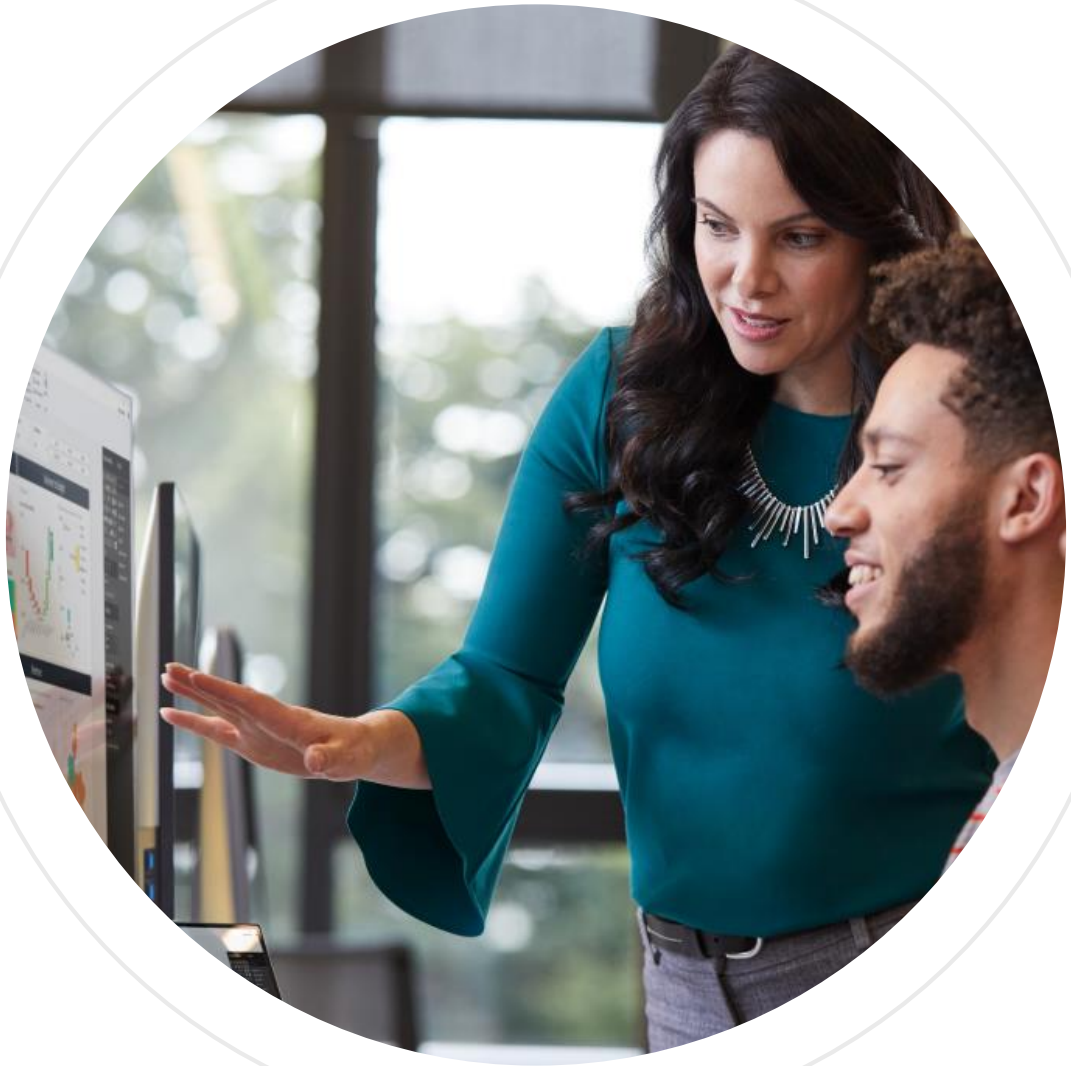
Statistics

Descriptive



Basic Statistics

Why Statistics?

1. Understand what is being Presented

2. Knowledge of Statistics helps you to conduct your own study
   - How should you analyze the information you get
   - Which method should you employ

Why Statistics?

We'll make a distinction between two types of statistics, descriptive and inferential statistics

Descriptive

Descriptive statistics we mean methods of **summarizing the information** we have collected for an analysis.

We can summarize information by means of **graphs**. Such as a pie chart or a bar graph or **numbers** such as a mean, percentage, or correlation coefficient

Inferential

Inferential statistics is about **drawing conclusions about a population** on the basis of only a limited number of cases.

An example is saying something about all citizens of France on the basis of a sample of relatively few French citizens.

Data and Visualization

First we introduce to the basics of **descriptive statistics**. We'll tell you why it makes sense to think about your data in terms of **cases** and **variables**, and we'll show you that the best way to order your cases and variables is by means of a **data matrix**. There are many different kinds of variables out there. To avoid confusion when we analyze them, we distinguish different **levels of measurement**.

When we present our data to others, we often summarize them by means of tables and/or graphs such as **frequency tables**, **pie charts**, **bar graphs**, **dot plots** and **histograms**. We'll also discuss various types of **distributions** of data.

Data and Visualization

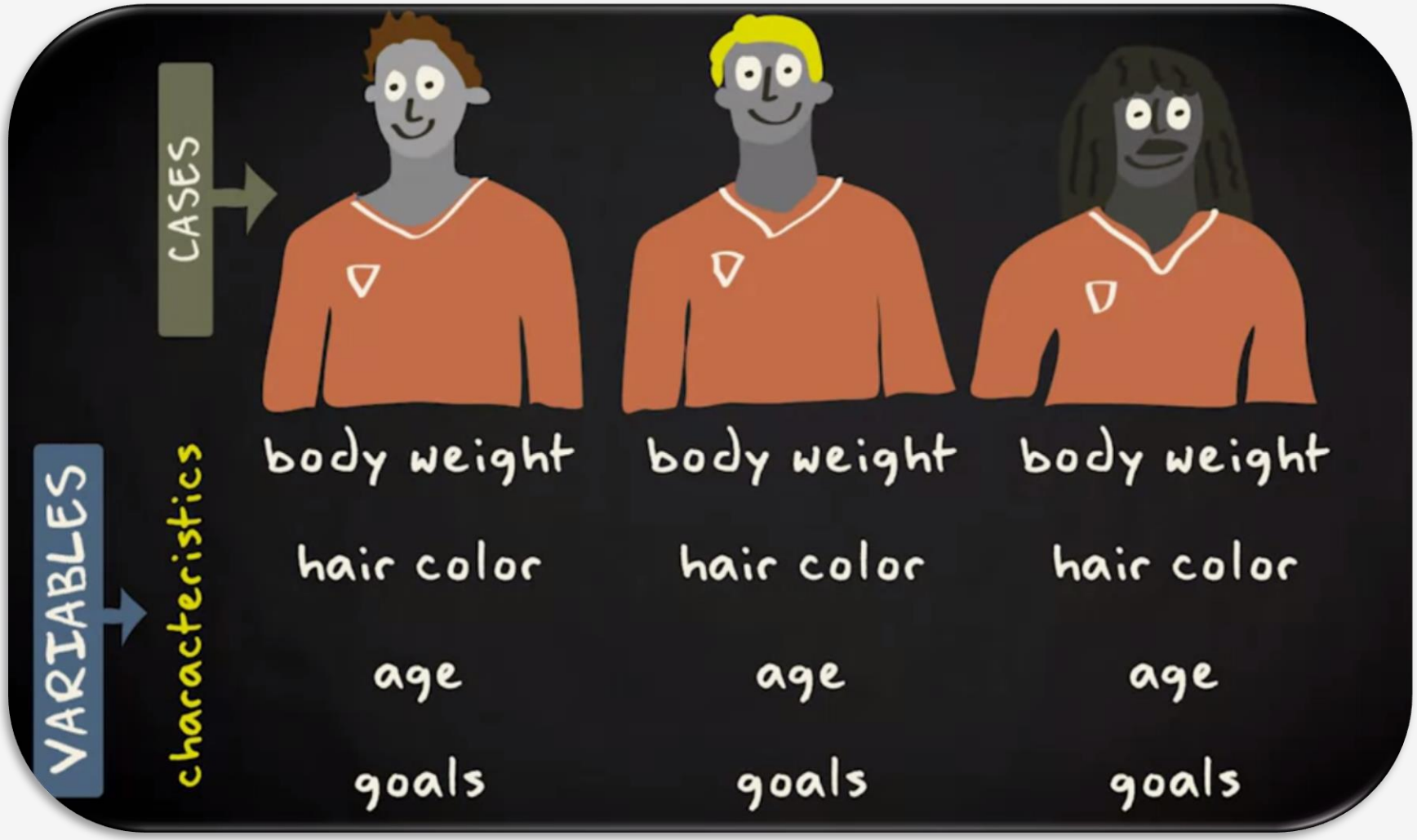
Imagine you're very, very interested in football.

You are the person who wants to know all the **details**, like how many goals were scored by some player? How many games were won by a particular team? Or how many penalties were stopped in a certain football competition.

The number of scored goal, won games, and stopped penalties are all pieces of information that can be thought of in terms of variables and cases.

Variables are features of something or someone. **And cases are that something or someone.**

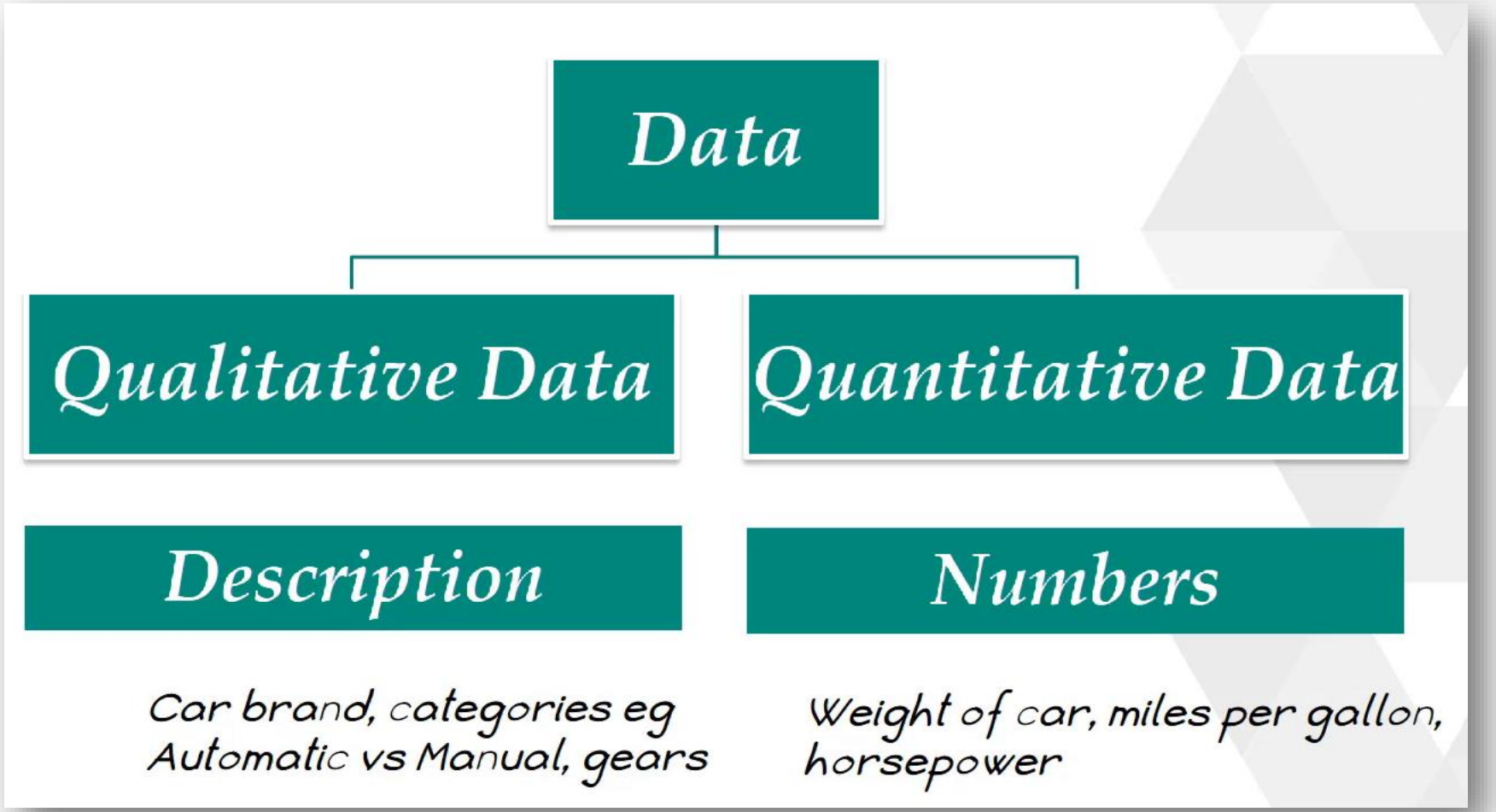
Data and Visualization



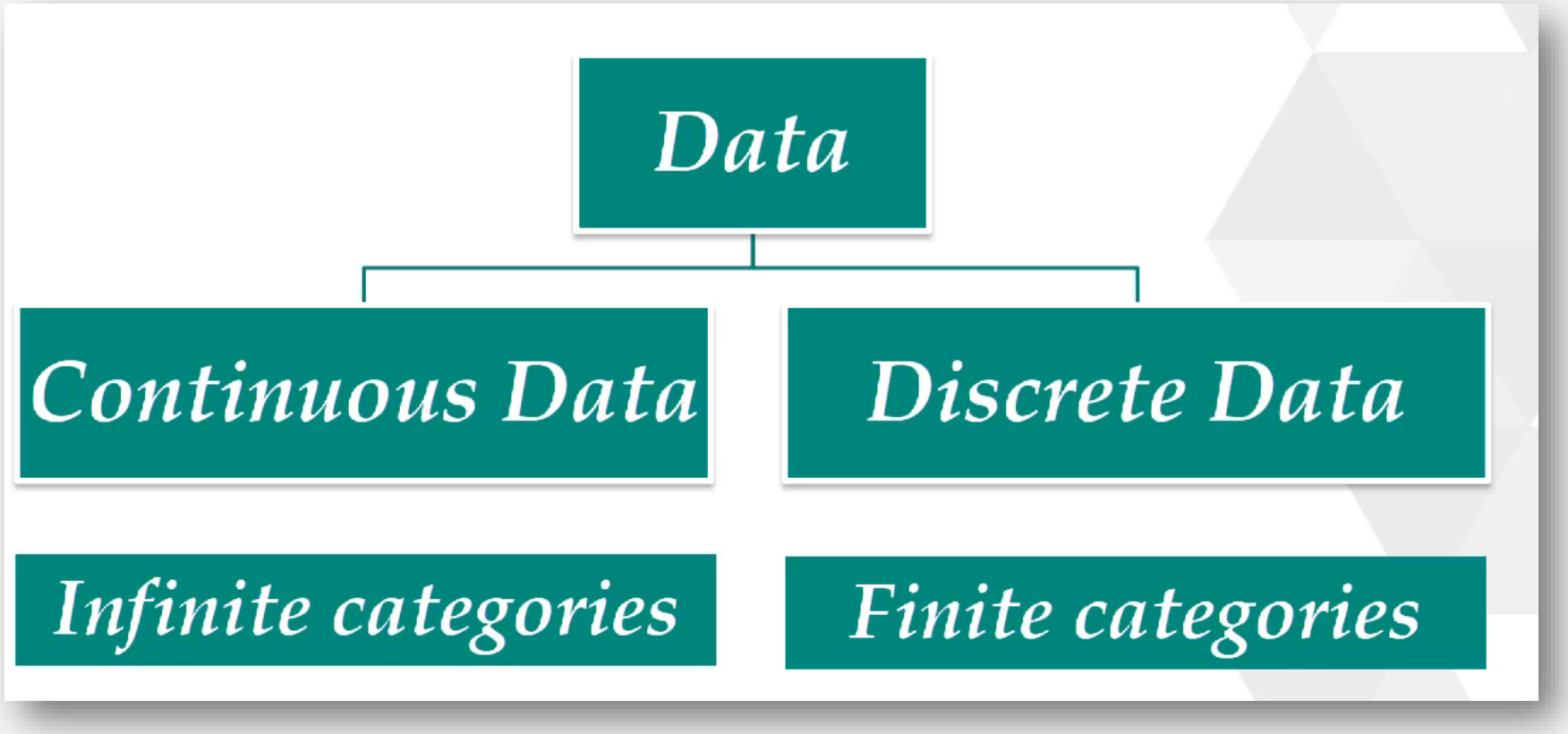
Data and Visualisation

Data

Data


Data
Qualitative Data
Quantitative Data
Description
Numbers
Car brand, categories eg Automatic vs Manual, gears
Weight of car, miles per gallon, horsepower

Data



[Apply Normal Distribution] [Apply Binomial or Poisson Distribution]

Levels of Measurement

Levels of Measurement

You can probably imagine that we can have many, **many different kinds of variables**, representing strongly **divergent characteristics**. For this reason, and also for other reasons that I will discuss later. It is of essential importance to distinguish different levels of measurement.

Nominal

A nominal variable is made up of **various categories** that differ from each other. There is **no order**, however. This means that it's not possible to argue that one category is better or worse, or more, or less than another.

An example is the nationality of the football players. The various categories, for instance Spanish, French, or Mexican differ from each other, but there is no ranking order.

Ordinal

There is not only the difference between the **categories** of the variable, there is also **an order**.

An example is the order in a football competition. You know who is the winner. You know who came second, and third, etc.

[Both nominal and ordinal levels can be called categorical variables]

Levels of Measurement

The next level of measurement is the interval level and ratio level.

Interval

With interval variables, we have different **categories** and **an order**, but also **similar intervals** between the categories.

An example is the age of a football player. We can say that a player of 18 years old differs from a player of 16 years old, in terms of his or her age.

In addition, we can say that this player is older. But we can also say that in terms of age, the difference between a 18 year old player and a 16 year old player, is similar to the difference between a 14 year old player and a 12 year old player.

Ratio

It is similar to the interval level but has, in addition, a **meaningful zero point**.

An example is a player's body height, measured in centimeters. There are differences between the categories. There is an order, there are similar intervals, and we have a meaningful zero point.

[Interval and ratio variables are what we call quantitative variables.
Because the categories are represented by numerical values.]

Levels of Measurement

Quantitative variables can also be distinguished in **discrete** and **continuous variables**.

Discrete

A variable is discrete if it's possible categories form a set of separate numbers.

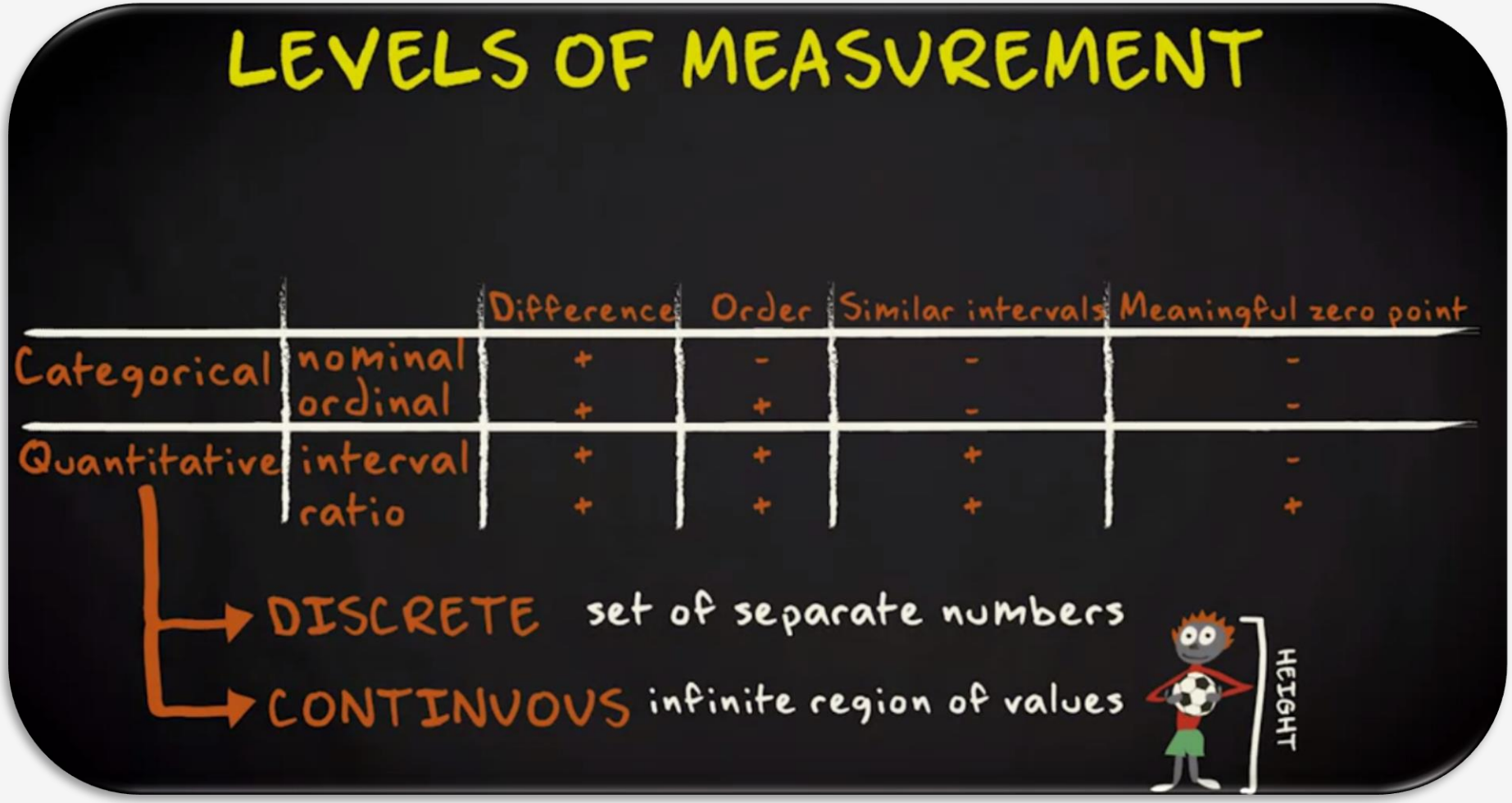
For instance, the number of goals scored by a football player. A player can score, for instance, one goal or two goals, but not 1.21 goals.

Continuous

A variable is continuous if the possible values of the variable form an interval.

An example is again, the height of a player. Someone can be 170 centimeters, 171 centimeters tall. But also for instance, 170.2461 centimeters tall. We don't have a set of separate numbers, but an infinite region of values

Levels of Measurement



Why is it so important to distinguish these various levels of measurement?

Well, because the methods we employ to analyze data depend on the level on which the variables are measured. Depending upon the measurement scales, you can use the most appropriate measure of central tendency.

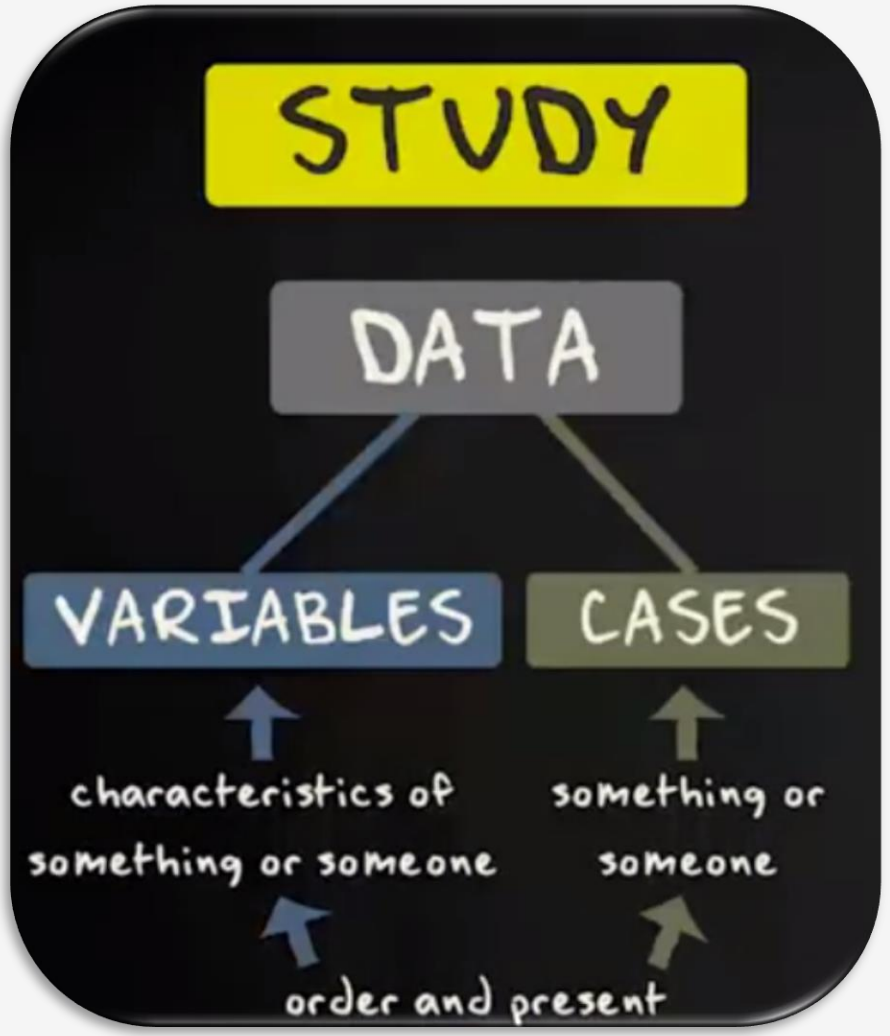
- For Nominal data, use Mode as the measurement of central tendency.
- For Interval data, you can use Mode or Median as the measurement of central tendency.
- For Interval and Ratio scale, you can use any of three measurements of central tendency (Mean, Mode or Median)

Data matrix and frequency table

Data matrix and frequency table

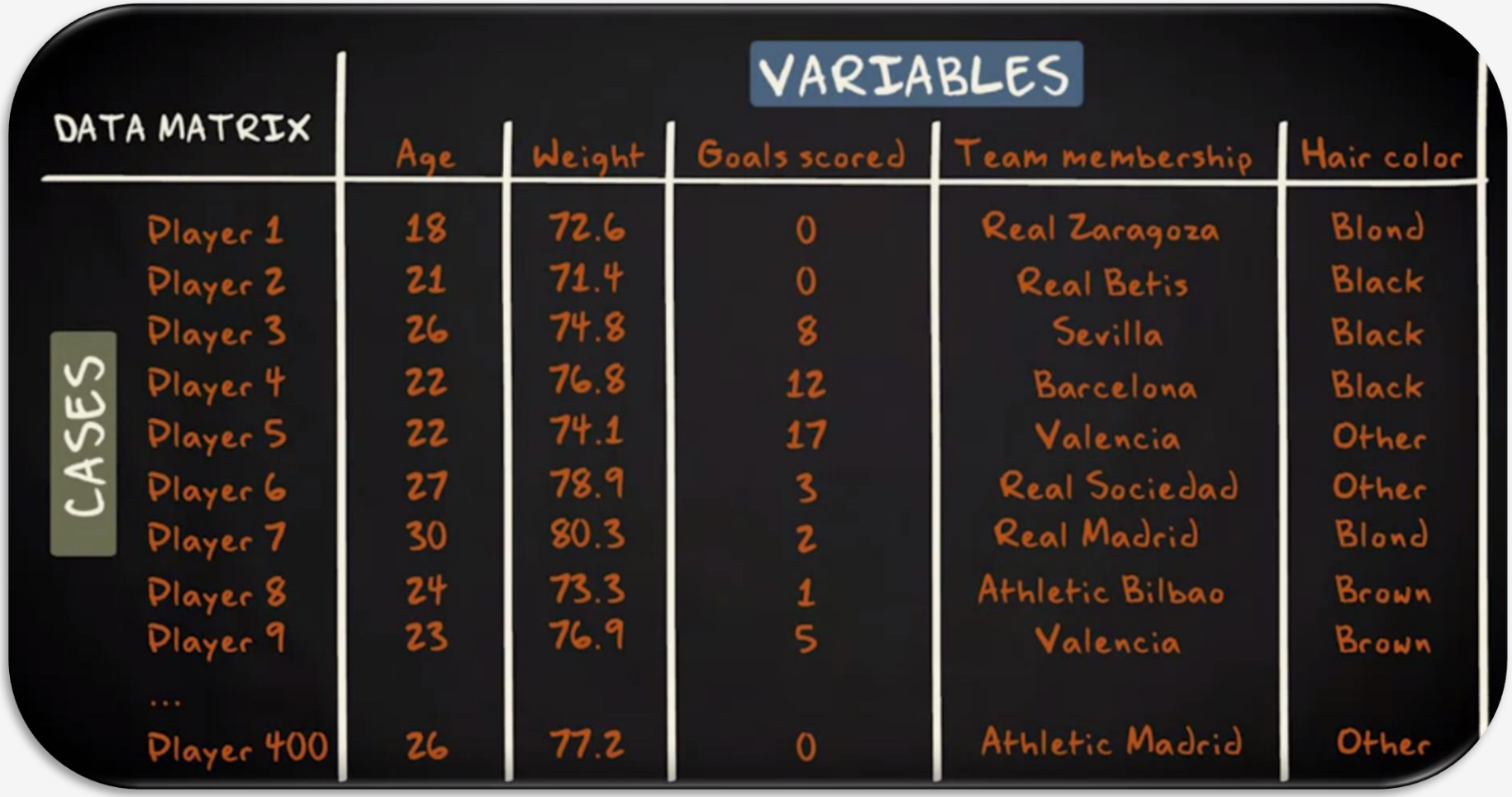
If you're conducting a study/Business Case. It makes sense to think about your data in terms of cases and variables. Cases are the persons, animals, or things you're studying (Use Case) , and variables are the characteristics of interest.

On this part, we will discuss how you can order and present your cases and variables.



Data matrix and frequency table

Data matrix and frequency table



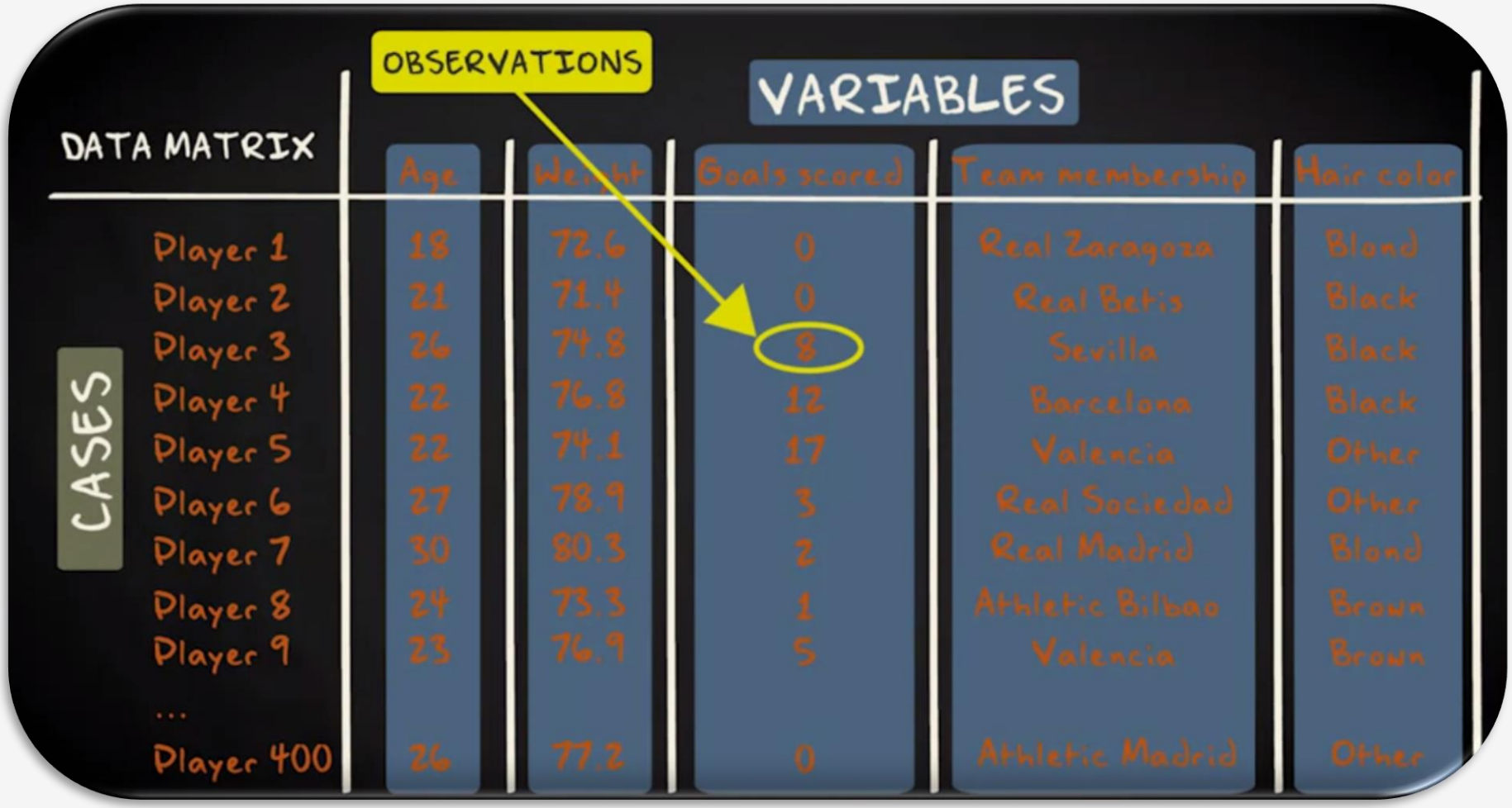
DATA MATRIX

VARIABLES

CASES

	Age	Weight	Goals scored	Team membership	Hair color
Player 1	18	72.6	0	Real Zaragoza	Blond
Player 2	21	71.4	0	Real Betis	Black
Player 3	26	74.8	8	Sevilla	Black
Player 4	22	76.8	12	Barcelona	Black
Player 5	22	74.1	17	Valencia	Other
Player 6	27	78.9	3	Real Sociedad	Other
Player 7	30	80.3	2	Real Madrid	Blond
Player 8	24	73.3	1	Athletic Bilbao	Brown
Player 9	23	76.9	5	Valencia	Brown
...					
Player 400	26	77.2	0	Athletic Madrid	Other

Data matrix and frequency table



Data matrix and frequency table

You need the **data matrix** for all your **statistical analyzes**.

However, you usually do not present your complete data matrix to other people. The reason is that a data matrix is often huge. In our case we have 400 rows. And doesn't give a clear overview of the statistical information contained within the data matrix.

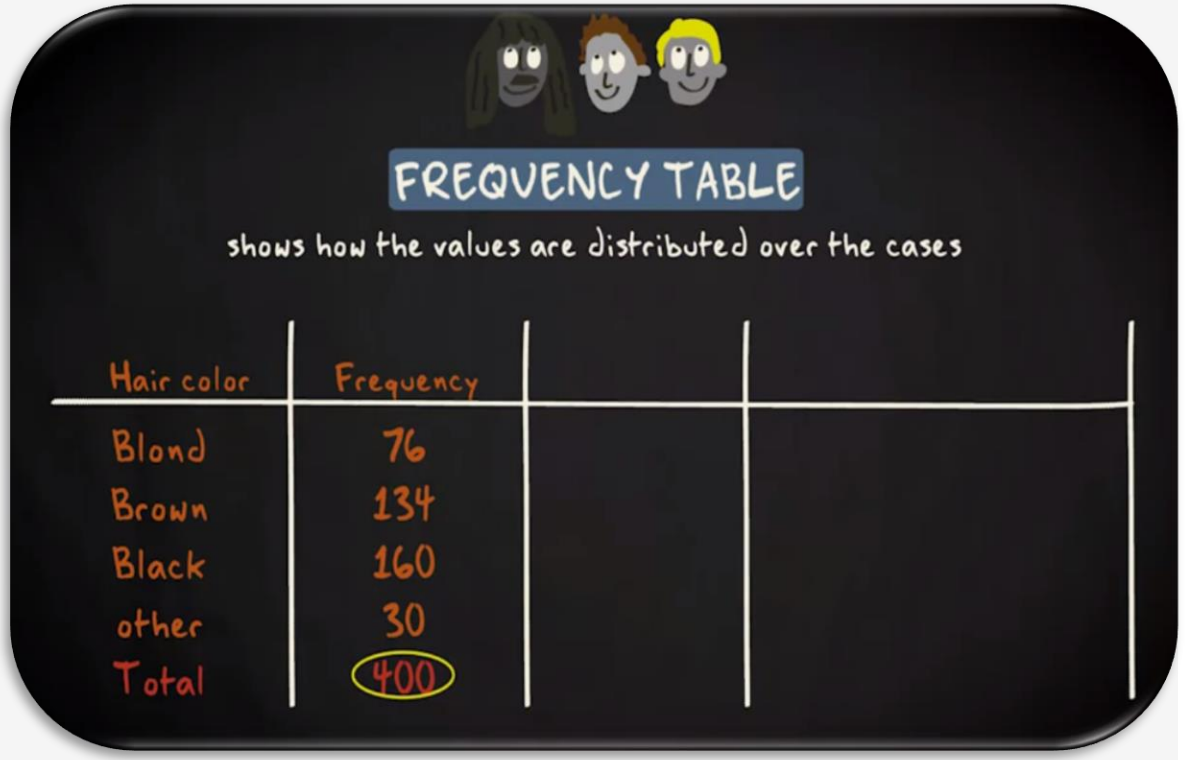
When we present the information in our data matrix to others. **We therefore often make use of summaries of data in the forms of tables and graphs.**

Data matrix and frequency table

Imagine you want to **summarize the information** you've got about the **hair color of the players** in the Spanish football competition.

A good way to do that is to make a frequency table.

A frequency table shows you have the values of a variable are distributed over the cases. The frequency table is nothing more than a list of all possible values of a variable. Together with the number of observations for each value.

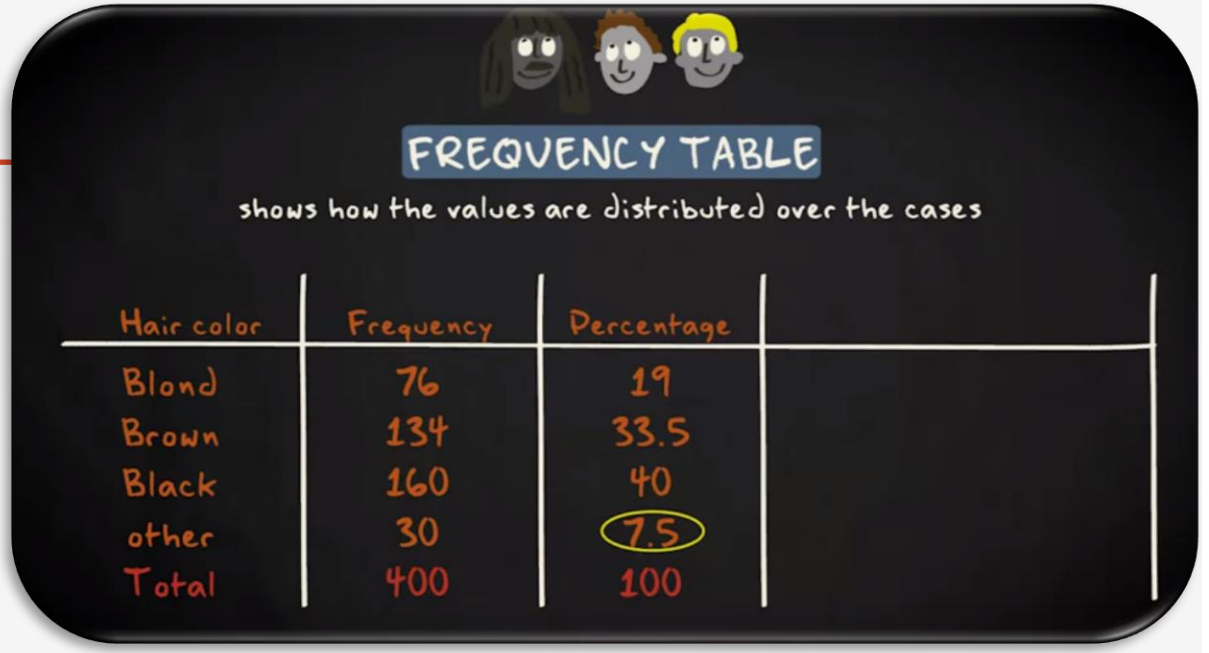


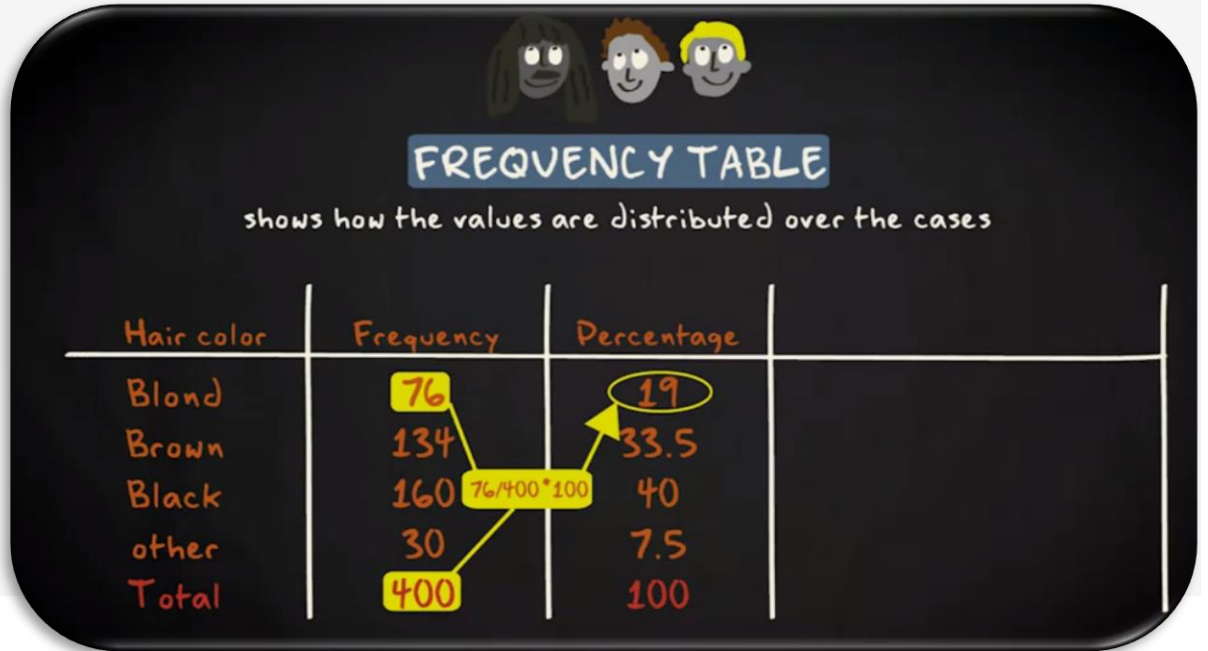
Data matrix and frequency table

We can also express the **relative frequencies** by means of percentages.

In the second column you see the percentages. You can see at a glance that 7.5% of all players has another hair color than blond, brown, or black. 19% of the players has blond hair.

You get the value 19 here by dividing 76 by 400 and multiplying that with 100.



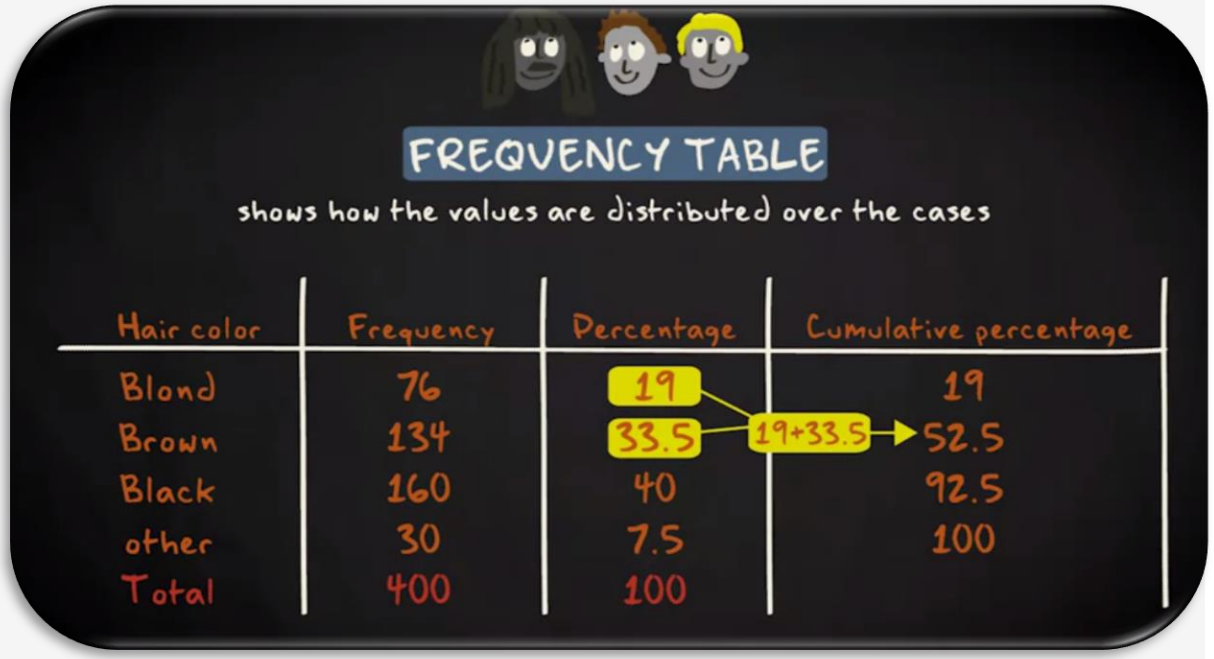


Data matrix and frequency table

Sometimes, researchers use **cumulative percentages**. It is easy to compute them.

Cumulative percentages are nothing more than the percentages in every category added up.

So you can see here that 19 plus 33.5% equals 52.5% of all players have blond, or brown hair.

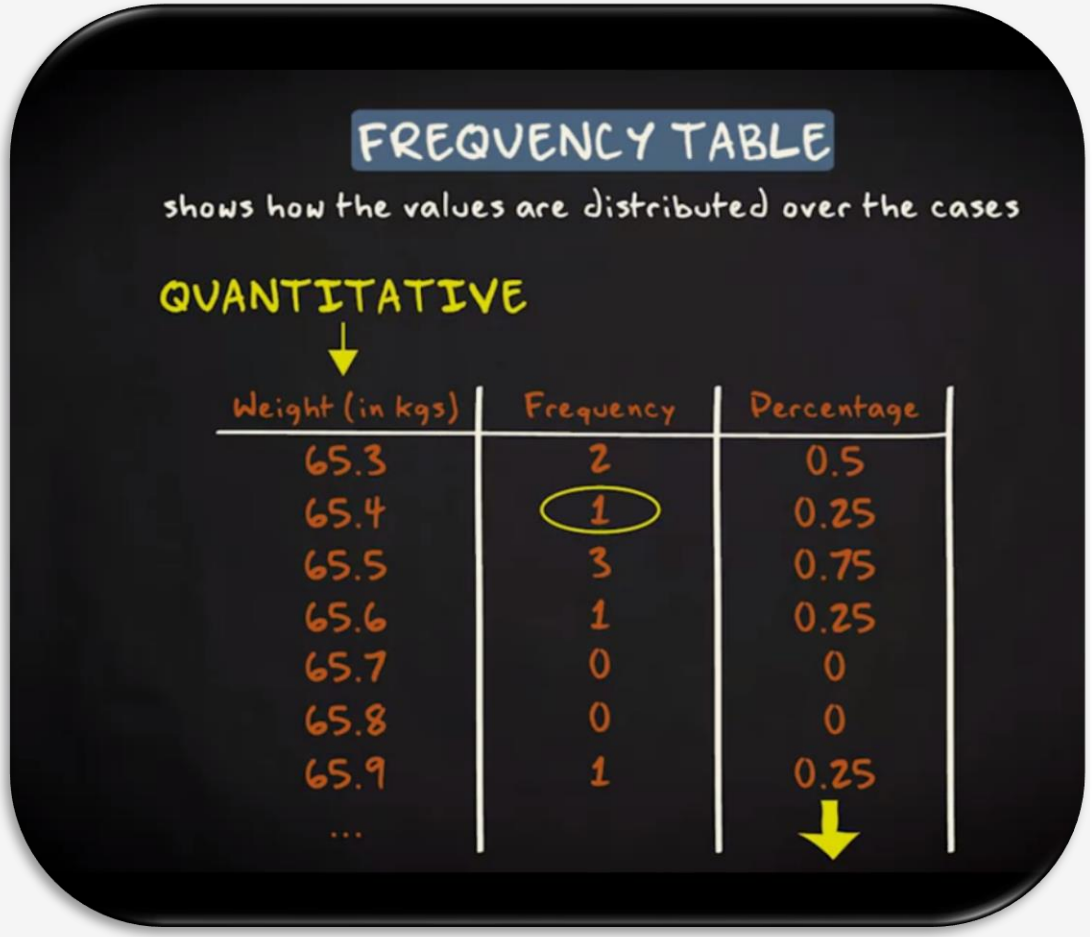


Data matrix and frequency table

In previous example, we talked about a categorical variable, hair color.

What if we are dealing with a quantitative variable?

Take weight for instance. It doesn't make sense to compute percentages for every specific value of weight. Because then we would end up with a countless number of categories.



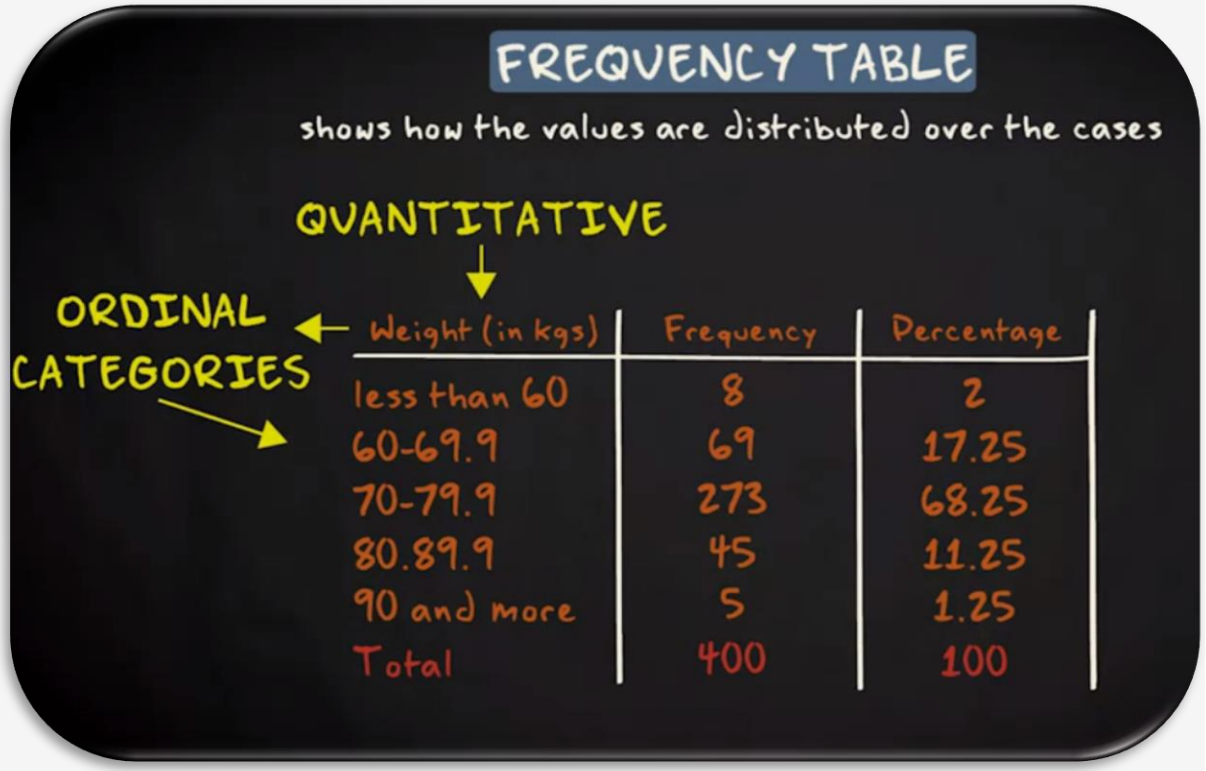
Data matrix and frequency table

What researchers usually do to solve that problem is building **new ordinal categories by using intervals**.

You could say for instance, that the first category contains those players who way less than 60 kilograms. The second, those who weigh between 60 and 69.9 kilograms. The next one, between 70 and 79.9. The following one between 80 and 89.9. And the final one, 90 and more kilograms.

This way you lose information. But the advantage is that you get a much better overview.

The variable weight was a quantitative variable that you've turned into an ordinal variable with only five categories.



Data matrix and frequency table

So, what do you know now?

❑ You use a **data matrix** as the source of all your statistical analyzes.

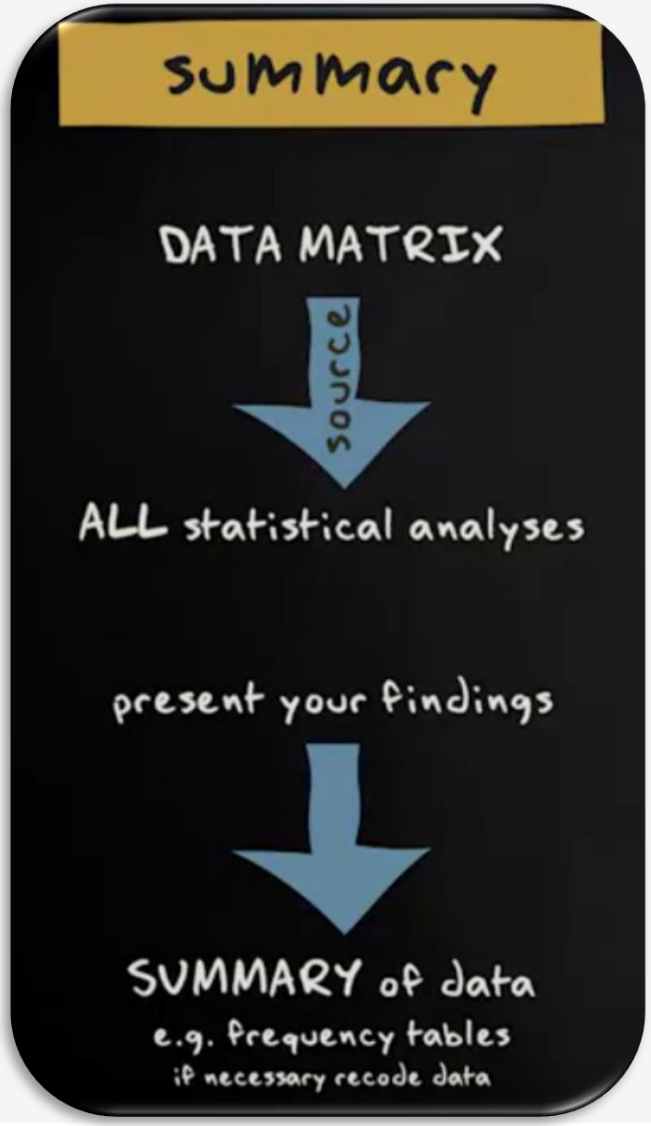
It is the overview of your data.

However, if you want to **present your findings** to other people.

❑ You make use of **summaries** of your data.

❑ One very good way to summarize is by making **frequency tables**.

❑ If necessary you can **recode** your quantitative variables into ordinal ones.

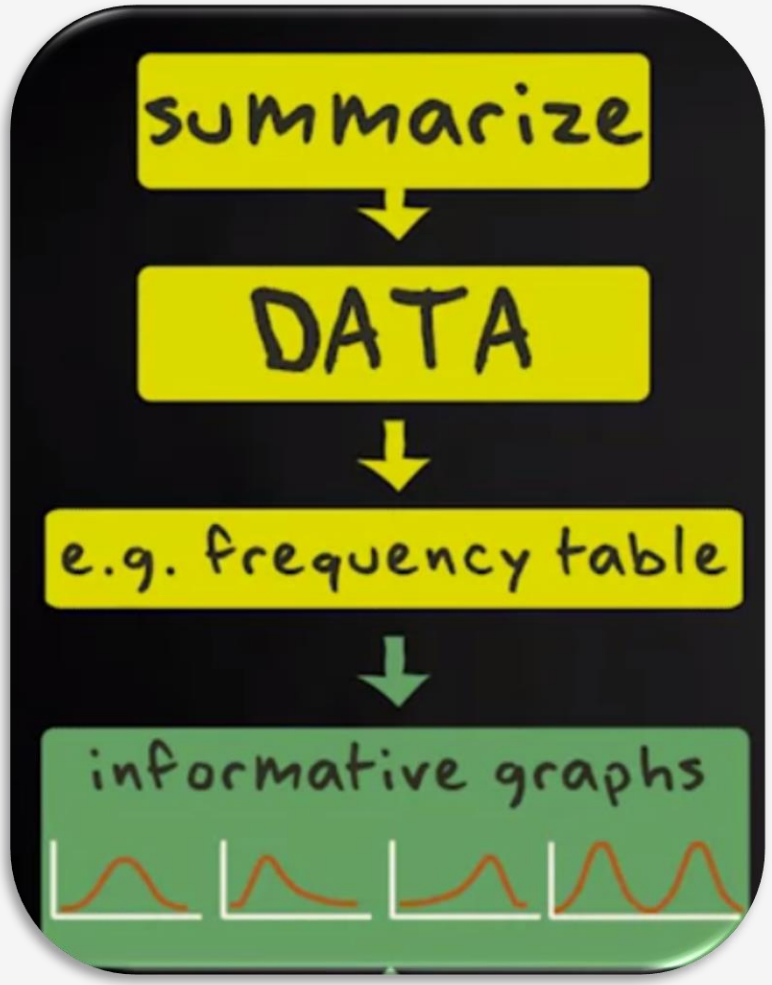


Graphs and shapes of distributions

Graphs and shapes of distributions

Researchers often want to summarize the data they have. They can do that, for instance, by means of a **frequency table**.

In this section, I will show you how you can use a frequency table to build **informative graphs**. I will also discuss the **possible shapes** that the data distributions in these graphs could take.



Graphs and shapes of distributions

Imagine a study where football players in the main football competition in Spain come from.

This frequency table could be the result.

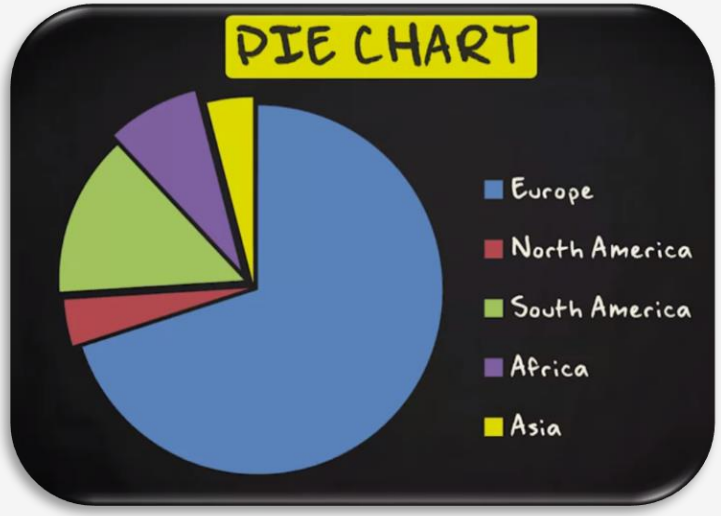
I've also added the relevant percentages. You might want to present these results by means of a **graph**.

Let me show you two possible ways in which you could do that.

Graphs and shapes of distributions

What you see here is a **pie chart**.

The categories of the variable you would like to summarize are displayed by means of slices of a pie. In a pie chart, the surface of the slices represent percentages of observations in each category. You can see at a glance that almost three-quarters of all the football players come from Europe.

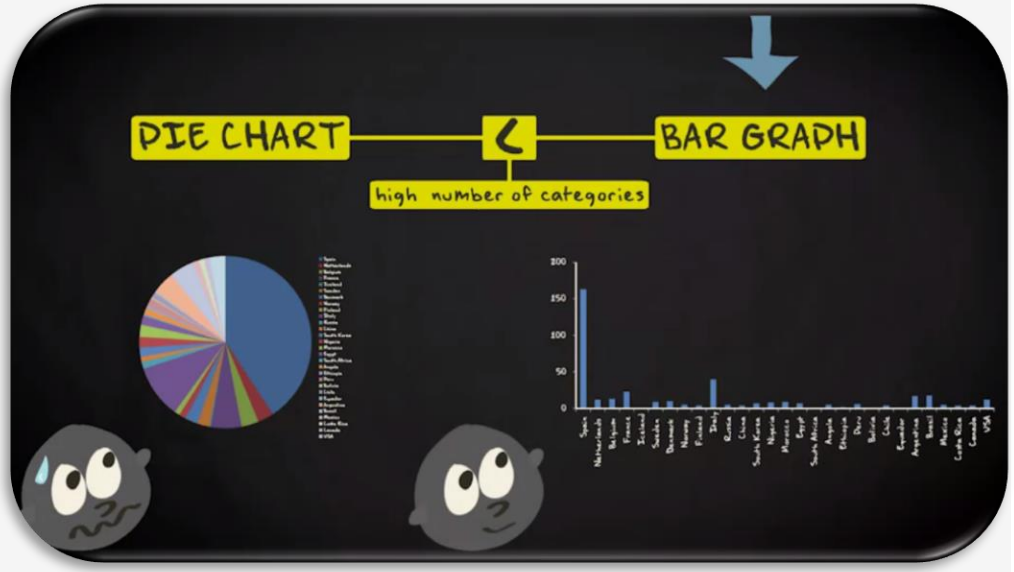


Another way to summarize the same data is with a **bar graph**, which also shows you very clearly how the data are distributed over the various categories of the variable.

Graphs and shapes of distributions

A bar graph has advantages over a pie chart if the number of categories of a variable increases.

Imagine, for instance, that you don't want to know which continent the football players come from, but that you want to know in which particular country they were born.

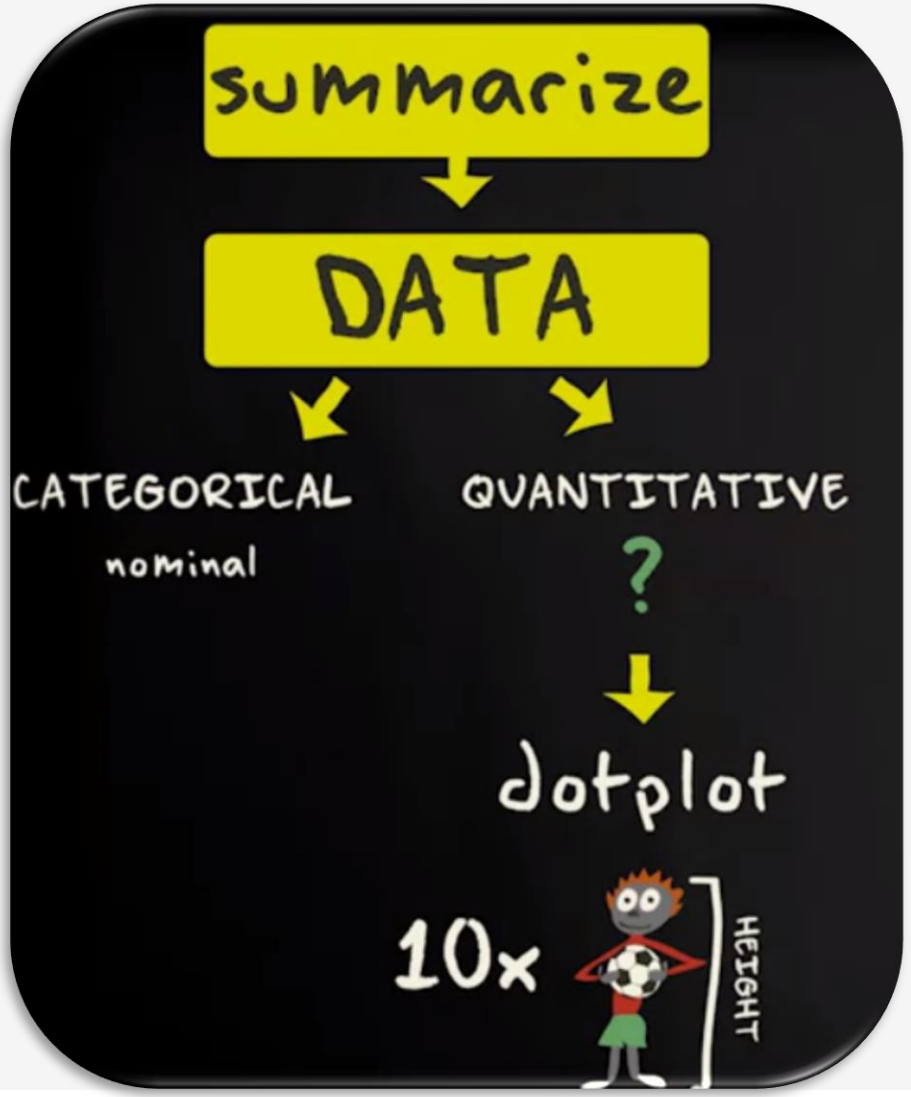


Graphs and shapes of distributions

Up till now, we have talked about a categorical or more precisely, a nominal variable. How can we summarize data if we are dealing with a **quantitative variable?**

One possibility is with a **dot plot**.

The idea is easy, imagine you have information about the physical height of ten football players expressed in centimeters.

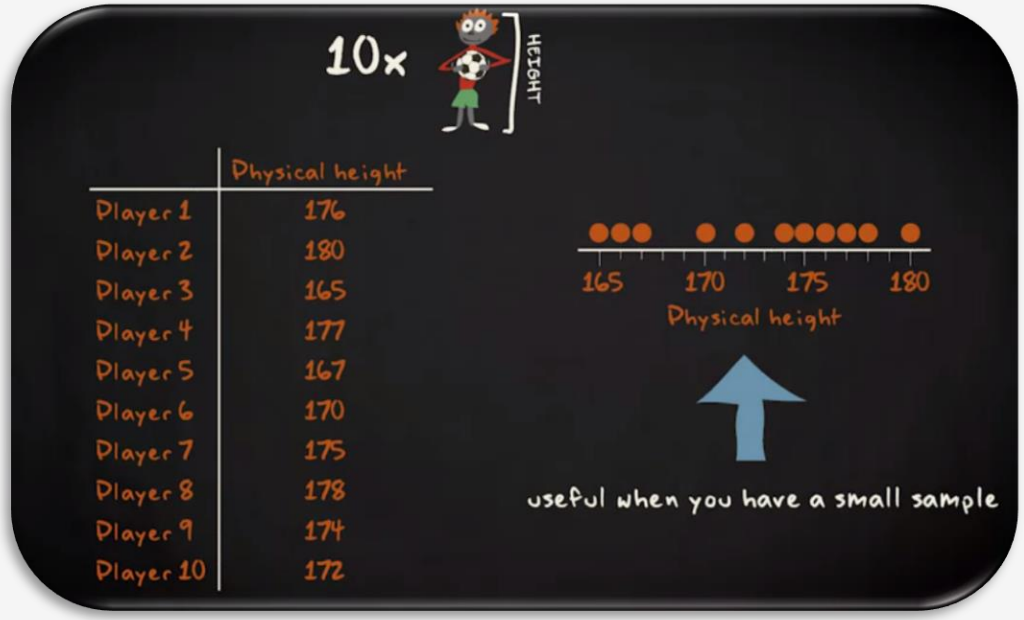


Graphs and shapes of distributions

First, you **draw a horizontal line** and **label the possible values** on it in regular intervals, like this. Next, for each observation, you place a dot above its value on the horizontal line.

Dot plot is useful when you have only a couple of observations. However, it becomes messy when you have a large sample.

If we have many observations, researchers, therefore, usually make use of another type of graph. **The histogram**.

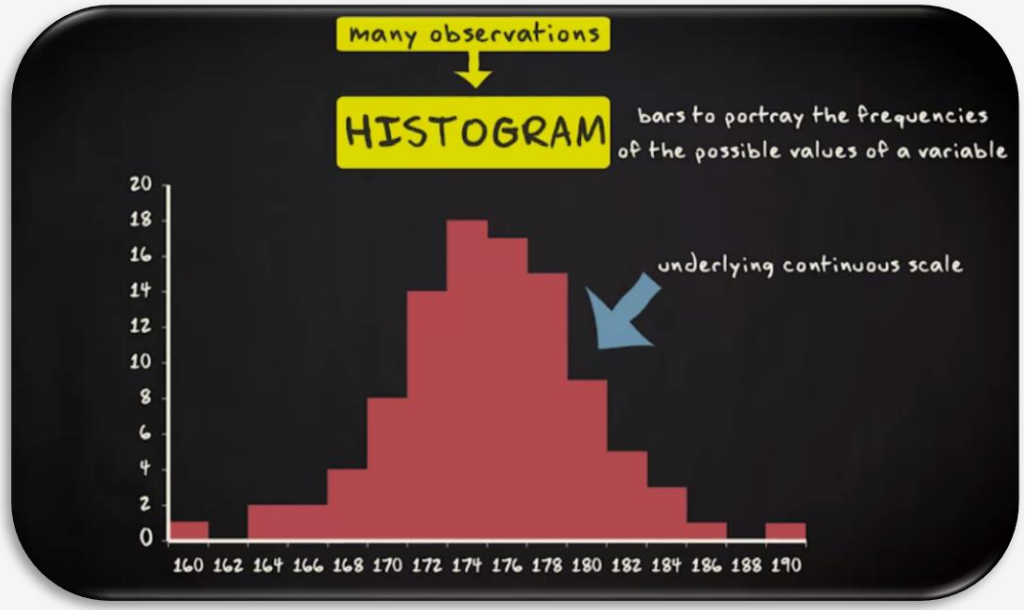


Graphs and shapes of distributions

A histogram is similar to a bar graph in the sense that it uses bars to portray the frequencies or relative frequencies of the possible values of a variable.

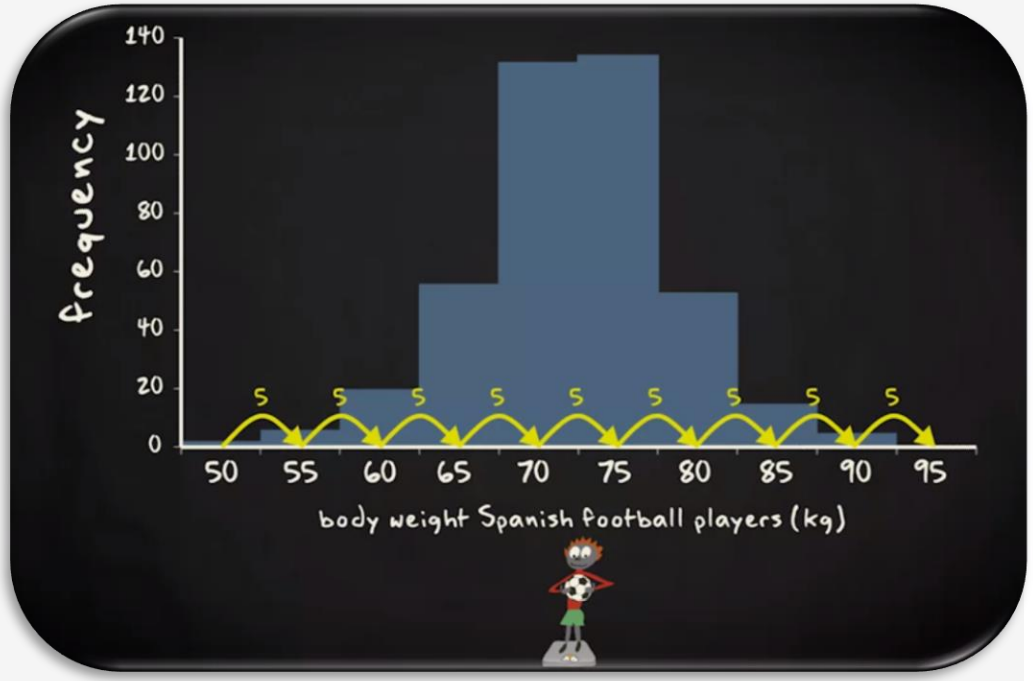
However, there is one **important difference**. The difference is that the **bars in a histogram touch each other**.

This touching represents that the values of an interval ratio variable represent an underlying **continuous** scale.



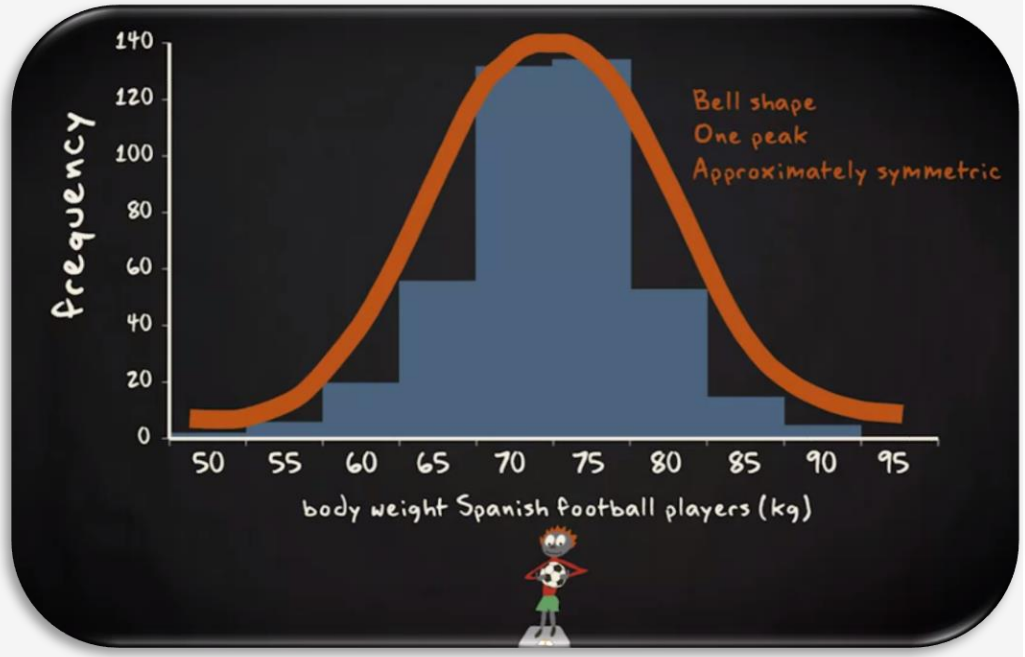
Graphs and shapes of distributions

Say, we are interested in the body weight of Spanish football players. If we have very detailed measure of weight like 83.9 or 74.5 kilograms, it doesn't make sense to draw a separate bar for every single value. Instead, **we construct intervals**. In this graph, we have ten intervals of five kilograms. The first interval ranges from 47.5 kilograms to 52.5 kilograms. 50 is displayed, because it is the middle of that interval. There are no fixed rules for how many intervals to make.



Graphs and shapes of distributions

As you can see, this histogram has a particular shape. It has the shape of a **bell**, has **one peak** and is **approximately symmetric**. You will encounter such distributions very often, but not all histograms have this shape.



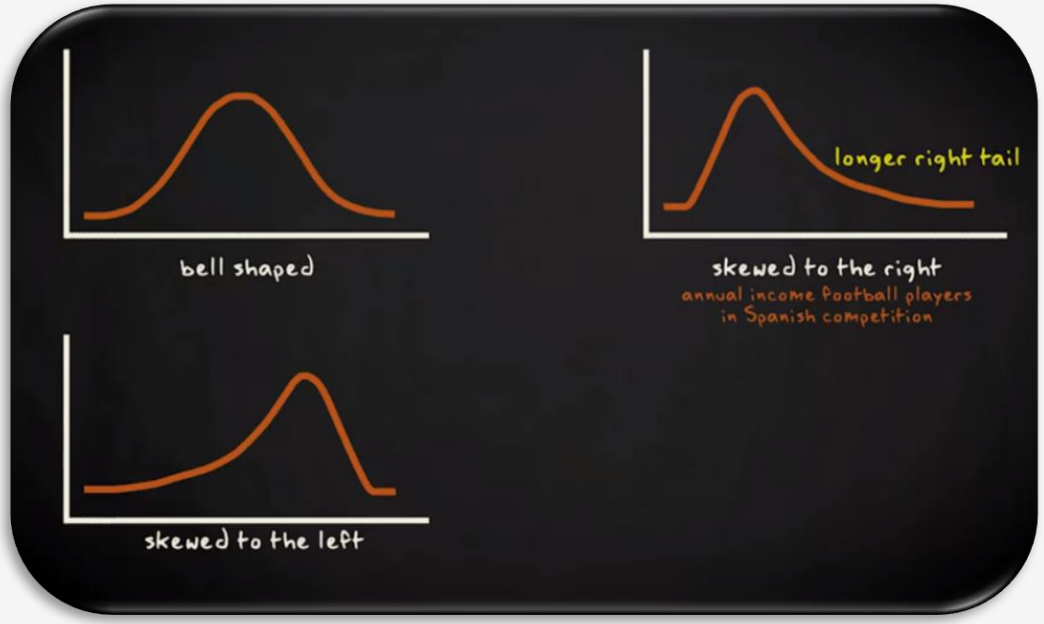
A histogram can also be skewed to the left or to the right.

The skewed histogram is not symmetric, because the one side of the distribution stretches out further than the other.

Graphs and shapes of distributions

A variable that might have a right skewed histogram is the annual income of the football players in the Spanish competition. There won't be many players with a very low income compared to the average income of the players.

However, there will be some players who earn much more money than the majority of the players. For that reason, there is a longer right tail.

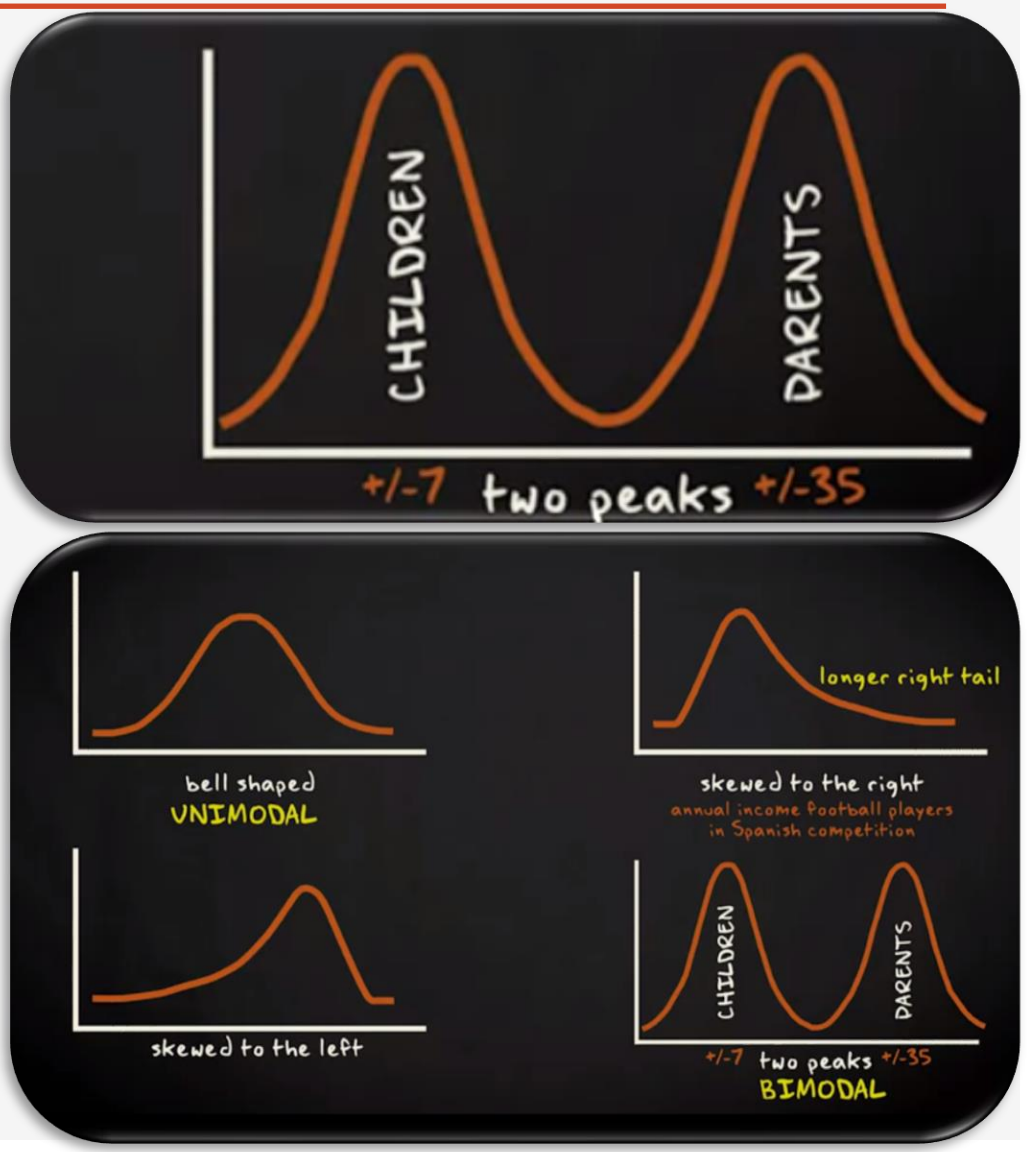


Graphs and shapes of distributions

A histogram could also have **two peaks instead of one**.

Imagine a football match between two teams of six to eight-year old players. After the match, all children and the parents go for a drink in the canteen. You are interested in the question, how old the people in the canteen are?

Well, the histogram of the variable age might, in this case, well have two peaks. After all, those present in the canteen are children between 6 and 8 years old and their parents, which are most likely somewhere between 30 and 40 years old.



Graphs and shapes of distributions

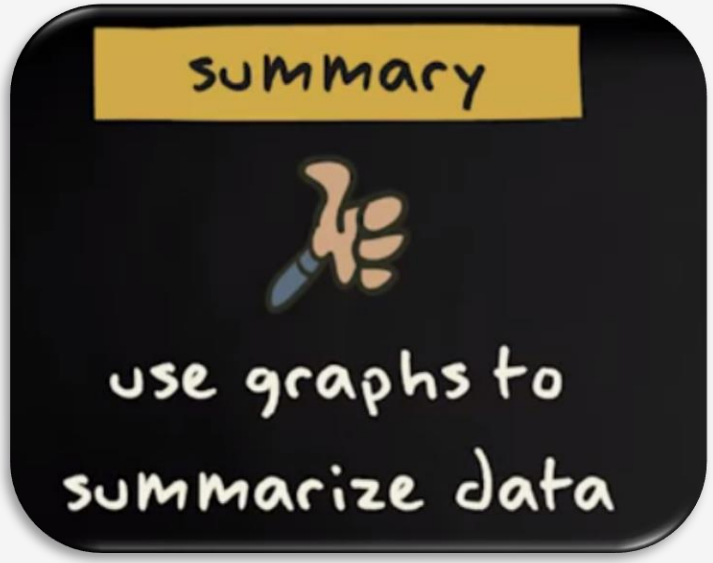
The most important lesson to take home from this section is that it's always a good idea to summarize your data by means of graphs.

- ❑ If we're dealing with nominal or ordinal variables, you should make a pie chart or a bar graph.
- ❑ If you're variable of interest is an interval ratio variable, you should make a histogram.

And never forget to look at the shape of your variable.

Is it a bell shape and symmetric? Is it unimodal? Or bimodal? Is the distribution skewed?

Assessing the shape of a distribution is of essential importance, because it could effect the statistical methods you are going to employ later on.



Measures of central tendency and dispersion

Measures of central tendency and dispersion

Besides summarizing data by means of tables and/or graphs, it can also be useful to describe the **center of a distribution**. We can do that by means of so-called measures of **central tendency**: the **mode**, **median** and **mean**.

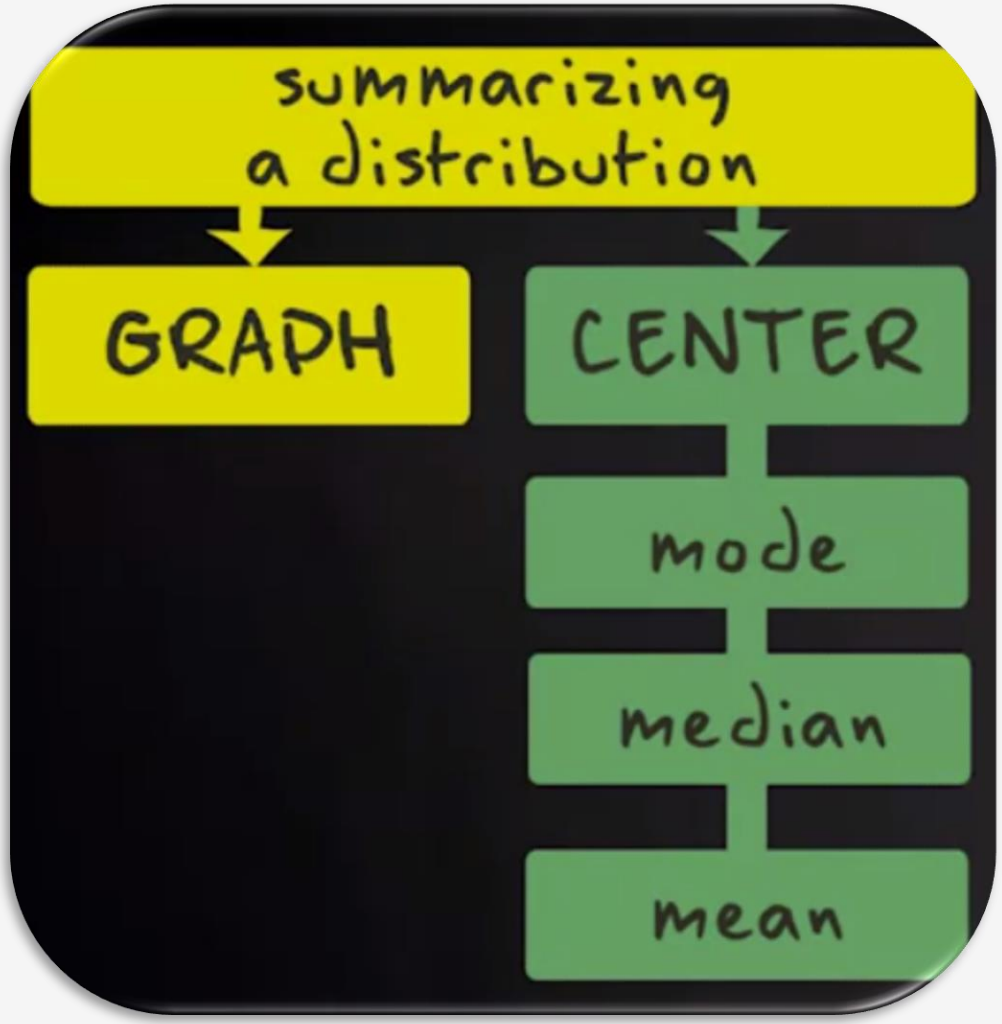
Yet to adequately describe a distribution we need more information. We also need information about the variability or dispersion of the data. We need, in other words, **measures of dispersion**. Well-known measures of dispersion are the **range**, the **interquartile range**, the **variance** and the **standard deviation**. A graph that nicely presents the variability of a distribution is the **box plot**.

In this section of the module we'll not only discuss how you should interpret these measures of central tendency and dispersion, we'll also show you how you can compute them yourself.

Mode, median and mean

Next to summarizing a distribution by means of graphs. It can also be useful to describe the center of your distribution. There are three main ways in which you can do that. By means of the mode, the median and by means of the mean.

These three m's are often referred to as measures of central tendency.



Mode, median and mean

The mode is often used as a measure of central tendency if a variable is measured on a nominal or ordinal level.

In this pie chart, you can see which continent players in the main Spanish football competition come from.

The pie chart makes immediately clear what the mode is, it is Europe. 70% of the players was born in Europe.

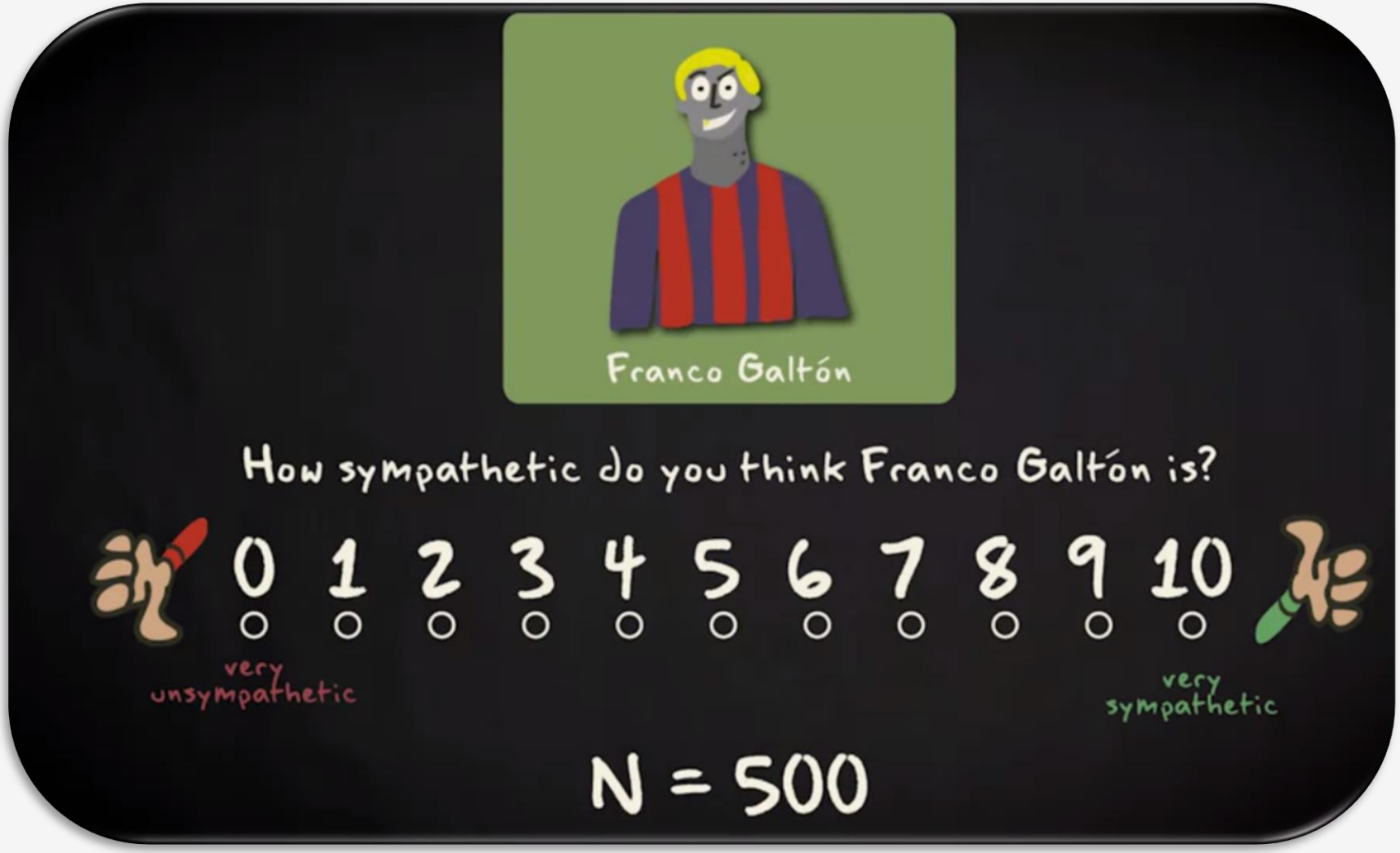
Mode, median and mean

You can also have more than one mode.

Imagine that there exists a football player that strongly divides football fans. Some people find him very sympathetic, while others find him strongly unsympathetic.

Imagine you have asked the representative sample of the Spanish population of 500 respondents, what they think of Franco Galton? Your respondents could indicate on a scale from 0 to 10, how sympathetic they think he is. 0 refers to very unsympathetic, and 10 refers to very sympathetic.

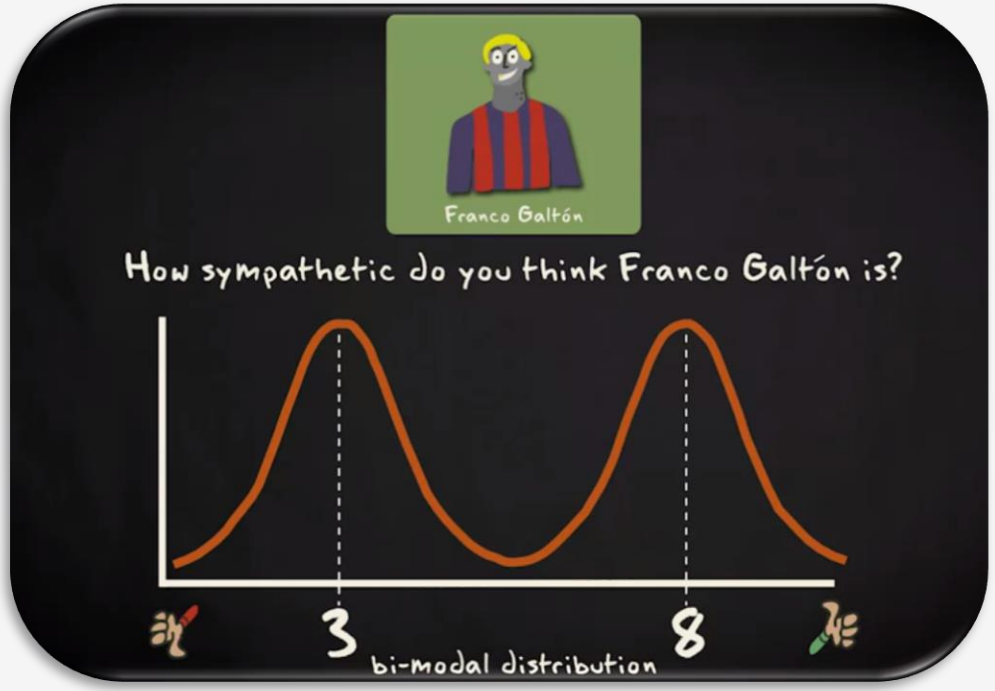
Mode, median and mean



Mode, median and mean

Let's say that this is the shape of the histogram resulting from this study. You can see that the Spanish population is strongly divided. Some find Galton very unsympathetic, and some find him very sympathetic.

As you can see the distribution has two modes, 3 and 8. This is clearly a bi-modal distribution



Mode, median and mean

The second measure of central tendency is the median. The median is nothing more than the middle value of your observations when they are ordered from the smallest to the largest.

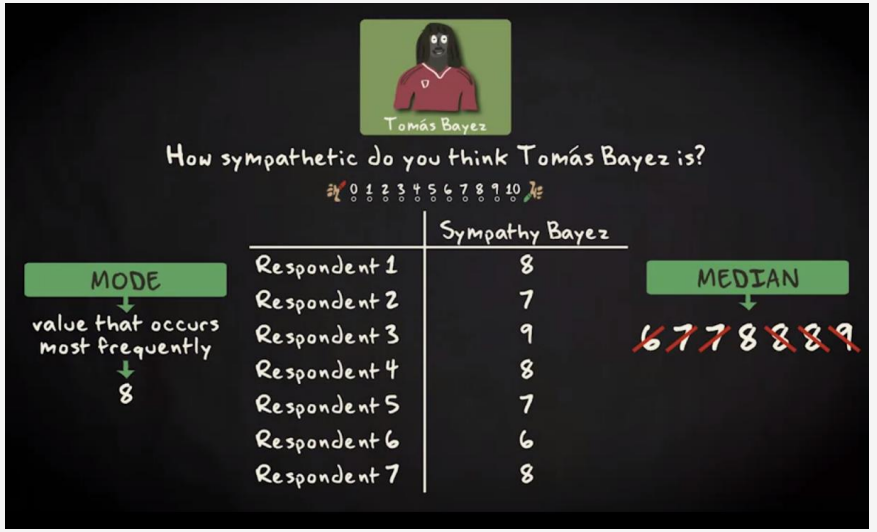
Mode, median and mean

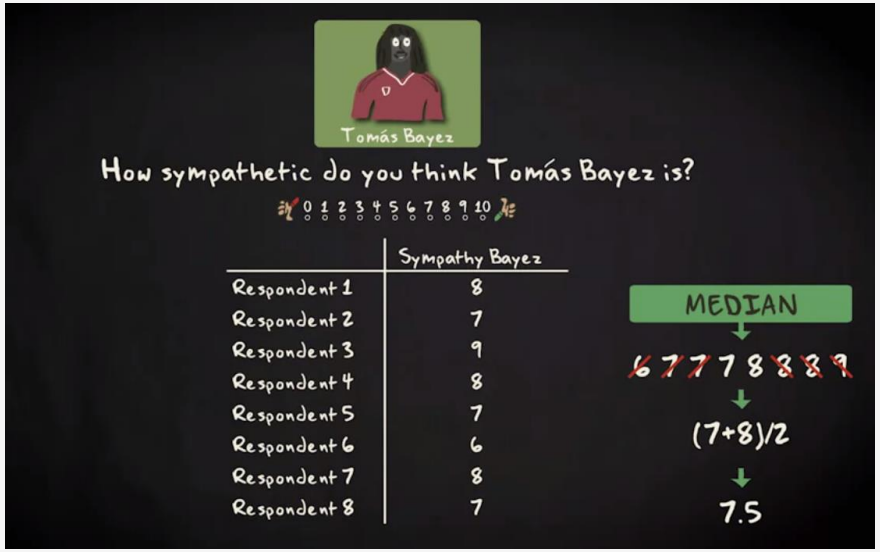
Imagine you have also asked seven of your respondents what they think of another famous football player named Tomas Bayez.The mode here is 8, the value that occurs most often.

To compute a median, we first have to order all values from low to high. Then we have to pick the middle value.

So, the median is 8. It is slightly more complicated if we have an even number of cases instead of an odd number of cases.

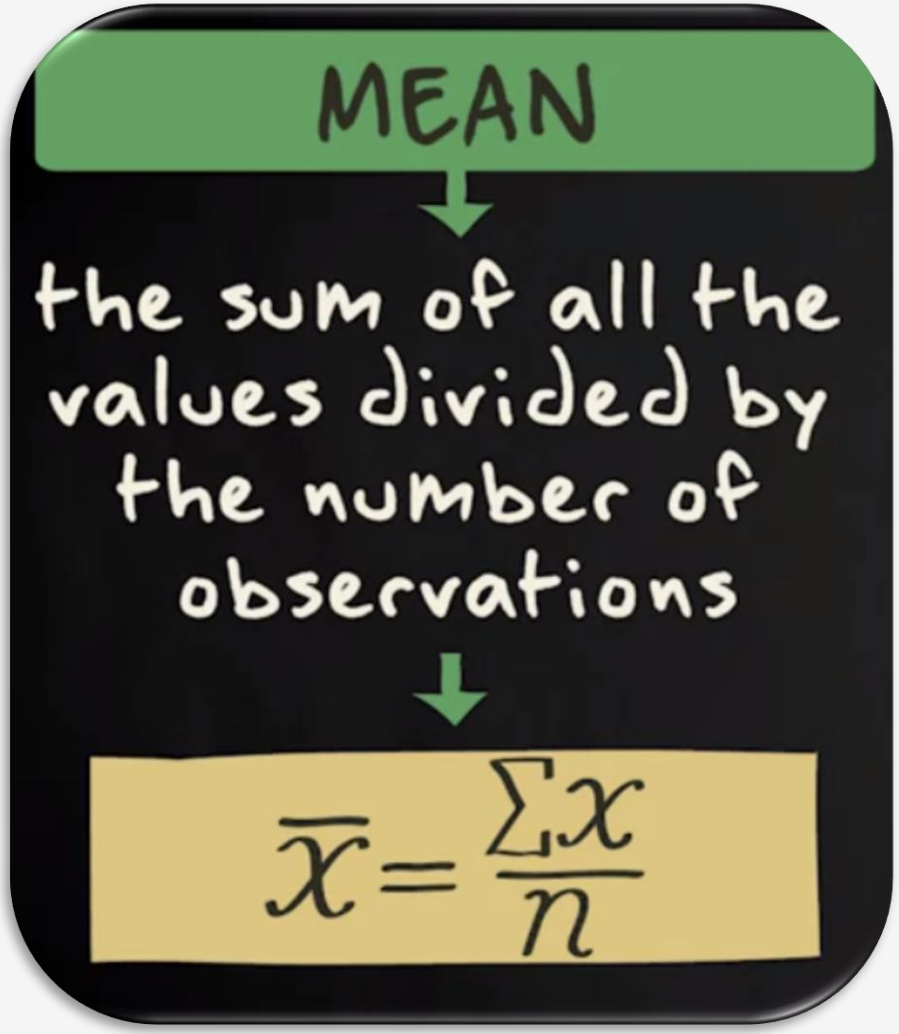
How do we solve that problem? Well, we just take the average of the two middle values. That 7 and 8, dived by 2 equals 7.5. The median in this case is 7.5





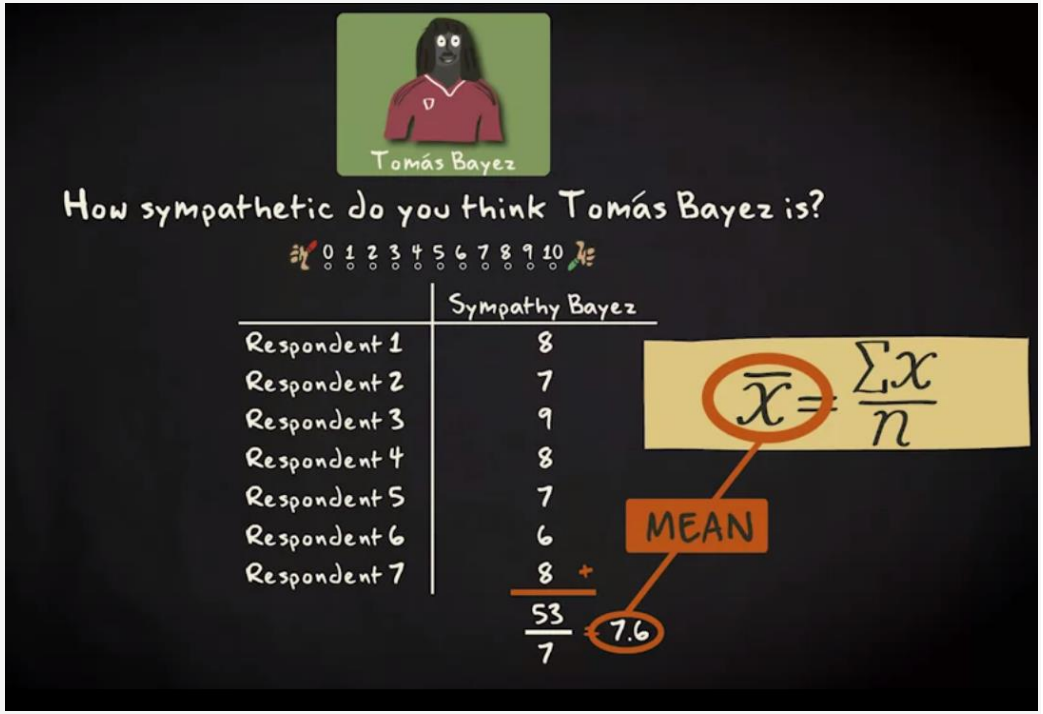
Mode, median and mean

The second measure of central tendency is the median. The median is nothing more than the middle value of your observations when they are ordered from the smallest to the largest.



Mode, median and mean

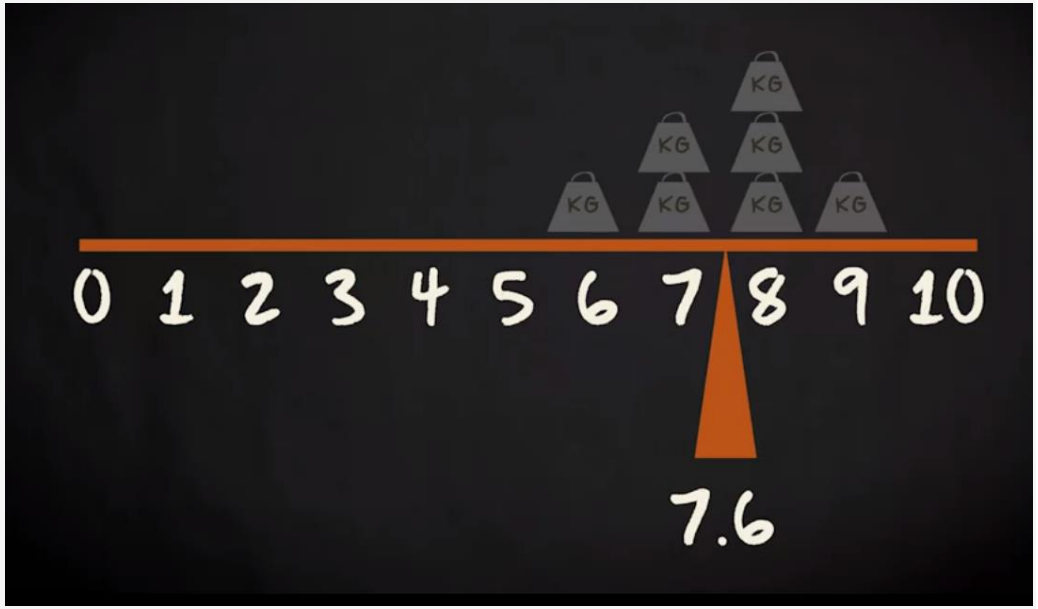
To give an example, let's again use the study on Tomas Bayez



Mode, median and mean

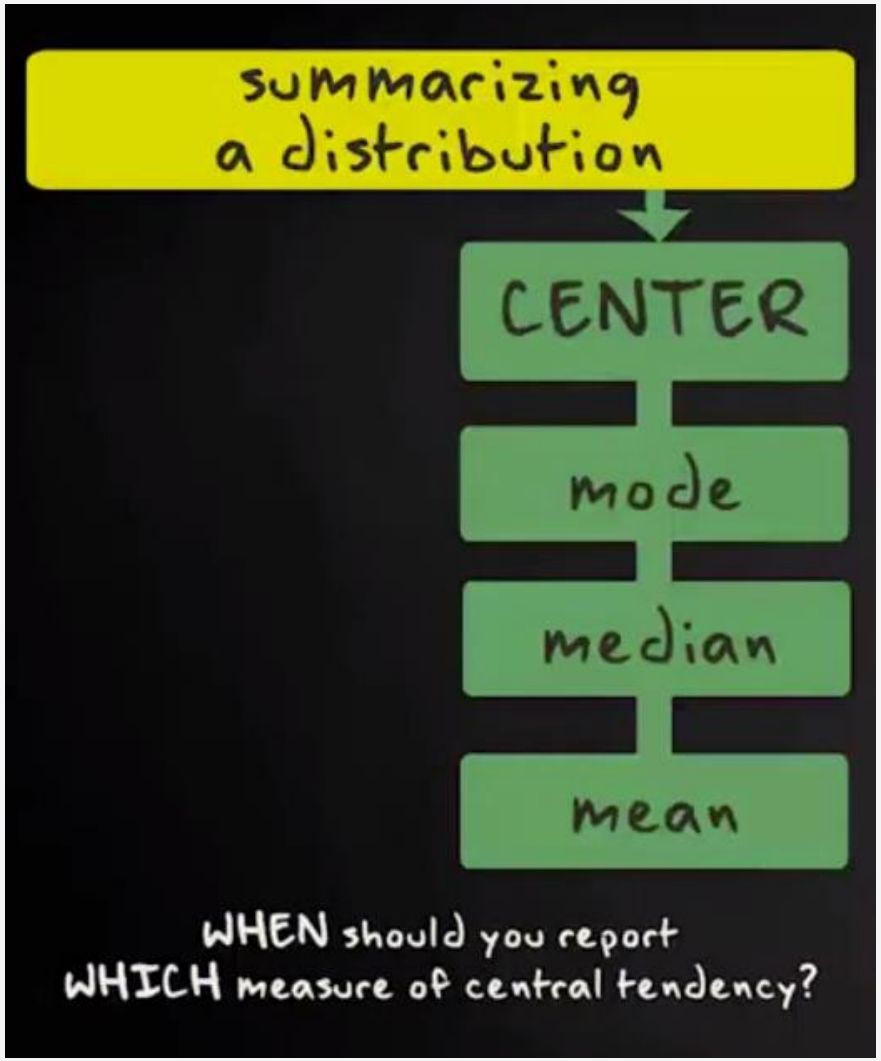
You can think of the **mean as the balance point of your data**.

Imagine we would place weights on a balance. One for each observation. Then the mean is the point on the balance where the total weight on the one side exactly equals the weight on the other side.



Mode, median and mean

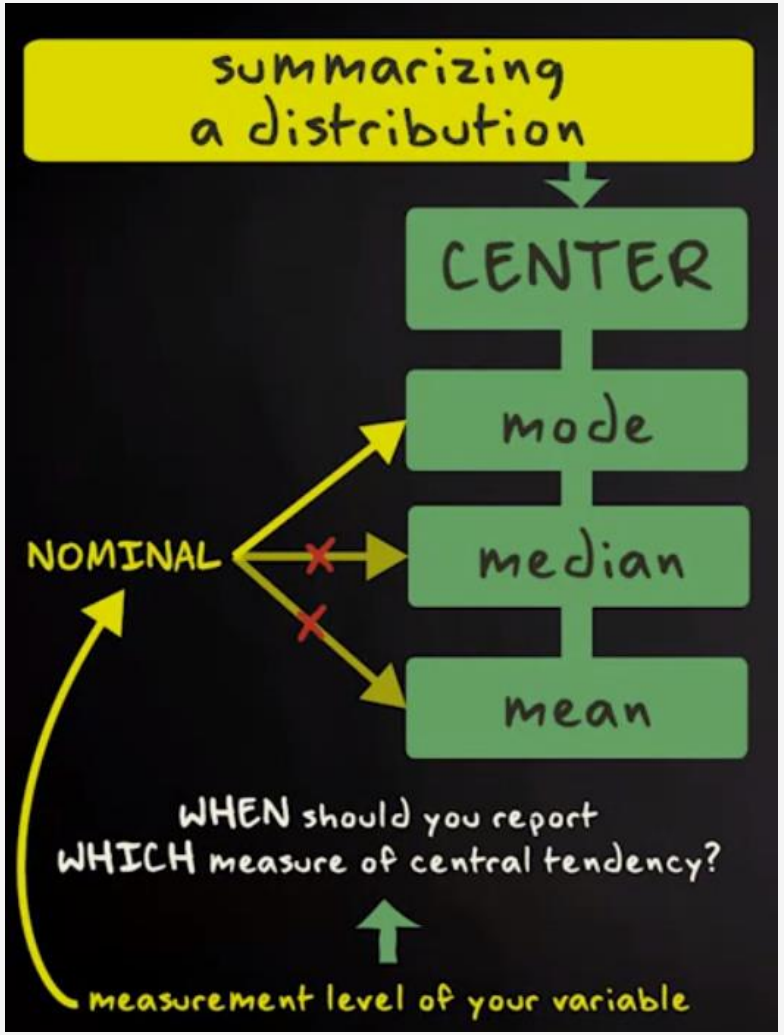
But when should you report which measure of central tendency?



Mode, median and mean

That partially depends on the the measurement level of your variable.

If it's nominal, it is impossible to compute the median, or the mean. Think about it, you cannot apply numerical operations on nominal variables, nor can you order them. The only appropriate measure of central tendency, **when a variable is nominal, is the mode**.

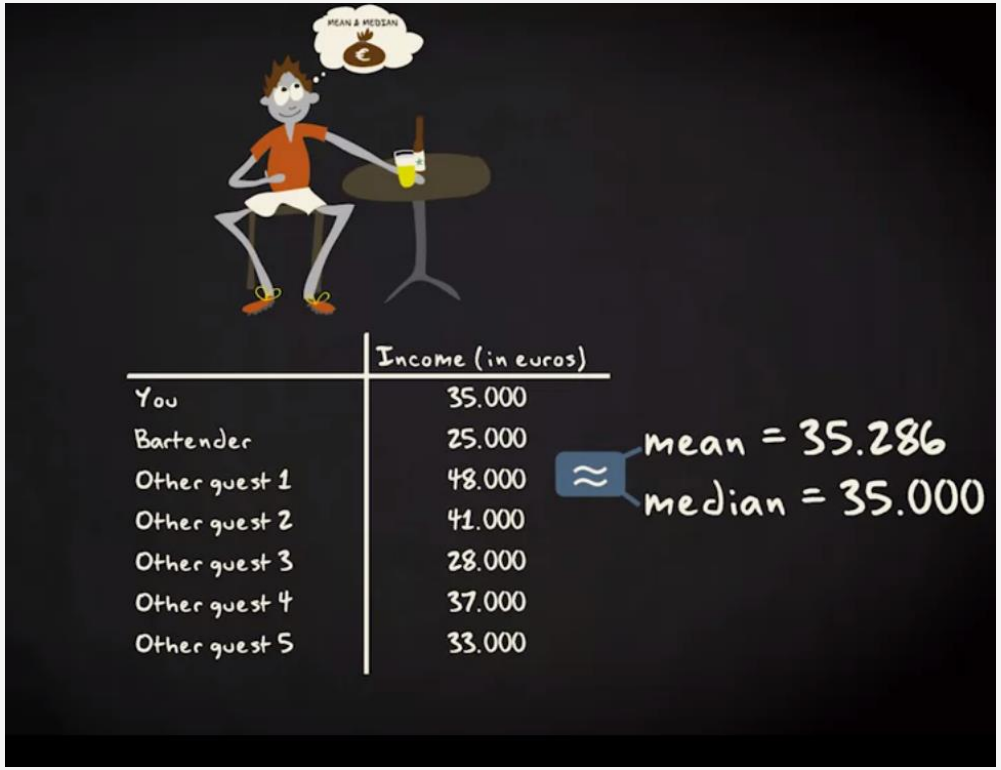


Mode, median and mean

But what to do in case of a quantitative variable?

Imagine you're sitting in a canteen of a football club in your hometown and you would like to compute the mean and median income of all persons present. That's you, 5 other guests and the bartender.

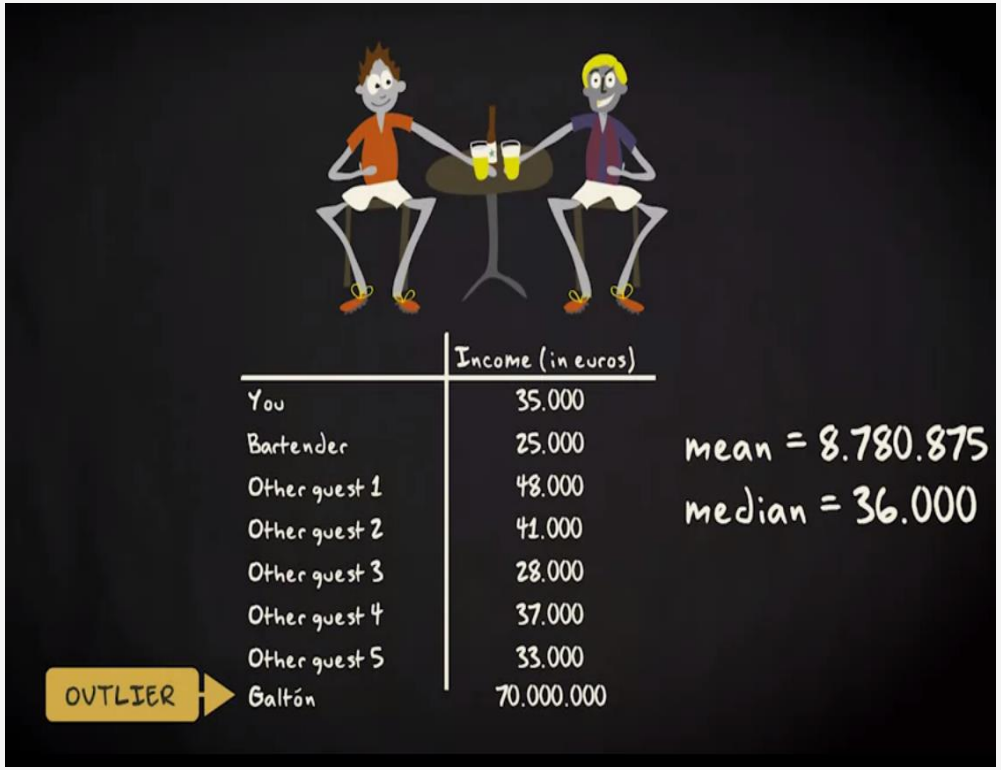
This is the data matrix, the mean is around 35.286. The median is exactly 35. Their pretty close to each other, and it doesn't matter which one you use to describe the center of your distribution.



Mode, median and mean

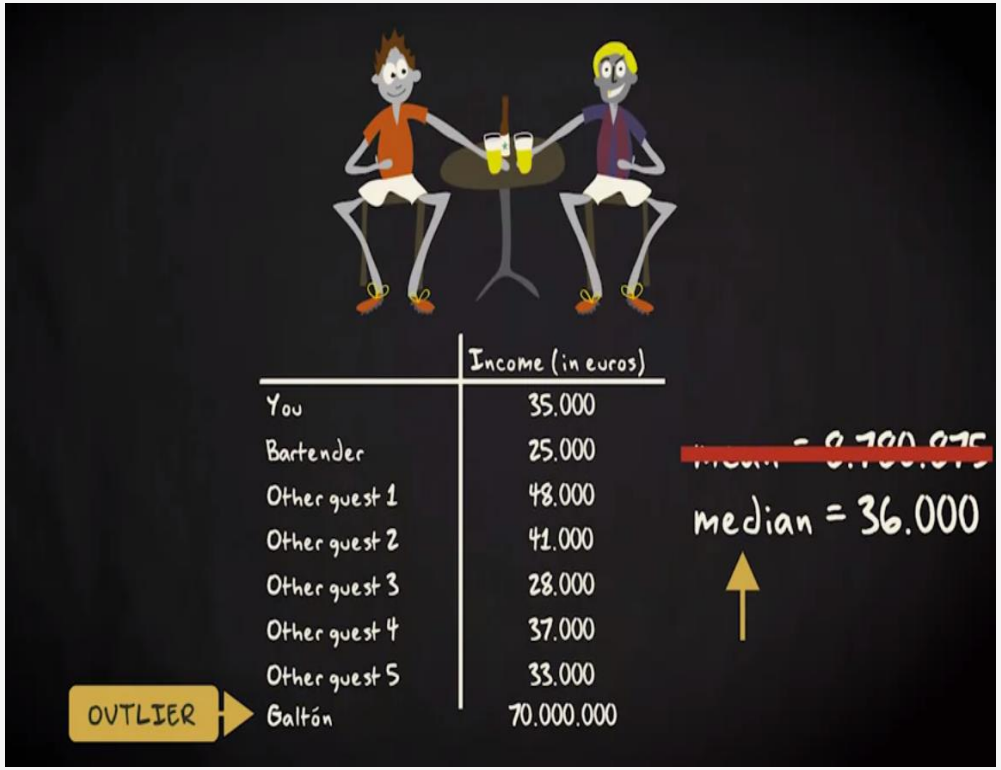
But now image the famous football player Franco Galton walks into the canteen. Say he gets about 70 million per year, the median increases slightly to 36. The mean however becomes more than 8 million now.

We say that Franco Galton is an **outlier** in this distribution. He earns much more than all the other people present and his income exerts a disproportional effect on the mean income.



Mode, median and mean

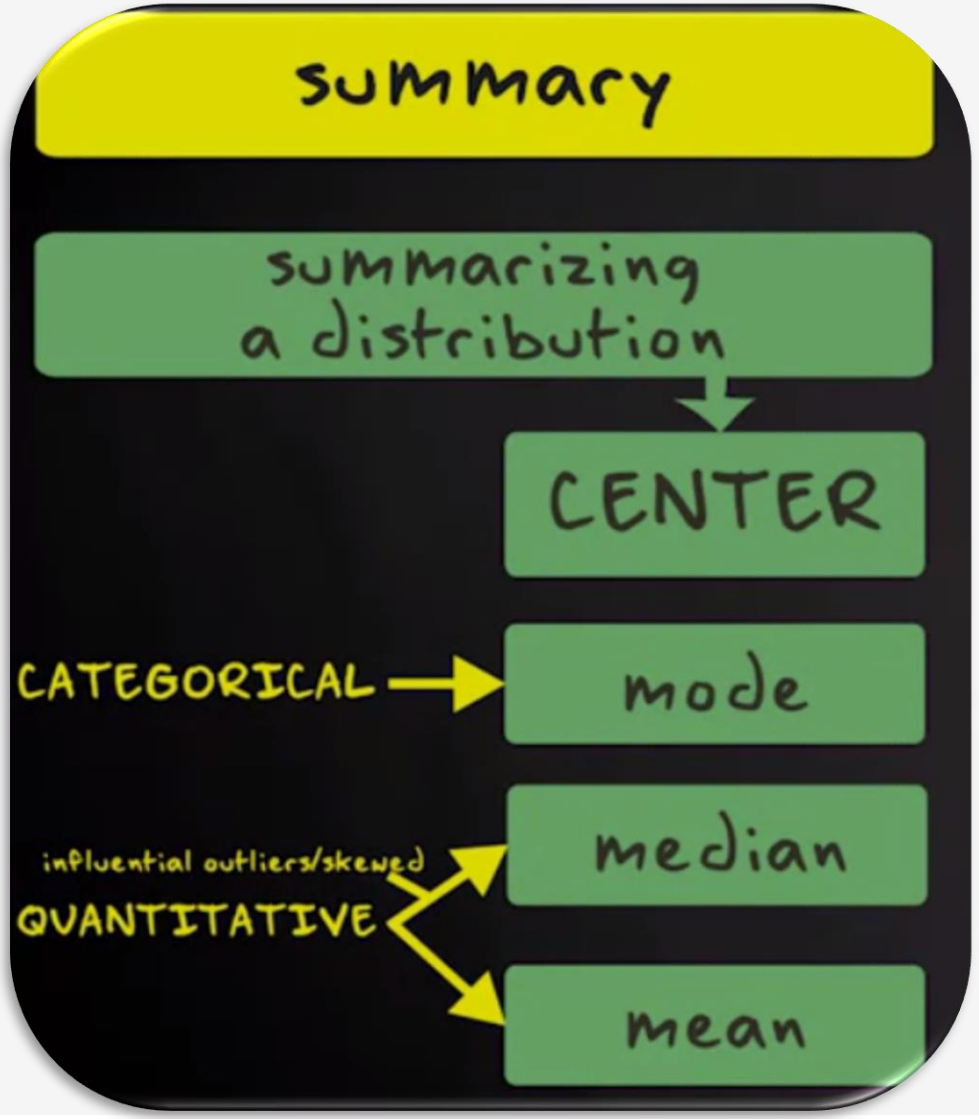
In this case, it might be argued that it makes more sense to compute a **median than** the **mean** to describe the center of the distribution.



Mode, median and mean

Let me briefly summarize what we've learned

- To describe the center of a distribution you can use three measures of center tendency, the mode, the median, and the mean.
- If your variable is categorical, you use the mode, and if it's quantitative, you employ the median or the mean.
- Go for the median if you have influential outliers or if the distribution is highly skewed, and if that's not the case, go for the mean.



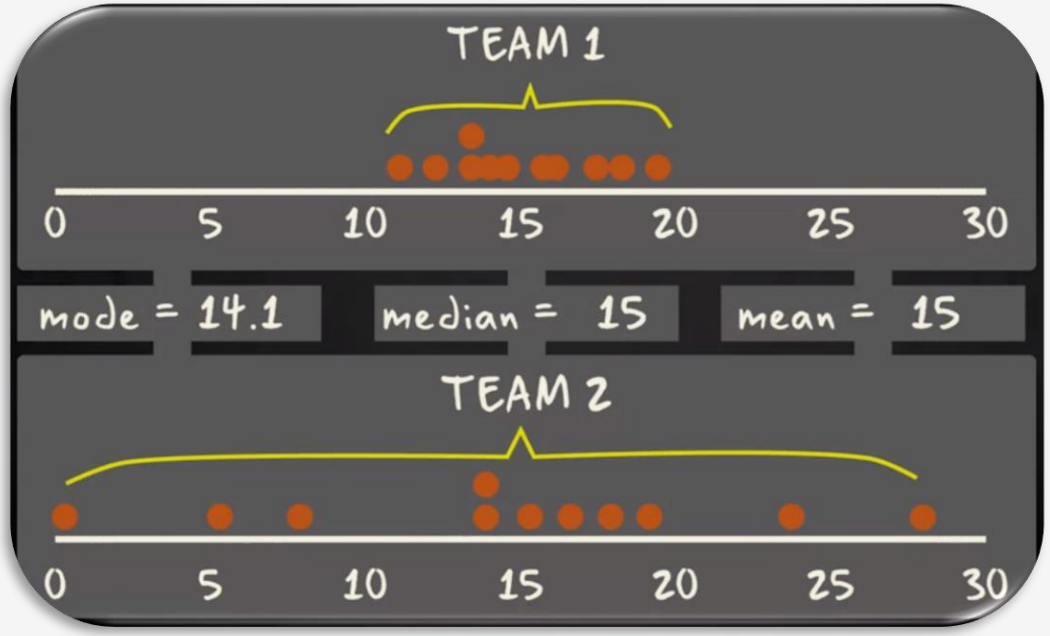
Range, interquartile range and box plot

As you might have noticed tattoos are increasingly popular among football players. The so-called tattoo sleeve in particular is rising on the football fields. A tattoo sleeve is what the name suggests, a sleeve of tattoos.

You are interested in the question to what extent football players have covered their bodies with tattoos?

Range, interquartile range and box plot

Imagine two football teams. What you see here, are dot plots representing the distribution of the variable percentage of body covered with tattoos in these two teams. The horizontal line represents this variable. And the dots stand for the 11 individuals in each team.

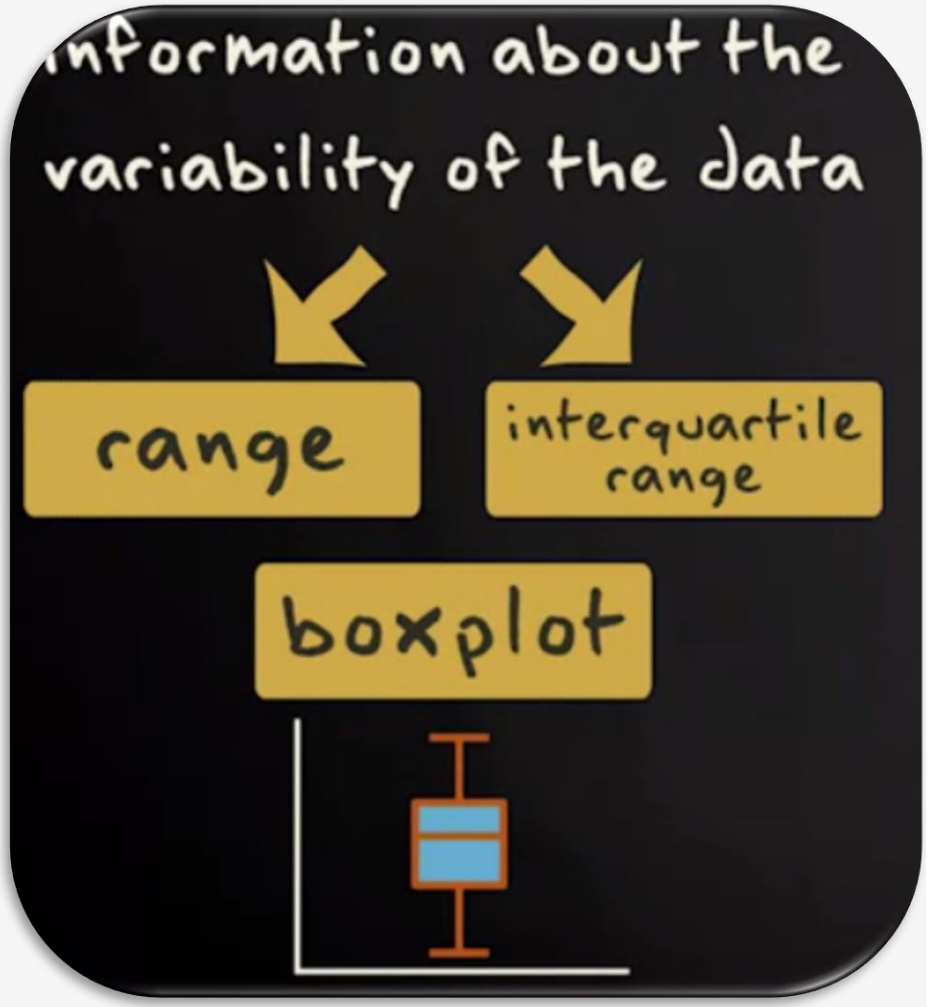


The players of team one have covered about 10 to 20% of their bodies with tattoos. In the second team, the players differ much more from each other in terms of their tattoo density. The percentage ranges from 0 to about 30%. However, mode, median and mean are the same.

This indicates that in order to adequately describe a distribution we need more information than the measures of central tendency.

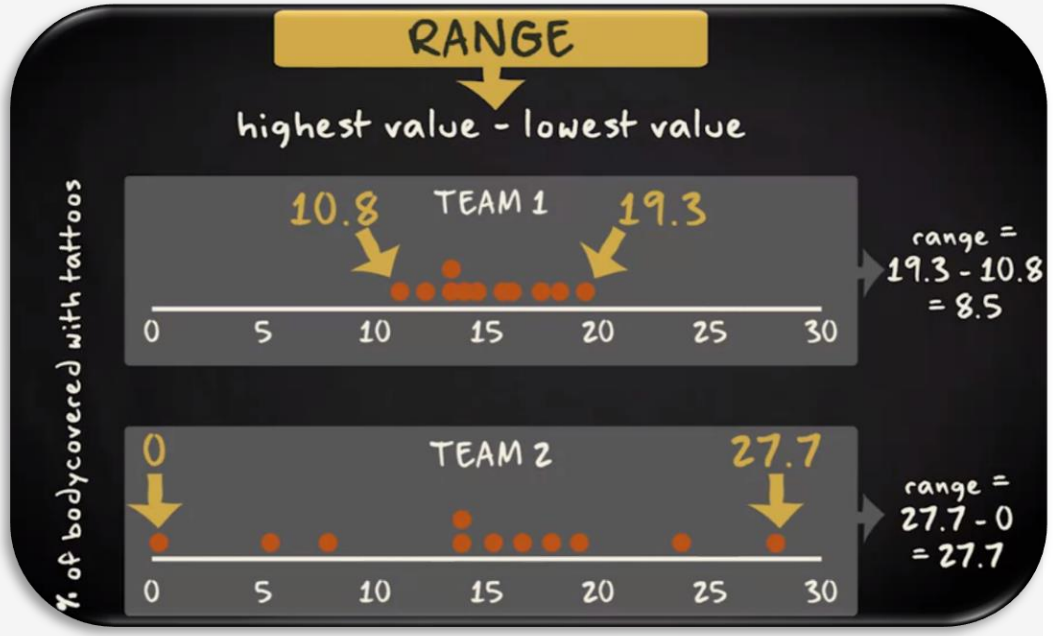
Range, interquartile range and box plot

In this section we will see information about the variability or dispersion of the data. We will discuss two measures of variability: the range and the interquartile range. We also discuss the so-called boxplot. A very useful graph that gives a good indication of how the values in a distribution are spread out.



Range, interquartile range and box plot

The most simple measure of variability is the range. It is the difference between the highest and the lowest value. Let's look at our two teams again.

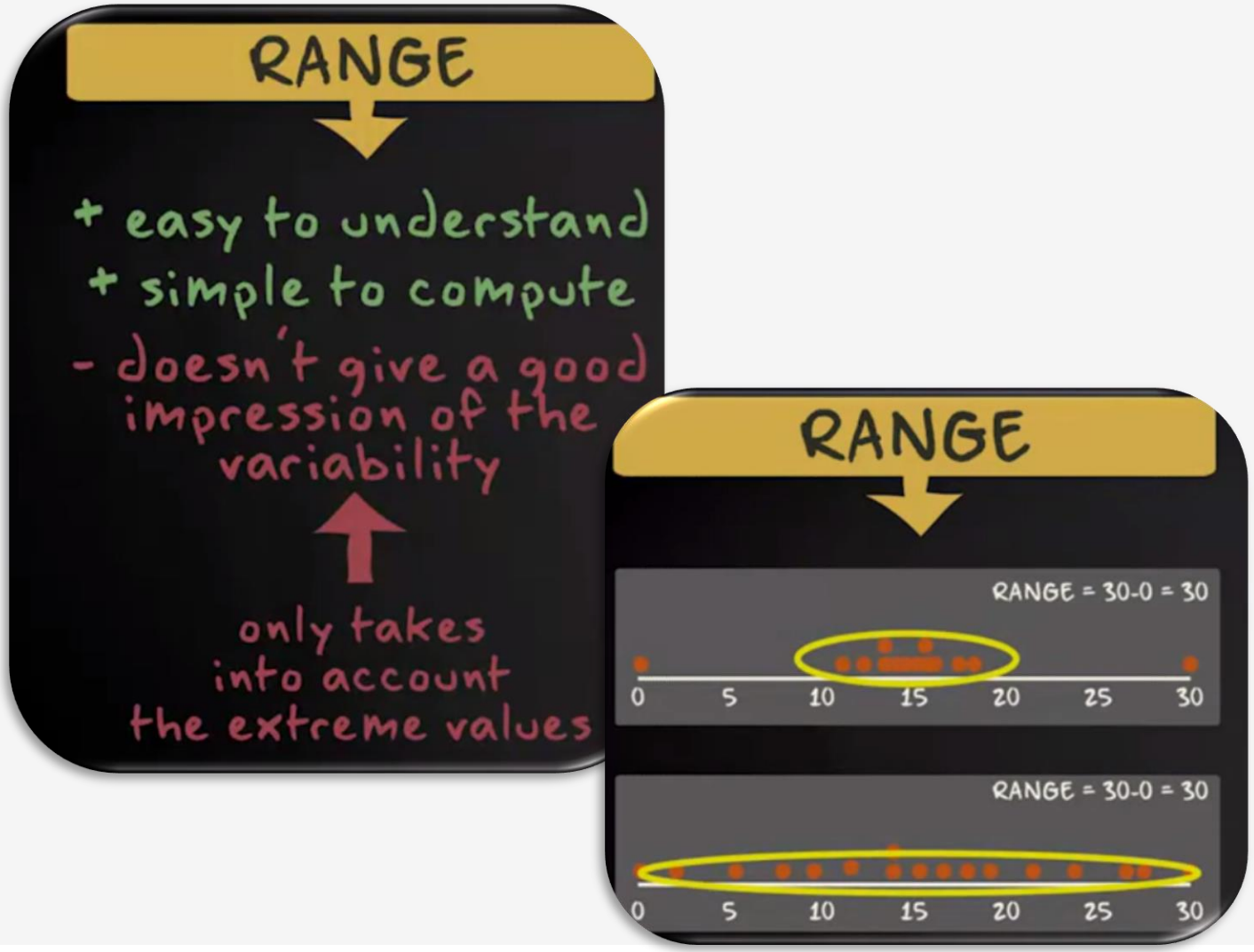


The player in Team 1 with the largest tattoo density has covered 19.3 percent of his body with tattoos. The player with the smallest tattoo density has covered 10.8 percent of his body. The range 19.3 minus 10.8 equals 8.5. In Team 2 the player with the largest tattoo density has covered is body for 27.7 percent with tattoos, and the player with the smallest density for 0 percent. The range is therefore 27.7 minus 0 is 27.7.

The range thus shows us at a glance that there is much more variability in Team 2 than in Team 1.

Range, interquartile range and box plot

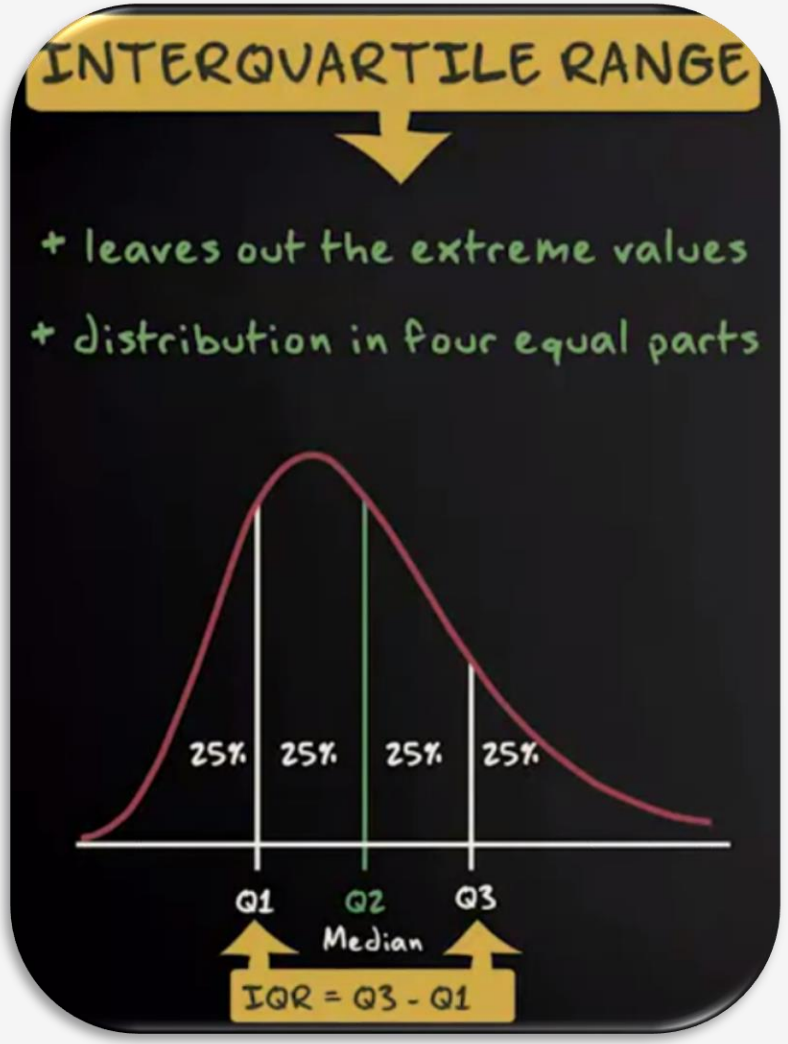
Look at these two distributions. They have the same range, but you can see immediately the variability in the second distribution is very different from the variability in the first graph



Range, interquartile range and box plot

Another measure of variability – the interquartile range – is a better measure of dispersion because it leaves out the extreme values. It basically divides your distribution in 4 equal parts.

The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1.



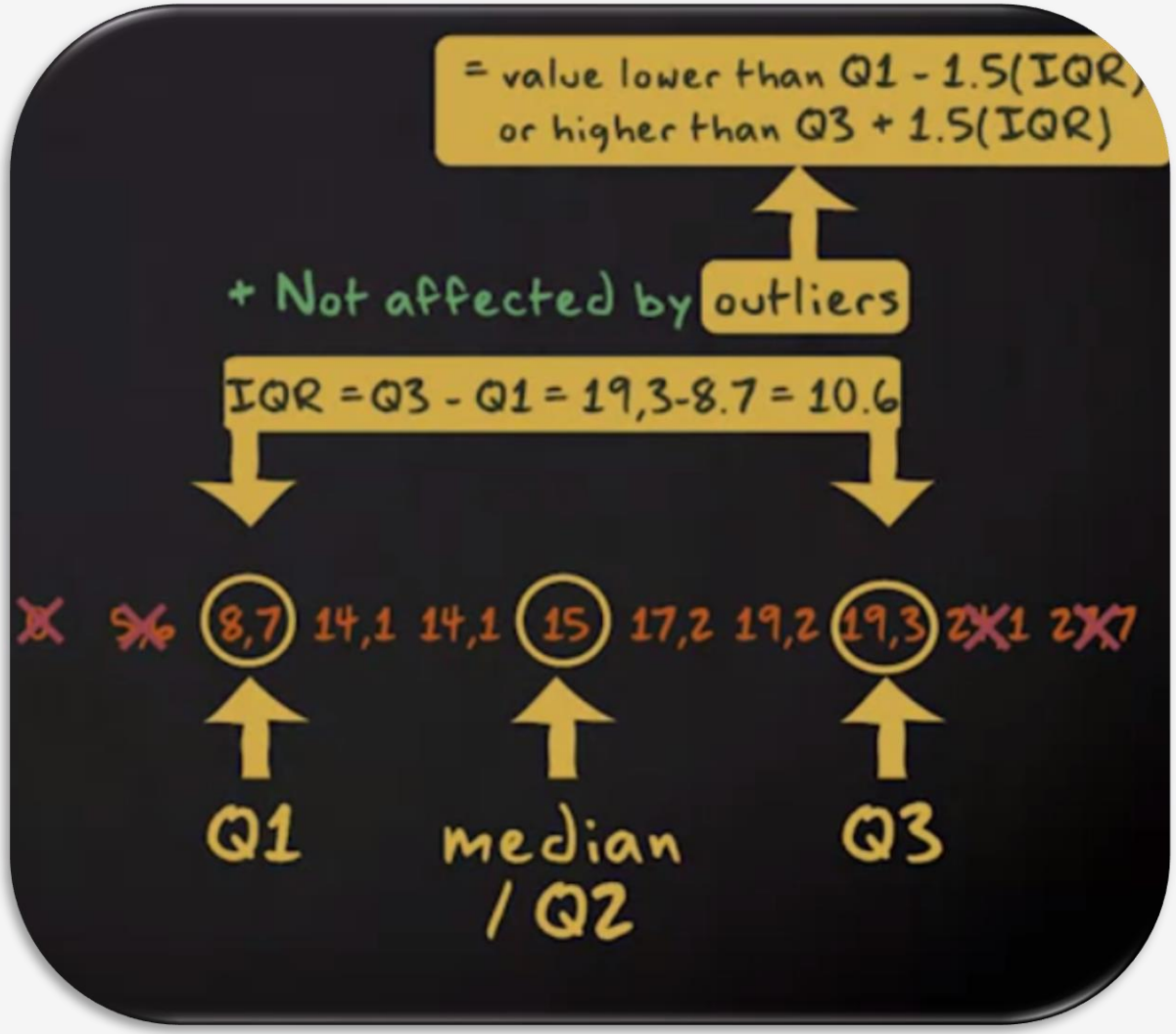
Range, interquartile range and box plot

How to compute IQR by going back to the tattoo density example. This is what the distribution of Team 2 looked like and calculate IQR from this data matrix.

Range, interquartile range and box plot

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. Yet, it might still be useful to look for possible outliers in your study.

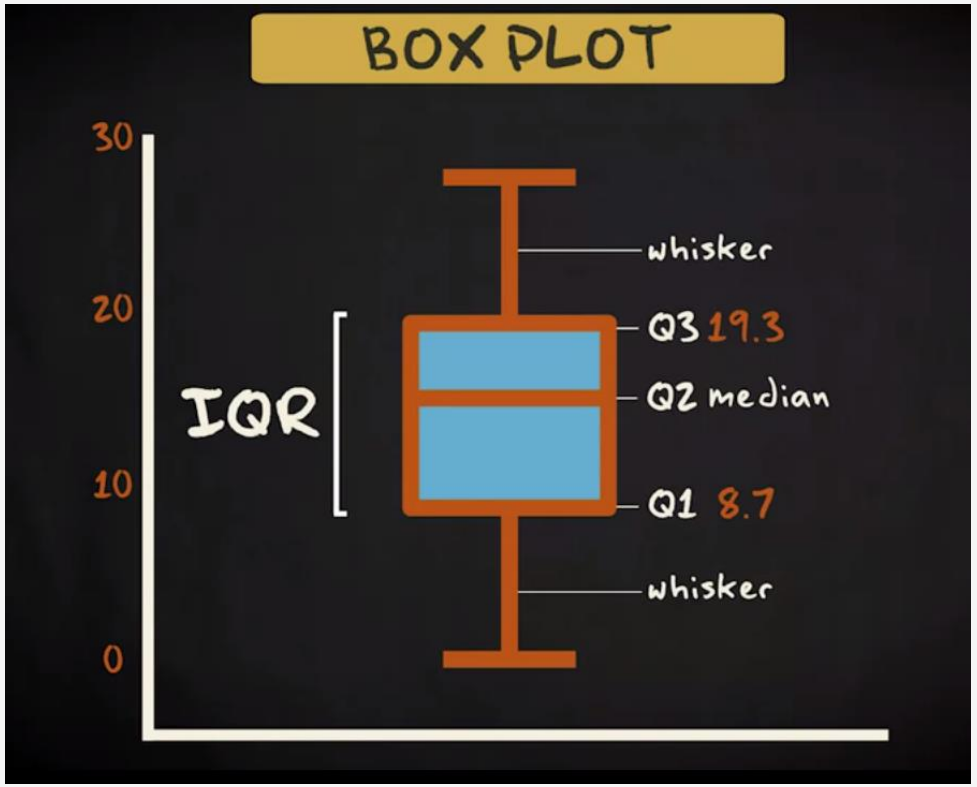
As a rule of thumb, observations can be qualified as **outliers** when they lie **more than 1.5 IQR** below the **first quartile** or **1.5 IQR above** the **third quartile**.



Range, interquartile range and box plot

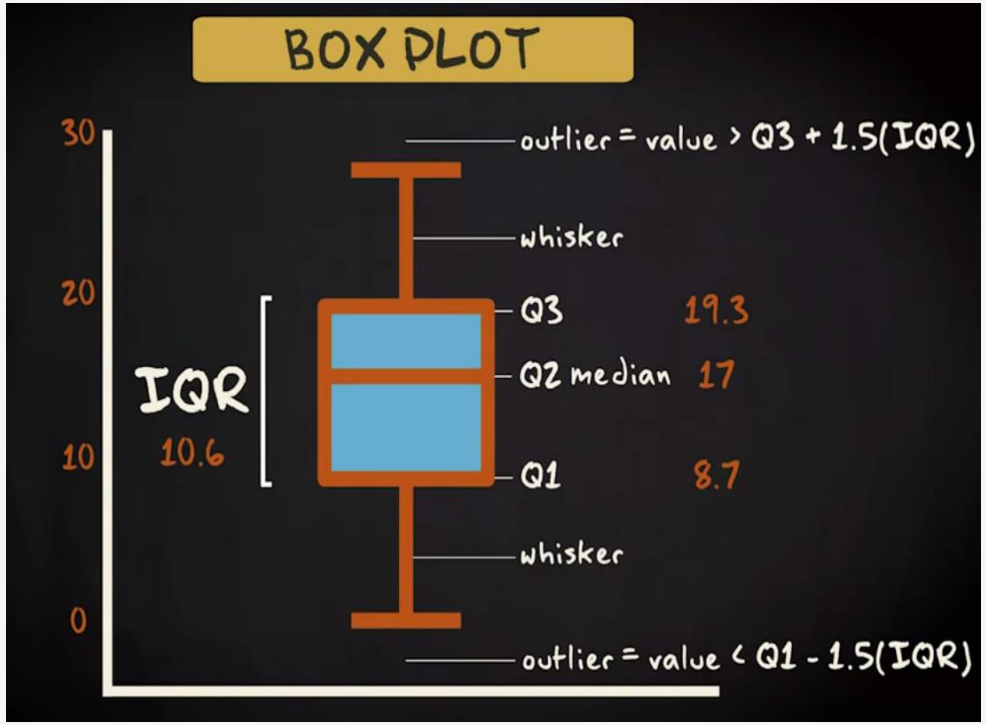
There is one specific type of graph that is very useful when it comes to describing center and variability and detecting outliers. That graph is the called **box plot**.

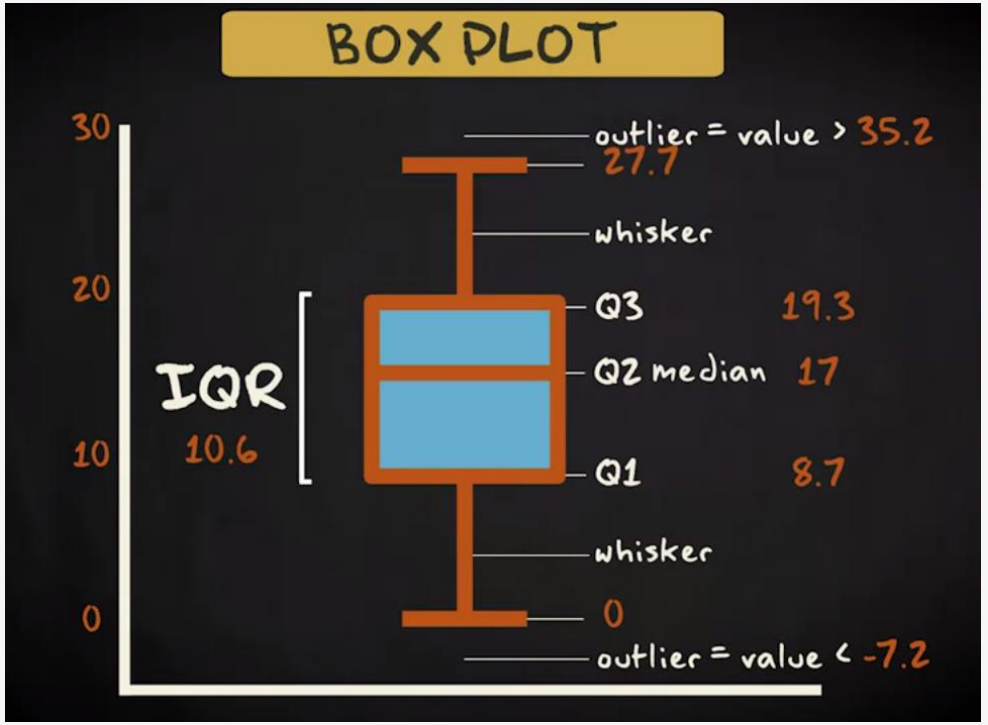
The box plot shows you at a glance Q1, Q2 and Q3, the minimum value that's not an outlier, the maximum value that's not an outlier, and the outliers.



Range, interquartile range and box plot

How do you decide how long the whiskers should be?

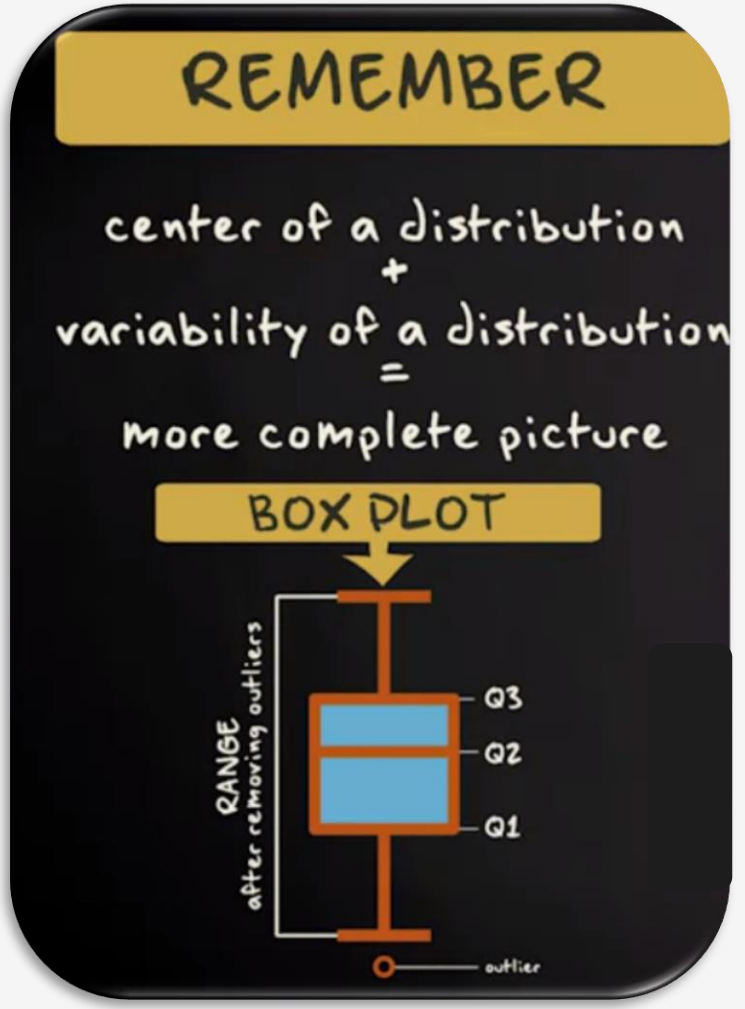




Range, interquartile range and box plot

In summary, the center of a distribution only tells you one part of the story. For a more complete picture, also assess the variability of a distribution!

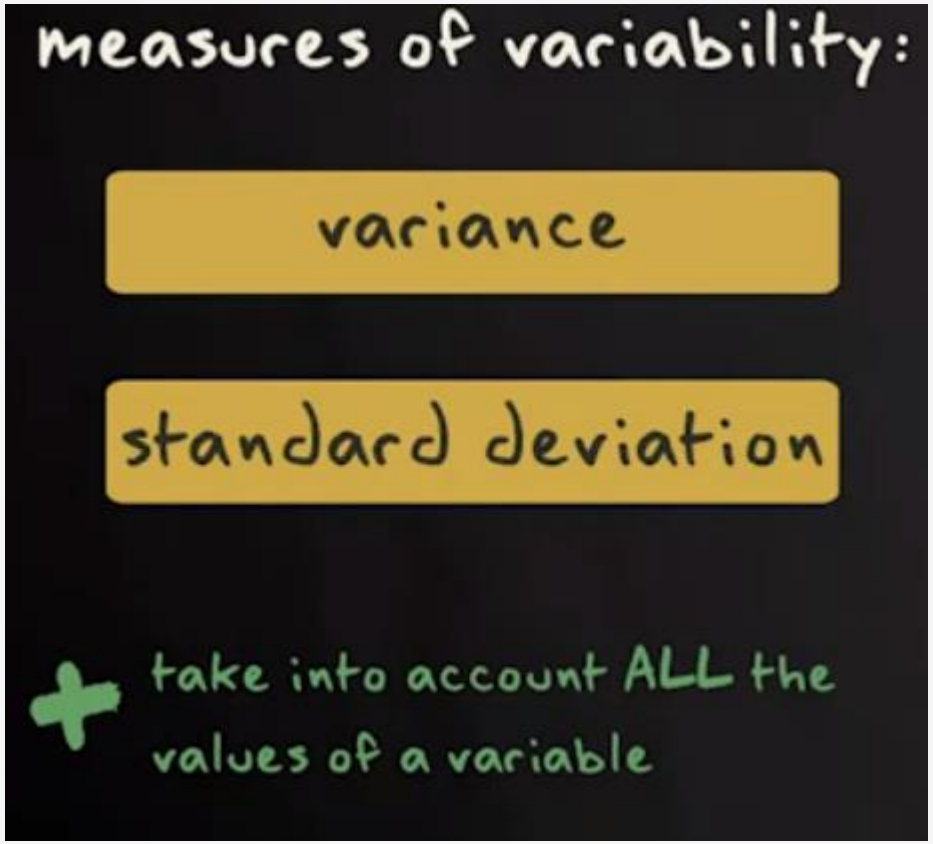
A box plot shows important aspects of a distribution in a compact way, using the three quartiles, the outliers, and the range of the data after removing the outliers.



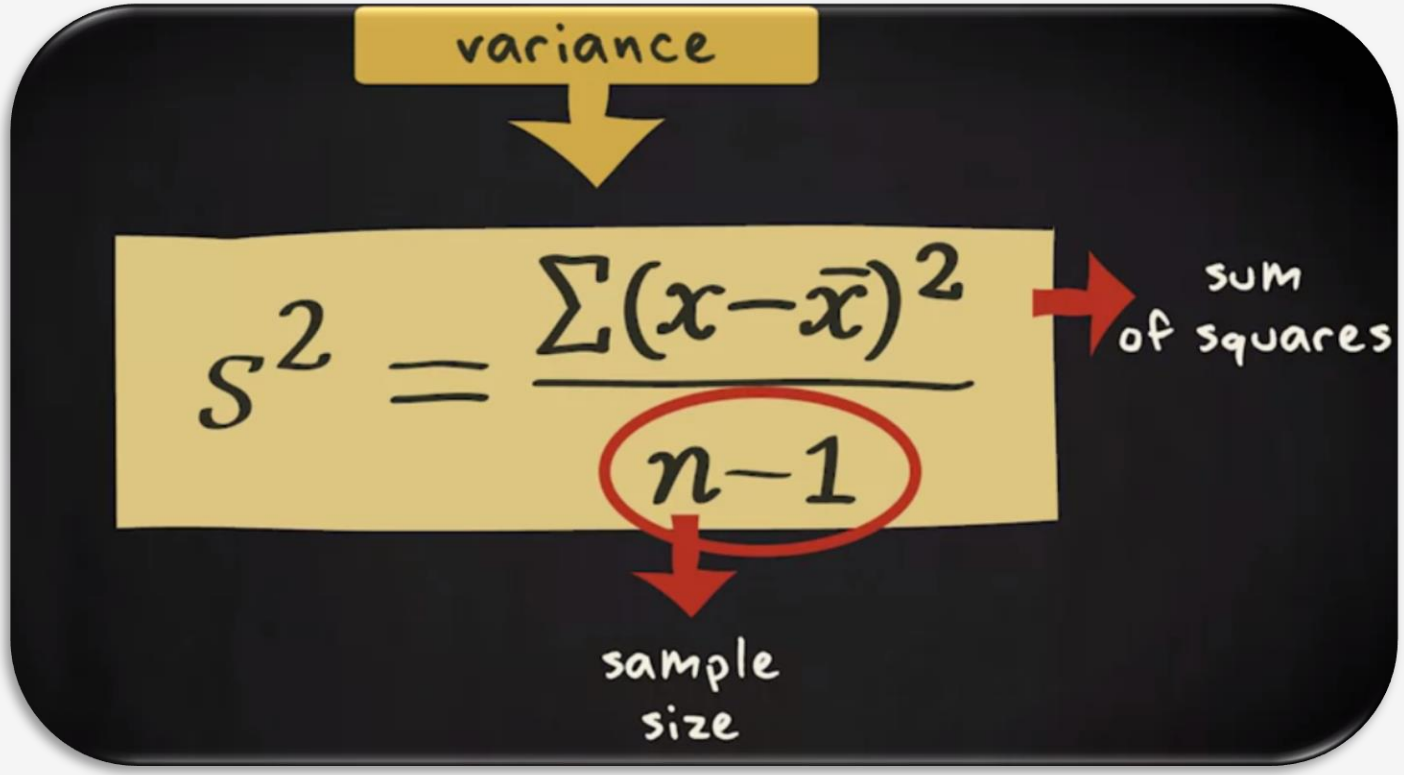
Variance and standard deviation

In this section we'll discuss two other measures of variability that are used very often in statistical studies: the **variance** and the **standard deviation**.

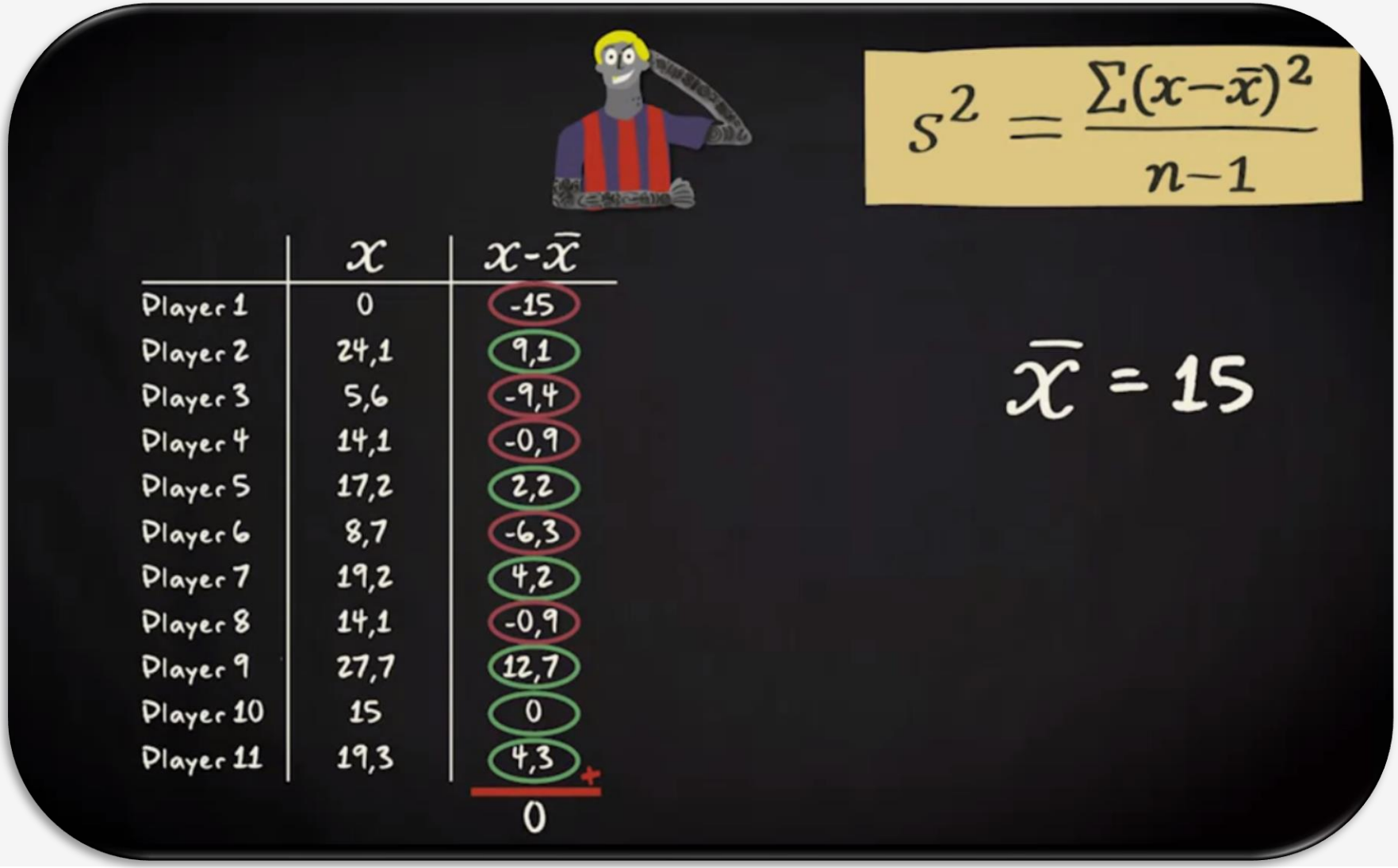
The huge advantage of the variance and standard deviation over many other measures of variability is that they take into account all the values of a variable.



Variance and standard deviation



Variance and standard deviation



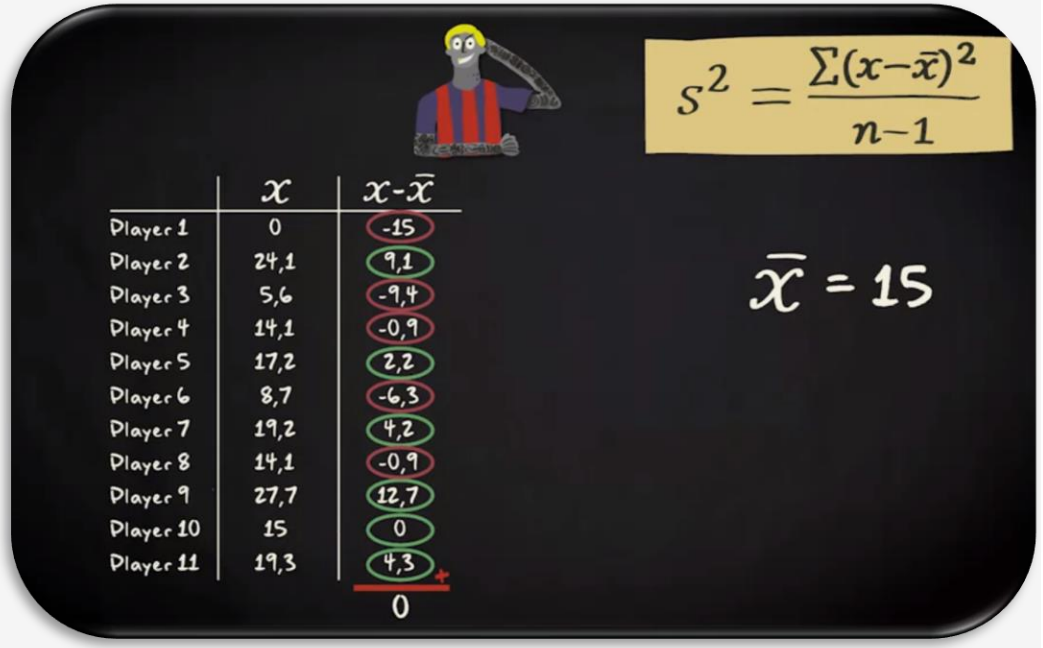
	x	$x-\bar{x}$
Player 1	0	-15
Player 2	24,1	9,1
Player 3	5,6	-9,4
Player 4	14,1	-0,9
Player 5	17,2	2,2
Player 6	8,7	-6,3
Player 7	19,2	4,2
Player 8	14,1	-0,9
Player 9	27,7	12,7
Player 10	15	0
Player 11	19,3	4,3

Variance and standard deviation

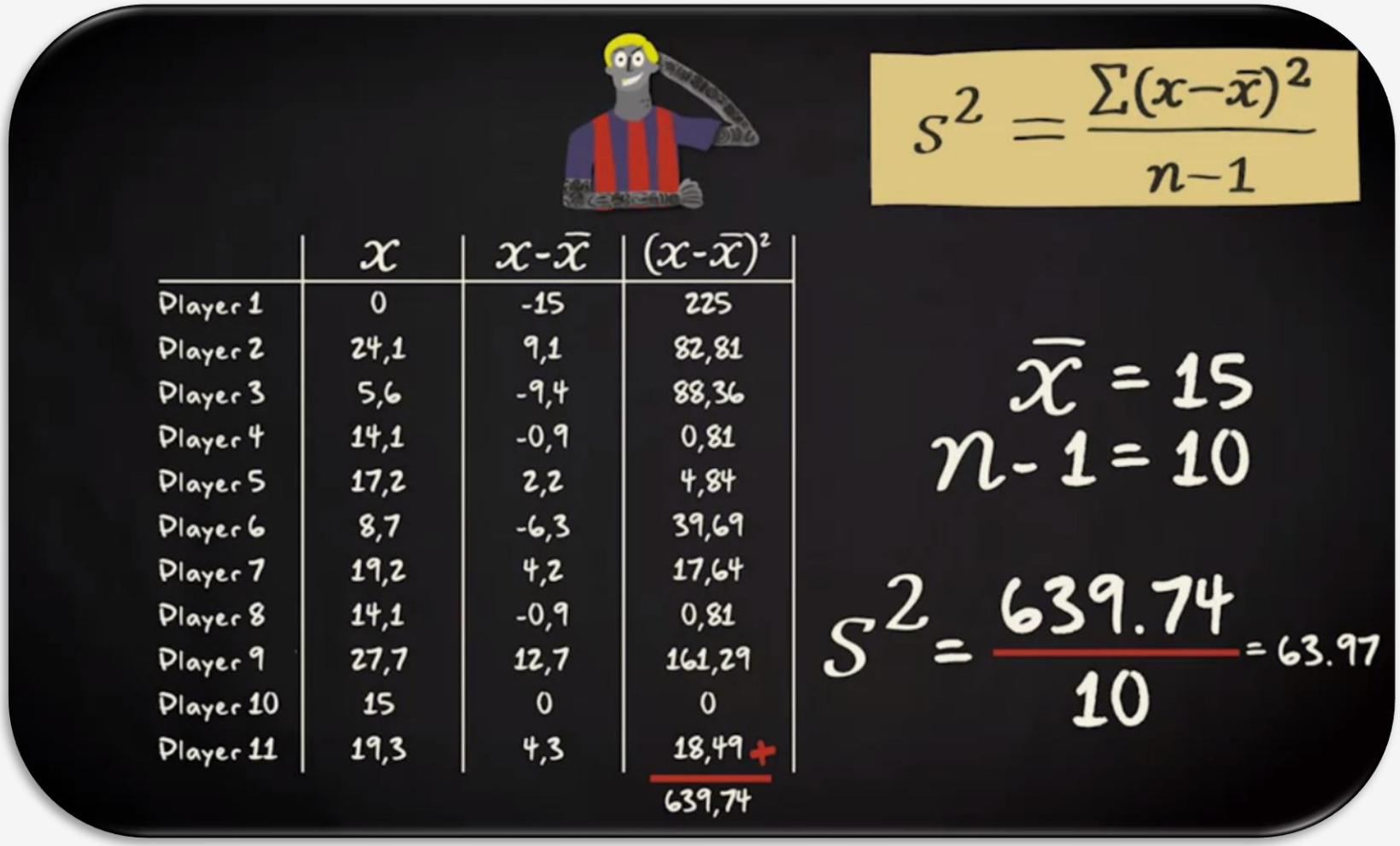
Notice that you now have negative and positive numbers. This is not strange as the mean is the middle, or the balance point of these values.

In fact, because the mean lies exactly in the middle, the negative deviations from the mean counterbalance the positive deviations from the mean, as a result of which the sum of the deviations equals 0. In other words: the sum of these values equals 0.

For that reason we don't use the original deviations but the squared deviations.



Variance and standard deviation



	x	$x-\bar{x}$	$(x-\bar{x})^2$
Player 1	0	-15	225
Player 2	24,1	9,1	82,81
Player 3	5,6	-9,4	88,36
Player 4	14,1	-0,9	0,81
Player 5	17,2	2,2	4,84
Player 6	8,7	-6,3	39,69
Player 7	19,2	4,2	17,64
Player 8	14,1	-0,9	0,81
Player 9	27,7	12,7	161,29
Player 10	15	0	0
Player 11	19,3	4,3	18,49
			639,74

Variance and standard deviation

The larger the variance, the larger the variability. That means: the larger the variance, the more the values are spread out around the mean.

The first team, displayed here, has a variance of about 6.33.

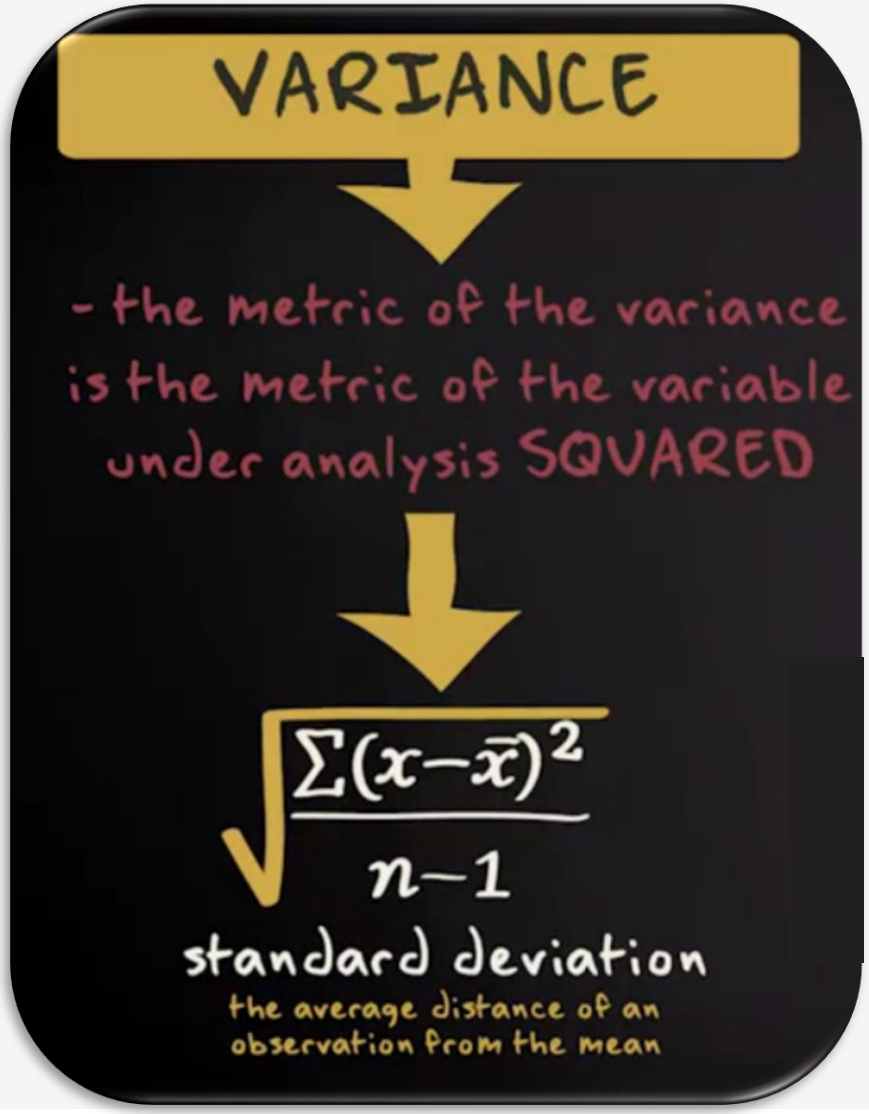
You can see that the larger variability of tattoo density in Team 2 that was already visible from the dot plots and the box plots is also represented by the larger variance.

Variance and standard deviation

An important **disadvantage** of the variance is that the metric of the variance is the metric of the variable under analysis squared. After all, we have squared the positive and negative deviations so that they don't cancel each other out.

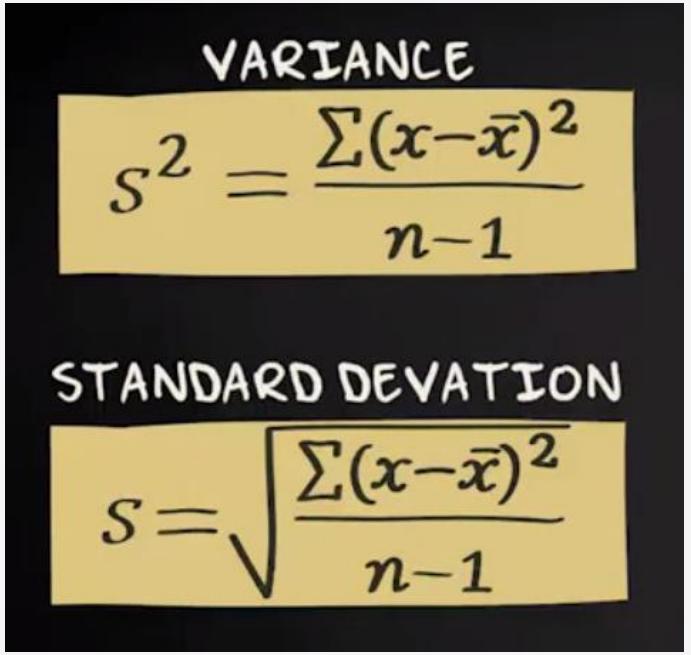
There is a very simple solution to get rid of this problem: we just take the square root of the variance.

We call what we get the **standard deviation**.



Variance and standard deviation

Standard Deviation can be seen as the average distance of an observation from the mean.



Variance and standard deviation

The standard deviation is the measure of dispersion that is used most often. However, in many statistical methods the variance plays an important role as well. In this section, we have learned that they are closely related, and that we can easily derive the one from the other.

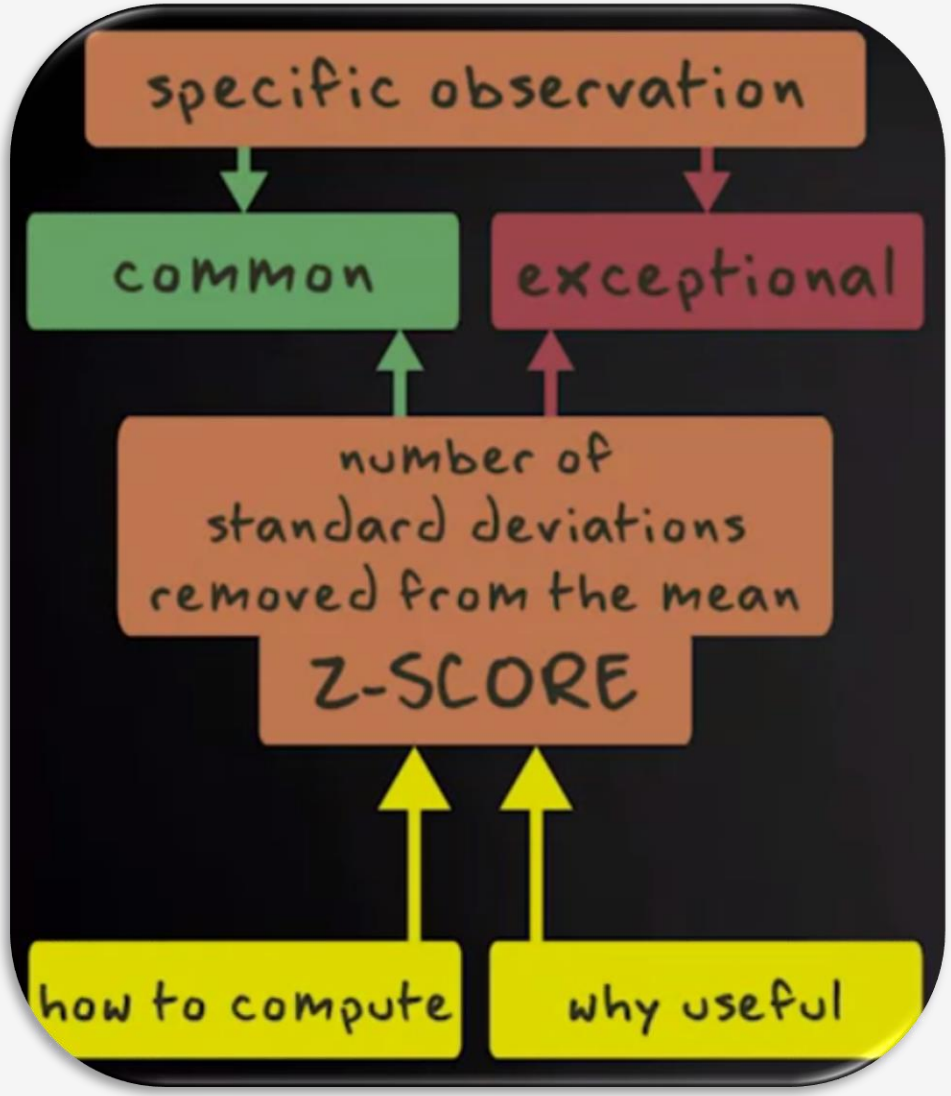
Z-scores

Sometimes researchers want to know if a specific observation is common or exceptional. To answer that question, they express a score in terms of the number of standard deviations it is removed from the mean. This number is what we call a **z-score**. If we recode original scores into z-scores, we say that we **standardize** a variable.

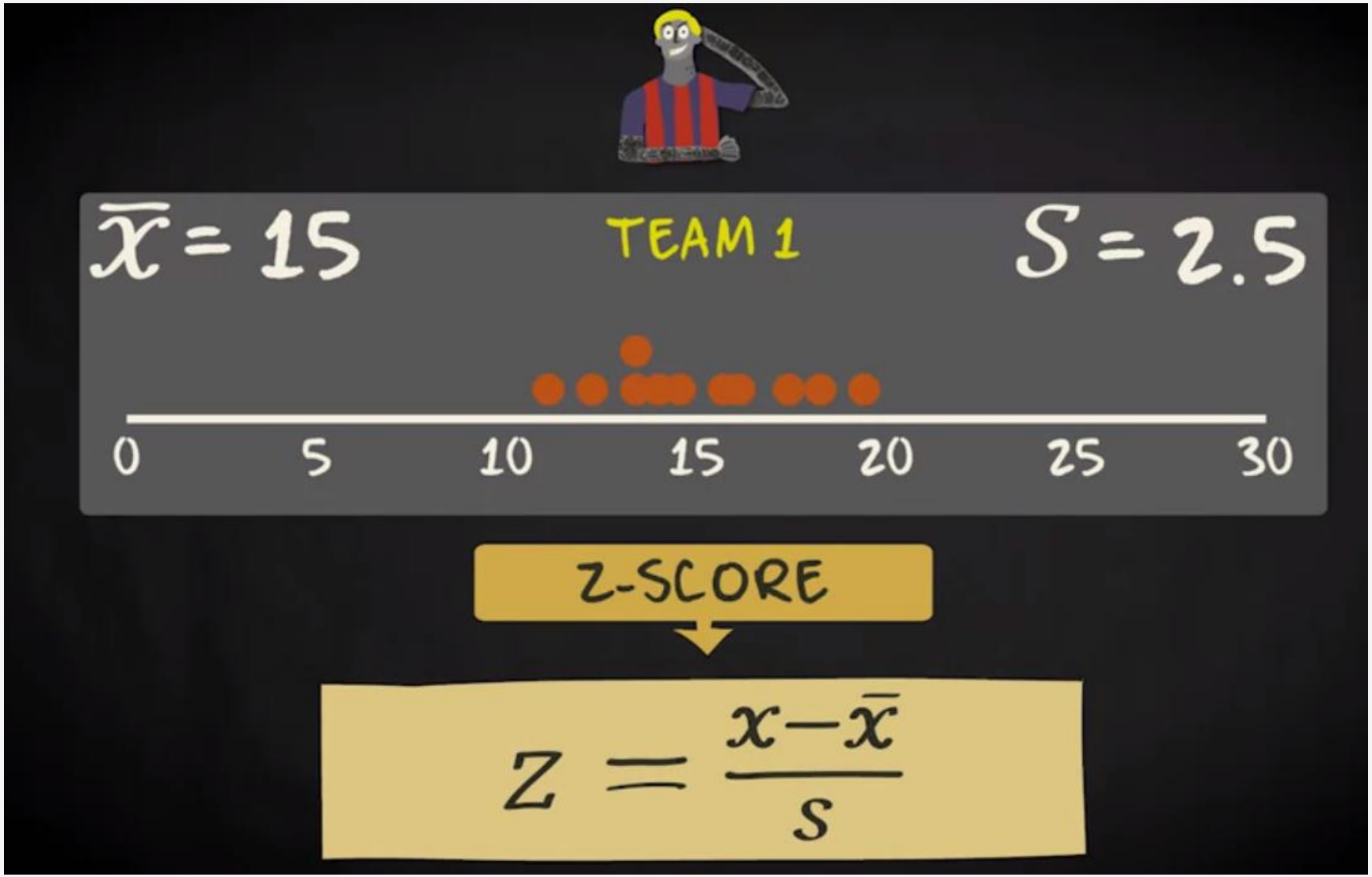
Z-scores

Sometimes, analyst ask the question if a specific observation is common or exceptional. To answer that question, they express a score in terms of the number of standard deviations it is removed from the mean. This number is what we call a z-score.

In this section we'll explain how we can compute z-scores and why they can be useful.



Z-scores



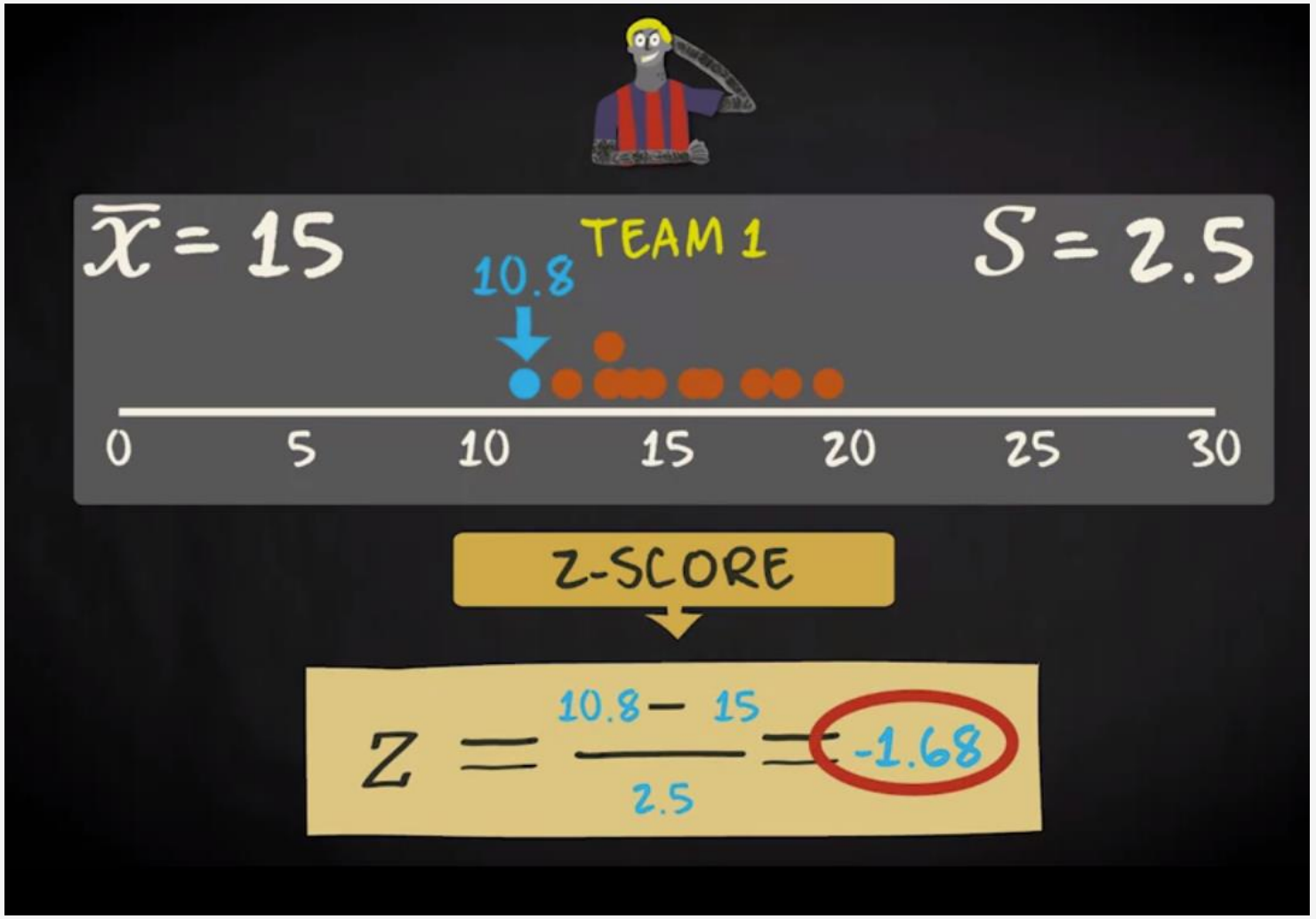
Z-scores

Let's see what that means for a tattoo density of 10.8 percent.

The z-score of that value is 10.8 minus 15 divided by 2.5. That equals -1.68.

So, the z-score is -1.68.

You can do that for all the values in your distribution.



Z-scores

Notice that you end up with negative z-scores and positive z-scores.

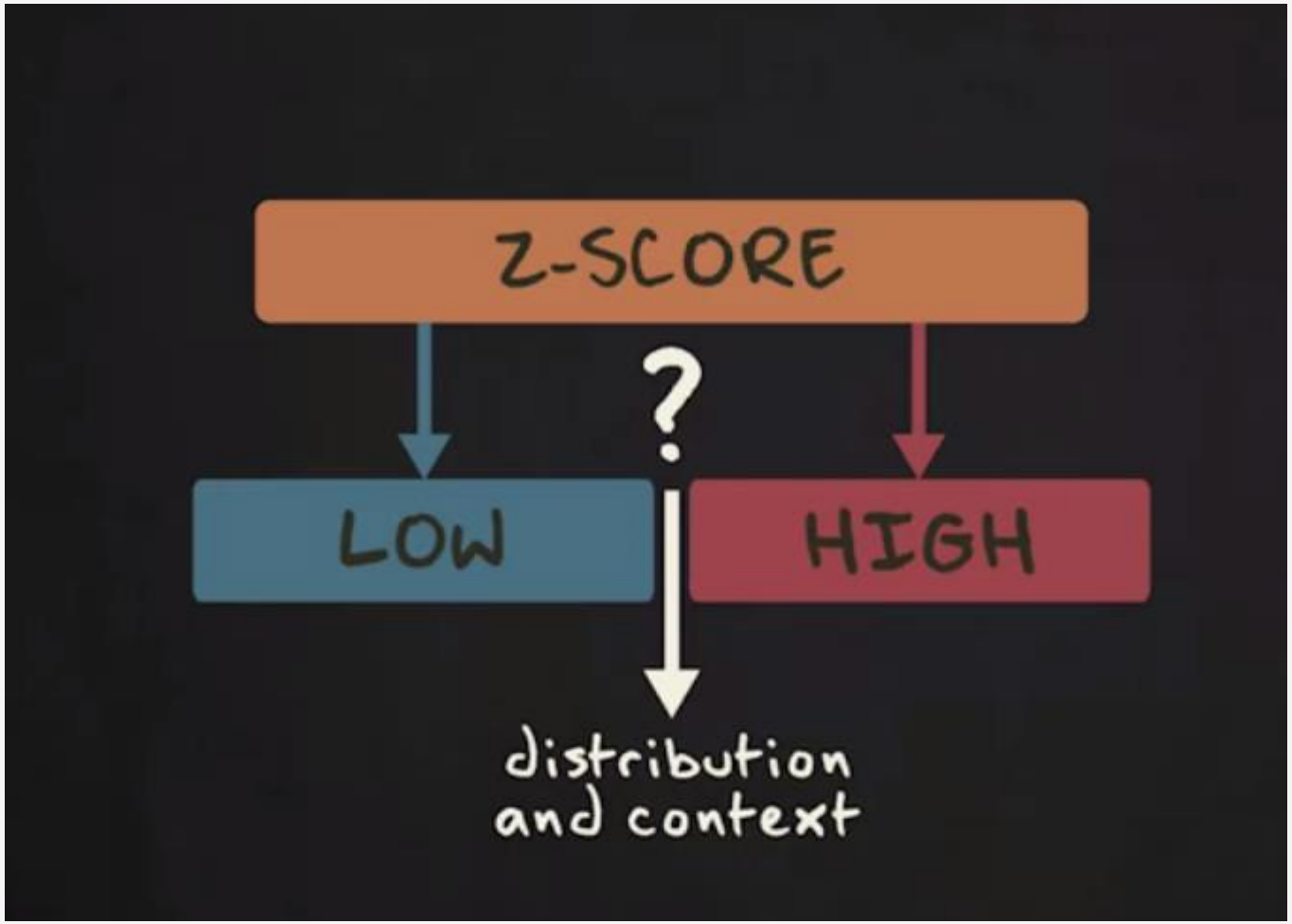
Negative z-scores represent values below the mean.

Positive z-scores represent values above the mean.

Because the mean is the balance point of your distribution, the negative and positive z-scores cancel each other out. In other words, if you add up all z-scores you will get a value of 0.

Z-scores

How do you know if a certain z-score is low or high? Well, that depends on your distribution and on context

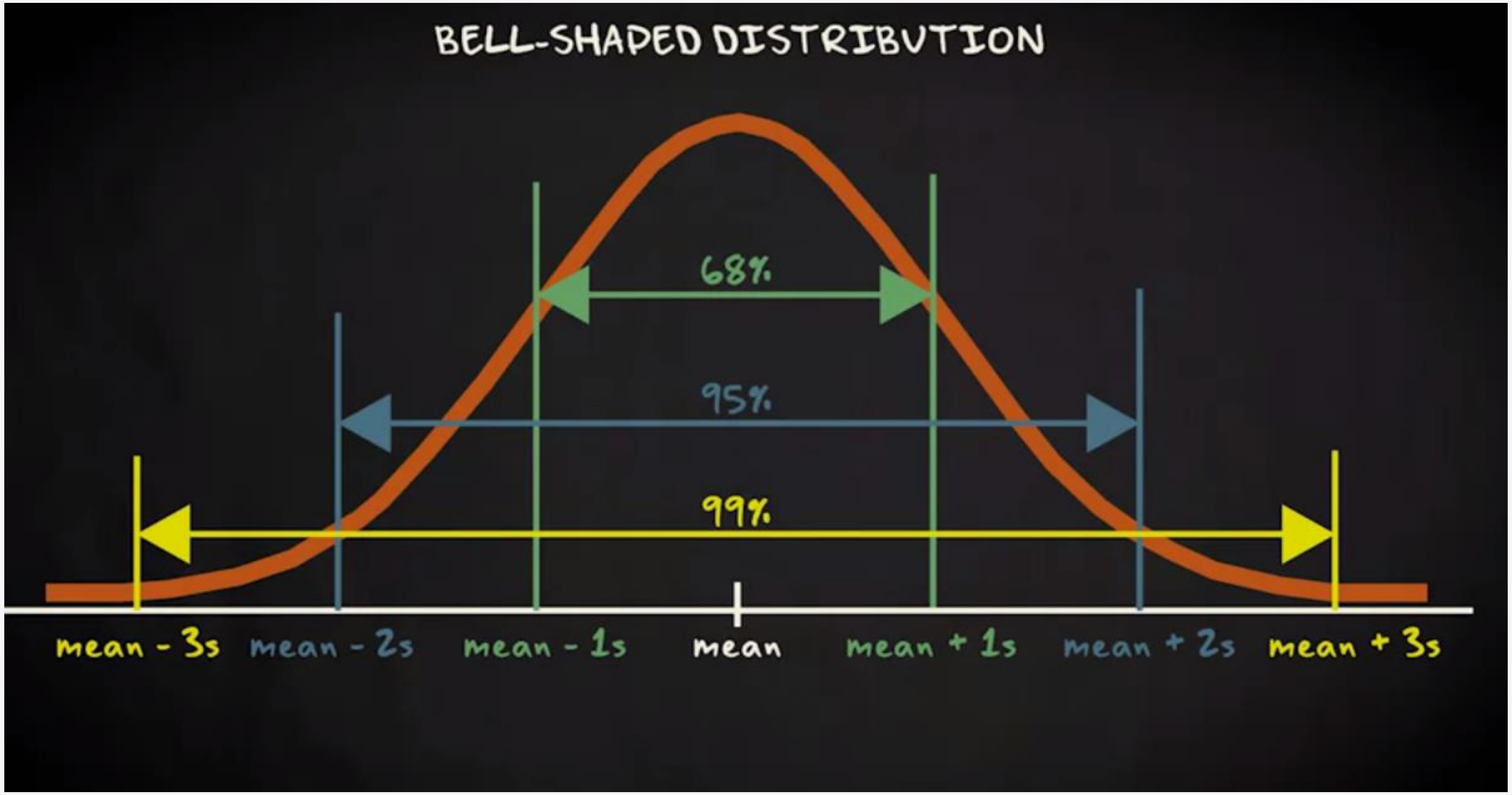


Z-scores

A good rule is that IF the histogram of your variable is bell-shaped,

- ✓ **68** percent of the observations fall between z-scores of **minus 1 and 1**;
- ✓ **95** percent between z-scores of **minus 2 and 2**; and
- ✓ **99** percent between z-scores of **minus 3 and 3**.

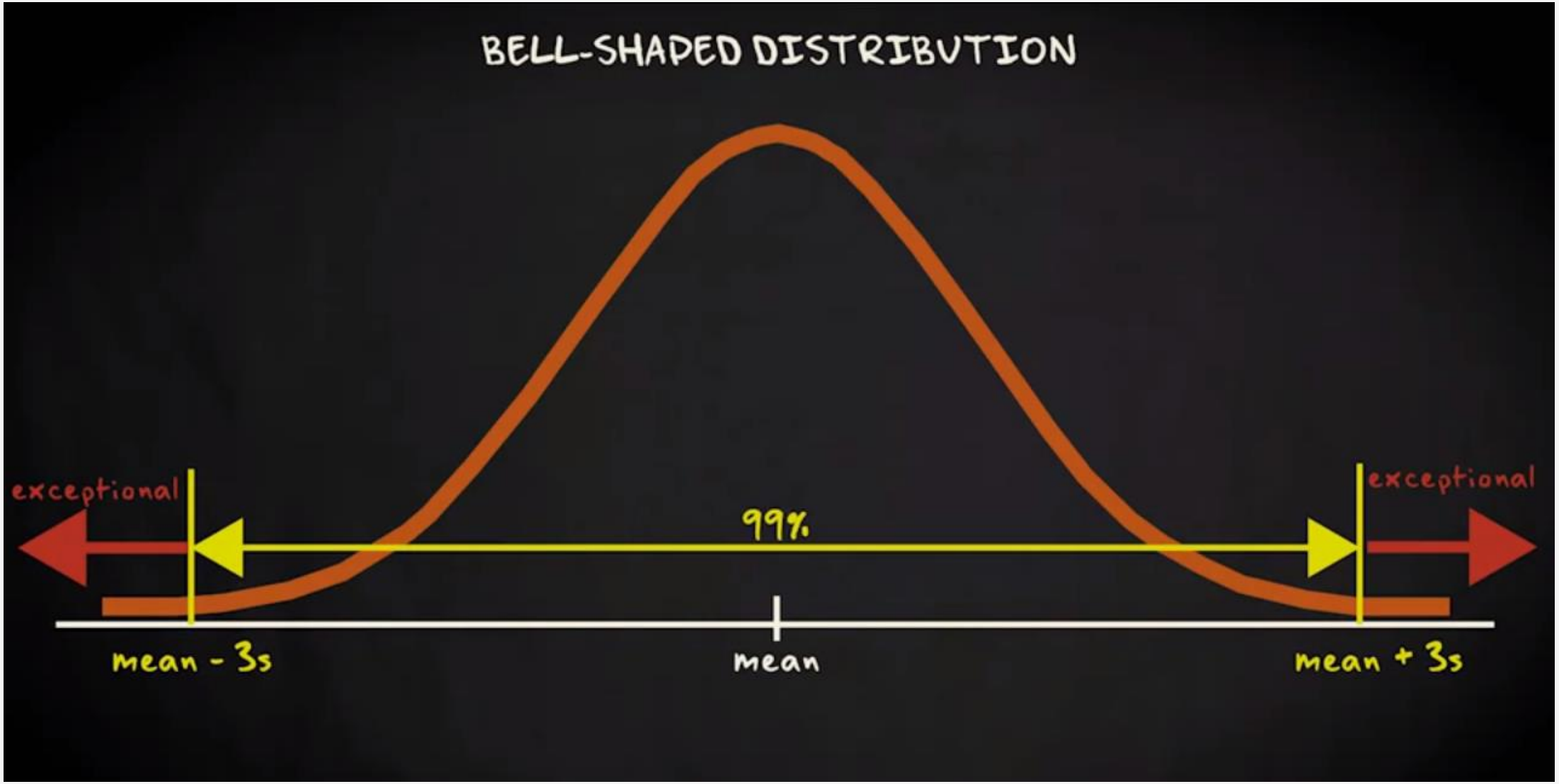
Z-scores


BELL-SHAPED DISTRIBUTION
68%
95%
99%
mean - 3s
mean - 2s
mean - 1s
mean
mean + 1s
mean + 2s
mean + 3s

Z-scores

This means that for this type of distribution, a z-score of more than 3 or less than minus 3 can be conceived of as rather **exceptional**.

Z-scores

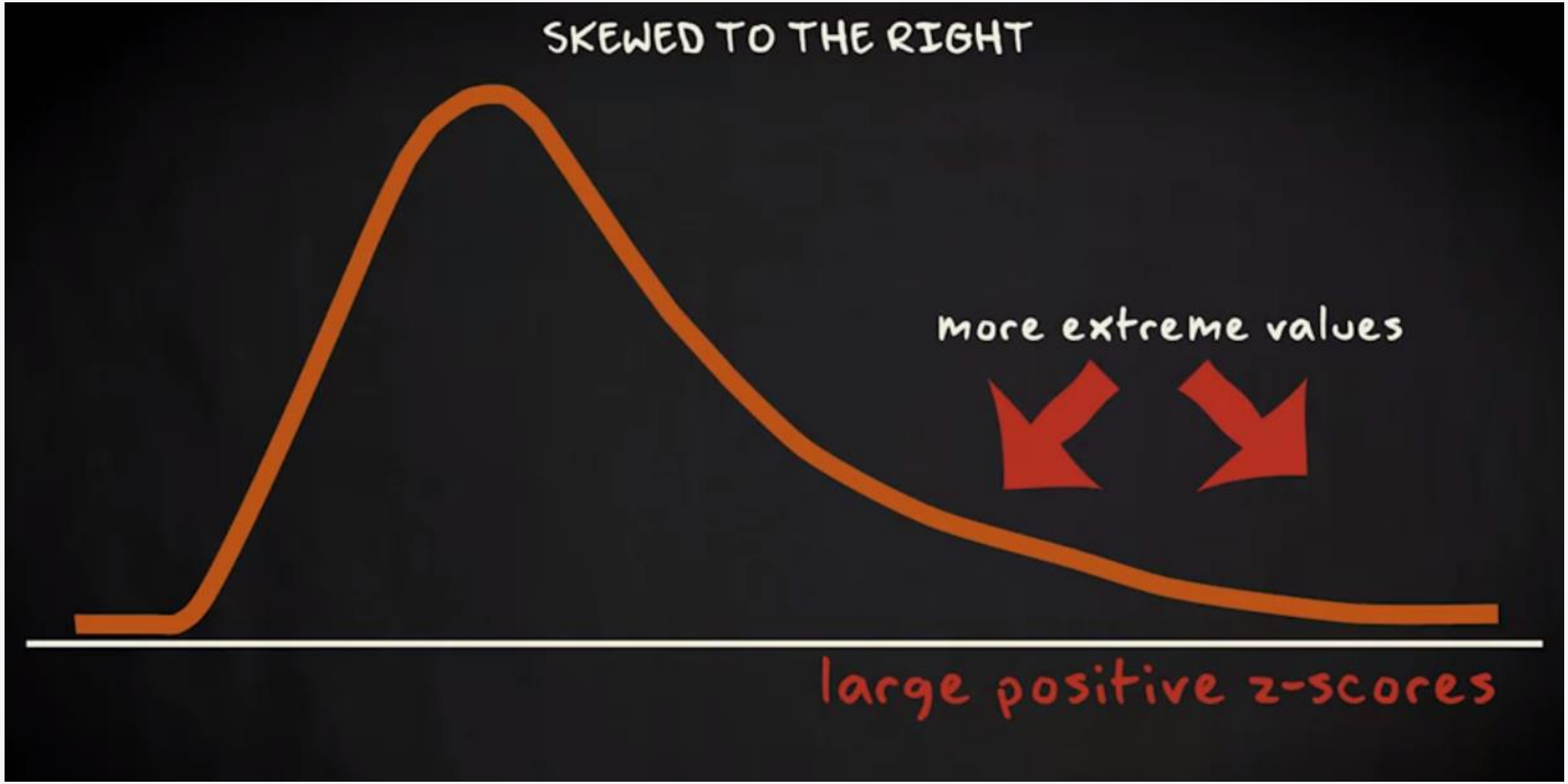

BELL-SHAPED DISTRIBUTION
exceptional
99%
exceptional
mean - 3s
mean
mean + 3s

Z-scores

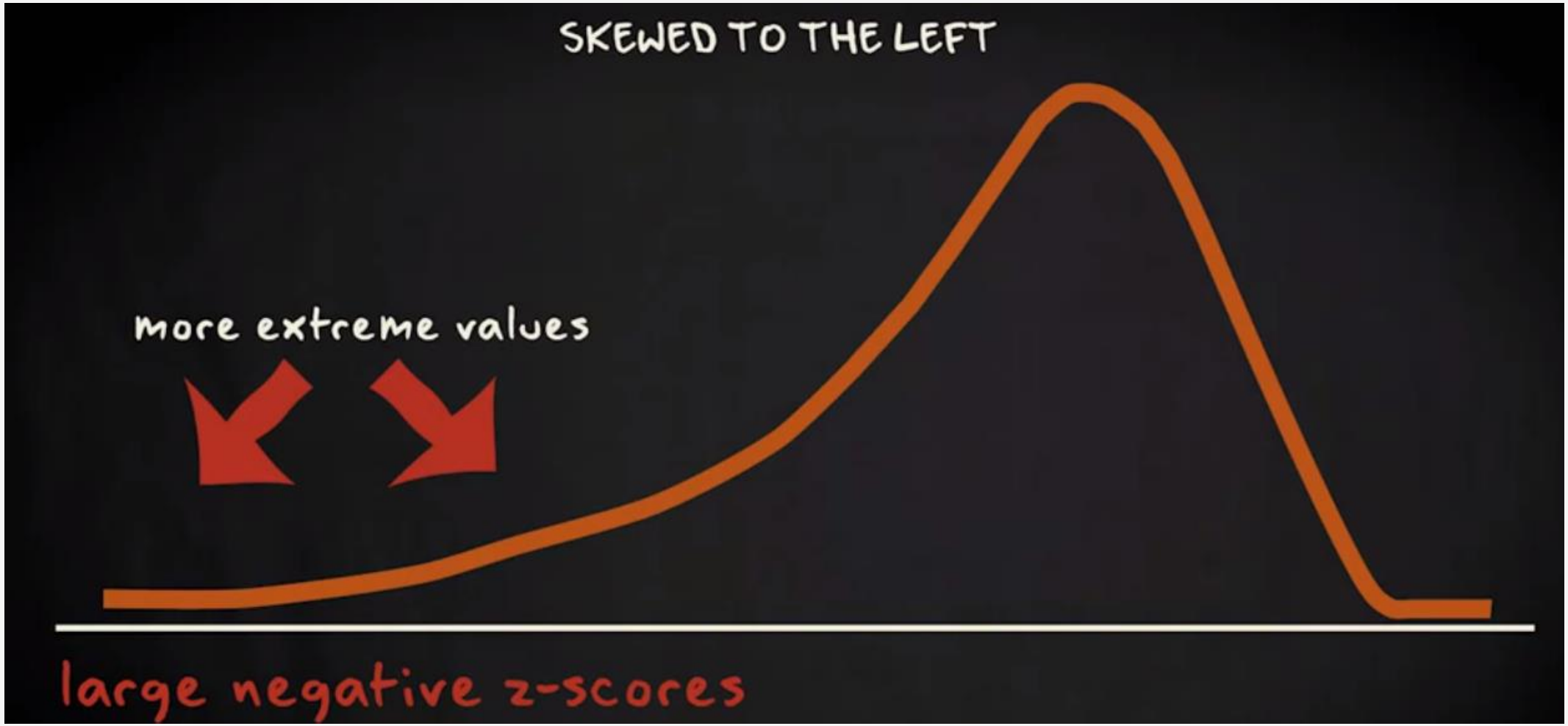
However, if a distribution is **strongly skewed to the right**, as in this graph, **large positive z-scores are more common**, because there are more extreme values on the right side of the distribution.

Similarly, if a distribution is **strongly skewed to the left**, then **large negative z-scores are more common**, because there are more extreme values on the left side of the distribution.

Z-scores


SKEWED TO THE RIGHT
more extreme values
large positive z-scores

Z-scores

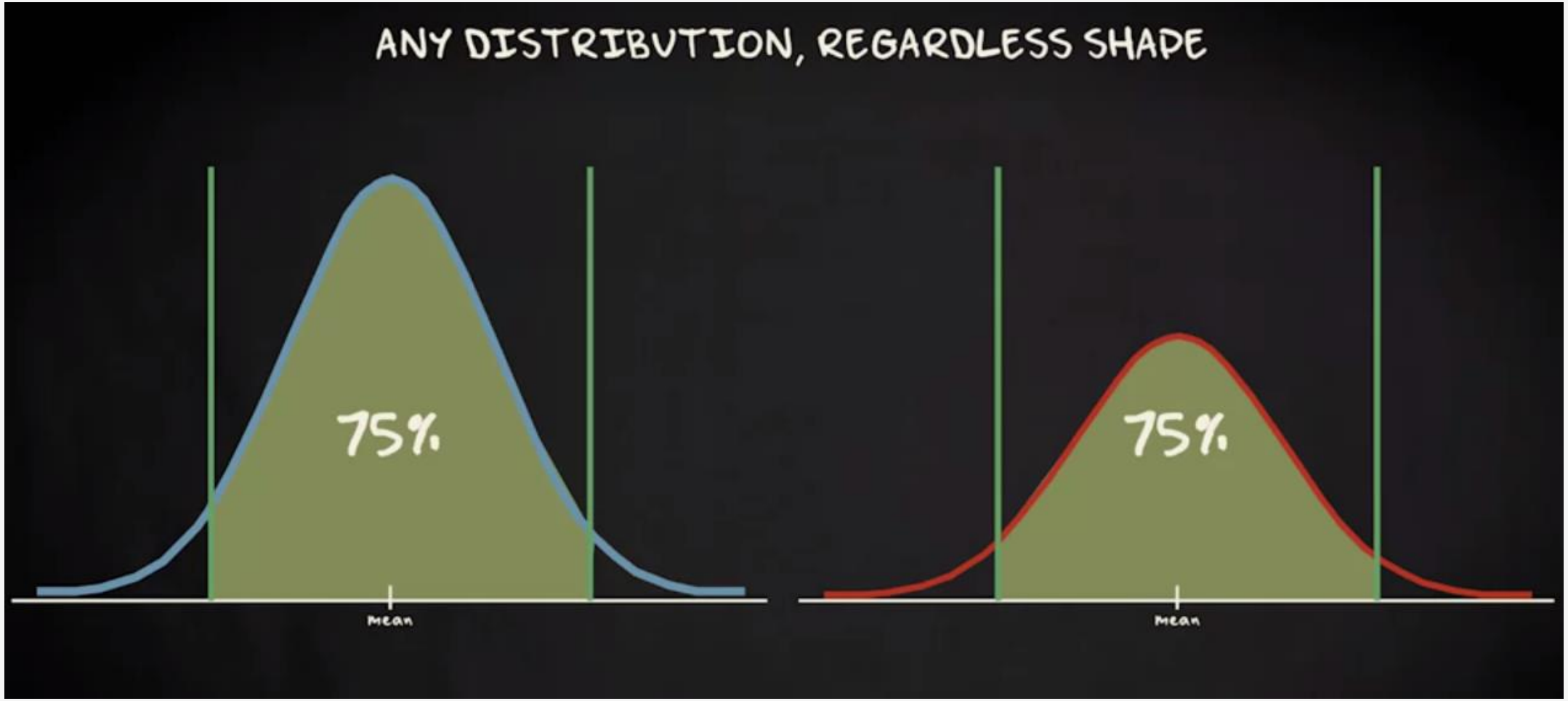

SKEWED TO THE LEFT
more extreme values
large negative z-scores

Z-scores

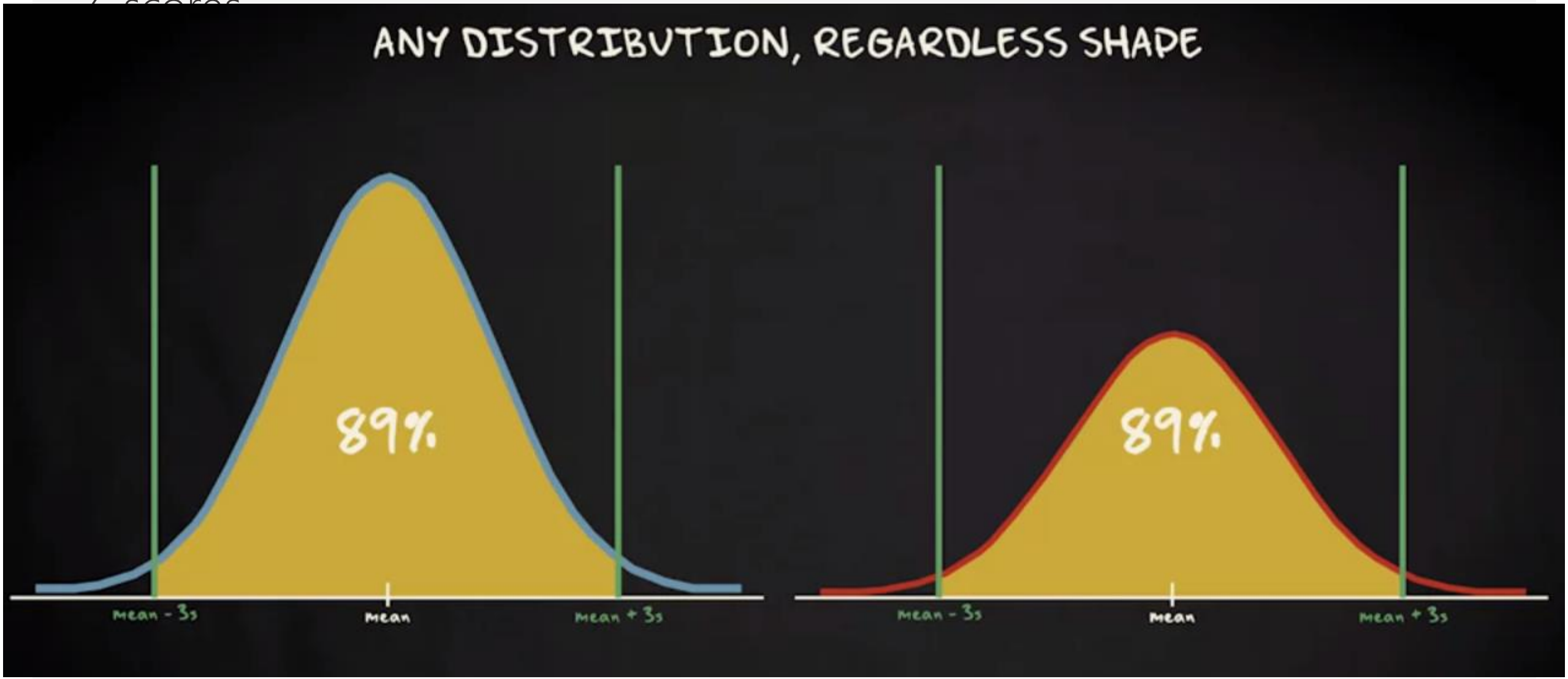
A rule that applies to any distribution, regardless shape, is that

- ✓ **75** percent of the data **must lie within** a z-score of **plus or minus 2** and
- ✓ **89** percent **within** a z-score of **plus or minus 3**

Z-scores



Z-scores

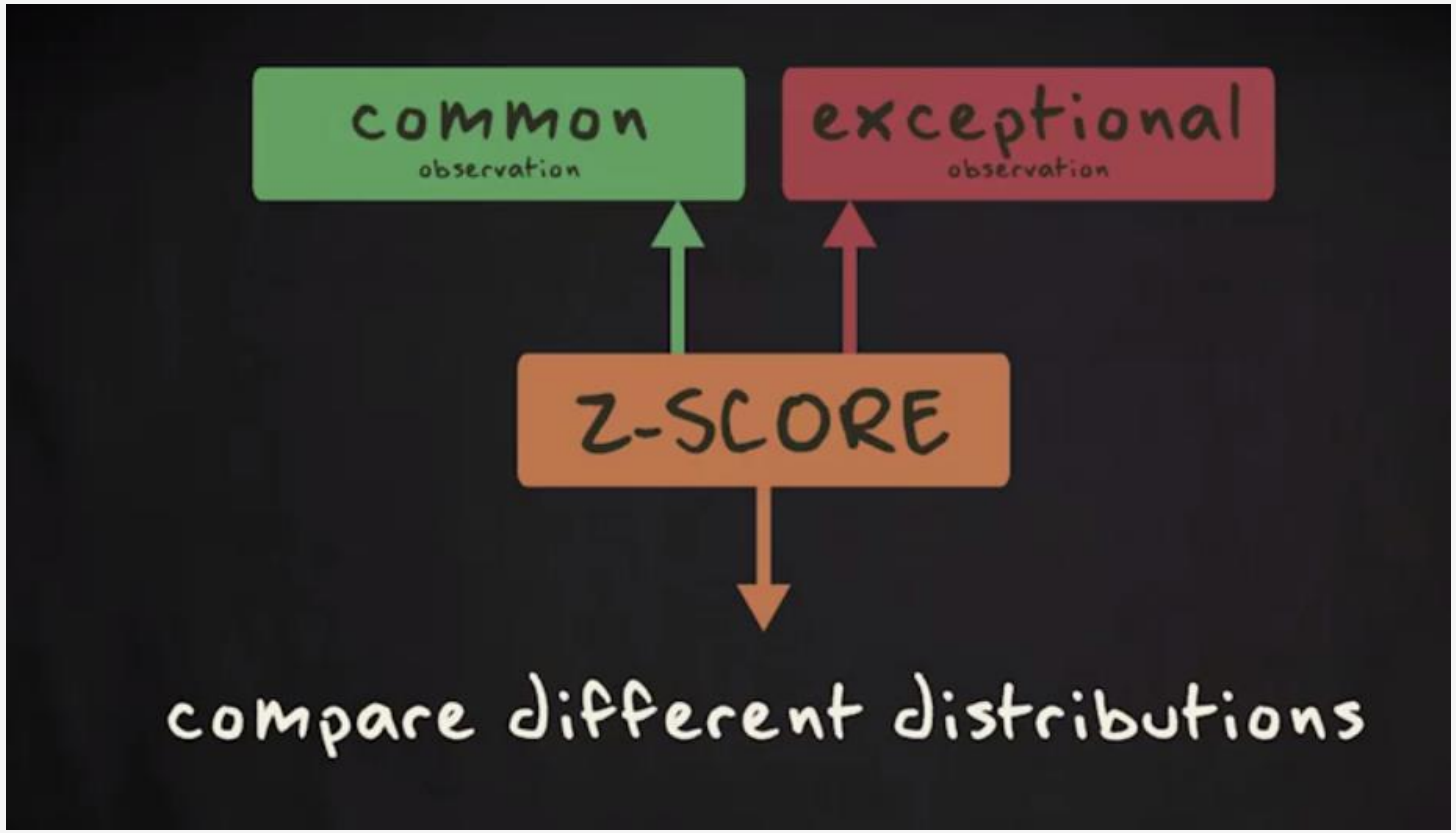

ANY DISTRIBUTION, REGARDLESS SHAPE
89%
mean - 3s
mean
mean + 3s
89%
mean - 3s
mean
mean + 3s

Z-scores

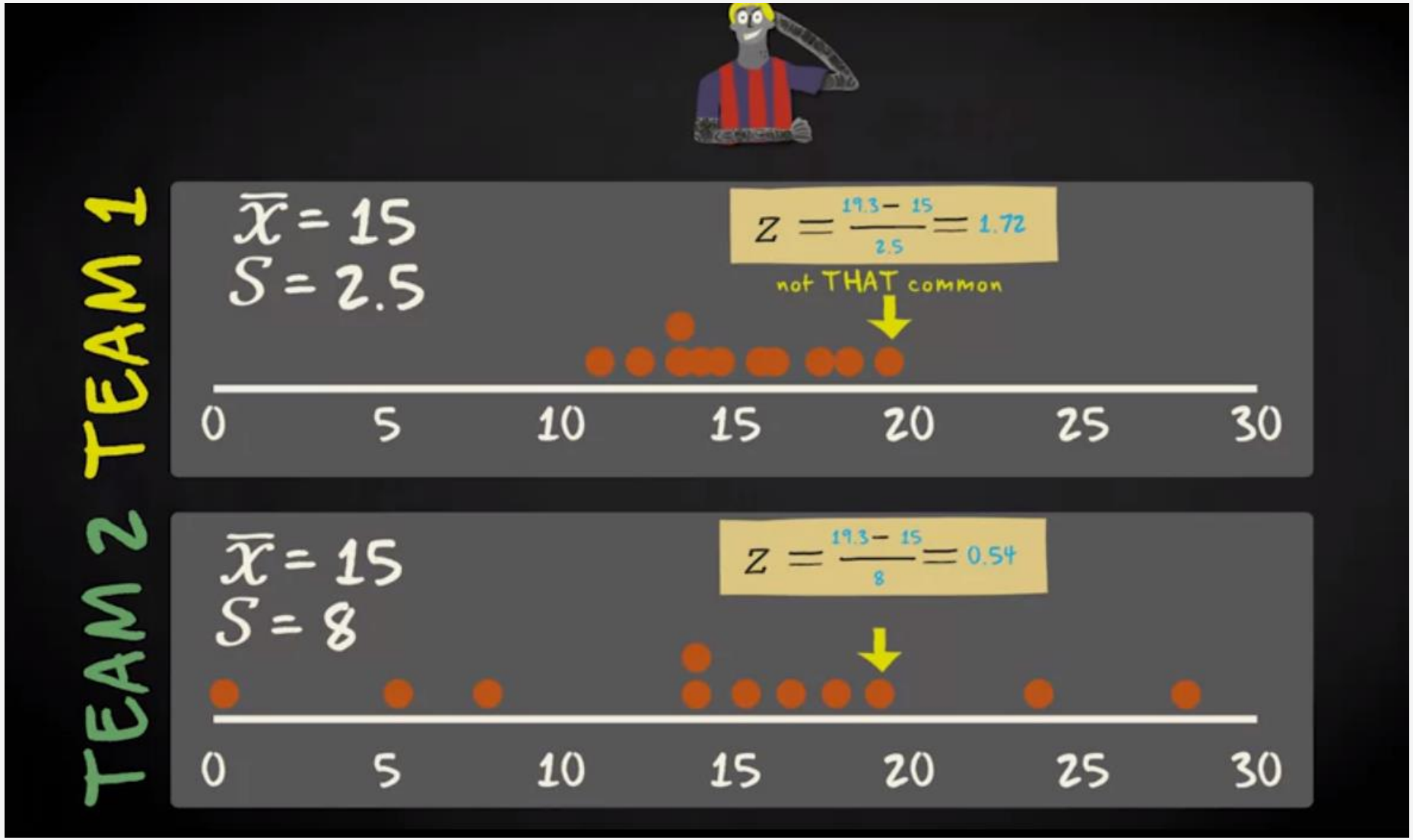
So, in itself a z-score gives you, to a certain extent, information about how extreme an observation is. Z-scores are even more useful if you want to compare different distributions.

Let's, for example,

look at the question whether a body weight of 19.3 is common or not.



Z-scores

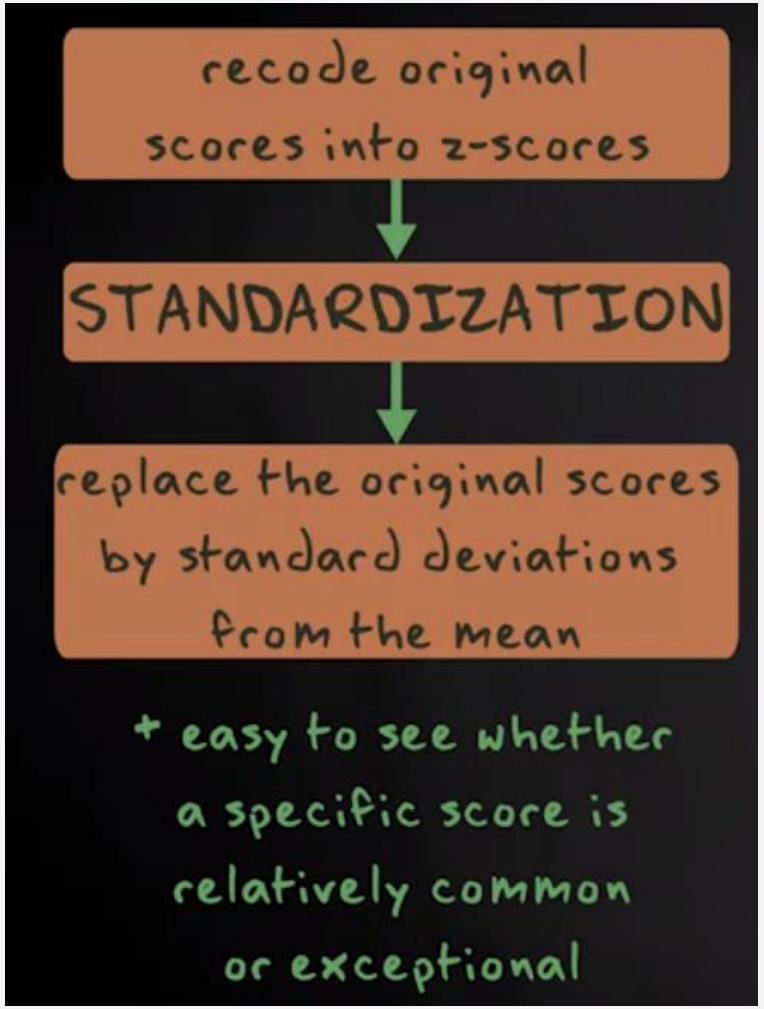

TEAM 1
$\bar{x} = 15$
$S = 2.5$
$z = \frac{19.3 - 15}{2.5} = 1.72$
not THAT common
0 5 10 15 20 25 30
TEAM 2
$\bar{x} = 15$
$S = 8$
$z = \frac{19.3 - 15}{8} = 0.54$
0 5 10 15 20 25 30

Z-scores

If we recode original scores into z-scores, we say that we standardize a variable.

Standardization means that we replace the scores measured in the original metric by scores expressed in standard deviations from the mean.

The advantage is that we can see at a glance whether a specific score is relatively common or exceptional.

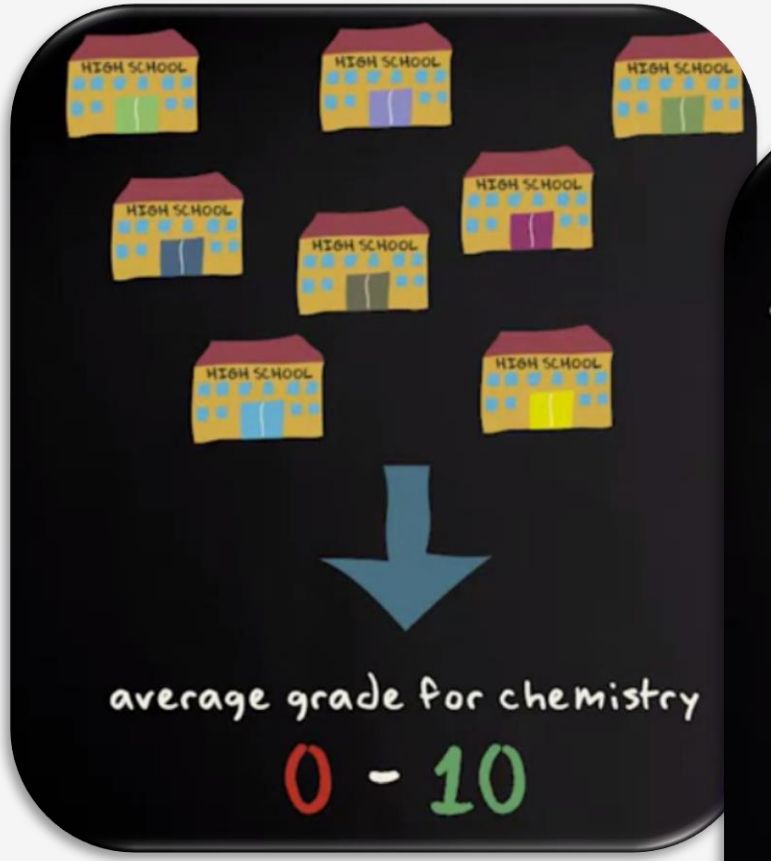


Exercise

Say I live in a city with 8 high schools. I want to know what, per high school, the average grade for chemistry is. The lowest possible grade is a 0 and the highest possible grade is a 10.

See the data matrix.

You can see that the **cases** studied here are not individual students, but schools. The **variable** of interest is the average grade for chemistry.



	Average grade chemistry
School 1	7,4
School 2	7,9
School 3	4,1
School 4	8,1
School 5	6,2
School 6	7,1
School 7	7,4
School 8	6,7

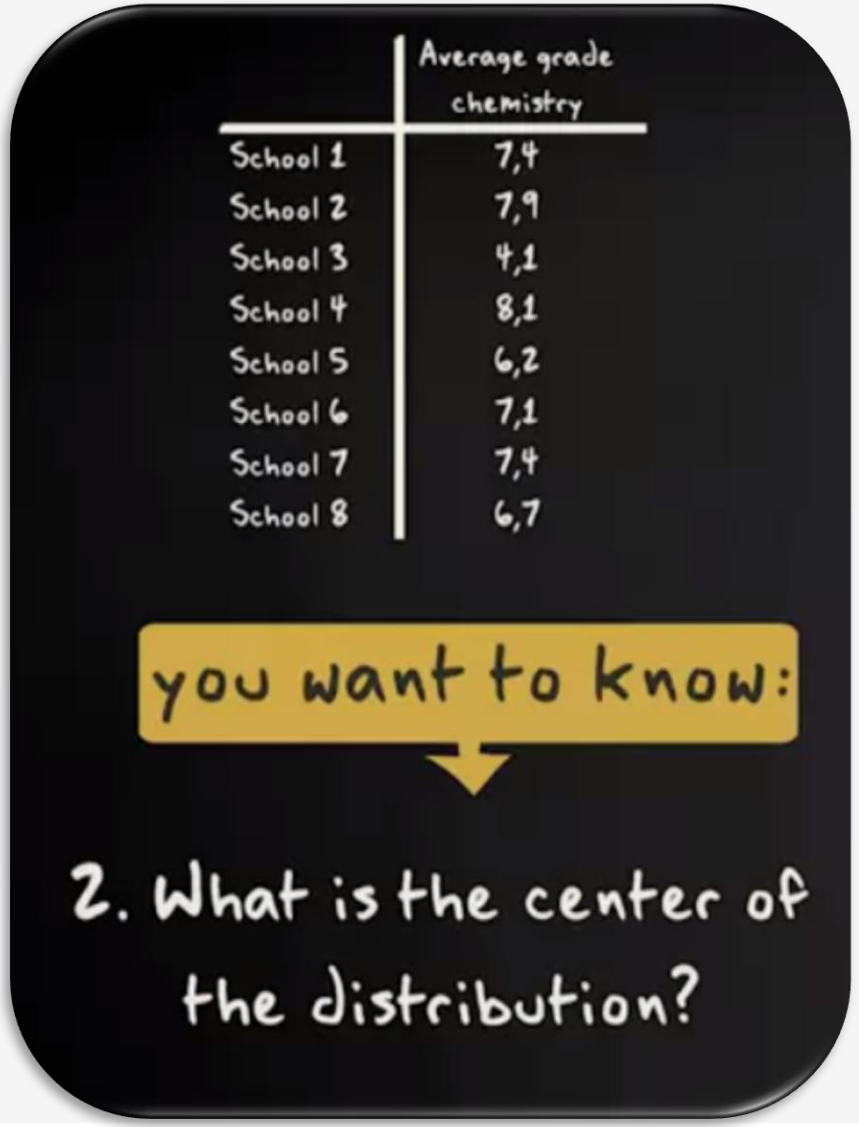
Exercise

	Average grade chemistry
School 1	7,4
School 2	7,9
School 3	4,1
School 4	8,1
School 5	6,2
School 6	7,1
School 7	7,4
School 8	6,7

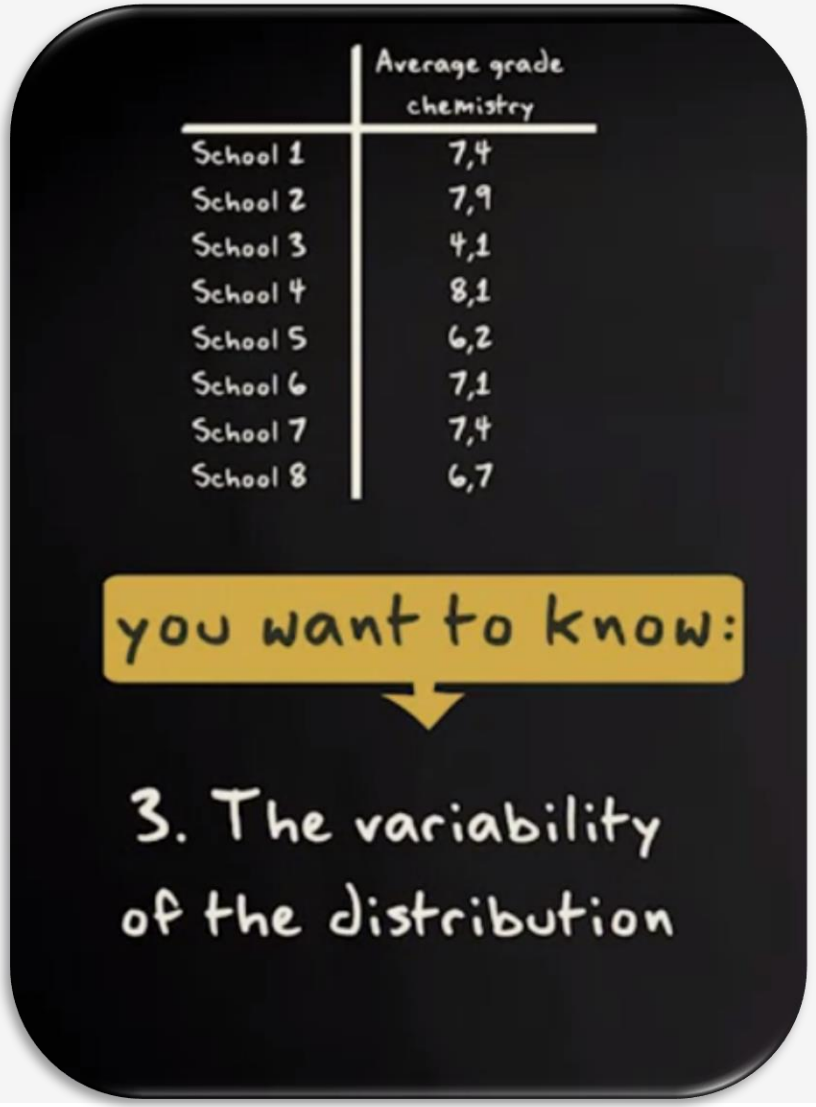
you want to know:

1. What does the distribution of the variable "average grade for chemistry" looks like?

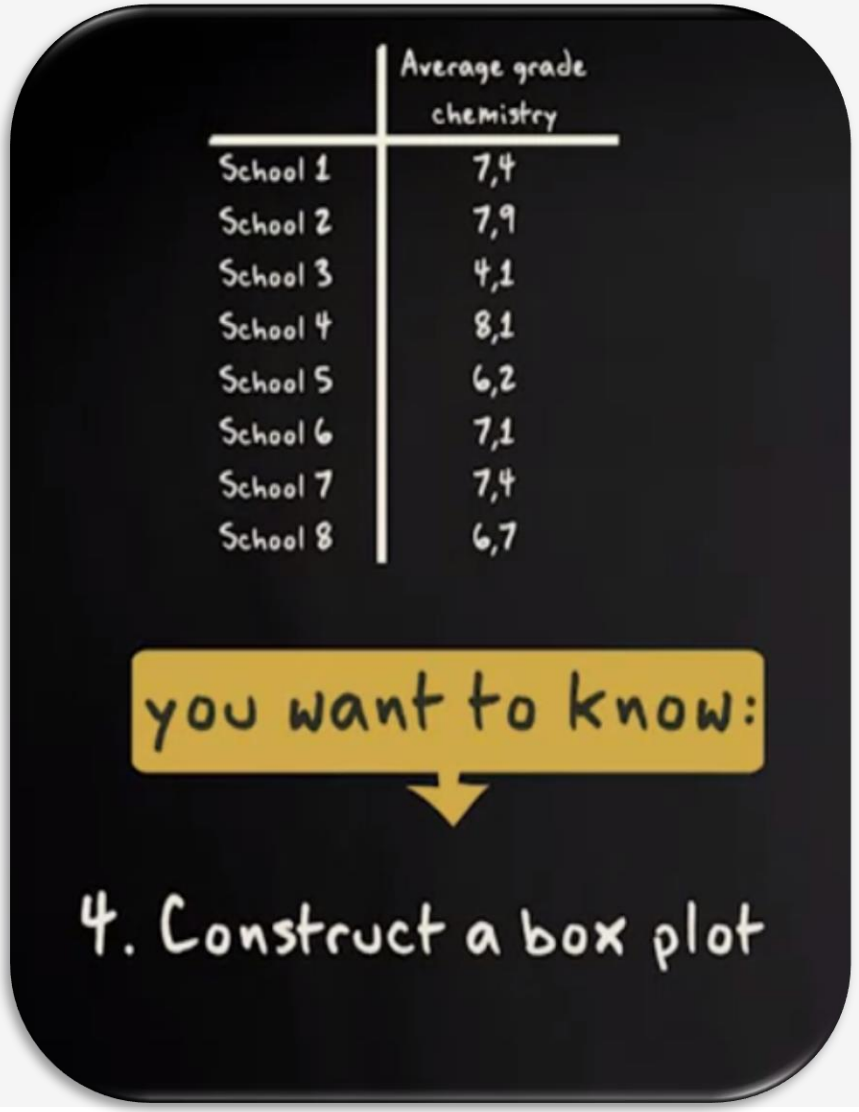
Exercise


Average grade chemistry
School 1 7,4
School 2 7,9
School 3 4,1
School 4 8,1
School 5 6,2
School 6 7,1
School 7 7,4
School 8 6,7
you want to know:
2. What is the center of the distribution?

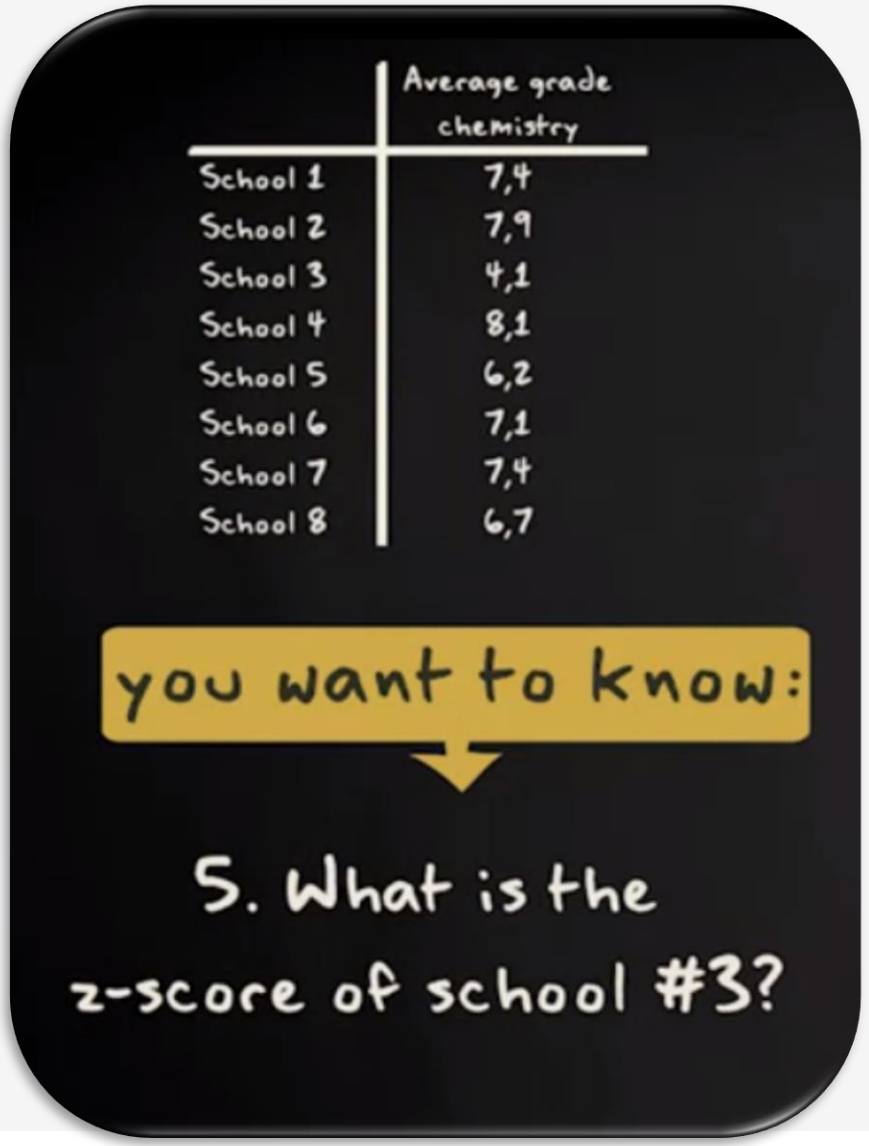
Exercise


Average grade chemistry
School 1 7,4
School 2 7,9
School 3 4,1
School 4 8,1
School 5 6,2
School 6 7,1
School 7 7,4
School 8 6,7
you want to know:
3. The variability of the distribution

Exercise


Average grade chemistry
School 1 7,4
School 2 7,9
School 3 4,1
School 4 8,1
School 5 6,2
School 6 7,1
School 7 7,4
School 8 6,7
you want to know:
4. Construct a box plot

Exercise


Average grade chemistry
School 1 7,4
School 2 7,9
School 3 4,1
School 4 8,1
School 5 6,2
School 6 7,1
School 7 7,4
School 8 6,7
you want to know:
5. What is the z-score of school #3?