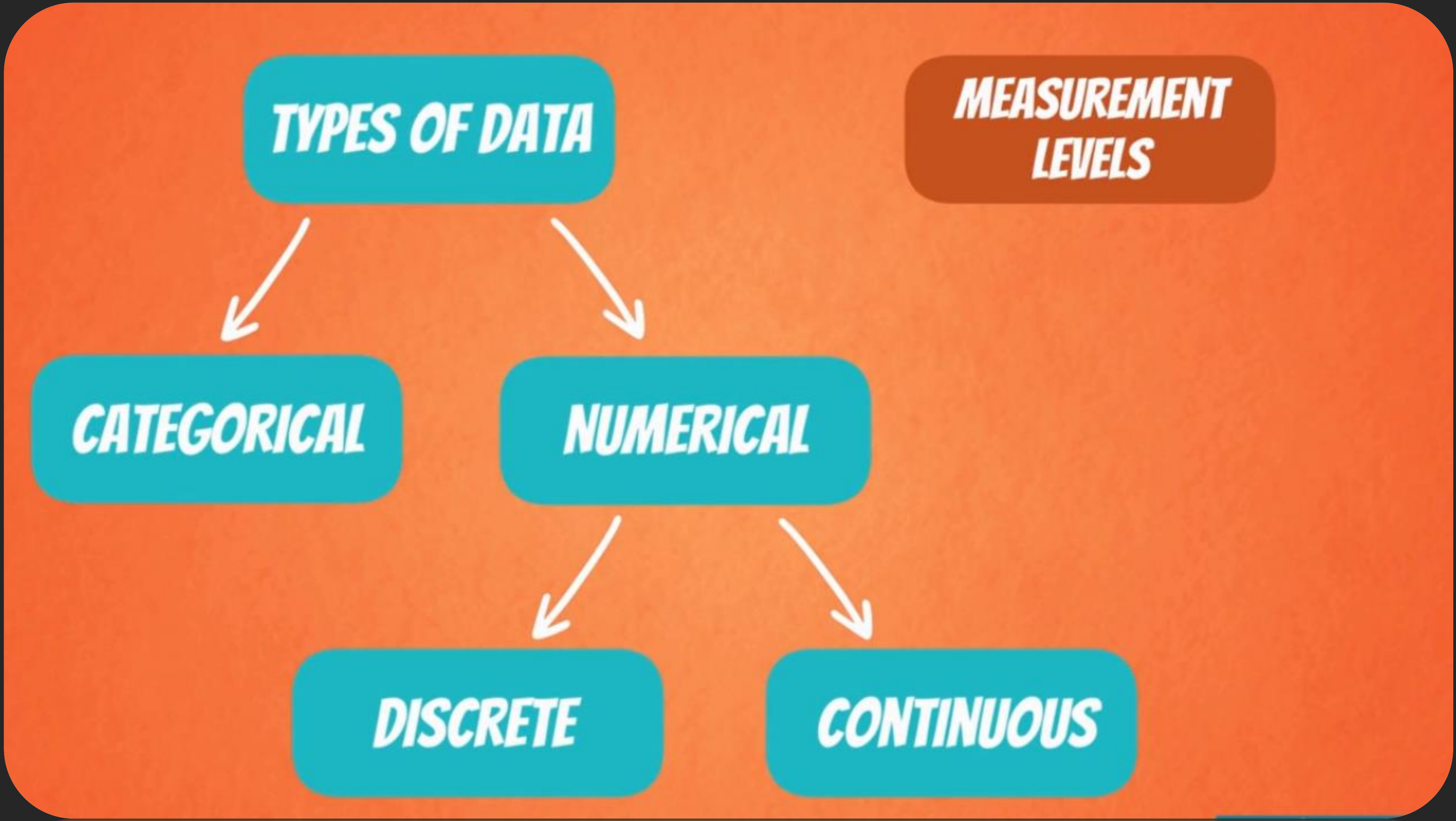
Realm of STATISTICS


POPULATION
SAMPLE
Collection of all items of interest
N
parameters
A subset of the population
n
statistics



RANDOMNESS
A random sample is collected when each member of the sample is chosen from the population strictly by chance.
REPRESENTATIVENESS
A representative sample is a subset of the population that accurately reflects the members of the entire population.


How to draw a sample that is both random and representative?


TYPES OF DATA
MEASUREMENT LEVELS
CATEGORICAL
NUMERICAL
DISCRETE
CONTINUOUS

Types of data

Categorical

Categorical data represents groups or categories.

Examples:

1. Car brands: Audi, BMW and Mercedes.

2. Answers to yes/no questions: yes and no

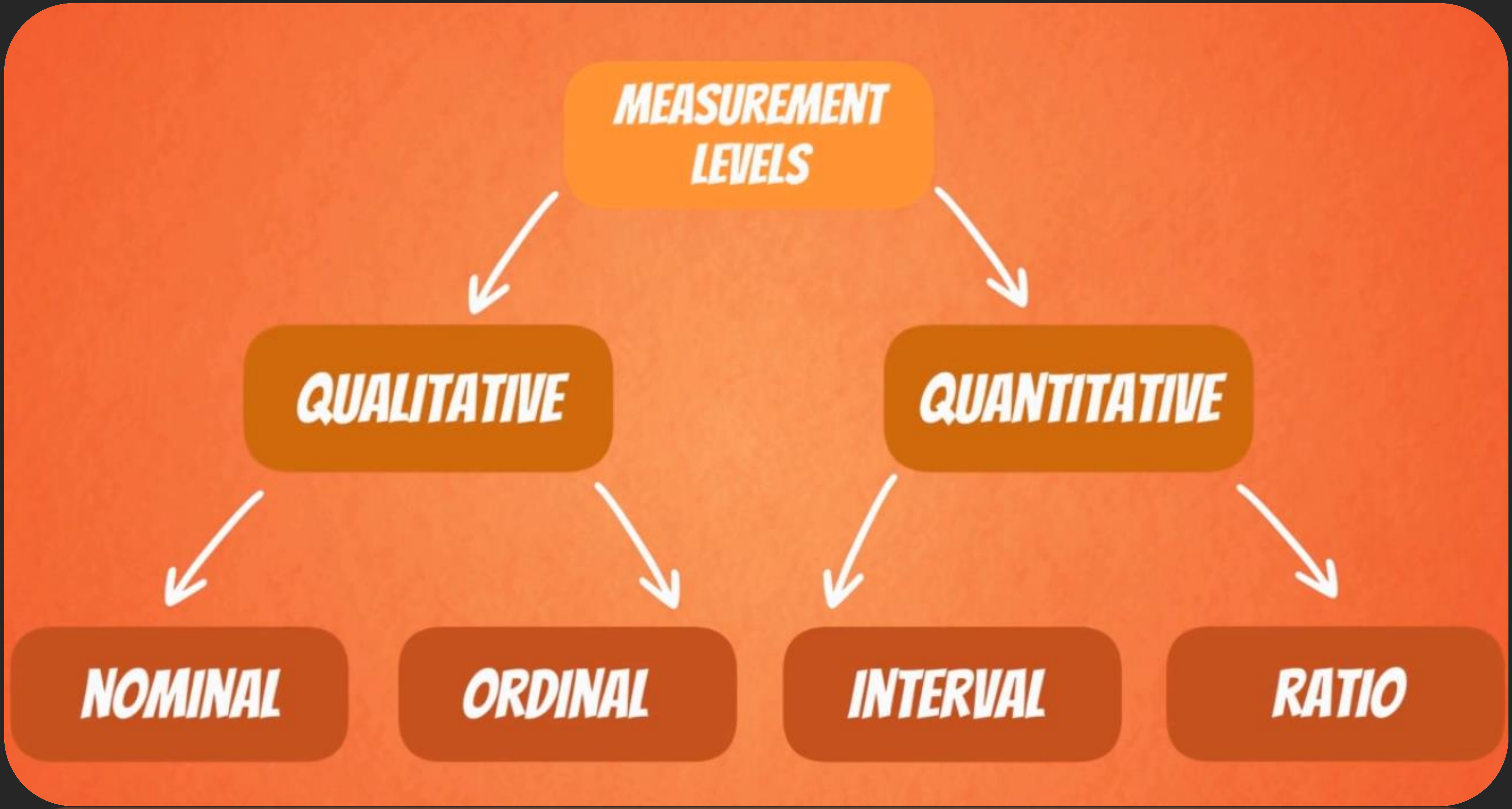
Numerical

Discrete

Continuous

Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

Examples:
Discrete: # children you want to have, SAT score
Continuous: weight, height


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO

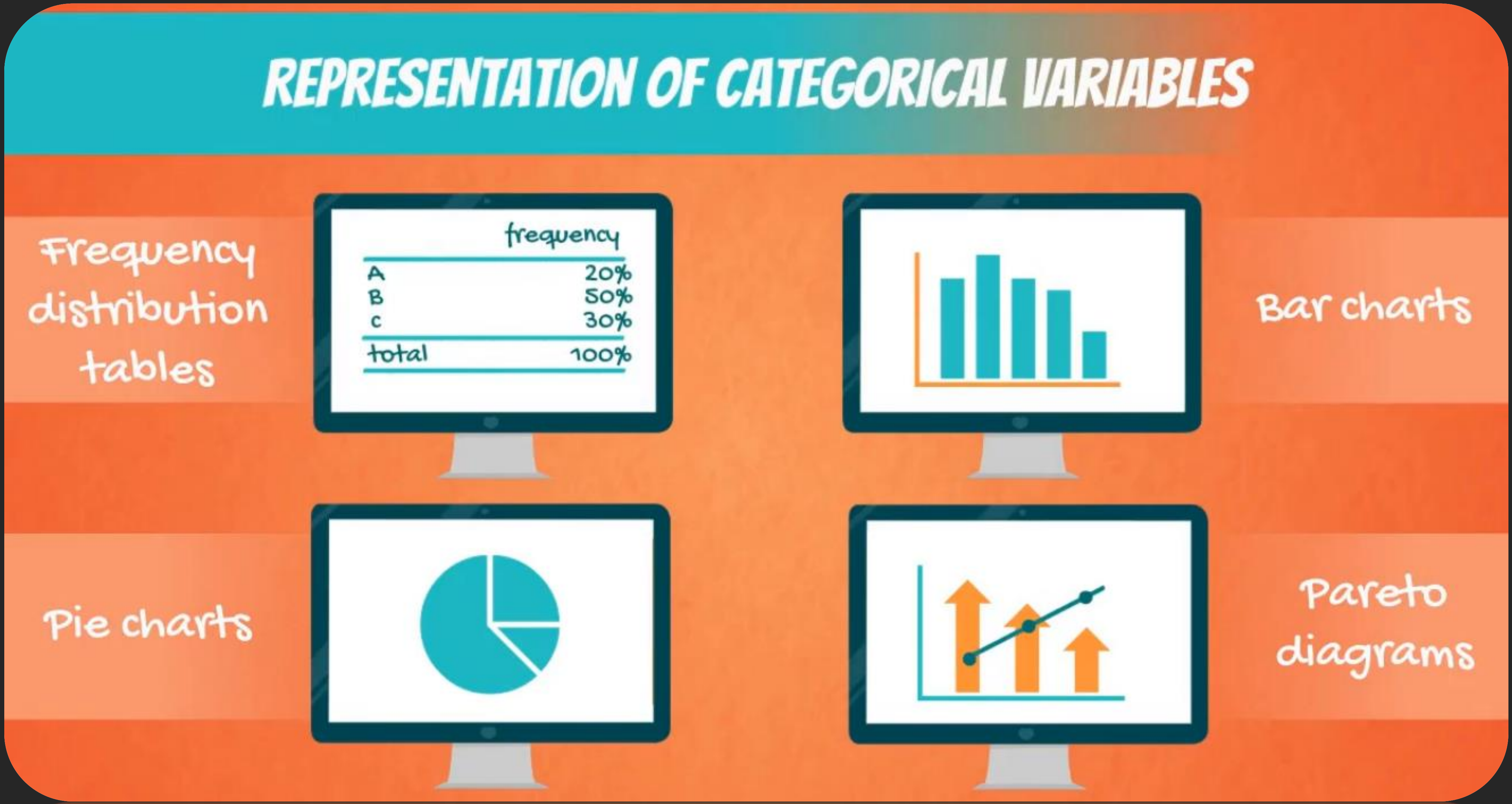
Levels of measurement

There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that **can** be ordered.

Examples:
Nominal: four seasons (winter, spring, summer, autumn)
Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios **have a true zero,** while intervals don't.

Examples:
Interval: degrees Celsius and Fahrenheit
Ratio: degrees Kelvin, length


REPRESENTATION OF CATEGORICAL VARIABLES
Frequency distribution tables
frequency
A 20%
B 50%
C 30%
total 100%
Bar charts
Pie charts
Pareto diagrams

Frequency distribution tables

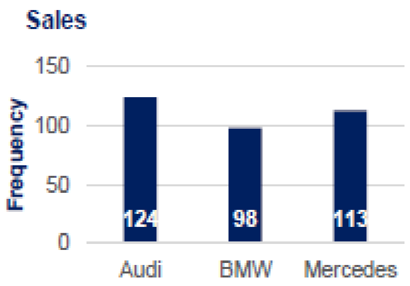
Bar charts

Pie charts

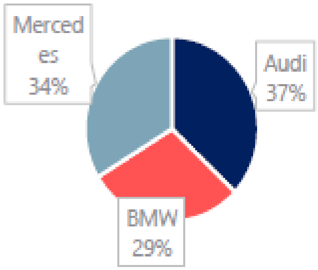
Pareto diagrams

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

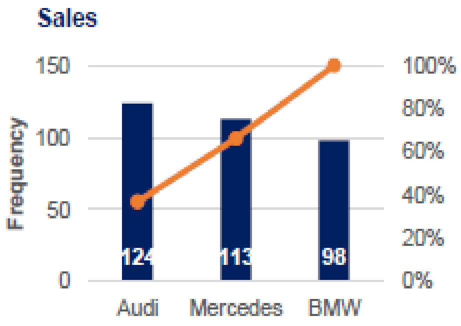
Frequency distribution tables show the category and its corresponding absolute frequency.



Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.

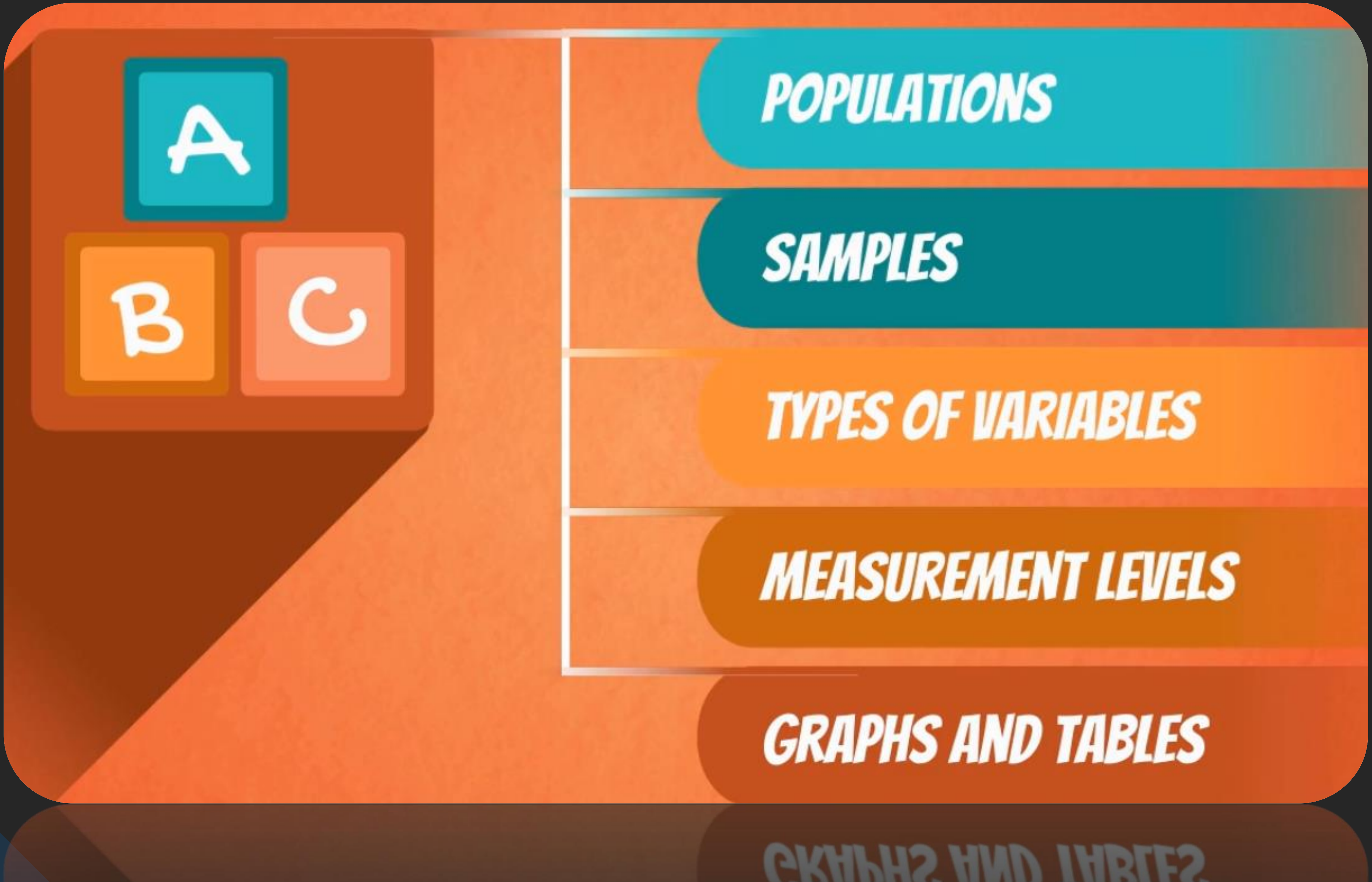


Pie charts are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.

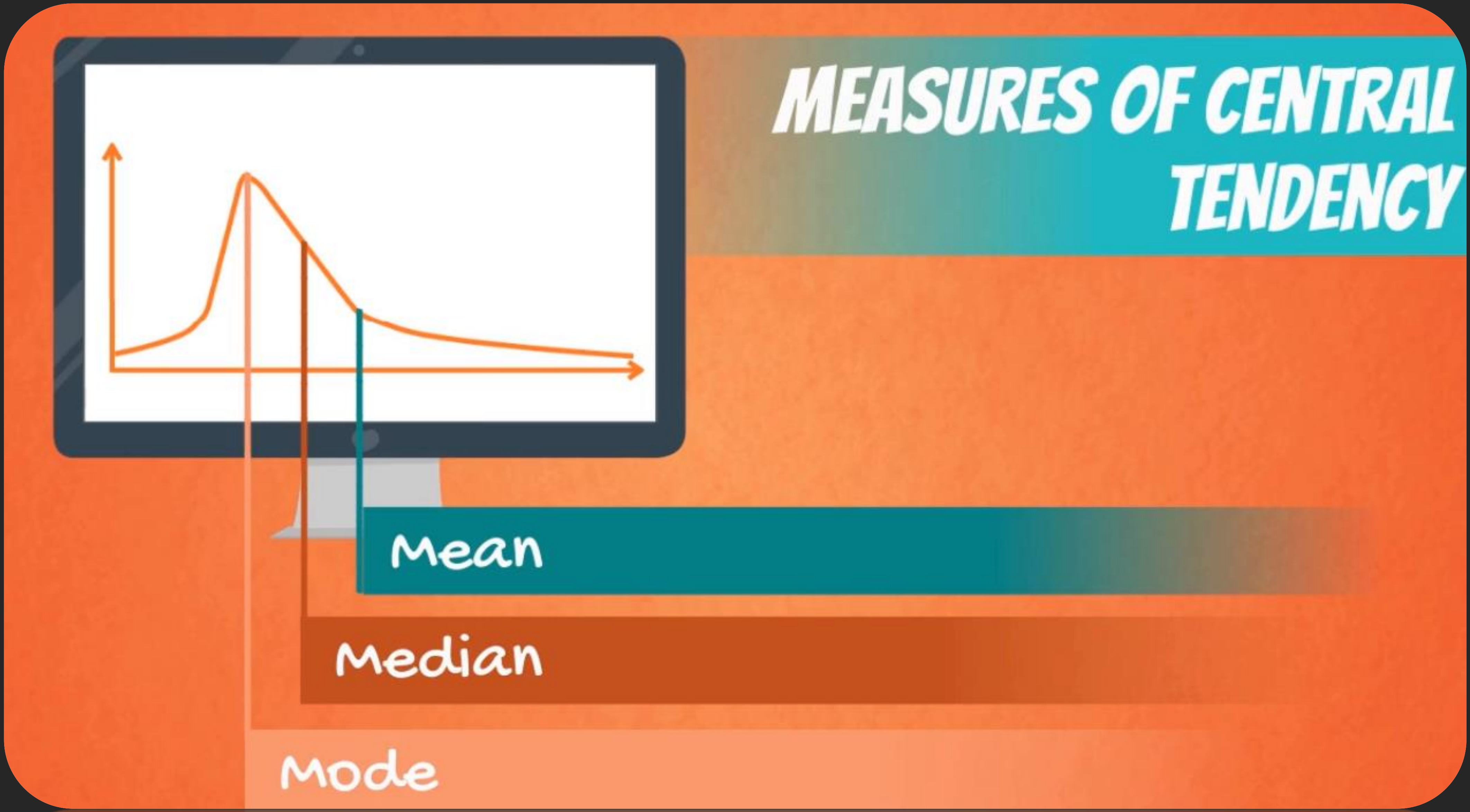


The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

www.m-fozouni.ir
Data Science Course



www.m-fozouni.ir
Data Science Course


MEASURES OF CENTRAL TENDENCY
Mean
Median
Mode

Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset.

Note: easily affected by outliers

The formula to calculate the mean is:

$$\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{or}$$

$$\frac{x_1 + x_2 + x_3 + \cdots + x_{N-1} + x_N}{N}$$

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.

In an ordered dataset, the median is the number at position $\frac{n+1}{2}$.

If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.

The mode is calculated simply by finding the value with the highest frequency.

Mean, median, mode

Pizza prices example

Position	New York City		Los Angeles	
1	$	1.00	$	1.00
2	$	2.00	$	2.00
3	$	3.00	$	3.00
4	$	3.00	$	4.00
5	$	5.00	$	5.00
6	$	6.00	$	6.00
7	$	7.00	$	7.00
8	$	8.00	$	8.00
9	$	9.00	$	9.00
10	$	11.00	$	10.00
11	$	66.00		

	New York City		Los Angeles	
Mean	$	11.00	$	5.50
Median	$	6.00	$	5.50
Mode	$	3.00		-

Which measure is best?

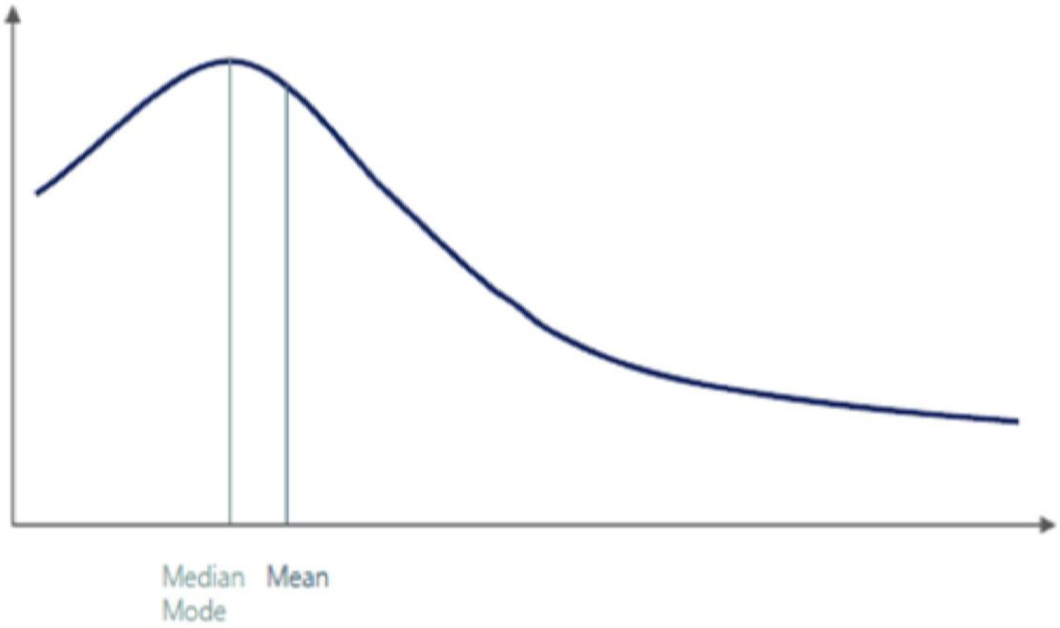
There is no best, but using only one is definitely the worst!

one is definitely the worst!

SAMPLE SKEWNESS FORMULA

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}^{\,3}}$$

Skewness



Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side.

Right (positive) skewness looks like the one in the graph. It means that the **outliers** are to the right (long tail to the right).

Left (negative) skewness means that the outliers are to the left.

Usually, you will use software to calculate skewness.

ewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

Zero (no skew)

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

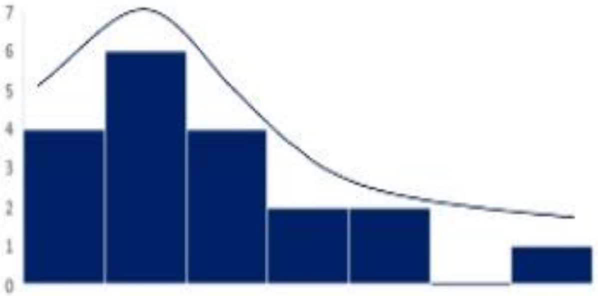
Negative (left)

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
6
7
7
7

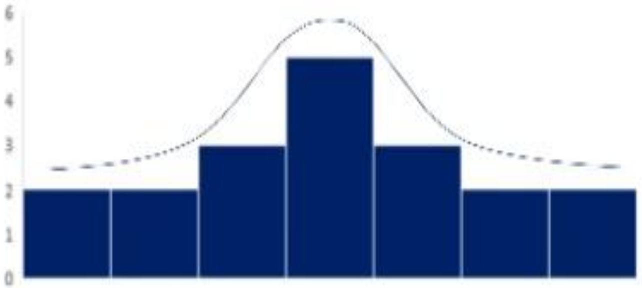
Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00

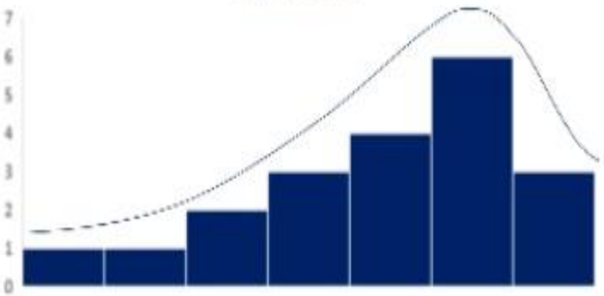
Positive skew



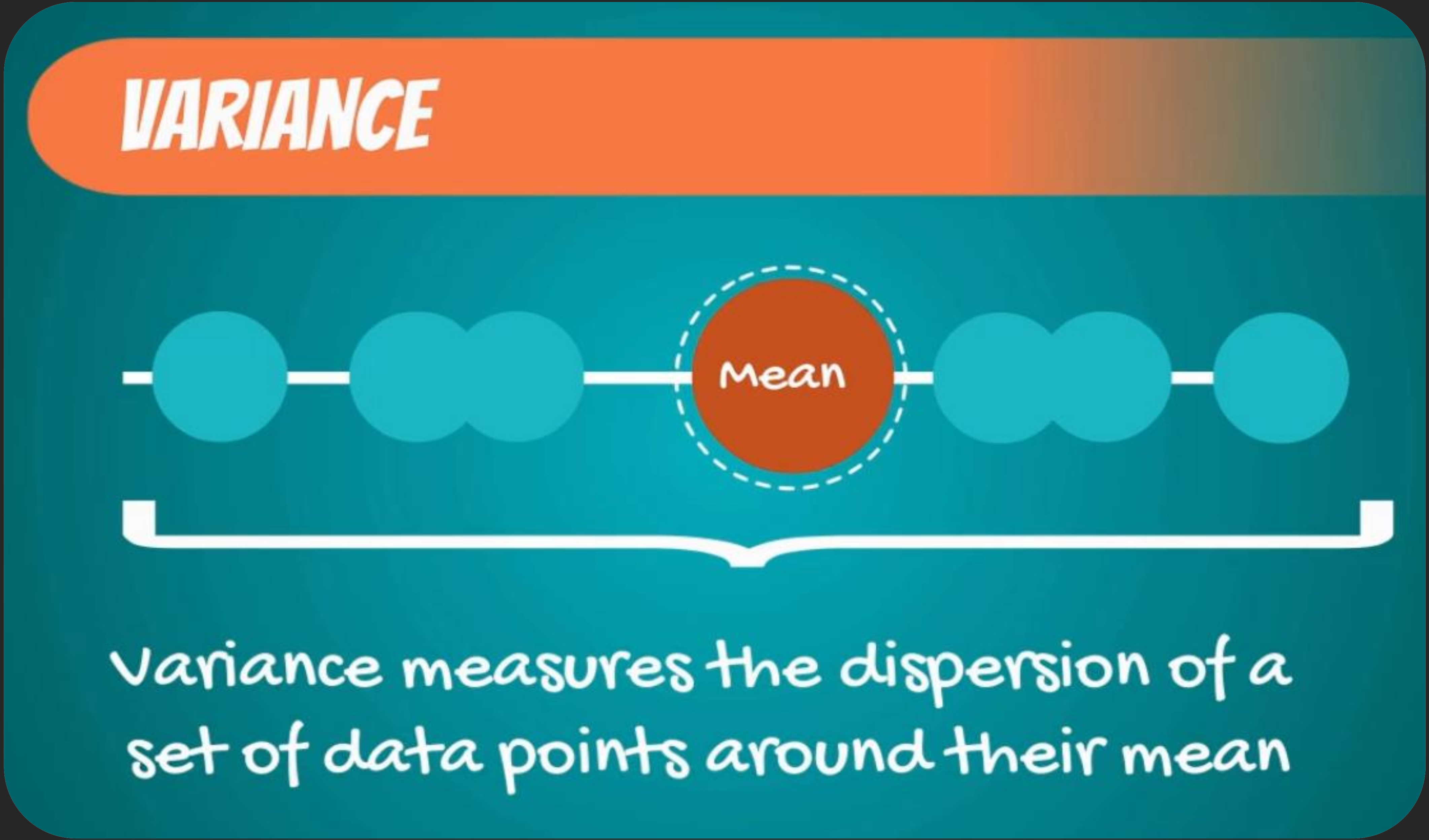
Zero skew



Negative skew



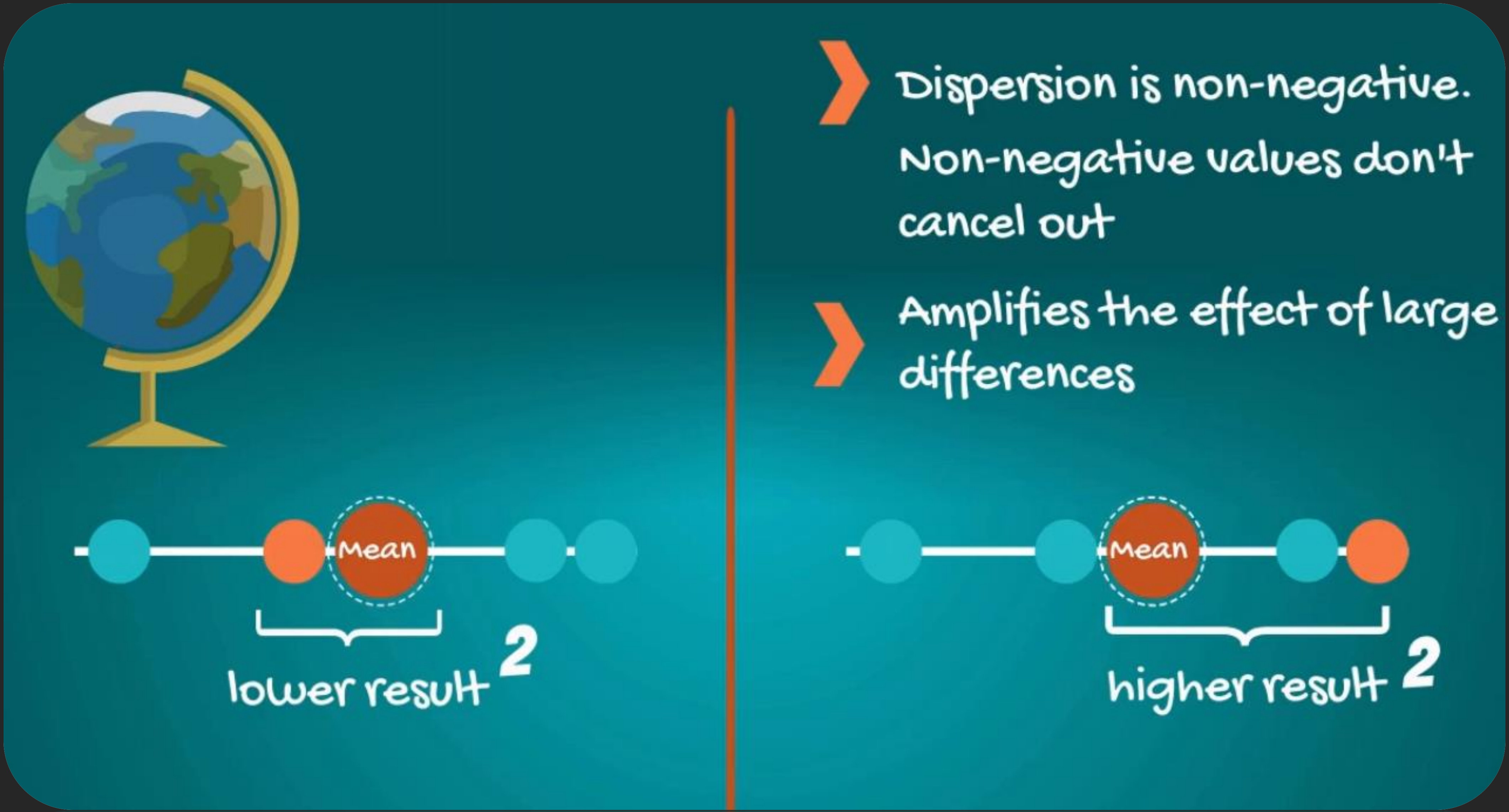
www.m-fozouni.ir
Data Science Course



VARIANCE

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

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


Dispersion is non-negative.
Non-negative values don't cancel out
Amplifies the effect of large differences
Mean
lower result 2
Mean
higher result 2

Variance and standard deviation measure the dispersion of a set of data points around its mean value.

There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas. More on the mathematics behind it.

Sample variance formula: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Population variance formula: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Sample standard deviation formula: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Population standard deviation formula: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

Standard deviation and coefficient of variation

Pizza price example

NY Dollars		Pesos	
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

	Dollars		Pesos	
Mean	$	5.50	MXN	103.46
Sample variance	$\$^2$	10.72	MXN^2	3793.69
Sample standard deviation	$	3.27	MXN	61.59
Sample coefficient of variation		0.60		0.60

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

Covariance and correlation

Covariance

Covariance is a measure of the joint variability of two variables.

- A positive covariance means that the two variables move together.
- A covariance of 0 means that the two variables are independent.
- A negative covariance means that the two variables move in opposite directions.

Covariance can take on values from -∞ to +∞ . This is a problem as it is very hard to put such numbers into perspective.

Sample covariance formula: $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$

Population covariance formula: $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$

Correlation

Correlation is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result.

- A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other.
- A correlation of 0 means that the variables are independent.
- A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

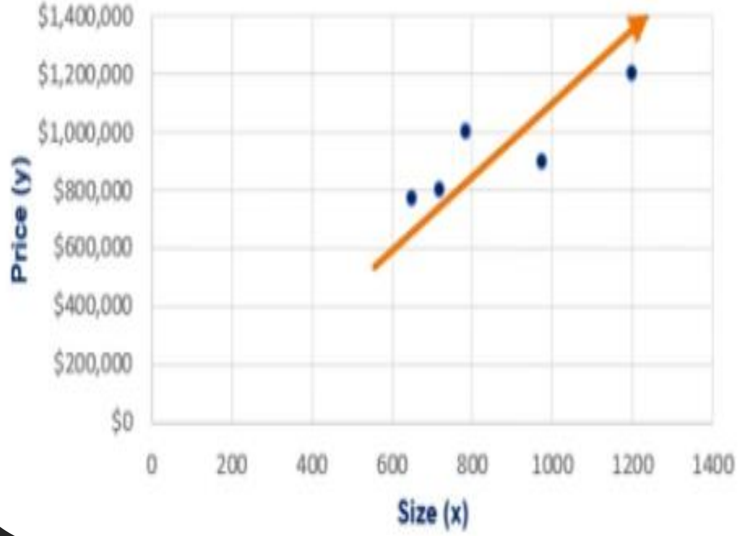
Sample correlation formula: $r = \frac{s_{xy}}{s_x s_y}$

Population correlation formula: $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Covariance

Housing data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000



The two variables are correlated and the main statistic to measure this correlation is called covariance

Sample formula

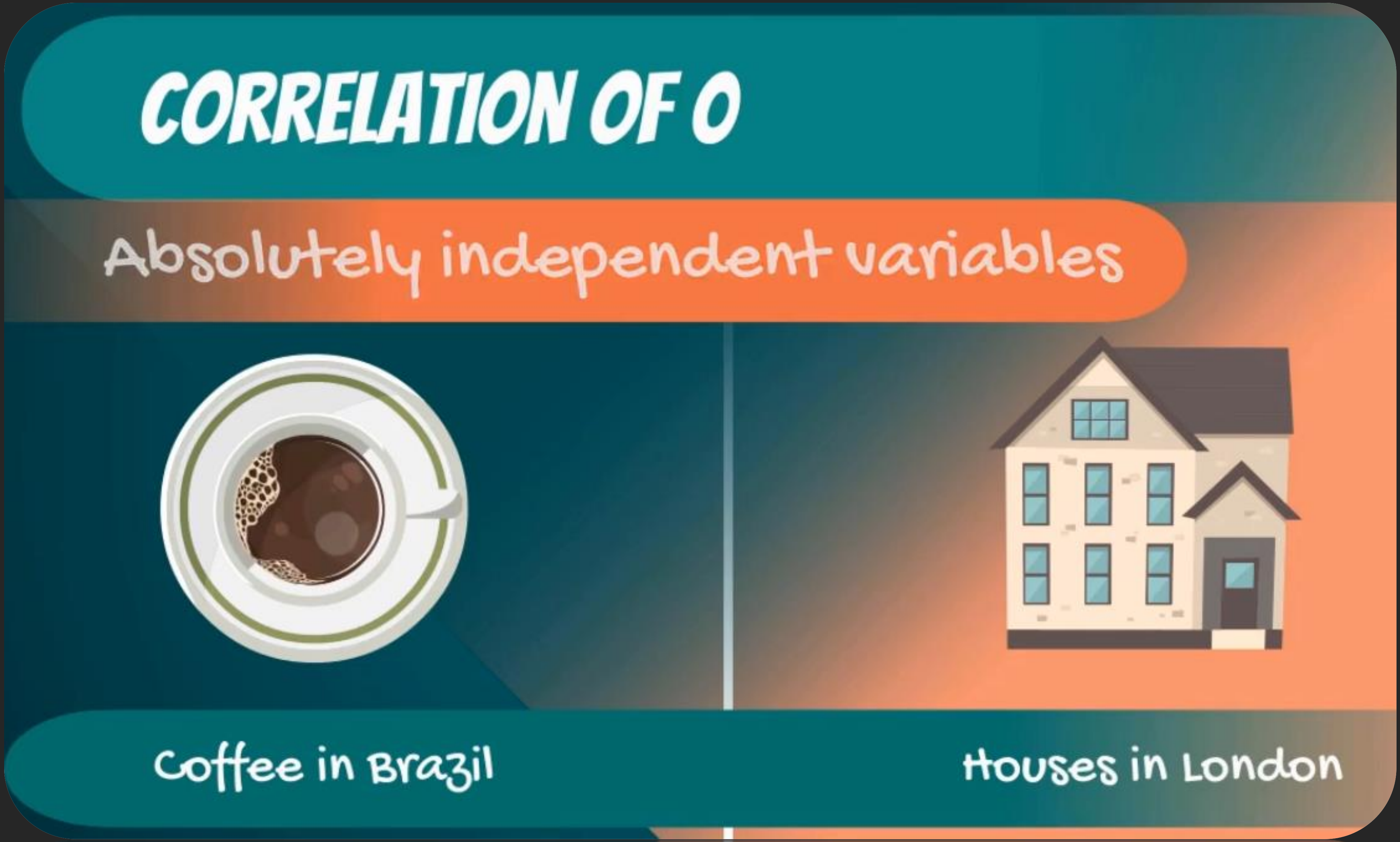
$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$

Covariance gives a sense of direction

> 0, the two variables move together
< 0, the two variables move in opposite directions
$= 0$, the two variables are independent


CORRELATION OF 0
Absolutely independent variables
Coffee in Brazil
Houses in London


NEGATIVE CORRELATION

Thanks for Watching

In the next video I’m going to show you some applications of these notions